



**TÜRKİYE CUMHURİYETİ
ANKARA ÜNİVERSİTESİ
SAĞLIK BİLİMLERİ ENSTİTÜSÜ**



**TÜRKİYE’DE İNGİLİZ YARIŞ ATLARININ KAZANÇ
DURUMLARINA ETKİLİ FAKTÖRLERİN
SINIFLAMAYA DAYALI BAZI MAKİNE ÖĞRENMESİ
ALGORİTMALARI İLE İNCELENMESİ**

Volkan TÜRKMEN

**VETERİNERLİK BİYOİSTATİSTİK ANABİLİM DALI
DOKTORA TEZİ**

**DANIŞMAN
Doç. Dr. Doğukan ÖZEN**

**ANKARA
2024**

TÜRKİYE CUMHURİYETİ
ANKARA ÜNİVERSİTESİ
SAĞLIK BİLİMLERİ ENSTİTÜSÜ

TÜRKİYE'DE İNGİLİZ YARIŞ ATLARININ KAZANÇ
DURUMLARINA ETKİLİ FAKTÖRLERİN SINIFLAMAYA
DAYALI BAZI MAKİNE ÖĞRENMESİ ALGORİTMALARI İLE
İNCELENMESİ

Volkan TÜRKMEN

VETERİNERLİK BİYOİSTATİSTİK ANABİLİM DALI
DOKTORA TEZİ

DANIŞMAN
Doç. Dr. Doğukan ÖZEN

ANKARA

2024

ETİK BEYAN

Ankara Üniversitesi
Sağlık Bilimleri Enstitüsü Müdürlüğü'ne,

Doktora tezi olarak hazırlayıp sunduğum “Türkiye’de İngiliz Yarış Atlarının Kazanç Durumlarına Etkili Faktörlerin Sınıflamaya Dayalı Bazı Makine Öğrenmesi Algoritmaları ile İncelenmesi” başlıklı tez; bilimsel ahlak ve değerlere uygun olarak tarafımdan yazılmıştır. Tezimin fikir/hipotezi tümüyle tez danışmanım ve bana aittir. Tezde yer alan araştırma tarafımdan yapılmış olup, tüm cümleler, yorumlar bana aittir.

Yukarıda belirtilen hususların doğruluğunu beyan ederim.

Öğrencinin Adı Soyadı: Volkan TÜRKMEN

Tarih: 11.12.2024

İmza:

KABUL VE ONAY

Ankara Üniversitesi Sağlık Bilimleri Enstitüsü

Veterinerlik Biyoistatistik Anabilim Dalında

Volkan TÜRKMEN tarafından hazırlanan

“Türkiye’de İngiliz Yarış Atlarının Kazanç Durumlarına Etkili Faktörlerin Sınıflamaya Dayalı Bazı Makine Öğrenmesi Algoritmaları ile İncelenmesi” adlı tez çalışması aşağıdaki jüri tarafından DOKTORA TEZİ olarak OY BİRLİĞİ ile kabul edilmiştir.

Tez Savunma Tarihi:

İmza
Prof. Dr. Bülent ÇELİK
Gazi Üniversitesi
Jüri Başkanı

İmza
Prof. Dr. Necmettin ÜNAL
Ankara Üniversitesi
Üye

İmza
Prof. Dr. Aytaç AKÇAY
Ankara Üniversitesi
Üye

İmza
Doç. Dr. Doğukan ÖZEN
Ankara Üniversitesi
Üye

İmza
Doç. Dr. Ufuk KAYA
Hatay Mustafa Kemal Üniversitesi
Üye

Tez hakkında alınan jüri kararı, Ankara Üniversitesi Sağlık Bilimleri Enstitüsü Yönetim Kurulu tarafından onaylanmıştır.

İmza
Prof. Dr. Fügen AKTAN
Sağlık Bilimleri Enstitüsü Müdürü

ÖZET

Türkiye’de İngiliz Yarış Atlarının Kazanç Durumlarına Etkili Faktörlerin Sınıflamaya Dayalı Bazı Makine Öğrenmesi Algoritmaları ile İncelenmesi

At yarışı endüstrisi, yarattığı ekonomik kazanç ve istihdam ile birlikte oldukça önemli bir sanayi dalıdır. Endüstrinin sürdürülebilirliği ve büyümesi, yarış atı yetiştiriciliğinin gelişmesi ve iyi safkanların yetiştirilmesine bağlıdır. Yarış atlarının kazançları, yarış endüstrisinde bir atın başarı seviyesinin en önemli göstergelerinden biridir. Bir atın kazancı, sadece o atın yarış kariyerini değil, aynı zamanda yetiştiriciliğinden sağlanan ekonomik getirileri de doğrudan etkiler. Bu durum, yarış endüstrisinde yatırım kararlarından seleksiyon stratejilerine kadar geniş bir yelpazede önemli sonuçlar doğurma potansiyelindedir. Bu çalışmada, sınıflandırmaya dayalı k-EYK, C5.0, RO ve GAM makine öğrenmesi (MÖ) algoritmaları yardımı ile Türkiye’deki safkan İngiliz atlarının yarış hayatı boyunca elde ettiği kazançta etkili olabilecek faktörleri belirlemek ve bu faktörleri içeren bir tahmin modeli geliştirmek amaçlanmıştır. Bu amaçla iki ve üç sınıflı kazanç değişkenini içeren iki farklı veri seti ile iki farklı veri bölme tekniği kullanılmıştır.

Araştırmanın materyalini Türkiye Jokey Kulübü’nün düzenli olarak tutmuş olduğu kamuya açık kayıtlar oluşturmuştur. Çalışmada her bir at bağımsız bir birey olarak kabul edilmiştir. 2007-2020 yılları aralığında doğup 30 Haziran 2022 tarihi itibarıyla yarış hayatı sonlanmış ve bu tarihe kadar düzenlenen yarışlara en az bir kez katılan 14.233 adet safkan İngiliz yarış atı çalışmaya dahil edilmiştir. Yapılan çalışmada, üç sınıflı kazanç değişkeni, hem orijinal, hem de rastgele aşırı örnekleme yöntemi (RAÖ) kullanılarak dengelenmiş hali ile; iki sınıflı kazanç değişkeni ise orijinal hali ile modele alınmış ve %70-%30, %75-%25 ve %80-%20 eğitim ve test veri setlerine bölünmüştür. Veri setleri 10 katlı çapraz doğrulama yöntemi 10 tekrarlı olarak yürütülerek algoritmalar için hiper parametre en iyilemesi yapılmıştır. Ayrılan test veri seti ile tahmin modeli test edilmiştir. Çalışmada performans değerlendirme ölçütü olarak, doğruluk, dengeli doğruluk, seçicilik ve negatif tahmin değeri ortalaması ile kesinlik, anma ve F1 skorunun makro ortalaması, MKK/ÇSMKK, G-ortalama ve EAKA kullanılmıştır.

Çalışma sonucunda karşılaştırılan modellerin içerisinde en yüksek doğruluk oranı %80-20 eğitim test veri seti ayrımı ile gerçekleşmiştir. İki sınıflı ve üç sınıflı dengelenmiş test veri setlerinde en yüksek doğruluk oranları sırasıyla GAM ve RO algoritmaları ile elde edilmiştir. Çalışmada elde edilen sınıflandırma tahmin modellerinde, farklı önem derecelerinde “Yarışa Başlama Yaşı”, “Koşu Sayısı”, “Koşu Parkuru”, “Baba Kazanç Durumu”, “Anne Kazanç Durumu”, “Sahiplik Durumu” ve “Cinsiyet” özniteliklerinin yarış hayatı boyunca elde edilen kazançta etki eden önemli birer faktör oldukları belirlenmiştir.

Bu çalışma, aynı zamanda iki sınıflı sınıflandırıcıların üç sınıflılara göre daha iyi tahmin performansı sergilediklerini, RAÖ ile dengelenmiş verilerin dengelenmemişlerden birçok performans ölçütü yönünden daha iyi performans sergilediğini göstermiştir. Çok sınıflı tahmin problemlerinin çözümünde hızlı ve etkili çözümler sunan k-EYK algoritmasının kullanımının etkinliği görülmüştür. Dengeli ve dengesiz veri setlerinde çok sınıflı sınıflandırıcı performansını değerlendirirken doğruluk oranlarının yanında, F1 skoru, G-ortalama ve ÇSMKK değerleri ile karışıklık matrisinin de beraber düşünülmesi gerektiği ancak EAKA ölçütünün kullanımında daha dikkatli olunması gerektiği görülmüştür.

Anahtar Sözcükler: Çok Sınıflı Dengesiz Veri, Kazanç, Makine Öğrenmesi, Safkan İngiliz Yarış Atı, Sınıflandırma

SUMMARY

Investigation of the Factors Affecting the Earning Status of Thoroughbreds in Türkiye with Some Machine Learning Algorithms Based on Classification

The horse racing industry represents a significant contributor to the national economy, providing substantial economic benefits and employment opportunities. The long-term viability and expansion of the industry depend on the advancement of racehorse breeding and the production of high-quality Thoroughbreds. The earnings of racehorses serve as a primary indicator of a horse's success within the racing industry. A horse's earnings have a direct impact on not only its racing career but also the economic returns derived from its breeding. This has the potential to have significant implications for a multitude of issues within the racing industry, including investment decisions and selection strategies. This study was carried out with the help of classification-based k-EYK, C5.0, RO and GAM machine learning (ML) algorithms to determine the factors that may affect the earnings of thoroughbred British horses in Türkiye throughout their racing life and to develop a prediction model that includes these factors. In order to achieve this goal, two different data sets containing two and three classes of earnings variables and two different data division techniques were used.

The material of the study consisted of public records maintained by the Turkish Jockey Club on a regular basis. Each horse was regarded as an independent subject in the study. The study included 14,233 Thoroughbred racehorses born between 2007 and 2020 and whose racing careers ended by June 30, 2022. These horses had participated in at least one race organized until this date. In this study, the three-class earning variable is modeled in both its original and balanced forms by employing a random oversampling method. Additionally, the two-class earning variable is modeled in its original form and divided into training and test data sets with varying proportions, including 70%-30%, 75%-25%, and 80%-20%. The data sets were subjected to 10-fold cross-validation with 10 repetitions, and hyper-parameter optimization was performed for the algorithms. The prediction model was tested with the separated test data set. The performance of the model was evaluated using a range of metrics, including accuracy, balanced accuracy, specificity, negative predictive value, precision, recall, F1 score, MCC/MMCC, G-mean, and AUC.

The highest accuracy rate among the models compared in the study was achieved with an 80-20% training-test data set separation. In two-class and three-class balanced test datasets, the highest accuracy rates were obtained with the GBM and RF algorithms, respectively. In the classification prediction models obtained in the study, it was determined that the "Age at Race Start," "Number of Runs," "Track Type," "Father's Earnings Status," "Mother's Earnings Status," "Ownership Status," and "Gender" attributes are important factors affecting the earnings obtained during the racing life, with varying degrees of importance.

The findings of this study indicate that two-class classifiers exhibit superior prediction performance compared to three-class classifiers. Additionally, data that has been balanced using the ROS method demonstrates enhanced performance across a range of metrics when compared to non-balanced data. The efficacy of the KNN algorithm, which offers rapid and effective solutions for addressing multi-class prediction challenges, has been substantiated. When assessing the performance of multiclass classifiers on balanced and unbalanced datasets, it is essential to consider additional metrics beyond accuracy rates, such as F1 score, G-mean and MMCC values, and a confusion matrix. However, greater caution should be exercised when utilising the AUC criterion.

Keywords: Classification, Earnings, Machine Learning, Multi-class Imbalanced Data, Thoroughbred Race Horse

İÇİNDEKİLER

Etik Beyan	ii
Kabul Onay	iii
Özet	iv
Summary	v
İçindekiler	vi
Önsöz	viii
Simgeler ve Kısaltmalar	x
Şekiller	xii
Çizelgeler	xiii

1. GİRİŞ	1
1.1. Türk Toplumlarında At Yetiştiriciliği ve Kısa Tarihçesi	2
1.2. Safkan İngiliz (Thoroughbred) Atı Yetiştiriciliği ve Kısa Tarihçesi	8
1.3. Safkan İngiliz (Thoroughbred) Atının Özellikleri	10
1.4. Atlı Sporlar ve At Yarışları	11
1.5. Türkiye’deki At Yarışları ve İngiliz Atlarının Yarışlardaki Yeri	13
1.6. Türkiye’de İngiliz Yarış Atı Yetiştiriciliği	16
1.7. At Yarışlarının Ekonomik Boyutu	17
1.8. Yarış Atlarının Performansını Etkileyen Faktörler	19
1.9. Makine Öğrenmesi ile İlgili Genel Bilgiler	25
1.9.1. İstatistiksel Öğrenme Yöntemi	26
1.9.2. Makine Öğrenmesi (MÖ) (Machine Learning, ML)	29
1.9.3. Büyük Veri	30
1.10. Sınıflamaya Dayalı Denetimli Öğrenme Algoritmaları	31
1.10.1. K-En Yakın Komşu (k-EYK) (K-Nearest Neighbor, KNN) Algoritması	32
1.10.2. C5.0 Karar Ağacı Algoritması	36
1.10.2.1. En Uygun Bölünme Nasıl Seçilir?	38
1.10.2.2. Budama (Pruning)	39
1.10.3. Topluluk (Ensemble) Öğrenme Algoritmaları	40
1.10.3.1. Torbalama (Bagging) Algoritması	40
1.10.3.2. Rastgele Orman (RO) (Random Forest, RF) Algoritması	41
1.10.3.3. Artırma (Boosting) Algoritması	47
1.10.3.4. Gradyan Artırıcı Makine (GAM) (Gradient Boosting Machine) Algoritması	48
1.10.3.4.1. GAM Nasıl Çalışır?	49
1.11. Çok Sınıflı (Multi-class) Sınıflandırma Yöntemleri	52
1.12. Sınıf Dengesizliği Tanımı ve Çözüm Yöntemleri	58

2. GEREÇ VE YÖNTEM	64
2.1. Veri Ön İşleme Süreci	64
2.2. Model Performans Değerlendirme	67
2.3. Hiper Parametre En İyilemesi (Tuning)	68
2.4. Öznitelik Önem Dereceleri	68
2.5. Model Karşılaştırması ve Performans Değerlendirme Metrikleri	68
2.5.1. F1 Skoru (F-Measure, F-ölçütü)	72
2.5.2. G-ortalama (G-Mean)	72
2.5.3. Eğri Altında Kalan Alan (EAKA) (Area Under Curve, AUC)	73
2.5.4. Matthews Korelasyon Katsayısı (MKK) (Matthews Correlation Coefficient, MCC)	74

2.6. Makro ve Mikro Ortalamalar	75
2.7. Dengesiz Veri İçin Rastgele Aşırı Örnekleme (RAÖ) Yöntemi	76
2.8. Veri Analizi İçin Kullanılan Program ve Paketler	77
3. BULGULAR	78
3.1. Üç Sınıflı Dengesiz Veri Seti Bulguları	78
3.1.1. Tanımlayıcı İstatistikler	78
3.1.2. MÖ Sınıflandırma Algoritma Sonuçları	79
3.1.3. Eğitim (%70) ve Test (%30) Veri Seti Sonuçları	79
3.1.4. Eğitim (%70) Veri Seti Performans Sonuçları	80
3.1.5. Test (%30) Veri Seti Performans Sonuçları	81
3.1.6. Eğitim (%75) ve Test (%25) Veri Seti Sonuçları	84
3.1.7. Eğitim (%75) Veri Seti Performans Sonuçları	84
3.1.8. Test (%25) Veri Seti Performans Sonuçları	85
3.1.9. Eğitim (%80) ve Test (%20) Veri Seti Sonuçları	88
3.1.10. Test (%20) Veri Seti Performans Sonuçları	88
3.1.11. Test (%20) Veri Seti Performans Sonuçları	89
3.2. RAÖ ile Dengelenmiş Veri Seti Bulguları	92
3.2.1. Tanımlayıcı İstatistikler	92
3.2.2. MÖ Sınıflandırma Algoritma Sonuçları	92
3.2.3. Eğitim (%70) ve Test (%30) Veri Seti Sonuçları	93
3.2.4. Eğitim (%70) Veri Seti Performans Sonuçları	93
3.2.5. Test (%30) Veri Seti Performans Sonuçları	94
3.2.6. Eğitim (%75) ve Test (%25) Veri Seti Sonuçları	97
3.2.7. Eğitim (%75) Veri Seti Performans Sonuçları	97
3.2.8. Test (%25) Veri Seti Performans Sonuçları	98
3.2.9. Eğitim (%80) ve Test (%20) Veri Seti Sonuçları	101
3.2.10. Eğitim (%20) Veri Seti Performans Sonuçları	102
3.2.11. Test (%20) Veri Seti Performans Sonuçları	103
3.3. İki Sınıflı Veri Seti Bulguları	106
3.3.1. Tanımlayıcı İstatistikler	107
3.3.2. MÖ Sınıflandırma Algoritma Sonuçları	107
3.3.3. Eğitim (%70) ve Test (%30) Veri Seti Sonuçları	108
3.3.4. Eğitim (%70) Veri Seti Performans Sonuçları	108
3.3.5. Test (%30) Veri Seti Performans Sonuçları	109
3.3.6. Eğitim (%75) ve Test (%25) Veri Seti Sonuçları	111
3.3.7. Eğitim (%75) Veri Seti Performans Sonuçları	112
3.3.8. Test (%25) Veri Seti Performans Sonuçları	113
3.3.9. Eğitim (%80) ve Test (%20) Veri Seti Sonuçları	115
3.3.10. Eğitim (%20) Veri Seti Performans Sonuçları	115
3.3.11. Test (%20) Veri Seti Performans Sonuçları	116
3.4. Tüm Test Veri Seti Performans Ölçüm Sonuçları	119
3.4.1. Test (%30) Veri Seti Performans Ölçüm Sonuçları	119
3.4.2. Test (%25) Veri Seti Performans Ölçüm Sonuçları	120
3.4.3. Test (%20) Veri Seti Performans Ölçüm Sonuçları	121
4. TARTIŞMA	122
5. SONUÇ VE ÖNERİLER	141
KAYNAKLAR	147

ÖNSÖZ

Türk kültürü ve tarihinde çok önemli bir yeri olan at ve at yetiştiriciliği, bütün eski Türk milletleri ve devletlerinin birbirine aktardığı gelenekler ile gelişmiş ve dönüşüme uğramış olup, bu süreç boyunca at, hem bir eğlence, hem bir şeref ve onur aracı, hem besin kaynağı, hem savaş aracı, hem tarım aracı, hem ulaşım aracı, hem dini ritüel aracı, hem de spor aracı olarak kullanılmıştır.

At, tarih boyunca toplulukların ihtiyacına uygun bir şekilde hayatın her alanında kullanılmıştır. Günümüzde ise attan en çok spor ve eğlence aracı olarak yararlanılmaktadır.

1920'lerde safkan İngiliz at ırkıyla tanışan Türkiye atçılığı, 1922'de Veliefendi'de haftalık koşu programına yurt dışından getirilen safkan İngiliz atlarının dahil edilmesi ile yeni bir döneme başlamıştır. Bu atların Türkiye'deki at severlere satılması sonucunda Türkiye'deki safkan İngiliz atçılığının temeli atılmıştır. Türkiye'de safkan İngiliz atı yetiştiriciliği 1928 yılında Karacabey Harası'nda devlet tarafından başlamış, günümüzde ise İngiliz koşu atı yetiştiriciliği özel yetiştiriciler tarafından yapılmaktadır.

At yarışı endüstrisi ve yarış atı yetiştiriciliği birbirini tamamlayan ve destekleyen iki bileşendir. Türkiye'de, uluslararası standartlara uygun ve daha modern bir yapıda at yarışlarının devam ettirilebilmesi ve standardın daha da geliştirilmesi için at yetiştiricilerinin sektöre daha iyi atlar kazandırması gerekir. Bu endüstri daha fazla gelişip büyüyerek ülke ekonomisine farklı türde iş imkanları ve dolayısıyla daha fazla istihdam sağlayabilir.

Yoğun emek isteyen bir iş olan iyi bir safkan yarış atı yetiştiriciliğinde harcanan para, zaman ve emek tasarrufu açısından bir yarış atının kazanç performansına etki eden öngörülemeyen risk faktörlerinin belirlenmesi ekonomik kayıpları önlemek açısından oldukça önemlidir. Endüstrinin sürdürülebilirliği ve büyümesi, yarış atı yetiştiriciliğinin gelişmesi ve iyi safkanların yetiştirilmesine bağlıdır.

Hayvan yetiştiriciliğinde, hayvan sağlığını doğrudan ve dolaylı yollardan etkileyen yetiştiricilik ile ilişkili olduğu düşünülen risk faktörlerinin bilinmesi çok önemlidir. Bu çalışma, Türkiye'deki safkan İngiliz yarış atlarının yarış hayatları boyunca kazanç performanslarına etki eden çevresel faktörlerin öneminin iki ve çok sınıflı makine öğrenmesi

(MÖ) (Machine Learning, ML) sınıflandırma algoritmaları kullanılarak modellenebileceğini göstermiştir.

İlk olarak, oluşturdıkları bilimsel ortamda bilim nosyonunu almama ve akademik dünyaya açılmama vesile olan iyi bir akademisyen olma motivasyonunu veren eski GATA hocalarıma teşekkür ederim.

Akademik gelişimim iki ayak üzerine kurulmuştur. İlki yüksek lisans aşaması Ankara Üniversitesi Tıp Fakültesi Biyoistatistik Anabilim dalında görev yapan değerli hocalarım Prof. Dr. Atilla Halil ELHAN, Prof. Dr. Yasemin GÖKMEN, Doç. Dr. Serdal Kenan KÖSE, Prof. Dr. Derya GÖKMEN ve Doç. Dr. Beyza DOĞANAY'a eğitimim aşamasında verdikleri emek ve motivasyon için teşekkür ederim.

İkinci ayağını Ankara Üniversitesi Veteriner Fakültesi Biyoistatistik Anabilim dalında görev yapan değerli hocalarım sayın Prof. Dr. İsmayıl Safa GÜRCAN' a ve Prof. Dr. Aytaç AKÇAY'a doktora eğitimim aşamasında verdikleri emek ve motivasyon için özellikle teşekkür ederim

Doktora eğitimimin başından itibaren en değerli katkıları sağlayan ve her türlü durumda desteğini üzerimden eksik etmeyen danışman hocam Doç. Dr. Doğukan ÖZEN ve değerli eşi Dr. Öğrt. Üyesi Hülya ÖZEN hocalarıma çok teşekkür ederim.

Üzerimde ödenmez hakkı bulunan merhum annem Yeter TÜRKMEN'e ve hayatta olan babam Mustafa TÜRKMEN'e, on yıldan uzun süren eğitimim sırasında bütün sıkıntıları beraber omuzladığım biricik eşim Zehra TÜRKMEN'e ve evimin neşesi oğullarım Enes Taha ve Mustafa Emir'e sonsuz ve en içten teşekkürler.

SİMGELER ve KISALTMALAR

AGAM	Aşırı Gradyan Artırıcı Makine (Extreme Gradient Boosting Machine, XGBoost)
AİK	Alıcı İşlem Karakteristiği (Receiver Operating Characteristic, ROC)
AKD	Anne Kazanç Durumu
BKB	Bire Karşı Bir (One-vs-One, OvO)
BKD	Baba Kazanç Durumu
BKH	Bire Karşı Hepsi (One-vs-All, OvA)
ÇD	Çapraz Doğrulama (Cross Validation)
ÇSMKK	Çok Sınıflı Matthews Korelasyon Katsayısı (Multiclass Matthews Correlation Coefficient, MMCC)
DVM	Destek Vektör Makineleri (Support Vector Machine, SVM)
EAKA	Eğrinin Altında Kalan Alan (Area Under Curve, AUC)
ESA	Evrişimli Sinir Ağları (Convolutional Neural Network, CNN)
GRA	Gözlemsel Risk Azaltımı (Empirical Risk Minimization, ERM)
GA	Güven Aralığı
GAM	Gradyan Artırıcı Makine (Gradient Boosting Machine, GBM)
HDÇK	Hata Düzeltme Çıkış Kodları (Error-Correcting Output Codes, ECOC)
HGAM	Hafif Gradyan Artırma Makine (Light Gradient Boosted Machine, LightGBM)
KA	Karar Ağacı
k-EYK	K-En Yakın Komşu (K-Nearest Neighbor, KNN)
KGAM	Kategorik Gradyan Artırma (Category Gradient Boosting, CatBoost)
KP	Koşu Parkuru
KS	Koşu Sayısı
KSÜ	Kariyer Süresi
LR	Lojistik Regresyon
Mİ	Meyilli İniş (Gradient Descent)
MKK	Matthews Korelasyon Katsayısı (Matthews Correlation Coefficient, MCC)
MÖ	Makine Öğrenmesi (Machine Learning, ML)
NB	Naïve Bayes
RAÖ	Rastgele Aşırı Örnekleme (Random Over Sampling, ROS)
RO	Rastgele Orman (Random Forest, RF)
RAU	Rastgele Alt Uzay (Random Subspace)
RÖA	Rastgele Örneklem Azaltma (Random Under Sampling, RUS)
SA	Sinir Ağları (Neural Network, NN)
SAÖAT	Sentetik Azınlık Örneklem Arttırma Tekniği (Synthetic Minority Oversampling Technique -SMOTE)
SD	Sahiplik Durumu

SRKA	Sınıflandırma ve Regresyon Karar Ağaçları (Classification and Regression Trees, CART)
TDH	Torba Dışı Hata (Out-Of-Bag, OOB)
Tİ	Torba İçi (In-Bag, inBag)
TİGEM	Tarım İşletmeleri Genel Müdürlüğü
TKB	Tarım ve Köyişleri Bakanlığı
TJK	Türkiye Jokey Kulübü
VC	Vapnik-Chervonenkis
YBY	Yarışa Başlama Yaşı
YSA	Yapay Sinir Ağı (Artificial Neural Network, ANN)
YTY	Yarış Terketme Yaşı
YZ	Yapay Zeka (Artificial Intelligence, AI)



ŞEKİLLER

Şekil 1.1. k-EYK algoritmasında çok sınıflı sınıflandırma gösterim şekli	33
Şekil 1.2. En sık kullanılan uzaklık ölçme yöntemleri	35
Şekil 1.3. Kökten son düğüme kadar giden yolları gösteren karar ağacı yapısı	37
Şekil 1.4. İki ve çok sınıflı karar ağacı bölünmesini gösteren şekil	39
Şekil 1.5. Torbalama algoritması eğitim aşaması görsel gösterim şekli	41
Şekil 1.6. RO sınıflandırma algoritmasında orijinal veri setinin “Tİ” ve “TDH” olarak bölünmesi ile oluşan çalışma şekli gösterimi	45
Şekil 1.7. RO sınıflandırma algoritmasının çalışma şekli gösterimi	45
Şekil 1.8. Boosting metodolojisinin işleyiş şeklini gösteren şekil	47
Şekil 1.9. Meyilli (Gradyan) inişin gösterim şekli	48
Şekil 1.10. k GAM algoritması çalışma adımları	49
Şekil 1.11. İki ve üç sınıflı sınıflandırma çözüm şekli	54
Şekil 1.12. Üç sınıflı sınıflandırma probleminde BKH yönteminin çözüm şekli	56
Şekil 1.13. Üç sınıflı sınıflandırma probleminde BKB yönteminin çözüm şekli	56
Şekil 3.1. Üç sınıflı %70 eğitim veri seti tahmin modeli öznitelik önem dereceleri	83
Şekil 3.2. Üç sınıflı %75 eğitim veri seti tahmin modeli öznitelik önem dereceleri	87
Şekil 3.3. Üç sınıflı %80 eğitim veri seti tahmin modeli öznitelik önem dereceleri	91
Şekil 3.4. RAÖ ile dengelenen üç sınıflı %70 eğitim veri seti tahmin modeli öznitelik önem dereceleri	96
Şekil 3.5. RAÖ ile dengelenen üç sınıflı %75 eğitim veri seti tahmin modeli öznitelik önem dereceleri	101
Şekil 3.6. RAÖ ile dengelenen üç sınıflı %80 eğitim veri seti tahmin modeli öznitelik önem dereceleri	106
Şekil 3.7. İki sınıflı %70 eğitim veri seti tahmin modeli öznitelik önem dereceleri	111
Şekil 3.8. İki sınıflı %75 eğitim veri seti tahmin modeli öznitelik önem dereceleri	114
Şekil 3.9. İki sınıflı %80 eğitim veri seti tahmin modeli öznitelik önem dereceleri	118

ÇİZELGELER

Çizelge 1.1. Yıllara göre Türkiye’de koşan ferdi at sayıları	15
Çizelge 2.1. Öznitelik, öznitelik adı ve açıklamaları, öznitelik tipleri ve tanımlayıcı bilgileri	64
Çizelge 2.2. Çok sınıflı karışıklık matrisi	70
Çizelge 3.1. Üç sınıflı kazanç veri seti açıklayıcı öznitelik dağılımları	78
Çizelge 3.2. Üç sınıflı %70 eğitim veri seti hiper parametre değerleri	80
Çizelge 3.3. Üç sınıflı %70 eğitim veri seti performans metrikleri	80
Çizelge 3.4. Üç sınıflı %30 test veri seti performans metrikleri	81
Çizelge 3.5. GAM karışıklık matrisi	82
Çizelge 3.6. %30 test veri seti GAM tüm sınıflara ait performans metrikleri	82
Çizelge 3.7. Üç sınıflı %75 eğitim veri seti hiper parametre değerleri	84
Çizelge 3.8. Üç sınıflı %75 eğitim veri seti performans metrikleri	85
Çizelge 3.9. Üç sınıflı %25 test veri seti performans metrikleri	85
Çizelge 3.10. GAM karışıklık matrisi	86
Çizelge 3.11. %25 test veri seti GAM tüm sınıflara ait performans metrikleri	86
Çizelge 3.12. Üç sınıflı %80 eğitim veri seti hiper parametre değerleri	88
Çizelge 3.13. Üç sınıflı %80 eğitim veri seti performans metrikleri	89
Çizelge 3.14. Üç sınıflı %20 test veri seti performans metrikleri	89
Çizelge 3.15. GAM karışıklık matrisi	90
Çizelge 3.16. %20 test veri seti GAM tüm sınıflara ait performans metrikleri	90
Çizelge 3.17. RAÖ’lü %70 eğitim veri seti hiper parametre değerleri	93
Çizelge 3.18. RAÖ’lü %70 eğitim veri seti performans metrikleri	94
Çizelge 3.19. RAÖ’lü %30 test veri seti performans metrikleri	94
Çizelge 3.20. GAM karışıklık matrisi	95
Çizelge 3.21. RAÖ’lü %30 test veri seti GAM performans metrikleri	95
Çizelge 3.22. RAÖ’lü %75 eğitim veri seti hiper parametre değerleri	97
Çizelge 3.23. RAÖ’lü %75 eğitim veri seti performans metrikleri	98
Çizelge 3.24. RAÖ’lü %25 test veri seti performans metrikleri	98
Çizelge 3.25. GAM karışıklık matrisi	99
Çizelge 3.26. RO karışıklık matrisi	99
Çizelge 3.27. RAÖ’lü %25 test veri seti GAM ve RO performans metrikleri	100
Çizelge 3.28. RAÖ’lü %80 eğitim veri seti hiper parametre değerleri	102
Çizelge 3.29. RAÖ’lü %80 eğitim veri seti performans metrikleri	103
Çizelge 3.30. RAÖ’lü %20 test veri seti performans metrikleri	103
Çizelge 3.31. RO karışıklık matrisi	104
Çizelge 3.32. GAM karışıklık matrisi	104
Çizelge 3.33. RAÖ’lü %20 test veri seti GAM ve RO performans metrikleri	105
Çizelge 3.34. İki sınıflı kazanç veri seti açıklayıcı öznitelik dağılımları	107

Çizelge 3.35. İki sınıflı %70 eğitim veri seti hiper parametre değerleri	108
Çizelge 3.36. İki sınıflı %70 eğitim veri seti performans metrikleri	109
Çizelge 3.37. İki sınıflı %30 test veri seti performans metrikleri	109
Çizelge 3.38. GAM karışıklık matrisi	110
Çizelge 3.39. İki sınıflı %75 eğitim veri seti hiper parametre değerleri	112
Çizelge 3.40. İki sınıflı %75 eğitim veri seti performans metrikleri	112
Çizelge 3.41. İki sınıflı %25 test veri seti performans metrikleri	113
Çizelge 3.42. GAM karışıklık matrisi	114
Çizelge 3.43. İki sınıflı %80 eğitim veri seti hiper parametre değerleri	115
Çizelge 3.44. İki sınıflı %80 eğitim veri seti performans metrikleri	116
Çizelge 3.45. İki sınıflı %20 test veri seti performans metrikleri	117
Çizelge 3.46. RO karışıklık matrisi	117
Çizelge 3.47. %30 test veri setleri tüm performans metrikleri	119
Çizelge 3.48. %25 test veri setleri tüm performans metrikleri	120
Çizelge 3.49. %20 test veri setleri tüm performans metrikleri	121

1. GİRİŞ

Yerleşik veya göçebe Türk topluluklarının hepsinde sosyal, ekonomik, askeri ve dini yönden önemli yere sahip olan at ve atlı yaşam kültürü Türk medeniyetinin oluşumunda çok büyük rol oynamıştır (Kafesoğlu, 1991).

Dünyadaki en değerli at ırklarından biri olan İngiliz atı 300 yıldan daha uzun bir süre önce üç temel aygır ile başlamış diğer birçok at ırkının gelişimini sağlamıştır. Türkiye'deki varlığı yüz yıllık geçmişe dayanan safkan İngiliz atları Cumhuriyetin kurulmasından kısa süre önce düzenlenen at yarışlarında boy göstermiştir. Cumhuriyetin kurulmasından sonra ülkenin ihtiyaçları doğrultusunda Karacabey harası atların titizlikle yetiştirilmesi için tahsis edilmiş ve ıslah faaliyetlerine büyük önem verilmiştir (Paksoy, 2016).

At yarışı endüstrisi ve yarış atı yetiştiriciliği birbirini tamamlayan ve destekleyen iki bileşendir. Türkiye'de, uluslararası standartlara uygun ve daha modern bir yapıda at yarışlarının devam ettirilebilmesi ve standardın daha da geliştirilmesi için at yetiştiricilerinin sektöre daha iyi atlar kazandırması gerekir. Bu endüstri, daha fazla gelişip büyüyerek ülke ekonomisine farklı türde iş imkanları ve dolayısıyla daha fazla istihdam sağlayabilir.

At yarışı endüstrisinin günümüzde yarattığı ekonomik kazanç ve istihdam çok büyük boyutlardadır. Bu endüstrinin doğrudan, dolaylı, tetiklenen ve toplam ekonomik etkileri oldukça önemli ölçeklerdedir. Endüstrinin sürdürülebilirliği ve büyümesi, yarış atı yetiştiriciliğinin gelişmesine ve iyi safkanların yetiştirilmesine bağlıdır.

Yarış atı endüstrisinde yetiştirilecek ırkın tespiti ve damızlık seçimi, hayvan yetiştiriciliğinde başarıyı yakalayabilmek ve kazanç değeri yüksek hayvanlar elde edebilmek için büyük önem arz eder. Düz koşu müsabakalarında yer alan safkan İngiliz atı yetiştiriciliğinde başarı, kazanç, hız ve başarı sırası gibi derecelerle ölçülen yarış performansı olup, cinsiyet, yarışa başlama yaşı ve yarış hayatı boyunca koştuğu pist tipi gibi birçok çevresel faktörler ile doğrudan ilişkilidir (Paksoy, 2016).

Yoğun emek isteyen bir iş olan iyi bir safkan yarış atı yetiştiriciliğinde harcanan para, zaman ve emek tasarrufu açısından bir yarış atının kazanç performansına etki eden öngörülemeyen risk faktörlerinin belirlenmesi ekonomik kayıpları önlemek açısından oldukça önemlidir.

Düşük kalıtsallığa sahip ancak değeri yüksek olan bu çevresel özelliklerin yarış performansının iyileştirmesinde önemli faydalar sağlayabileceği düşünüldüğü için bu alanda yapılan araştırmalar önem kazanmıştır.

Bu çalışma, safkan İngiliz yarış atlarının yarış hayatları boyunca kazanç performanslarına etki eden çevresel faktörlerin büyük veri seti ile makine öğrenmesi (MÖ) (Machine Learning, ML) sınıflandırma algoritmaları kullanılarak modellenilebileceğini göstermek amacıyla tasarlanmıştır. Bu bağlamda yapılan tez çalışmasının kapsamını, yarış hayatı boyunca elde edilen kazanç durumuna etki eden yedi farklı çevresel faktöre ait bilgiler oluşturmaktadır. Çalışmada ilk olarak, Türk toplumlarında at yetiştiriciliği ve kısa tarihçesi, safkan İngiliz atı yetiştiriciliği ve kısa tarihçesi, safkan İngiliz atının özellikleri, atlı sporlar ve at yarışları, Türkiye'deki at yarışları ve İngiliz atlarının yarışlardaki yeri, at yarışlarının ekonomik boyutu ve yarış atlarının performansını etkileyen faktörlere ait bilgiler verilmiştir. Bir sonraki bölümde MÖ kavramları olan istatistiksel öğrenme, makine öğrenmesi, denetimli öğrenme ve büyük veri açıklanmış, çalışmada kullanılan K-En Yakın Komşu Algoritması (k-EYK) (K-Nearest Neighbor, KNN) ve ağaç tabanlı algoritmalar olan C5.0, Rastgele Orman (RO) (Random Forest, RF) ve Gradyan Artırma Makinesi (GAM) (Gradient Boosting Machine, GBM) hakkında bilgiler verilmiştir.

1.1. Türk Toplumlarında At Yetiştiriciliği ve Kısa Tarihçesi

Elli milyon yıl öncesinde vahşi bir tabiata sahip atın, evrim geçirerek evcilleştirilme süreci ve yarışabilecek yapıya kavuşması uzun yıllar almıştır. At-insan etkileşimi her kültürde olduğu gibi Türk kültüründe de tarih boyunca değişimler geçirerek günümüze kadar gelmiştir (Köstem, 2018a).

Atın M.Ö. 4000'li yıllarda eski Türk boyları tarafından Orta Asya bozkırlarında evcilleştirildiği belirtilmiştir (Yılmaz, 2012). Evcilleştirilip insanlığın hizmetine sokulması ile at yarışlarına kadar uzanan tarihsel süreç başlamıştır. Bugünkü at ırklarının kökeninin M.Ö. 3000'li yıllarda yaşayan *Equus przewalski* (Moğol yabani atı), *Equus tarpan* (Doğu Avrupa yabani atı) ve *Equus robustus* (Orman yabani atı) isimli üç yabani ata dayanmakta olduğu bildirilmiştir (Akçapınar ve Özbeyaz, 1999).

At yetiştiriciliği, dünyadaki bütün milletlerin kültüründe yer alır. Türk toplumlarında ise at ile etkileşimin boyutları ve hayatın içindeki yeri diğer toplumlardan farklıdır. Türk

toplumları atların kendine mahsus dilini en iyi anlayan ve onlar ile iletişim kurabilen nadir milletlerdendir (Endicott, 2018). Bu yüzden at yetiştiriciliği ve kullanımı Türklerde çok daha özel bir yere sahip olmuştur.

Dünya tarihi incelendiğinde, Türkler atı evcilleştirip diğer toplumlardan farklı olarak ondan çok yönlü olarak istifade etmişlerdir. Atlı bozkır kavimleri kavramı dünya tarihine Türkler tarafından hediye edilmiştir (Taşağıl, 2018).

At, diğer yerleşik veya göçebe toplulukların gündelik hayatında Türklerde olduğu kadar önemli tesirlerde bulunmamıştır. Türk toplumlarının sosyal ve dini hayatlarında atın etkisi diğer milletlerden oldukça farklılık göstermiştir. Türk topluluklarının ayrılmaz bir parçası haline gelmiş, gücünden ve süratinden, et, süt, kıl ve derisinden faydanılmıştır (Kafesoğlu, 1991).

At Türk kültür tarihinin her alanında kendine yer bulmuştur. Sosyal hayatta ve günlük yaşamda binek hayvanı olarak kullanılan at, hayvan sürülerinin güdülmesi ve süre avlarında en önemli vasıta olmuştur (Taşağıl, 2018). Asya bozkırlarındaki zorlu iklim şartlarına uyum sağlayan Türkler hayatlarını sürdürmek için binlerce baş hayvanın güdülmesi, hayvanların en uygun iklim koşullarına sahip otlaklara zamanında sevk ve su başlarına süratle yetiştirilmesi için hızlı ve dayanıklı vasıtalar olan atlara ihtiyaç duymuşlardır (Kafesoğlu, 1991). Bunun yanında, eski Türk uygarlıklarında at, şeref ve şan kaynağı olup at sürüleri zenginlik belirtisi olarak görülmüştür (Sivrikaya, 2017).

Eski Türklerin atı kutsallaştırması ve gökten indiğine inanması dini inançlarının içinde yer bulmuştur. Gök Tanrıya sunulmak için en değerli kurban ve ataların ruhları için ise en değerli hediye olarak görülmüştür. Atlar matem alameti olarak bazen törenle sahibinin yanına ya da tekbaşına ayrı mezarlara gömülmüştür. Binicisi savaşta ölen atlarında mezarına konulmak üzere bazen kuyrukları kesilmiştir (Kafesoğlu, 1991). Türk töresi içinde ölülerin, atları ile birlikte toprağa verilmesi ve atların kurban olarak sunulması yer almıştır. Onuncu ve onbirinci yüzyıllarda Türk toplumlarında İslamiyetin yayılması ile insanların ata atfettikleri kudsiyet yani dini inanç sistemindeki yeri azalmış, fakat sosyo-kültürel hayattaki önemi devam etmiştir (Köstem, 2000).

Türk halk destanı, masal, şiir ve şarkılarında çok önemli rol üstlenen at Türk kültürü ile gelenek ve göreneklerinde kendine önemli bir yer edinmiştir. At için kaside yazıp övgüler

düzen ilk ve tek millet Türklerdir (Kafesoğlu, 1991). Türklerin süvari birliklerini kurması ile at ve atçılığın önemi daha da artmış, at hastalığı ve tedavisi konularını içeren baytarname ile biniciliğe, okçuluğa ve silahşörlüğe ait konuları içeren “*fürûsiyye*”, “*haylnâme*”, “*feresnâme*” gibi adlar altında eserler yazılmıştır (Eliacık, 2018). Diğer yandan, Türkler tarih boyunca şarkılarına eşlik edecek çalgı olan kopuzu atın kuyruk kılından imal etmişlerdir. Atı cinsine, cinsiyetine, yaşına ve rengi vb. özelliklerine göre farklı adlar ile adlandırmışlar çok değişik ve bol sayıda deyimler üreterek Türk dilindeki kelime zenginliğine katkıda bulunmuşlardır.

Eski Türk sportif kültüründe at yarışları, at güreşleri, cirit ve diğer atlı sporların yeri büyüktür. Eski Türklerde ilkbahar ile güz dönemlerinde bozkırlarda bayram eğlencesi olarak düzenlenen atlı top oyunları, çevgan ve küre oyunları halkın eğlence kültüründe önemli bir yere sahiptir (Kafesoğlu, 1991; Taşağıl, 2018).

Atlı sporların farklı milletlerin eğlence kültürlerine yaptığı katkı çok önemlidir. Her kültürde atın kullanım şekli değişik formlarda görülmektedir. Çevgan Orta Asya’nın güneyinde oynanan atlı takım oyunlarından biri iken atlı oyun türü olan buzkaşı yine Orta Asya’nın farklı ülkelerinde sevilerek oynanmaktadır (Koçkar, 2018).

Türkler tarih boyunca atı bir beslenme kaynağı olarak da görmüşlerdir. At etini yemiş ve bu eti uzun süre saklayabilmek için farklı teknikler geliştirmişlerdir. Sütünden kımız yaparak eğlence aracı olarak kullanmışlardır.

Atın evcilleştirilmesi ve sonrasında binek hayvanı olarak kullanılması ile sürat kavramı insanlık için farklı şekilde algılanmıştır. Uzak mesafelere kısa sürelerde ulaşmaya başlanmış ve böylece birbirinden habersiz yaşayan insan topluluklarının ilişki kurmaları kolaylaşmıştır. Bu sayede, kültür alışverişleri yaygınlaşarak insanların dünyaya bakışı değişmiştir. Sınırlar ve devlet yönetim anlayışları farklılaşmış ve sonuç olarak medeniyetler gelişmiştir (Kafesoğlu, 1991).

Atlı yaşam kültürü Türkleri tarih sahnesinde önemli işler başaran millet haline getirmiş, geniş alanlara yayılan Türk topluluklarının büyük devletler kurmalarına sebep olmuştur. Türk medeniyetinin ortaya çıkışında at ile olan etkileşimden doğan kültürün rolü çok büyüktür (Taşağıl, 2018).

Göçler ve savaşlar göçebe Türk toplumlarının atlardan büyük ölçüde yararlanmasının en önemli sebeplerindendir. Hareket ve vurucu güce sahip atlar savaş meydanlarında binicisi kadar cengaverlik göstermişlerdir. Türk atının asıl üstünlüğünü sürati ve dayanıklılığı ile atın emniyet içinde kullanımını kolaylaştıran teçhizata sahip olması olduğunu ifade eden hikayeler bulunmaktadır (Kesik, 2018). Düzenlenen süreklilik avları bir çeşit askeri manevra taktiği gibi yürütülmüş ava katılan binlerce kişinin vazgeçilmez vasıtası olmuştur. M.Ö. 2000 ortalarında ilk defa hafif teçhizatlı süvari birlikleri Türk ordusunda kurulmuştur. Türkler, kendilerine özgü savaş metotlarına uygun bir şekilde farklı tipte savaş atları ve silahlara sahiptiler. Diğer yandan, ok imalatında atın kemiğini ok ucu olarak kullanmışlardır (Özen, 2014).

Eski Türk uygarlıklarında at etrafında şekillenen üst düzey bir ekonomik faaliyet alanı ortaya çıkmıştır. At koşumlarının güzelliğini ve zenginliğini artırmak için farklı meslekler veya zanaatlar doğmuştur. Atın binicilikte, seyahatte, avda ve savaşlarda kullanılabilmesi için farklı şekillerde donatılması ya da giydirilmesi birçok mesleğin ve farklı zanaatların doğmasına at çevresinde ticari faaliyetlerin belirmesine neden olmuştur. Eğercilik, semercilik, nalbantçılık, üzengicilik, yular takımı ustacılığı, raht isimli atlar için kullanılan meşin takım ustacılığı gibi çok farklı mesleklerin toplum hayatına dahil olmasını sağlamıştır. Çok değerli ihraç malı olarak canlı halde Çin'e ve diğer devletlere satılan atlar, eski Türk devletlerine büyük iktisadi destek sağlamıştır. Türkler tarih boyunca atın derisini işlemiş ve ticaretini yapmışlardır (Özen, 2014).

Hunlar ileri düzeyde at yetiştiriciliği yapan bir Türk kavmi olup uzun süren savaşlarda çok dayanıklı olan, seri hareket edebilen ve karmaşık hareketlere çok hızlı uyum sağlayabilen bozkır atlarını kendilerine özgü bir eğitim sistemi ile yetiştirmişlerdir (Taşağıl, 2018).

Eski dönemlerden itibaren köklü ve sürekli bir at kültürüne sahip oldukları bilinen Türklerin ata verdiği değer Selçuklu devrine ait tarihi kaynaklarda görülmektedir. Selçuklu sultanları ata karşı büyük ilgi duymuş sultanlara ait haralarda 10.000 kadar atın beslendiği bildirilmiştir. En iyi cins atların sultanlara hediye olarak sunulduğu kaydedilmiştir (Halaçoğlu, 1991).

Selçuklu Türk medeniyetinin temelinde atçılıkta çağdaşı olan diğer milletlere üstünlükleri ile ata Orta Asya'daki Türk devletlerinde olduğu gibi önem vermeleri yatmaktadır (Kesik, 2018).

At kültüründeki süreklilik, Türk at tipinde de kendini göstermiş “Türkmen atı” adıyla geçen Türk at tipi Selçuklu döneminde de korunmuş ve yaygınlaştırılmıştır. Türk atlarının en temel özellikleri, başlarının biçimli ve küçük olması ile boylarının orta ve kısa seviyelerde olmasıdır. Kulaklar dikkat çekecek kadar küçük, gözler ise son derece canlı ve tesir edicidir. Ağız yumuşak, göğüs ve sağırlar kuvvetli, yeleler ise oldukça uzundur. Bununla beraber, dizgin ve eyer takımları kendine has olup, üzengi ile göğüs ve kuyruk altından geçen kolon sistemi bulunmaktadır (Kesik, 2018).

Selçuklular Orta Asya’dan Türk atı ve atçılığına ait kültürü Orta ve Yakındoğu’ya taşımışlardır (Kesik, 2018). Selçuklu Türklerine Anadolu’nun kapısını açan savaş meydanlarında sevki kolay, hızlı ve çevik atların oluşturduğu süvari birlikleridir.

Selçukluların dağılmasıyla birlikte kurulan Anadolu beylikleri eskiden beri devam eden at kültürü ve at yetiştirme geleneğini sürdürmüşlerdir. Karamanoğulları ve Germiyanogulları en değerli atların yetiştirildiği topraklara sahipti. Karamanoğulları ve Germiyanogulları Osmanlı devletine katılınca yetiştiricilik geleneği de Osmanlılara geçmiştir (Halaçoğlu, 1991; Tan, 2015).

Atlara özel bir önem veren Osmanlı padişahları birçok çiftlikte at yetiştirilmesini sağlamışlardır. Osman Gazi ve Orhan Gazi dönemlerinde kurulan ‘Akıncılar’ ve ‘Sipahi’ süvarileri Osmanlı devletinde ilk atlı kuruluşun temellerini oluşturmuştur (Köstem, 2000).

Osmanlılar savaşlarda kullanımın yanı sıra günümüzdeki motorlu taşıtlardan önce kara üzerinde en hızlı hareket eden binek hayvanı olan atı resmi ulaşım ve haberleşme aracı olarak kullanmışlardır. Osmanlı devletinde, çağlar boyunca devlet hazinesine ait değerli eşyalar at ile hızlı ve güvenli bir şekilde taşınmış, resmi görevlilerin ulaşımı sağlanmış ve idareciler tarafından gönderilen değerli evrakların iletimi mümkün olmuştur. Lütfi Paşa sadrazam (1539-1541) olana kadar resmi ulaşım ve haberleşme ihtiyacı için binekler halktan temin edilmiştir. Resmi hizmetlerde kullanılan beygirlerin beslenmesine yönelik düzenleme ile halkın vergileri karşılığında menzil beygiri beslemesine 1691 yılına kadar devam edilmiştir. 1696 yılında yapılan değişiklik ile resmi ulaşım ve haberleşme kurumsal bir yapıya kavuşmuştur. Resmi ulaşım ve haberleşme “*menzilhâne*ler” vasıtasıyla 1840 yılında posta idaresi kurulana kadar devam etmiştir (Çetin, 2018).

XIV. ve XV. yüzyıllarda Osmanlı Devleti'nde “*hayvânat* ocakları” yani “*Yunt* ocakları” kurulmuş, atçılığın ileri düzeye ulaşabilmesi ve at soyunun gelişimi için uğraşmıştır. Karacabey, Edirne, Eskişehir, İnönü, Dimetoke, Akköprü, Selanik ve Filibe bölgelerinin çevresinde 19 ocak bulunmaktaydı. Bu ocaklar günümüzdeki haraların benzeri görev yapan kuruluşlardı. Devletin yükselme döneminde ihtiyaca uygun şekilde dayanıklı, atik, çevik ve güçlü atlar yetiştirilmiştir. Arap ve Acem atları yerli at ırkının ıslahı için kullanılmıştır. Devletin parlak dönemlerinde kusursuz görev yapan bu kuruluşlar devletin son zamanlarında bozulmuş ve işlevlerini yitirerek 1760 yılında kapatılmıştır (Köstem, 2000).

Bilimsel yöntemlerin eksikliği ve akılcı programların uygulanmaması yüzünden çağın gerisinde kalan padişah çiftliklerinden beklenen verim alınamamıştır. Atçılığın mevcut vaziyetini ıslah etmek için hayvanat ocakları haraya dönüştürülmüştür. Ordudaki hayvan yetersizliğine çözüm olması için 1887 yılında alınan karar ile Çifteler, Sultansuyu, Çukurova ve Elhamra çiftlikleri haraya dönüştürülmüştür (Çetinkaya, 2018).

At yetiştiriciliği faaliyetleri haraların kapatıldığı 1908 yılına kadar devam etmiştir. 1909-1913 yılları arasında ülke genelinde atların ıslahı için 12 aygır deposu açılmıştır. Orduların at ihtiyacını karşılamak için ise II. Meşrutiyet sonrası *Remont* müfettişlikleri ile *Remont* depoları kurulmuştur (Çetinkaya, 2018; Köstem, 2000).

Osmanlı ordusu 1850'li yıllara kadar at ihtiyacını çoğunlukla halkın yetiştirdiği hayvanlardan karşılamıştır. Maddi olanağa sahip olan hemen her Türk köylüsü orduya hep iyi yetiştirilmiş ve cüsseli hayvanlar göndererek hem onur kazanmış hem de ekonomik yönden önemli ve sürekli bir kazanç elde etmiştir. Devlet kurumları tarafından yetiştirilen atların yanında yerli üreticiden temin edilen hayvanlarda Osmanlı ordusu için asırlarca önemli bir kaynak olmuştur. Bunun sayesinde XIX. yüzyılın ikinci yarısına kadar Osmanlı ordusunda nitelikli süvari ve nakliye hayvanının dışarıdan teminine ihtiyaç kalmamıştır (Gedik ve Kolay, 2023).

Kırım Savaşı (1853-1856) sonrasında Türk atçılık geleneği ilk büyük darbeyi almış orduya hayvan sağlayan halk yetiştiriciliği büyük bir gerileme yaşamıştır. Savaş bitiminde, müttefiklerden (Fransa, İngiltere ve İtalya) satın alınan atlar ile uzun bir süre ordunun hayvan ihtiyacı karşılanmıştır. Bunun sonucunda, yerli yetiştiricilerden hayvan satın alınmadığı için yetiştiricilik gerilemiştir. Bu durum, Osmanlı Devleti'nin yıkılışına kadar çözümlenemeyen, orduya nitelikli hayvan temin edememe sorununu doğurmuştur. Halk yetiştiriciliğinin sekteye

uğratılmasının olumsuz etkisi 1877-1878 Osmanlı-Rus Savaşı'nda kendisini hissettirmeye başlamış Osmanlı hükümeti cephelerde mücadele eden süvari alaylarına hem süvari hayvanı hem de nakliye hayvanı sağlamada yetersiz kalmıştır (Gedik ve Kolay, 2023).

Balkan Savaşı ve I. Dünya Savaşı gibi olumsuz koşullar ile haraların faaliyetlerine son verilmesi at varlığını giderek azaltmış ve atçılığı olumsuz etkilemiştir (Özen, 2014). Cumhuriyetin kurulması ile at yetiştiriciliği ve ıslahı konusunda yasal düzenleme yapılmış *Islah-ı Hayvanat* Kanunu kabul edilmiştir. Haraların tekrar faaliyete geçirilmesi ile yetiştirme ve ıslah faaliyetlerine yön verilmiştir (Altuntaş, 2010).

1924 yılında Çifteler ve Uzunyayla aygır depoları açılmış ve Karacabey harası faaliyete geçmiştir. 1926 yılında Yarış ve Islah Encümeni'nin oluşturulması ile özel haralar kurulmaya başlamış ve at yetiştiriciliğine olan ilgi artmıştır (Özen, 2014).

Sultansuyu (1928), Çukurova (1931), Çifteler (1934), Konya (1934) Karaköy (1949) ve Altındere (1950) haraları sırasıyla faaliyete geçerek ülkenin ihtiyaçları doğrultusunda farklı verim yönlerinden farklı at tipleri üzerinde çalışmalar hızlanmıştır (Özen, 2014).

İngiliz koşu atı yetiştiriciliği özel yetiştiriciler tarafından, safkan Arap atı yetiştiriciliği ise Tarım İşletmeleri Genel Müdürlüğüne (TİGEM) ait üç işletme ve özel yetiştiriciler tarafından yapılmaktadır (Altuntaş, 2010).

1.2. Safkan İngiliz (Thoroughbred) Atı Yetiştiriciliği ve Kısa Tarihçesi

Safkan at yarışı, genellikle bahis amaçlı olarak dünya çapında yapılan en yaygın at yarışı şeklidir. İngiliz atları kısa mesafelerde en iyi yarışçılardır. Öncelikle orta mesafelerde hız için geliştirilmiş koşu atlarıdır fakat avcılık, polo ve tilki avı ile günümüzde Uluslararası Binicilik Federasyonunun resmi olarak kabul ettiği olimpik binicilik branşları olan engel atlama, at terbiyesi ve üç günlük yarışma vb. için de kullanılırlar. Düz yarış ve atlama yarışları yani engelli koşular at yarışlarının iki türüdür (Thiruvankadan vd., 2009).

Safkan atlar yüzyıllardır yarış atı olarak kullanılmakta olup hız ve dayanıklılık için seçici yetiştirme ile bugünkü bildiğimiz safkan İngilizleri geliştirilmiştir (Cochran, 2023).

Safkan İngiliz atları gibi asil bir cins ortaya çıktığı zaman genellikle ortaya çıkan bu cinsin atası ile ilgili kan hattı tartışmalara konu olmuştur. İngiliz atının geliştirilmesinde kullanılan üç atası *Byerley Turk*, *Darley Arabian* ve *Godolphin Arabian* adlı aygırların Osmanlı Devleti'nden İngiltere'ye götürülen atlar olduğu belirtilmektedir.

Darley Arabian 1704 yılında İngiltere'ye Halep'teki at piyasasından satın alınarak getirilmiştir. Bu atın, birçok Türkmen atın çaprazlaması ile oluşan kanı taşıyan *Maneghi* cinsi yarış atının kanını taşıdığı belirtilmiştir. *Darley Arabian* ile sadece Yorkshire bölgesindeki kısıraklar çiftleştirilmiştir. İlk safkan İngiliz yarış atı olan *Flying Childers*, *Darley Arabian* ile *Betty Leedes* çiftinin tayıdır. İngiliz atının gelişiminde çok büyük rol oynamıştır. *Bartlett's Childers*, *Betty Leedes*'in kardeşi olup kan damarlarında çatlama riski yüzünden yarışmamış fakat damızlık olarak kullanılarak bütün zamanların en iyi İngiliz atı olan *Eclipse*'in atası olmuştur (Bonnie, 1995).

Birçok tarihçi tarafından muhtemelen bir "Ahal-teke" olduğuna inanılan *Byerley Turk* İngiltere'ye 1689'da getirilmiştir. Yüzbaşı Byerley tarafından İngiltere'ye getirilen atın Türklerden Viyana kuşatmasında ganimet olarak alındığı aktarılmıştır. Az sayıda soylu kısırak ile çiftleştirilerek XX. yüzyılın önemli aygırlardan birisi olan ve yarış parkurunda hiç yenilgi yüzü görmeyen *Tetrach* elde edilmiştir. *Jigg*, *Tartar* ve *Herod* bu soyun en bilinen atlarıdır (Bonnie, 1995).

Lord Godolphin tarafından 1730'lu yıllarda Fas ya da Tunus'tan damızlık olarak getirilen *Godolphin Barb (Arabian)* bir Mağrip atıydı. Bu atın İngiliz saf kanı üzerindeki en büyük etkisi torunu *Matchem* aracılığıyla hissedilmiştir (Bonnie, 1995).

James Weatherby tarafından 1791 yılında yayınlanan "Genel Soy" kitabı safkan atlar için ilk cins kaydı yapılan soy kitabıdır. Büyük Britanya'daki üç temel aygırdan köken alan yarış atlarının soylarını düzenli bir şekilde toplayan bir kitaptır (Cochran, 2023).

Son zamanlarda yapılan genetik araştırmalar, günümüzde yaşayan modern safkan İngiliz atların çoğunun kökeninin 18. ve 19. yüzyıllardan kalma 27 aygırdan geldiğini göstermiştir. Atların baba soylarının %95'inden fazlası, *Darley Arabian* isimli aygırın büyük-büyük-torunu olan *Eclipse* isimli aygıra dayanmaktadır (Cochran, 2023; Cunningham vd., 2001).

Özetle, safkan İngiliz yarış atı, doğu atlarının ve Britanya Adaları'na özgü atların melezlerinden ve ardından boyut, güç ve dayanıklılık için seleksiyondan türetilmiştir.

1.3. Safkan İngiliz (Thoroughbred) Atının Özellikleri

İngiliz atları, safkan Arap atları gibi sıcakkanlı atlardır. Safkan İngiliz atları zariflik, dayanıklılık ve hızın bir bedende vücut bulmuş halidir. Geniş aralıklı büyük ve canlı gözler ile narin ve zarif görünümlü bir kafaya sahiptir. Hareketli ve orantılı kulaklar, geniş burun delikler ile diğer ırklara göre biraz daha uzun, ince ve daha hafif eğimli boyun safkan İngiliz atlarının yüz ve boyun görünüşlerini en iyi tarif eden kelimelerdir. Derileri çok incedir (Bonnie, 1995; Sevim, 2022).

Safkan İngiliz atları genellikle hıza uygun, zayıf vücutlara sahiptir. Yetişkin bir atın ortalama cidago yüksekliği 165-170 cm olup ortalama yetişkin bir attan daha yüksektir. Ortalama 500-550 kg ağırlığında hafif bir at ırkı olarak nitelendirilir. Azami hızlarına ulaştıklarında adım boyu altı metreyi aşabilmektedir. İngiliz atları büyüme ve gelişmeyi iki yaşında tamamlayıp yarışa hazır hale gelir (Paksoy vd., 2018). At sahiplerinin atının yarış kariyerini doğru yönetebilmesi amacıyla bütün safkan atların resmi kayıtlardaki doğum tarihi 1 Ocak'tır. Bu durum atın gelişim süresi için duyulan zamanın en iyi şekilde kullanılabilmesi için yapılır.

Safkan atlardaki, temiz, belirgin ve sağlam tendonlara sahip uzun ve düzgün bacaklar sağlamlığı ve adım uzunluğunu artırır. Üst arka bacağın uzun, güçlü kasları, uzun uyluk kemiği ve güç için geniş kalça açısı çok hızlı koşmayı sağlar (Cochran, 2023). Düz arazide yere yumuşak basan ve ahenkle yürüyen atlar, kısa, eşit şekilde kavisli bir sırta yol açan yüksek ve belirgin omuzlara sahiptir. Derin, yüksek ve geniş göğüsler ile eğimli, iyi kaslı bir omuz, uzun ve kaslı ön kollar, oldukça ince incik kemikleri ve temiz eklemler ile hızlı koşma becerisine yardımcı olacak beden yapısına sahiptir Cidagolar çıkıntılı, sırtlar ise genelde uzundur. Eğimli belleri sağrılarına bağlıdır. Donları çoğunlukla doru, koyu doru, kestane, siyah veya gridir. Yüz ve bacaklarda sıklıkla beyaz işaretler görülür (Bonnie, 1995; Sevim, 2022).

Hız ve performans için yetiştirilen atlar genellikle cins olarak iyi bir mizaca sahiptir. Canlı, cesur ve yüksek enerjili kişiliklere sahip, çok atik ve sıcak kanlı hayvanlardır.

Genellikle zeki ve hassas hayvanlar oldukları için yeni başlayan biniciler için uygun bir binek hayvanı değildir (Cochran, 2023).

Safkanların yukarıda sayılan birçok iyi yanları ve yarış atı olarak avantajlarına rağmen dezavantajlarında bulunmaktadır. Normalden küçük kalpleri ve vücut büyüklüklerine oranla küçük toynakları vardır. Safkanlar çok genç yaşlarda eğitime başladığı için egzersize bağlı akciğer kanaması görülme sıklığı yüksektir. Sıklıkla kabızlık çekerler, düşük doğurganlık oranına sahiptir. Hızlı bir metabolizmaya sahip olduğu için yarışlarda enerji ihtiyacını karşılamak için diğer at ırklarına oranla daha çok tahıllı diyetlerle beslenirler. Koşularda üst düzeyde verimliliğe ulaşabilmek için çok daha fazla yakıtı ihtiyaç duyarlar. Oldukça enerjik oldukları için düzenli egzersiz gerekir. Çoğu dikkat çekecek derecede duyarlı ve hassastır. Safkan İngilizlerin akrabaları yetiştirmeden kaynaklanan sağlık sorunları kapalı soy kütüğü nedeniyle daha da kötüleşebilir. Literatürde yapılan bazı araştırmalar, hıza yönelik seçici yetiştirme nedeniyle safkanlarda kalp ve iskelet sorunları riskinin artırmış olabileceğini öne sürmektedir (Cochran, 2023).

Bazı Safkan İngilizlerin kalplerinin anormal derecede büyük olması, yarış sırasında kalp debisini artırır ve üstün performansa sebep olur. Bu durum “X Faktörü” olarak adlandırılmıştır (Cochran, 2023).

Yarış atlarının emekliye ayrılmasının en yaygın nedenleri arasında kas-iskelet sistemi yaralanmaları, solunum sorunları, kalp rahatsızlıkları ve davranış sorunları yer almaktadır. Yarış atları çok yoğun strese maruz kaldıkları için safkan yarış atlarının pek çoğunda mide ülserine rastlanmıştır. Safkan İngilizlerin ırk standardı öncelikle yarış yeteneğine dayanmaktadır. Atların üreme değerleri, belirli konformasyonel özelliklerine yani vücudun yapı, biçim ve orantısına göre değil pistte gösterdikleri performanslara göredir. Seçici yetiştirme sebebiyle atların sağlamlık, hız ve dayanıklılıkla ilişkilendirilen uyum ve mizaç özellikleri ön plana çıkmıştır (Cochran, 2023).

1.4. Atlı Sporlar ve At Yarışları

Atlı sporların farklı milletlerin eğlence kültürlerine yaptığı katkı çok önemlidir. Her kültürde atın kullanım şekli değişik formlarda görülmektedir. Atlı sporlardan olan cirit, oğlak oyunu ve *Çevgan* ile at yarışlarına Türkler eskiden beri önem vermiştir. Atlı takım oyunlarından olan *Çevgan* Orta Asya’nın güneyi ve Azerbaycan taraflarında, atlı bir oyun olan

Buzkaşı, Orta Asya ülkeleri ile Afganistan’da ve çiftlerin güç ve uyumlarını gözlemlemek için düğün merasimlerinde oynanan *Kyz kuu* (kız kovalama) ise Kazakistan ve Kırgızistan’da oynanan tarihi çok eskiye dayanan oyunlardır (Koçkar, 2018).

M.Ö. VI. yüzyıldan beri Pers ülkesinde oynanan dünyanın en hızlı atlı top sporu olan polo Asya’daki göçebe kabileler vasıtasıyla Hindistan’da da yaklaşık 2.000 yıldır oynanmaktadır. İngilizler 1800’lü yıllarda Hindistan’ı işgal ettiklerinde bu oyunu görmüş ve çok sevmişlerdir. İngiliz soyluları arasında yaygınlaşan bu oyun yeni kurallar eklenerek “polo” adıyla ünlenmiştir.

At yarışlarının tarihi çok eski zamanlara kadar uzanmaktadır. At yarışlarının başlangıç noktası olarak Türkler tarafından evcilleştirildiği kabul edilen zamana gidip, at yarışlarının kökenini de Türk kültüründe aramak gerekir. At yarışları, ilk olarak *Dîvânu Lugâti’t-Türk*’te anlatılan Manas destanında yer alan at yarışlarıyla ilgili bölümde geçmektedir. Savaşlarda kullanılan atlı arabalar çeşitli evrelerden geçerek günümüzdeki arabalı yarış kültürünü doğurmuştur. Eski Yunan’da iki ya da üç atla, Roma’da ise dört atla çekilen arabalar “hipodrom” isimli devasa binalarda yarıştırmış bu yarışlar günümüzde Fransa’da büyük ilgi görmektedir (Köstem, 2018a).

Düz yarışlar 1174 yılında İngiltere’de ilk kez düzenlenmiş, kraliyet ailesinin desteği ile yarışlar devam etmiş özellikle *Stuart* Hanedanlığından II. *Charles* (1660-1685) döneminde yarışlar ve ıslah çalışmaları sürdürülmüştür. 18. yüzyılda at yarışları popülerleşmiştir (Paksoy, 2016). Ondokuzuncu yüzyılın başlarında ise Amerika Birleşik Devletleri’nde popüler hale gelmiştir (Cochran, 2023).

Son dönemde, at yarışlarında “organize yarışlar” ön plana çıkmıştır. Bu yarışlarda kendi içinde, engelli ve düz yarışlar gibi birkaç alt gruba ayrılmıştır. Dünyanın bütün ülkeleri soy kütüklerinde (stud-book) kaydı bulunan atları saf kan (thoroughbred) olarak kabul etmekte ve belirlenmiş yarış türleri için belirlenen uluslararası kurallara uygun olarak bu yarışlar yapılmaktadır (Köstem, 2018a).

At yarışlarının düzenlenmesindeki asıl amaç soyun iyileştirilmesi ile birlikte yüksek performans gösteren atları ortaya çıkarmaktır. At yarışlarında oynanan bahisler ve elde edilen kazançlar, ata olan ilgiyi artırmak ve elde edilen gelirler ile at yetiştiriciliğini ve ıslahını teşvik etmek içindir. Günümüzde uluslararası boyutlara ulaşan at yetiştiriciliğinde dünyaya uyum

sağlamak ve dünya ile rekabete girecek derecede yüksek performans gösteren atlar yetiştirmek için at yetiştirme zevkini yaygınlaştırmak gerekir (Köstem, 2018a; Özen, 2014).

Öncelikle genç atlara odaklanan modern at yarışları yüzünden safkan İngilizlerin çoğu pistte yalnızca birkaç yıl geçirir. Başarılı yarış kariyerlerinin ardından, safkan İngilizlerin çoğu için binicilik dönemi başlarken bazı safkan aygır ve kısıraklar ise üreme için kullanılır. İngiliz terbiyesi ile eğitilen safkan İngiliz atları geçtiğimiz yüzyılda “Engel Atlama” ve “Üç Günlük Yarışma” gibi gösteri amaçlı Olimpiyat disiplinlerinde zirveye yükselmiştir (Cochran, 2023).

Türkiyede düzenlenen yarışlar incelendiğinde yarışlara Türkiye’de doğmuş safkan İngiliz atları ile diğer ülkelerden Türkiye’ye doğduğu yıl içinde getirilenler katılabilmektedir. İngiliz atları için düzenlenen uluslararası koşullara yerli ve yabancı atlar katılabilmektedir. Yabancı İngiliz atları Uluslararası Soykütüğü Kitabı’nda kayıtlı ise yarışlara katılabilirler. Gazi, Cumhurbaşkanlığı ve Ankara koşullarına yalnızca Türkiye’de doğmuş İngiliz atları katılabilmektedir (Ünlüoğlu, 2022).

1.5. Türkiye’deki At Yarışları ve İngiliz Atlarının Yarışlardaki Yeri

Türkiye topraklarında günümüz kriterlerine uygun ilk düzenli at yarışı Osmanlı döneminde 1856 yılında İzmir’de düzenlenen organize yarışlardır (Köstem, 2000). Yarışlar Şirinyer Çayırı’ndaki tek tribünlü pistte 1856 yılında levantenlerin kurduğu İzmir Yarış Kulübü (Smyrna Races Club) tarafından düzenlenmiştir. Spora düşkünlüğüyle bilinen Sultan Abdülaziz’in İzmir’deki yarışları çok beğenmesi üzerine yarışlar 16-17 Ekim 1863 tarihleri arasında İstanbul’da Kağıthane Çayırı’nda koşulmuştur (Köstem, 2018a). 1877-1878 Rus harbi ile at yarışları da kesintiye uğramıştır. Ancak harbin doğurduğu olumsuz ortama rağmen Bursa, Aydın, Adana, Balıkesir, Amasya, Konya, Sivas, Manisa, Samsun, Urfa, Erzurum, Bağdat, Musul, Halep ve Şam gibi şehirlerde at yarışlarına devam edilmiştir (Akson, 1971; Akyüz, 2012; Ünlüoğlu, 2022).

Batı’daki yarış otoritelerinin görevini yapmak amacıyla 1909 yılında kurulan Osmanlı Jokey Kulübü 4 ay gibi kısa bir sürede tarih sahnesinden silinmiştir. Padişahın himayesinde at yarışları düzenlemek, cins atlar yetiştirmek ve biniciliği geliştirmek için 11 Nisan 1912’de At Soyunu İyileştirme Derneği (*Islah-ı Nesl-i Feres Cemiyeti*) kurulmuştur (Kurnaz Şahin, 2022). 22 Mart 1913’te “Sipahi Ocağı Binicilik Kulübü” kurulmuş ve *Islah-ı Nesl-i Feres* derneğinin

yariř ile ilgili faaliyetleri bu kulübe verilmiřtir. Bu dönemde Kadıköy'de at yariřları düzenlenirken Veliefendi yariř yeri olarak seçilmiřtir (Köstem, 2018a).

Düzenlenen yariřlara I. Dünya Savařı'nın başlaması nedeni ile ara verilmiřtir. Kurtuluř Savařı sırasında Ankara'da at yariřlarının düzenlenmesi için Mustafa Kemal Atatürk tarafından yariř pisti olarak tahsis edilen tren istasyonunun yanındaki alanda 1.600 metrelik kořu pisti yapılıp yariřlar burada devam etmiřtir (Köstem, 2000; Köstem, 2018a; Paksoy, 2016).

Britanya'daki "Yariř Yeri İřletmecileri Birlięi" bařkanı Sir Loftus Bates ile 1921 yılında yapılan anlařma ile 1922 yılında kurulan *Makrikuey Racing Syndicate* (Bakırköy Yariř Sendikası) Veliefendi hipodrumundaki fiziksel kořulları iyileřtirmiřtir. İngiltere'deki kurallara uygun olarak Veliefendideki yariřlar yeniden başlamıřtır (Köstem, 2000). Haftalık kořu programında Bakırköy Yariř Sendikası'nın yurt dıřından getirdięi safkan İngiliz atları için de kořu konulmuřtur. Bu atlar satıřa çıkarılarak Türkiye'deki safkan İngiliz atçılıęının temelini oluřturmuřtur. Nitekim 1927 yılındaki ilk Gazi Kořusu'nu kazanan Neriman'ın atası satın alınan *Quarrelsome* isimli bir safkandır. Bu yariřlar sadece bir yıl sürmüş bir yıl aradan sonra Sipahi Ocaęı Binicilik Kulübünün düzenledięi kořular Veliefendi'de 1924 yılında tekrar başlamıřtır (Köstem, 2018a; Paksoy, 2016).

At yariřlarıyla ilgili ilk yasal düzenlemeler 1923 ve 1926 yıllarında Türkiye Büyük Millet Meclisi (TBMM) tarafından kabul edilen iki kanun ile yapılmıřtır. İlki "İřlah ve Teksiri Hayvanatı Ehliye ve Kořu Cemiyetlerine, Kořu Müsabaka ve Sergilere Dair Kanun" ikincisi ise "İřlahı Hayvanat Kanunu" dur. Yapılan yasal düzenlemeler ile "Yüksek Yariř ve İřlah Encümeni" 1927 yılında kurulmuřtur. Türkiye'deki at yariřlarına ait yıllık programın ilk kez düzenlenmesi ve yalnızca safkan İngiliz atlarına özel yariřların kořulması Encümen dönemi ile gelen önemli yeniliklerdendir (Köstem, 2018a). Yariřlar İstanbul, Ankara, Bursa, İzmir, Adana, Konya, Sivas ve Samsun illerinde düzenlenmiřtir. Safkan İngiliz atı kořuları Adana Konya ve Bursa'da yapılmıřtır (Özen, 2014; Paksoy, 2016; Sivrikaya, 2017).

Aynı yıl ilk kez sadece İngiliz atlarının katılabildięi 2.000 m mesafeli Gazi Kořusu düzenlenmiřtir. Türkiye'nin en önemli klasik kořularından biri haline gelen Gazi Kořusunda yalnızca Türkiye'de doğmuş İngiliz atlar 1932 yılından itibaren katılmaya başlamıřtır. Bu durum, İngiliz atı yetiřtiricilięinin Türkiye'de gelişimine büyük katkıda bulunmuřtur (Özen, 2014; Paksoy, 2016).

Yarış Atları Yetiştiricileri ve Sahipleri Derneği 1943 yılında kurulmuş İstanbul ve Ankara'daki yarışlar 1948'da dernek kapanana kadar düzenlenmiştir. 1950 yılında Jokey Kulübü kurulana kadar Muvakkat Yarış Komitesi yarışları düzenlemiştir. Bakanlar Kurulu Jokey Kulübü'nün ismini 1953 yılında Türkiye Jokey Kulübü (TJK) olarak değiştirmiştir (Özen, 2014; Paksoy, 2016).

6132 sayılı “At Yarışları Hakkında Kanun”un birinci maddesinde “Türkiye sınırları içerisinde at yarışları düzenlemeye, yurt içinde ve yurt dışında düzenlenen yarışlar üzerine yurt içinden ve yurt dışından müşterek bahis kabul etmeye Tarım ve Köyişleri Bakanlığı (TKB) yetkilidir.” ifadesi geçmektedir. Kanunun 5. maddesine göre TKB'ye tahsis edilen arazilerin at yarışı düzenleme yetkisi olan kurum veya özel kuruluşa bedelsiz olarak kullanılabileceği belirtilmiştir. Bu maddelerde belirtildiği gibi at yarışları düzenlemeye yetkili olan TKB 20 senelik periyodlar halinde at yarışı düzenleme yetkisini TJK'ya devretmektedir. Bakanlık kulübün her türlü işlemlerini her zaman teftiş etme yetkisine sahiptir. 2022 yılında Antalya hipodromunun açılması ile yarış düzenlenen hipodrom sayısı İstanbul, Ankara, İzmir, Bursa, Adana, Kocaeli, Elazığ, Diyarbakır ve Şanlıurfa ile birlikte 10'a yükselmiştir (Özen, 2014; Ünlüoğlu, 2022).

Resmi Gazete'nin 27910 sayılı ve 19.04.2011 tarihli “At Yarışları Yönetmeliği” at yarışlarını düzenlemek için çıkartılmıştır. Yönetmelik üç defa değiştirilmiştir (Ünlüoğlu, 2022). At yarışlarına ait lisans hakları 01.01.2018 tarihi itibarıyla 49 yıllığına Türkiye Varlık Fonu'na devredilmiştir.

Dünyadaki at yarışı organizasyonlarında safkan İngiliz atları en çok tercih edilen atlardır. Türkiye'de düzenlenen koşularda koşan safkan İngiliz ve Arap atı sayılarının ve bütün koşan atlar içinde İngiliz atı oranlarının yıllara göre dağılımı Çizelge 1.1.'de görülmektedir (Özen, 2014).

Çizelge 1.1. Yıllara göre Türkiye'de koşan ferdi at sayıları

Yıl	Koşan İngiliz atı Sayısı (ferdi)	Koşan Arap atı Sayısı (ferdi)	İngiliz Atı/Toplam Koşan At Sayısı
1975	276	426	%39
1980	337	367	%48
1990	579	550	%51
2000	1374	928	%60
2010	2686	2001	%57
2015	3313	2668	%55
2020	3532	2834	%55

Çizelge 1.1. Devam Yıllara göre Türkiye’de koşan ferdi at sayıları

2021	3734	3085	%55
2022	3789	3130	%55
2023	3829	3062	%56

1.6. Türkiye’de İngiliz Yarış Atı Yetiştiriciliği

Akson (1971)’a göre, Türkiye’deki ilk safkan İngiliz yarış atları Rusya’dan İstanbul’a iltica eden Arcady ve Dr. Seferof isimli iki atsever kardeş tarafından tespit edilmiştir.

1909 yılı *Epsom* derbisinin galibi *Minoru* ile *Aboyeur* ve yanlarında *Saltpeter* ve *Tesei* isimli safkan İngiliz aygırları İngiltere’den Rusya’ya getirilmiş, bu aygırlar birinci dünya savaşı sırasında farklı yollardan Rusya’dan İstanbul’a kaçırılmışlardır. Arcady ve Dr. Seferof kardeşler gazetelerden takip ettikleri at yarışı fotoğrafları sayesinde İstanbul’da bulunan bu dört aygırı tanımış ve dönemin at severlerine tanıtmıştır (Akson, 1971).

Bundan sonra safkan İngiliz atları, at yarışlarına dahil edilmiş ve yarışlardaki üstünlüğü Bakırköy Yarış Sendikası tarafından Veliefendi’de düzenlenen yarışlarda görülmüştür (Köstem, 2018a). Yarışlardaki başarısından dolayı safkan İngiliz atı yetiştiriciliği önem kazanmış ve bunun üzerine devlet, 1924 yılına kadar Sultan Çiftliği olarak kullanılan ve Cumhuriyet’in ilk harası olup at soyunun iyileştirilme ve çoğaltılması için kullanılan Karacabey Harasını 1928 yılında safkan İngiliz atı yetiştiriciliği kullanımına açmıştır.

Yetiştiricilik için kullanılmak üzere dışarıdan toplam 25 kısarak ithal edilmiş bunların 21’i Yarış ve Islah Encümeni dördü ise Tarım Bakanlığı aracılığı ile olmuştur. Sonradan üç aygır daha getirilerek İngiliz atı yetiştiriciliğine başlanmıştır (Anonim, 2018). Böylece safkan İngiliz atı yetiştiriciliğine dair alt yapı oluşturulmuş ve devletin denetimi sağlanmıştır.

İngiliz atı yarışçılığının başlaması ile koşu programındaki safkan İngiliz at koşuları ile ikramiye miktarı arttıkça İngiliz atı sahibi olmak isteyen sayısı hızla artmıştır. Devletin denetiminde Karacabey Harası’nda 1928’de başlatılan safkan İngiliz atı yetiştiriciliği başarılı olamamış sonradan 1938 ve 1972 yıllarında Tarım Bakanlığı tarafından iki kez daha İngiliz atı yetiştiriciliği başlatılmış fakat yine başarı sağlanamamıştır. Sonrasında TJK 1973 yılında İzmit’teki yaklaşık 715 dönümlük araziye kurulmuş İzmit Merkez Aşım İstasyonunda safkan İngiliz atı yetiştiriciliği yapmıştır. Günümüzde ise İngiliz atları özel sektör tarafından yetiştirilmektedir (Paksoy, 2016).

İngiliz atlarına ait soykütüğü 1950’den itibaren tutulmaya başlanmış ve TJK’nın düzenlediği yarışlara ancak soykütüğüne kayıtlı atlar katılabilmektedir (Arpacık, 1999).

1.7. At Yarışlarının Ekonomik Boyutu

At yarışı endüstrisi günümüzde yarattığı ekonomik kazanç ve istihdam ile birlikte oldukça önemli bir sanayi dalı olarak karşımıza çıkmaktadır. Bu endüstrinin doğrudan, dolaylı, tetiklenen ve toplam ekonomik etkileri oldukça önemli ölçeklerdedir. At sahipleri, yetiştiriciler, antrenörler, seyisler, jockeyler, bakım işçileri, çiftlik çalışanları ve arka plandaki yarış personeli gibi binlerce kişiye geçim kaynağı sağlarken istihdamı doğrudan etkilemektedir. Yarış atı yetiştiriciliği, tarım, tekstil ve sağlık sektörlerinden mal ve hizmet alımı yoluyla dolaylı yollardan istihdama yol açmaktadır (Ünlüoğlu, 2022).

Dünyanın birçok ülkesinde düz koşu (flat racing), engelli koşu (jump racing), arabalı koşu (harness racing), tırıs (trot) ve rahvan (ambling) gibi birçok farklı türde at yarışları düzenlenmektedir. Bu koşu çeşitlerinden dünyada en yaygın olan şekli oval ve düz bir pistte gerçekleşen düz koşulardır. Düz koşu Türkiye’de ve dünyanın çoğu ülkesinde en çok ilgi gören ve ilgi ile takip edilen en eski ve en yaygın sporlardan biridir. Düz koşu endüstrisi de diğer at yarışları türlerinde olduğu gibi farklı türde meslekler olan veteriner hekim, antrenör, eğitmen, seyis, jockey ve yetiştirici için önemli bir istihdam kaynağı oluşturmaktadır. Bu endüstrinin sağladığı vergiler ve kesintiler kamu bütçesine büyük kaynak sağlamaktadır (Ünlüoğlu, 2022).

Dünyadaki önemli endüstrilerden biri olan at yarışçılığının en önemli gelir kaynağı bahislerdir. Bunun yanı sıra bu sektörün finansmanının çoğunluğu bahislerden sağlanmaktadır. Bahisler, diğer şans oyunlarından ayrı olarak yalnızca şansa dayalı olmayıp çoğu itibariyle atın ve binicinin emek ve mücadelesi sonucunda elde edilen kazanç sonucunda şekillenmektedir. Bunun yanında bahisler, bu sporu daha heyecanlı hale getirmekte olup izleyicilerin ilgisini çekmek için kullanılmaktadır (Ünlüoğlu, 2022).

Düz koşular, hem en değerli safkanların yarıştırıldığı hem de diğer atlı sporlardan daha yüksek para ödülünün dağıtıldığı yarışlardır. 2019 yılı verileri, Türkiye’nin dünya genelinde at yarışı düzenleyen ülkeler arasında altıncı, bahis geliri toplamında ise ilk 10 ülke arasında yer aldığını göstermektedir (Ünlüoğlu, 2022).

2009-2019 yılları arasında dünya genelinde at yarışı toplam bahis gelirleri ve dağıtılan toplam para ödülü miktarlarını gösteren Uluslararası At Yarışı Otoriteleri Federasyonu (International Federation Horseracing Authority, IFHA) verileri incelendiğinde, toplam bahis gelirleri bu yıllar arasında 72 milyar avrodan (euro, €) 115 milyar avroya (euro, €) yükselmiştir. ABD, Avustralya ve Japonya en fazla yarış düzenleyen ülkeler olup, Japonya bahis geliri ve dağıtılan para ödülünde en yüksek ülke olurken, Hong Kong ise yarış başına ortalama para ödülü en yüksek ülke olmuştur (Ünlüoğlu, 2022).

Literatürdeki at yarışı endüstrisinin ekonomiye etkisini araştıran bazı çalışmalar incelenmiş Tuck (2017) tarafından yapılan çalışmada, bu endüstrisinin 2015 yılındaki Minnesota eyaleti istihdamına katkısı 5.590 kişi, ekonomisine ise toplam 409,2 milyon dolar (US\$) civarında olduğu belirtilmiştir (Ünlüoğlu, 2022).

Lambert (2022) at yarışı endüstrisinin Kentucky eyaleti için ekonomik etkisini araştırdığı çalışmada, doğrudan ve dolaylı yoldan sağladığı istihdamın yaklaşık olarak 7.500 kişi olduğunu belirtmiştir. Doğrudan ekonomik getirisi yaklaşık 500 milyon dolar (US\$) iken, dolaylı etkisi yaklaşık 223 milyon dolardır (Ünlüoğlu, 2022).

Devlet tarafından at yarışçılığının denetlenmesi ve mevzuatlarla desteklenmesi sayesinde yarış atı yetiştiriciliğinin değeri ve önemi artmıştır. Bunun sonucunda birçok insana gelir ve istihdam olanağı sağlamıştır. Devlet haralarında ve özel yetiştiriciler tarafından yetiştirilen ve yarış hayatına kazandırılan binlerce yarış atından iyi genotipe sahip ve yarış hayatında başarılı olanlar, damızlığa ayrılıp tohumları yüksek fiyatlardan satılabilmektedir. TİGEM sahibi olduğu şampiyon safkan Arap aygırlarının donmuş spermalarını ihraç etmektedir (Ünlüoğlu, 2022).

Türkiyedeki “At Yarışları Yönetmeliği” incelendiğinde gerçek veya tüzel kişi olsun atların sahiplerine ve yetiştiricilerine koşu şartları ve yıllık yarış programı genel hükümlerine göre dereceye girenlerin kazandığı ikramiyeye göre farklı oranlarda prim ödenmektedir. Koşularda birinci gelen at birincilik ikramiyesinin tümünü alırken iki, üç ve dördüncü sıraları alanlar ise ikramiyenin sırasıyla %40, %20 ve %10 kadarı bir ikramiyeye sahip olmaktadır. Yetiştiricilik primi yıllık olarak her koşunun şartına göre belirlenmekte olup kazanılan ikramiyelerin %10’u ile %25’i aralığındadır (Paksoy, 2016).

Türkiye’de yıl içinde 10 hipodromda düzenlenen 5.000’den fazla at yarışından elde edilen bahis geliri önemli bir endüstri oluşturmuştur. Atatürk’ün isteğiyle 1927 yılından beri her yıl düzenlenen Gazi Koşusu Türk yarışçılığının en önemli koşusu olan ve en yüksek ikramiyeye sahip koşusudur (Ünlüoğlu, 2022).

Türkiye’deki yarış sektöründe 2017 yılında 5.012.208.097 TL (KDV dahil) müşterek bahis satışı gerçekleşmiş, 1.441.141.859 TL kamuya aktarılmıştır. 2017 itibariyle soy kütüğüne kayıtlı 42 bin at olduğu görülmüştür. Toplam 384 bin kişi doğrudan ve dolaylı yoldan istihdam edilmiştir. Bir yılda 2,3 milyar TL ekonomiye katkıda bulunurken kamu bütçesine 1,2 milyar TL aktarılmıştır. Deloitte isimli uluslararası bağımsız denetleme kuruluşu tarafından yapılan araştırma ile bir atın üç kişiye doğrudan iş imkanı sağlarken toplamda 11 kişiye geçim kaynağı sağladığı belirtilmiştir (Köstem, 2018b).

TJK’nın 2023 yılı istatistikleri incelendiğinde; 3.829 adet safkan İngiliz atı 3.481 adet yarışta koşmuş at başına 231.930,35 TL ortalama kazanç elde edilmiş birincilik ikramiyesi olarak 507.800.550,00 TL ve 888.061.325,00 TL toplam ikramiye dağıtılmıştır. 3.062 adet Safkan Arap atı 2927 adet yarışta koşmuş at başına 224.610,16 TL ortalama kazanç elde edilmiş birincilik ikramiyesi olarak 393.240.400,00 TL ve 687.756.325,00 TL toplam ikramiye dağıtılmıştır (TJK, 2024).

At yarışı endüstrisi ve yarış atı yetiştiriciliği birbirini tamamlayan ve destekleyen iki bileşendir. Türkiye’de, uluslararası standartlara uygun ve daha modern bir yapıda at yarışlarının devam ettirilebilmesi ve standardın daha da geliştirilmesi ve at yetiştiricilerinin de sektöre daha iyi atlar kazandırması ile sektör gelişebilir. Bu endüstrinin daha fazla gelişip büyüyerek ülke ekonomisine farklı türde iş imkanları ve dolayısıyla daha fazla istihdam sağlaması amaçlanmaktadır (Ünlüoğlu, 2022).

1.8. Yarış Atlarının Performansını Etkileyen Faktörler

Kötü performans, bir atın efor harcayarak antrenmanlarda çalışırken yaptığı egzersizleri tamamlayamaması ya da herhangi bir hastalığı olmadığı, fiziki yapısı da uygun ve tüm idmanları tamamladığı halde yarışlarda istenilen başarıyı yakalayamama olarak tanımlanır. Klinik yaklaşım düşük ya da yetersiz yarış performansının bireysel yaralanma veya idman yetersizliğinden kaynaklandığını düşünmekte fakat çevresel faktörleri değerlendirememektedir (Özen, 2014; Sevim, 2022).

Günümüzde, genellikle düz yarışlar için yetiştirilen safkan İngiliz atlarının yarış performanslarının en önemli göstergesi hipodromlarda farklı pist ve mesafelerde koştuğu yarışlarda ortaya koyduğu koşu kabiliyeti olan hızlarıdır (m/sn). Bu performans ölçütünün yanında diğer performans ölçütleri şunlardır; yıllık ikramiye, her yarış için ortalama ikramiye, belli bir sürede kazanılan toplam ikramiye, yarış sonu sıralaması, toplam kazanılan yarış sayısı, yarışları tamamlama süresi ve elde edilen en iyi süre, ortalama handikap ağırlığı, en iyi handikap ağırlığı, sıralamadaki yeri, kazanan ata uzaklığı yer almaktadır. Bunlar içerisinde hız (m/sn) ve kazanç, en fazla yararlanılan performans ölçütüdür (Bailey vd., 2022; Paksoy vd., 2018).

Yarış performansı kantitatif bir özellik olup genotip, vücut yapısı ve çevresel faktörler tarafından kontrol edilir. Safkan İngiliz atlarının vücut yapısına etki edip yarış performansını dolaylı yoldan olumlu veya olumsuz yönden etkileyen etmenler; atın bakım ve antrenmanları, beslenmesi, cinsiyeti, yaşı, anne yaşı ve hastalıklarıdır (Paksoy ve Ünal, 2010; Paksoy, 2016).

Atların vücut yapıları genetik faktörlerin elverdiği ölçüde çevre koşulları ve beslenmeden önemli ölçüde etkilenmektedir. Fiziksel olarak iyi bir beden yapısı yarış atlarının gelişiminin en iyi göstergesidir. Atların kas ve iskelet sistemleri birçok benzerlik göstermesine rağmen farklı at ırkları değişik vücut yapıları göstermektedir. Bu durum, atların yarış özelliklerini farklılaştırmakta ve performanslarını önemli şekilde etkilemektedir (Paksoy, 2016).

Zayıf bir vücut yapısı yarış atının sakatlıklarla karşılaşma ihtimalini artırır. Atların kemik uzunlukları, eklem açıklıkları ve fiziksel yapılarındaki oranlar hem atletik yeteneklerinin önemli göstergeleri hem de atların sakatlanmaya olan yatkınlıklarını gösteren önemli ipuçlarıdır. Bir atın vücut yapısının morfolojik özellikleri sonucu kazandığı sürati ve sağlıklı kalabilme durumu yarış hayatının uzun sürebilmesine yardımcı olur (Tepeoğlu, 2018).

Atlar yüksek düzeyde konvansiyonel yüksek karbonhidratla beslenmeye ihtiyaç duyarlar. Atın beslenmesinde yemin kalitesi ve hijyen atın sağlığını ve dolayısıyla performansını etkiler. Egzersiz sonrasında atların kas glikojeninin normal seviyesine dönmesi için 72 saate ihtiyaç vardır. Kısa ve yoğun şiddette egzersiz yaptırılan atların kas glikojeni tükenmeye başlar ve buna bağlı olarak da performans düşüşü görülür. Yarış atlarının yüksek performans gösterebilmesi için yüksek düzeyde enerji gerekir. Kasların bu enerjiyi elde edebilmesi için ihtiyaç duydukları yemleri (arpa, yulaf, mısır) tüketmeleri ve rasyonlarına

glikojen ve yağ ilavesi yapılması gerekir. Günde 3 defa temiz ve ılık su verilmeli bundan sonra da arpa, mısır, yulaf, yeşil otlar, havuç, şeker pancarı ve meyve posaları gibi maddeler yem olarak verilmelidir. Atlar eğitimleri sırasında sevdikleri kesme şeker, elma ve havuç gibi yiyecekler ile ödüllendirilirse istenen davranışı kazanması kolay olur. Yarış atları için belirlenen idman ve yarış programı besin maddesi ihtiyacını belirler (Paksoy, 2016; Tepeoğlu, 2018).

Atların bakım uygulamaları (barınak, tımar, tırnak bakımı gibi) ve antreman programlarının farklılığı performansı doğrudan etkiler. Uygun idman programları yarış atının potansiyelini göstermesine yardımcı olursa uygun olmayan veya yetersiz egzersizler sakatlanmayı artırır ve yarış performansını düşürür (Paksoy, 2016). Atlar yarışlardan önce sürat ve dayanıklılıklarını artırarak performanslarını yükseltmek için çok sayıda antrenör tarafından tempo, mesafe ve süreleri belirlenen idmanlara katılıp eğitilirler. Atletik yapıları farklı olduğu için her ata uygulanan farklı ve özel antrenman teknikleri bulunur. Basit koşu, üzerinde ağırlıkla koşma, üzerinde binicisi ile dolaşma, belli bir çap etrafında dönerek koşma ve nefes açma gibi teknikler sayılabilir. Atların atletik yapısını oluşturan kemikler, tendonlar ve ligament gibi hareket sistemini oluşturan uzuvların geliştirilmesi sırasında diğer organlara zarar vermemek gerekir (Tepeoğlu, 2018).

Atlar farklı antrenmanlara çok hızlı uyum sağlarken sürat, güç ve dayanıklılık aynı ölçüde gelişim göstermez. Bu nedenle katılacağı yarış türleri ve mesafelere göre antrenman programları hazırlanmalıdır. Yükselen egzersiz kapasitesi ile daha geç yorulur, sürat, güç ve dayanıklılık açısından yüksek performans gösterir ve sakatlık riskleri de düşer (Tepeoğlu, 2018).

Yarış performansını etkileyen faktörlerden bir diğeri ise cinsiyettir. Aygırlar daha fazla gelişmiş kas ve kemik yapısına sahip olduğu için adımları daha uzundur. Bu durum yarışlar için erkek atlara avantaj sağladığı için erkek taylar daha fazla fiyatlara satılmaktadır. Yarışlardaki erkek atların bu üstünlüğünü dengelemek için dişilere göre ağırlığı 1,5 kg fazla olmaktadır. Yaş, cinsiyet, koşu türü, mesafe, ırk ve biniciye (jokey veya apranti) göre tespit edilen handikap ağırlığı jokeyin, atın eyerinin ve üzerindeki ilave ağırlıkların toplamını gösterir (Özen vd., 2021; Paksoy, 2016).

İngiliz atlarında yarışa başlama yaşı ikidir. Yapılan çalışmalar yaştan yarış performansına etkisinin genellikle önemli olduğunu göstermiştir (Ekiz vd., 2005a; Moritsu vd., 1994; Mota

vd., 2005). Ana yaşının yarış performansına etkisini inceleyen araştırma sayısı çok azdır. Bunlardan birisi İngiltere'deki İngiliz atları üzerinde yapılmış, anaları genç olan taylardaki performansın ana yaşı 11'den büyük olan taylardan daha yüksek olduğu bulunmuştur (Paksoy, 2016).

Yarış atlarının yakalandığı çeşitli hastalıklar performansına olumsuz etki edip istenilen düzeyde performansa ulaşamamasına neden olmaktadır. Buna sebep olan en çok bilinen sağlık sorunları; üst solunum yolu, kas iskelet sistemi, kardiyovasküler sistem, sinir sistemi, sindirim sistemi hastalıkları ile endokrin sistem ve termoregülasyon bozukluklarıdır. Ayak ve tırnak sorunları da en fazla karşılaşılan hastalıklardan birisidir. Üst solunum yolunda meydana gelen hastalıklar yarış atlarındaki kötü performansın %48'inin sebebidir (Sevim, 2022). Akciğerlerin geniş ve sağlıklı olması, atın nefes alışverişinin rahat olmasını, kaslara ve vücudun tüm bölümlerine daha fazla oksijen taşınmasını sağlayarak atların performans için ihtiyaç duydukları enerjiyi kazanmasına yardımcı olur (Paksoy, 2016; Tepeoğlu, 2018).

Yarış atları fizyolojik, anatomik olarak yarış için gerekli ideal vücut ölçülerine sağlıklı bir biçimde ulaştıklarında beslenme ve antrenmanlar ile yeterli egzersizler yaptırıldığında bireysel olarak yarış atlarının tamamı aynı performans gösterecek şartları sağlamış olurlar. Performansa doğrudan etki edecek genetik faktörler haricinde, çevresel faktörler olan farklı mesafeler (1.000 m ile 2.400 m), farklı yaşlar, farklı koşu yüzeyleri (toprak, çim), sahip türleri (tek kişi veya hisseli), binicinin deneyimi, bilgisi ve ağırlığı, handikap ağırlığı, yarışların düzenlendiği hipodrom ve koşu tipi gibi değişkenlerin yanı sıra çeşitli hava koşulları yarışın sonucunu doğrudan ve dolaylı olarak etkilemektedir (Bailey vd., 2022; Paksoy, 2016; Paksoy vd., 2018).

Türkiye'de kısa (800-1.300 m arası), orta (1.400-1.700 m arası) ve uzun (1.800-3.000 m arası) olmak üzere üç farklı mesafede at yarışları düzenlenmektedir. Farklı mesafelerde yarış atları farklı performans sergilerler. Herhangi bir mesafede iyi performans gösteren atın bütün mesafelerde benzer başarıyı sergilemesi olanaklı değildir (Paksoy ve Ünal, 2010; Paksoy, 2016).

Yarışların çim veya kum pistte düzenlenmesinin atların yarış performansını etkilemediği İngiliz atların çim pistte kum piste göre daha iyi performans gösterdikleri bildirilmiştir. Atların farklı pist durumlarında performanslarının farklı olduğunda belirlenmiştir (Paksoy ve Ünal, 2010). Örneğin, normal, yumuşak ya da çok yumuşak durumda

çim ve kum pistin olması performansı etkilemektedir (Özen vd., 2021; Paksoy, 2016; Park vd., 2011).

Binicinin deneyimi ve bilgisi ile atların koştuğu mesafelerde yarış performansını etkileyen faktörlerdendir. Nitekim 1.200 ile 1.800 m arasındaki mesafelerin hepsinde birinci gelen İngiliz atlarının çim pisti kum piste göre daha kısa sürede tamamladığı görülmüştür (Paksoy ve Ünal, 2010). Bunun yanında İngiliz yarış atları ile yapılan bir araştırma binicinin ağırlığının yarış performansında doğrudan etkili olmadığını fakat binici ile atın ağırlığı arasındaki oransal değerin artmasının yarış süresini kısalttığını göstermiştir.

Park vd. (2011) İngiliz atları üzerinde yaptıkları çalışma ile gün içi yarış sırası (1-14 arası), pistin durumu (sekiz farklı durum), atın köken aldığı 14 farklı ülke, beş farklı hava durumu, yarışların düzenlendiği iki farklı şehir hipodromu, cinsiyet, atın yaşı, handikap ağırlığı ve yarış mesafesinin (1.000-2.300 m) yarış süresine anlamlı derecede etki ettiğini göstermiştir. Buxadera ve Mota (2008) yıllık kazanç ve en iyi yarış süresi gibi performans ölçütlerine yaş ile cinsiyetin etki ettiğini bulmuştur. Oki vd. (1995) yaş, jokey, kilo ve hatta yarış sezonunun yarış süreleri üzerinde önemli etkileri olduğunu bildirmiştir.

Yarış performansı kantitatif bir özellik olduğu için genotip tarafından kontrol edilir. Atletik fenotipler çevreden ve antrenmandan önemli derecede etkilenirken bunun yanında atletik performans yeteneklerini etkileyen genetik faktörlerin olduğu kabul gören bir gerçektir (Bower vd., 2012; Paksoy ve Ünal, 2010; Paksoy, 2016).

Kondisyonla ilgili 200'den fazla gen insanların sağlık ve egzersiz özellikleriyle ilişkili bulunmuştur (Bower vd., 2012). Bu durum safkan İngilizlerin yarış performansının da poligenik olabileceğini ve çok çeşitli anatomik, metabolik özelliklere katkıda bulunan genlerden etkilenmesinin muhtemel olabileceğini göstermiştir. Safkan atların atletizmi ve egzersizlerine yönelik fizyolojik adaptasyonlar iyi tanımlanmış fakat bu özelliklerini açıklayacak çok az gen tanımlanmıştır.

Safkanların yarış fenotiplerinin Miyostatin (MSTN) lokusundaki varyasyon tarafından etkilenmekte olduğu ve bunun da erken iskelet kası gelişimine ve kısa mesafelerde yarışma yeteneğine olumlu katkıda bulunduğu görülmüştür. Son yıllarda yapılan araştırmalar, MSTN lokusundaki varyasyonun diğer at ırklarında morfolojik tipe katkıda

bulunduđu, safkan yarış atlarında ise yarış mesafesi yeteneđini tahmin etmekte kullanılabilecek bir genetik potansiyel olduđunu göstermiştir (Bower vd., 2012).

Bu çalışmalar, atın hızına olumlu yönde etki eden MSTN C-alelinin artışı ve korunmasını teşvik etmektedir. Safkan popölasyonu içindeki çeşitliliđi artırmaya devam edecek üreme ve seleksiyon kararları verilirken genetik bilginin de bu işe dahil edilmesi gerektiđi görölmüştür. MSTN lokusundaki genetik varyasyon bilgisi ile tek bir safkan için optimum yarış mesafesi tahmin edilebilecekken diđer genomik lokuslardaki uygun alel kombinasyonları ile atların koşullarda en iyi performans gösterecekleri mesafeleri tahmin etmek kolaylaşacaktır (Bower vd., 2012).

Yarış performansının kalıtsallığı varsayımı, yarış atlarının seçimi ve yetiştirilmesini olanaklı hale getirmektedir (Bower vd., 2012). Atların yarışı tamamlama süresi ve hızı gibi performans değerlerinden atın damızlık değeri hesaplanabilmektedir. Safkan İngiliz atlarına ait performansın genetiđi Langlois (1996)'ın yaptıđı detaylı araştırma ile incelenmiş performansın kalıtsallık tahminleri elde edilmiştir (Paksoy, 2016; Thiruvankadan vd., 2009).

Yarış süresi, kazanılan para ve handicap özelliklerinin kalıtımla geçmesini inceleyen çalışmalar, düşük kalıtsal özellikleri içeren çevresel faktörlerin genetik faktörlerden daha kolay şekilde yönetilebileceđini göstermiştir. Düşük kalıtsallığa sahip ancak değeri yüksek olan özellikler, performansın iyileştirmesinde önemli faydalar sağlayabilir (Bailey vd., 2022).

Kalıtım tahminleri, fenotip seçimi, verilerin kalitesi ve büyüklüğü ile analiz yöntemine göre deđişmektedir. Bu tahminler, performansın ne kadarının (örneğin, yarış süresi, kazançlar) bireyin kendi soyuna geçebileceđi varyasyondan kaynaklanacađı konusunda yetiştiricilere rehberlik eder. Yüksek kalıtsallığa sahip özellikler, düşük kalıtsallığa sahip olanlara göre seçilime daha olumlu yanıt verecektir (Bailey vd., 2022).

Bireysel olarak yarış atının yarış değeri, gösterdiđi performans ile ve soyağacı indeksleme yöntemleriyle oluşturulur. Yarış atı popölasyonları ise nispeten daha uzun süreler boyunca seçilen türlerin örneklerini sağlar. Yarış atı ırkları gelişimi, kısa bir sürede belirli bir mesafeyi kat etme yeteneđinin kısmen genetik bileşenler tarafından belirlendiđi varsayımı üzerine kurulmuştur (Tolley vd., 1985). Bazı yarış performansı ölçümlerinin makul kalıtsallığı ile bazı faktörler birleştinde performans testlerine dayalı kitle seçiminin, safkan atların yarış performansını iyileştirmek için tercih edilen seçim prosedürü olabileceđini göstermiştir

(Thiruvankadan vd., 2009). Diğer taraftan benzer koşullarda yarışan atlardan oluşan popülasyon odaklı yaklaşıma sahip büyük veri içeren araştırmalar, kapsamlı bir şekilde bu atlarda istenen yüksek yarış performansına etki edebilecek faktörleri incelemede etkili olabilecek bir model sunabilirler (Özen, 2014).

Kalıtım tahminlerinin mesafeye göre değiştiği daha kısa mesafeler için yarış süresi kalıtsallığının daha yüksek olduğunu belirtilmiştir (Bailey vd., 2022). Mota (2006) yaptığı bir araştırmada atların 1.000 metrede gösterdiği performansa ait kalıtım derecesi 1.300, 1.400 ve 1.500 metrelerden yüksek bulunmuştur. Bu yüzden bu mesafedeki yarış performansının seleksiyon kriteri olarak değerlendirilebileceğini rapor etmiştir (Paksoy, 2016).

Thiruvankadan vd. (2009) yaptığı derlemede, İsveç, Fin, Alman, Norveç ve Fransız Trotter ve Standardbred atlarının yarışlarda elde ettiği kazançların kalıtsallığının değerlendirildiği çalışmaları incelemiştir. Toplam ve yıllık kazancın kalıtsallığının tahmin edildiği dokuz çalışma, toplam kazancın logaritmasının kalıtsallığının tahmin edildiği altı çalışma, yıllık kazancın logaritmasının ve kare kökünün kalıtsallığının tahmin edildiği on iki çalışma ve yarış başına ortalama kazancın kalıtsallığının tahmin edildiği on bir adet çalışma bildirmişlerdir.

Farklı at türleri (trotter ve standartbred gibi) için farklı koşu türlerinde (düz koşu, engelli ve zıplama koşuları gibi) yıllık kazanca etki eden faktörlerin incelendiği çalışmalar (Hintz, 1980; Langlois, 1980; Langlois, 1996; Langlois and Hernu, 2003; Tolley vd., 1985) çok olmasına karşılık safkan İngiliz atlarının yarış hayatı boyunca ikramiye kazanma durumunu etkileyen çalışma sayısı çok azdır.

Aşağıdaki bölümde MÖ, istatistiksel öğrenme ve büyük veri hakkında kısa bilgilere yer verilecek sonrasında tez kapsamında kullanılacak MÖ sınıflandırma algoritmaları hakkında kısa açıklamalar yapılacaktır.

1.9. Makine Öğrenmesi ile İlgili Genel Bilgiler

MÖ büyük veri setlerini analiz etmek için kullanılan algoritmalar bütünüdür. Yapay zekanın (YZ) bileşenlerinden birisi olan MÖ istatistiksel öğrenme yöntemlerini kullanarak sınıflandırma, tahmin ve kümeleme gibi problemlerin çözümünde kullanılır. Denetimli,

denetimsiz, yarı denetimli ve pekiştirmeli öğrenme gibi eğitim yöntemlerini kullanarak büyük veri setlerindeki model desenlerini ortaya çıkartmaya çalışır (Balta, 2018).

Tezin bu bölümünde MÖ temel kavramlarından kısaca bahsedilmiş, bu çalışmada veri setinin analizinde kullanılan denetimli öğrenme sınıflandırma algoritmalarından; C5.0 karar ağaçları, k-EYK, Torbalama (Bagging), RO, Artırma (Boosting) ve GAM hakkında kısa bilgi verilmiştir.

1.9.1. İstatistiksel Öğrenme Yöntemi

İstatistiksel öğrenme teorisi “Vapnik-Chervonenkis (VC) Teorisi” olarak da adlandırılmakta olup, algoritmanın başarısı için ihtiyaç duyulan temel özelliklerin neler olduğu, başarılı MÖ için hangi varsayımların gerekli olduğu, bilgisayarların genelde hangi öğrenme görevlerini yerine getirilebileceği ve algoritma sonuçlarının nasıl genellenebileceği temel sorularına cevaplar aramaktadır (Vapnik, 1998).

Öğrenme problemi araştırmalarında başlangıçta Rosenblatt (1962) tarafından “*perceptron*” isimli öğrenme makinesinin ilk modeli önerilmiştir. Perceptron fikri nöro-fizyoloji literatüründe yıllardan beri tartışılan bir kavram iken Rosenblatt “*perceptron*” modelini bilgisayarlar için bir program olarak tanımlamış ve bu modelin genelleştirilebileceğini basit deneylerle göstermiştir (Rosenblatt, 1962). Rosenblatt (1962) “*perceptron*”un en basit nöron modelinin uyarlanabilir özelliklerini kullanabilmesi için bir kural geliştirmiştir. 1962 yılında “*perceptron*” için ilk teorem Novikoff tarafından ispatlanmıştır (Vapnik, 1998). Bu teorem öğrenme teorisinin başlaması ve oluşumunda önemli derecede rol oynamıştır. Sonrasında, 1986 yılında geriye yayılma (back-propagation) tekniği geliştirilerek öğrenme teorisinin genel tipi oluşturulmuştur (Vapnik, 1998).

Rosenblatt (1962) tarafından geliştirilen “*perceptron*” modelinden 1986 yılında geliştirilen geriye yayılma (back-propagation) tekniğine kadar istatistiksel öğrenme teorisinde gelişme kaydedilmiştir. İstatistiksel öğrenme teorisinin felsefesi 1968 yılına kadar geliştirilmiştir (Vapnik, 1998).

VC entropi ve VC boyut kavramları, desen hatırlama problemlerinin çözümünde kullanılmış, fonksiyonel uzayda büyük sayılar kanununa ulaşılmış ve öğrenme işlemi ile

ilişkisi tanımlanmıştır. Yakınsama oranının gerçek asimptotik olmayan sınırları elde edilmiştir (Vapnik, 1998).

Doğru kestirimler için büyük sayıda örneklemelere ihtiyaç olduğu düşüncesi istatistik biliminde genel kabul gören bir durum iken küçük örneklemin doğru tahminler için istatistiksel bir problem olduğu düşünülmüştür. Öğrenme teorisi temelli, yeni fonksiyon tahmin metotları yüksek boyutlu bilinmeyen fonksiyonların doğru tahminler yapabilmesi için büyük sayıda örnekleme ihtiyaç olmadığı Vapnik tarafından gösterilmiştir (Vapnik, 1998).

Yine bu teorenin ışığında bulunan yeni metotlar ile genelleştirme, kapasite faktörü ile kontrol edilmiş ve yüksek uzaya (hyperplane) gerek duymamıştır. Bu durumun VC teorisi ile çözüldüğü sanılırken Destek Vektör Makineleri (DVM) (Support Vector Machine, SVM) öğrenme yöntemi ile bu sorun çözülmüştür.

Öğrenme teorisinin çözüm getirdiği sorunlardan bir diğeri ise sınırların belirlenmesi kavramıdır. Klasik istatistik sorunu boyut azaltarak çözerken DVM ise çözümü yüksek boyutlu uzayda bulmuş ve ismine büyük marjın faktörü denmiştir. Kernel fonksiyonları vasıtasıyla yüksek boyutlu uzaya taşınarak çözümler üretilmiştir (Vapnik, 1998).

Parametrik olmayan yaklaşımların bulunması parametrik varsayımlara ve olasılıksal yoğunluk fonksiyonu bilinmeden de doğru çözümlere ulaşılabileceği gösterilmiştir. Bu sayede, dağılımdan bağımsız yöntemler ile sınıflandırma, regresyon ve yoğunluk tahmini görevleri için çözümler üretilabileceği düşüncesi öğrenme teorisinin gelişiminde önemli rol oynamıştır (Vapnik, 1998).

Regülerizasyon kavramı öğrenme teorisi içinde önemli bir yere sahiptir. İstatistik problemleri çözümünde en çok üzerinde durulması gereken yoğunluk fonksiyonudur. Ancak bu kavram, istatistiksel öğrenme dağılımdan bağımsız çözümler üretebilmek için yoğunluk fonksiyonunu değil regülerizasyon formüllerini kullanır (Vapnik, 1998).

Yukarıda sayılan yeni yaklaşımlar istatistiksel öğrenmeden makine öğrenmesine geçişi kolaylaştırmıştır. Öğrenme probleminin kapsadığı üç ana konu; desen öğrenme veya sınıflandırma, regresyon ve yoğunluk tahminidir. Öğrenme sorunu genel bir durum olduğu için istatistik biliminde tartışılan her konunun öğrenme teorisinde bir benzeri vardır. Ayrıca,

öğrenme teorisi çerçevesinde ele alınan her konu istatistik terimleriyle yeniden formüle edilmiş ve MÖ'nün matematiksel altyapısı oluşturulmaya çalışılmıştır (Vapnik, 1998).

Öğrenme probleminin en önemli amaçlarından birisi verilen fonksiyon setlerinden doğru cevaba en çok yaklaştıracak fonksiyonu seçmektir. Bu fonksiyondan elde edilen veriler ile gerçek gözlemler arasındaki kaybı (loss) ya da ayrışmayı ölçmek için risk fonksiyonu kullanılır ve gözlemlenen veriler temelinde en küçük değeri alıncaya kadar işlem devam eder (Vapnik, 1998). Gözlemsel risk azaltımı (GRA) (Empirical Risk Minimization) tümevarımsal prensibi öğrenme teorisinin gelişiminde önemli role sahip olmuştur.

Öğrenme teorisi dört ana sorunun çözümüne odaklanmıştır. Bunlar;

1. GRA ilkesine dayalı öğrenme sürecinin tutarlılığı için gerekli ve yeterli koşullar nelerdir?
2. Öğrenme sürecinin yakınsama hızı ne kadar?
3. Öğrenme sürecinin yakınsama oranı (genelleme yeteneği) nasıl kontrol edilebilir?
4. Genelleme yeteneğini kontrol edebilecek algoritmalar nasıl oluşturulabilir?

Klasik istatistik, belirli bir modelin yapısının bilindiğini kabul eder ve bu modelin parametrelerini veri yardımıyla tahmin ederek en doğru model ve tahminlere ulaşmayı amaçlar. İstatistiksel öğrenmede ise tahmin yapısı başlangıçta bilinmiyor olup, farklı olası modeller arasından en iyi modelin bulunması hedeflenir. Bu süreçte modelin parametreleri de öğrenilir (Keskin, 2018).

Yapay öğrenme, genellenebilir yapılar ortaya koyabilmek için istatistik biliminden yararlanmaktadır. Bilgisayar bilimleri, istatistiksel öğrenmeyi kullanan öğrenme yaklaşımları ile elde edilen modellerin en iyilenmesi (optimizasyonu) ve tahmin performansının artırılması ile uğraşır. Öğrenme modellerinin birçoğunun çıkış noktası istatistik iken, tahmin başarısı yüksek en iyi modellere bilgisayar bilimleri yardımı ile ulaşılmaktadır (Keskin, 2018).

İstatistiksel öğrenme istatistiğin alt alanı olduğu için tahmin modelleri varsayımlar üzerine kurulur. Elde edilen tahmin fonksiyonları doğrusal olduğu için modellerin içinde

kullanılan bağımsız değişkenlerdeki değişimlerin yanıt değişkeninde meydana getirmesi beklenen değişimleri ile ilgili yorum yapılabilir. Kurulan tahmin modellerinin yorumunda varsayımlar dikkate alınır. Gerçek dünyada doğrusallık kolay sağlanamadığı için doğrusallığı sağlamak beraberinde fazlaca varsayım gerektirir. Bağımlı değişkende (yanıt değişkeni) meydana gelen değişikliklere hangi bağımsız değişkenin ne düzeyde katkıda bulunduğu istatistiksel öğrenme modelleri ile öğrenilebilir.

İstatistiksel öğrenme ile oluşturulan doğrusal ya da doğrusallaştırılabilen tahmin fonksiyonları yardımı ile elde edilen tahmin modellerinin MÖ ile elde edilenlerin aksine yorumlanabilirliği ve açıklanabilirliği yüksektir (Keskin, 2018).

1.9.2. Makine Öğrenmesi (MÖ) (Machine Learning, ML)

Endüstri devrimi ile hayatımıza giren makineler ilk zamanlarda fabrikalarda ve üretim tesislerinde kullanılan araç gereçler iken zamanla hayatımızın her alanına girmeye başlamıştır. 1943'te II. Dünya Savaşında ürettiği elektromekanik cihaz ile çok zor kriptanalizlerini başarı ile yapıp dünya savaşının seyrini değiştiren Alan Mathison Turing'in üzerinde çalıştığı en önemli iki kavram, düşünebilen ve zekaya sahip makinelerin üretilmesidir. Bu kavramlar üzerine yapılan çalışmalar ile bilgisayar bilimi ve YZ doğmuştur.

YZ, insan zekasından beklenen davranışları sergilemesi için programlanan yapay işletim sistemi olarak tanımlanabilir. Bilgisayarın, robotun veya farklı tipte akıllı nesnelerin çözümü için akıl gerektiren çeşitli faaliyetleri zekice davranan canlılar gibi yerine getirme kabiliyetidir. Öğrenme, düşünme, çıkarımda bulunma, sorun çözme, fikir yürütme ve karar verme davranışlarını sergileyebilen bilişsel fonksiyonlara ya da otonom davranışlara sahip olmak YZ'nin göstergesidir (Çağlayan Akay, 2018).

YZ, makine öğrenmesi, yapay sinir ağları, doğal dil işleme ve derin öğrenme gibi alt bileşenlerden oluşmaktadır. Bu bileşenlerden birisi olan MÖ kavramı 1959 yılında Arthur Samuel'in "IBM Journal of Research and Development" isimli dergide yazdığı yazıda ilk defa bilimsel olarak ifade edilmiştir (Samuel, 1959). Çalışmada MÖ'den bilgisayar bilimlerinin alt alanı olarak bahsetmiş ve MÖ'yü programlanmamış bilgisayarlara öğrenme kabiliyeti kazandırmak olarak belirtmiştir.

MÖ, çevresi ile uyumu gerçekleştirebilmek için sistemin kendini yenileyebilmesi olarak tarif edilebilecek öğrenme işinin bilgisayarlar tarafından gerçekleştirilmesini sağlar. Edinilen bilgiler vasıtasıyla problem çözümünde bulunabilecek ve tekrar eden olaylardan elde ettiği tecrübelerin yardımıyla benzeri sorunlar için daha doğru kararlar verebilecek sistemler geliştirmesi gelecekte MÖ'den beklenen görevlerdendir (Kıral Özkan, 2015).

Temel amacı doğru kestirimler yapmak olan MÖ ile elde edilen tahmin modelleri istatistiksel öğrenme modellerinin aksine açıklanabilirliği ve yorumlanabilirliği düşüktür (Çağlayan Akay, 2018). İstatistiksel modeller model hipotezleriyle başlar, varsayımları belirler ve çıkarımlara odaklanır. MÖ öncelikle tahminle ilgilidir, en iyi performans gösteren tahmin modeli seçilir. Geleneksel yöntemlerin aksine, MÖ'nün hipotezsiz, az varsayımli ve veri dağılımından bağımsız doğası onu karmaşık veri analizi için çekici bir seçim haline getirmiştir (Gouda vd., 2022).

Artan veri sayısı ve büyük verilerin karmaşıklığı geleneksel veri işleme araçları yerine kullanılacak MÖ'ye olan ihtiyacı artırmıştır. MÖ karmaşık yapıli büyük miktardaki, yüksek boyutlu veya yapılandırılmamış veri setlerini kullanarak anlamlı ilişkiler bulan model ve örüntü geliştiren yöntemler ailesidir (Çağlayan Akay, 2018). Denetimli, denetimsiz, yarı denetimli ve pekiştirmeli öğrenme gibi eğitim yöntemlerini kullanarak büyük veri setlerindeki model desenlerini ortaya çıkartmaya çalışır.

Bundan sonraki bölümlerde, büyük verinin özellikleri, denetimli öğrenme stratejisi ve bu araştırma kapsamında kullanılacak sınıflandırma algoritmalarından bahsedilmiştir.

1.9.3. Büyük Veri

Büyük veri, internet ve nesnelerin interneti gibi farklı veri kaynaklarından çeşitli formlarda elde edilen çok büyük miktarlardaki verilere verilen isimdir. Farklı kaynaklardan elde edilen hacmi, hızı ve karmaşıklığı yüksek büyük verilerin depolanması, yönetimi ve analizini yapabilmek için YZ'nin gelişmiş tekniklerinden faydalanılır. İşlenecek verinin boyut, yapı ve karışıklık düzeyi dikkate alınarak tanımlanan büyük veri yaygın olarak kullanılan geleneksel verilerden farklılık göstermektedir (Çağlayan Akay, 2018).

Büyük veriyi tanımlayan 10 temel özellik, İngilizce terimlerin baş harfleri ile tanımlanan 10V ile ifade edilebilir. Terabayt, petabayt, ekzabayt, zettabayt veya yottabayt ile

ölçülen Hacim (Volume) özelliği, verinin sıklığını ifade eden Hız (Velocity) özelliği, yapısal heterojenliği belirten Çeşitlilik (Variety) özelliği, aşırı belirsiz ve öngörülemezliğini belirten Değişkenlik (Variability) özelliği, güvenilirlik, hatasızlık ve kesinliği belirten Doğruluk (Verification) özelliği, verilerden elde edilen değeri ve onun önemini belirten Değer (Value) özelliği, verilerin kullanım amaçlarının geçerlilik ve doğruluğunu belirten Gerçeklik (Veracity) özelliği, verilerin ne kadar süre geçerli olduğunu belirten Oynaklık (Volatility) özelliği, birçok veri değişkenini içerebilen karmaşık grafikleri belirten Görselleştirme (Visualization) özelliği ve bir kaynağın bilinen zayıflığına işaret eden Güvenlik Açığı (Vulnerability) özelliğidir.

Karışıklık (Complexity) özelliği büyük verinin bu 10V özelliğine ilave edilebilir. Verilerin işlenmesinde karşılaşılan zorlukları ve çözümündeki karışıklıkları ifade eder. Büyük veriler için sayılan bütün özellikler araştırmacıların bu tür verilerin zorluk ve avantajlarını anlamalarına yardımcı olmaktadır. Çözümleme aşamasında karşılaşma olasılığı yüksek zorluklarla başa çıkmak için araştırmacıların önlemler almasına yardımcı olmaktadır (Çağlayan Akay, 2018).

Karışıklık düzeyi yüksek ve 10V özellikleri olan büyük verili problemlerin çözümünde geleneksel veri yönetimleri yetersiz kalmaktadır. Bu verilerin geliştirilen yeni istatistiksel yaklaşımlar ve MÖ algoritmaları gibi zenginleştirilmiş yöntemlerle işlenmesine gerek duyulmaktadır. Büyük veri analizlerinin günümüzde yaygınlaşma sebeplerinden en önemlisi yüksek boyutlu ve yapılandırılmamış verilerin az maliyetle sürekli olarak üretilebilmesidir (Çağlayan Akay, 2018).

1.10. Sınıflamaya Dayalı Denetimli Öğrenme Algoritmaları

Denetimli öğrenme yönteminde veri setindeki değişkenler, özellik veya öznitelik isimleri ile nitelendirilir. Bağımsız öznitelik olarak bilinen veriler girdi verisi, hedef öznitelik yani bağımlı değişkene ait veriler de çıktı verisi olarak etiketlenir. Amaç, denetimli öğrenme algoritmalarının etiketli veriler üzerinden sınıflandırma veya regresyon tahmin modelini eğitim veri setindeki bilgiler ile öğrenebilmesi sonra da test veri seti olarak ayrılan veriler için doğru tahminlerde bulunabilmesidir.

Örneğin, gül ve papatya görselleri kategorize edilmek istenildiğinde kullanılacak algoritmaya gülün ve papatyanın özellikleri öğretilmelidir. Bunun için algoritma “gül” ve

“papatya” olarak etiketlenmiş binlerce gül ve papatya görseli ile eğitilir. Algoritmanın amacı; gülü papatyadan ayırt edecek özellikleri doğru bir şekilde öğrenip, eğitim setinden bağımsız test veri setindeki girdileri başarılı bir şekilde “gül” ve “papatya” olarak sınıflandırabilmesidir (Ceylan, 2018).

Bu öğrenme yöntemini kullanan algoritmalar sınıflandırma ve regresyon problemlerinin analizlerinde çözüm sunmaktadır.

Bu kısımda, çalışmada kullanılan sınıflandırma algoritmaları k-EYK, C5.0, RO ve GAM ile birlikte RO ve GAM algoritmalarının temelini oluşturan Torbalama (Bagging) ve Artırma (Boosting) yöntemleri hakkında bilgi verilecektir.

1.10.1. K-En Yakın Komşu (k-EYK) (K-Nearest Neighbor, KNN) Algoritması

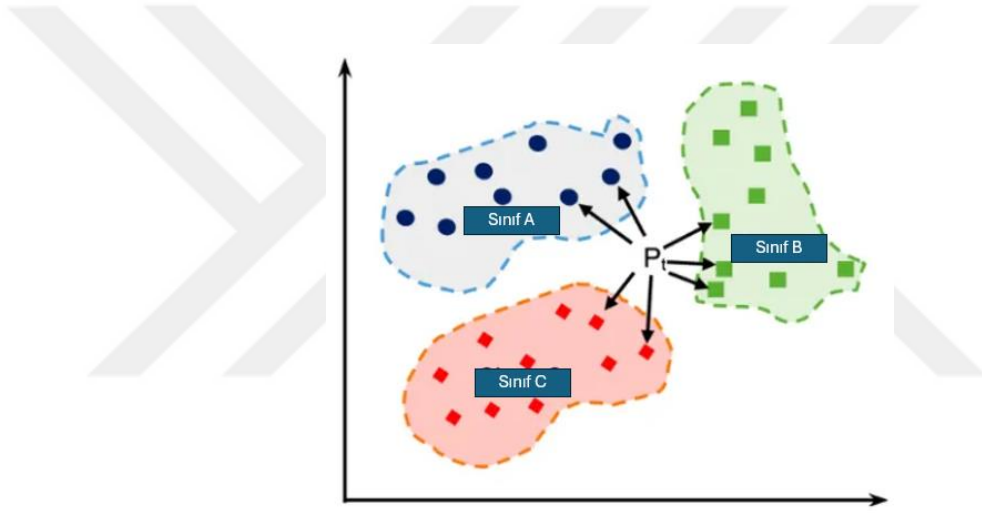
Evelyn Fix ve Joseph Hodges tarafından 1951'de geliştirilen en yakın komşu algoritması Cover ve Hart (1967) tarafından k adet komşuluk için genişletilerek k-EYK bulunmuştur. Basit, anlaşılması kolay parametrik olmayan bir denetimli öğrenme yöntemidir. Sınıflandırma ve regresyon problemlerinin çözümünde kullanılır. K-EYK, öğrenme şekline göre tembel öğrenme (lazy learning) algoritması, veri setinden herhangi bir parametre öğrenmediğinden dolayı parametrik olmayan bellek tabanlı bir MÖ yaklaşımı olarak da tanımlanır. Veri seti içinden bir ayırma fonksiyonu öğrenmekten ve model inşa etmekten çok eğitim veri kümesini bütünüyle hafızaya alıp test veri seti için tahminde bulunur (Lantz, 2015).

K-EYK algoritmasının güçlü olduğu yönler; hem doğrusal hem doğrusal olmayan durumlarda çözüm üretebilmesi, sınıf dağılımı hakkında herhangi bir varsayıma gerek duymaması, aykırı değerlerden etkilenmemesi, eğitim aşamasını hızla tamamlaması ve en önemlisi de hem iki hem de çok kategorili sınıflandırma problemlerine etkili çözümler üretebilmesidir. K-EYK algoritmasının zayıf yönleri ise; öğrenmeyip hafızaya kaydetmesi, bir model geliştirmemesi ve değişkenlerin önemlilik derecelerinin belirlenememesidir (Lantz, 2015).

Algoritma analize başlamadan önce, her biri farklı ölçüm türlerinde olan özellikler için veri ön işleme olarak standardizasyon, normalizasyon gibi dönüşüm işlemleri uygulanır. Veriler arasındaki uzaklıkların ölçümü için Öklid ve Manhattan gibi yöntemler belirlenir. Sonrasında, algoritma eğitim sürecinde k-katlı (k-fold) çapraz doğrulama yöntemi kullanılarak

“k” hiper parametresi için en iyi değere karar verilir. Algoritmanın eğitimi sırasında eğitim veri setinden model geliştirilmez yani ayırım fonksiyonu oluşturulmaz. Sadece veri setindeki örüntü ezberlenir. Öğrenme sürecinde ezberlenen örüntüye uygun olarak test veri setindeki yeni veriler doğru sınıflara atanmaya çalışılır.

Test veri setindeki örnekler, belirlenen “k” tane örneğin ait olduğu sınıftan kendilerine en yakın karakteristik özelliklere sahip olan çoğunluğun sınıfına atanır. Her bir öznitelik için test verisinde sınıflandırılmayı bekleyen her bir örneğin etrafındaki örneklerin uzaklıkları formül ile hesap edilir. Sonucunda, Şekil 1.1’de gösterilen “p”ye en yakın olan değerlere sahip “k” adet komşu belirlenir. Sınıflandırılmak istenen “p”nin yakın komşularının çoğunluğu hangi sınıfta ise o sınıfa dahil edilir (Bulut, 2019).



Şekil 1.1. k-EYK algoritmasında çok sınıflı sınıflandırma gösterim şekli (Sachinsoni, 2023)

Algoritmanın sınıflandırma başarısını direkt etkileyen iki önemli değer “k” en yakın komşu sayısı ve seçilen uzaklık fonksiyonudur. Algoritmanın sınıflandırma başarısını etkileyen tek hiper parametre olan “k” değerinin seçimi, algoritmanın eğitimi ile öğrenilen sınıflandırma deseninin genelleştirilmesi açısından önemli bir yer tutar. “k” sayısına karar verirken belirli bir standart yoktur (Lantz, 2015).

“k” büyük seçilirse gürültülü verilerin etkisi ve anlamsız örneklerden kaynaklanan varyansın etkisi azalır. “k” sayısının yüksek olması durumunda geliştirilecek modelin çoğunluğa ait sınıfa yönelik öngörüsü artacak ve ait olduğu sınıfa en yakın komşular dikkate alınmayacaktır. Bunun sonucunda, geliştirilmek istenen hedef model için katlanılan maliyet anlamsız ve karşılıksız kalacaktır (Bulut, 2019).

“k” sayı oldukça düşük belirlendiğinde ise anlamsız ve aykırı örnekler olması gerekenden daha fazla model üzerine etki eder. “k” sayının gerekenden çok düşük olması durumunda en yakın komşu aykırı bir örnek olduğunda sınıflandırma tahmini yanlış olacaktır. Küçük “k” değerleri ile sınıflar arasında kompleks sınırlar bulunduğu daha iyi sonuçlar elde etmek mümkün olabilmektedir. Sınıflar arasındaki sınırın kompleksliği artıyorsa ve bu sınır doğrusal olmayan bir fonksiyon ile ayrılıyorsa sınıflandırma probleminin çözümü zorlaşmaktadır (Lantz, 2015).

Algoritma, eğitim sırasında her ne kadar veri setinden bir fonksiyon öğrenemese de öne çıkan avantajlarından birisi, doğrusal olmayan fonksiyonlar ile ayrılmamış sınıflar için de doğru sınıflandırma yapabilmesidir (Bulut, 2019).

Literatürde, “k” sayısının çoğunlukla 3 ile 10 arasında bir değer olarak belirlendiği ve tek sayı olarak belirlenmesinde büyük yarar bulunduğu görülmüştür. “k” çift sayı olarak alınırsa, örneğin; üç sınıflı bir sınıflandırma problemi çözümünde $k=6$ olarak belirlendiğinde, en yakın komşudan ikisi A sınıfı, ikisi B sınıfı ise ve diğer ikisi ise C olduğu durumlarda, algoritma işlevini yitirir ve sonuç üretmez (Lantz, 2015). Sonuç olarak, “k” değerini belirlemek için geliştirilmiş ideal bir yöntem bulunmamaktadır. Sınıflandırma başarı oranı en yüksek “k” değerine ulaşabilmek için 10 kat çapraz doğrulama tekniğini kullanmak en doğru ve en iyi yöntem olarak görülmektedir. Test aşamasında gözlemlerin doğru sınıflara atanabilmesi için “k” değerinin doğru seçilmesinin etkisi modeldeki karar aşamasında daha fazla görülür (Lantz, 2015).

Algoritmanın model geliştirdiği veri setinde her bir özelliğin en küçük ve en büyük değer aralıkları farklı olabilmekte, örneğin veri setindeki bir özelliğin değer aralığı 1.000-12.000 arasında değişirken, diğer bir özelliğin en küçük ve en büyük değer aralığı 0,25-5 olabilmektedir. K-EYK algoritmasında yakın komşu uzaklıklarını belirlemek için kullanılan uzaklık fonksiyonları özniteliklerdeki bu ölçüm aralıklarına bağımlıdır. Bu yüzden analize başlamadan önce ön işlem olarak, öznitelikler genellikle standart bir aralığa dönüştürülür. Bu dönüşüm yapılmadığı durumda daha büyük değerlerde olanlar diğerleri ile kıyaslandığında uzaklık hesaplamasında daha ağır basacaktır (Bulut, 2019). Bu noktada yapılması gereken durum, bütün özniteliklerin yeniden ölçeklendirilip uzaklık hesaplamasında daha eşit etki etmesinin sağlanmasıdır. Yeniden ölçeklendirmede kullanılacak ilk metot “min-maks” normalizasyonu olarak isimlendirilen ve bütün özelliklere ait değerleri 0-1 aralığına indirgeyen yöntemdir. “Min-maks” normalleştirme işleminde X özelliğinin her bir değeri X

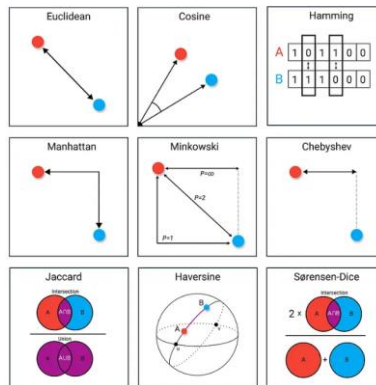
değişkeninin en küçük değerinden çıkarılmakta, bulunan sonuç X 'in değer aralığına bölünmektedir (Bulut, 2019).

En yaygın kullanılan dönüştürme yaklaşımlarından biri de z-skor standardizasyonudur. Dönüşüm şu şekilde yapılır; X özniteliğindeki her bir veri, ortalamadan çıkarılıp standart sapmaya bölünür (Lantz, 2015).

Algoritmanın sınıflandırma başarısını doğrudan etkileyen ikinci önemli durum ise uzaklık fonksiyonunun belirlenmesidir. Uzaklık formülü ile bütün özellikler için her bir gözlemin diğer gözlemlere olan uzaklıkları hesaplanmaktadır. Hangi sınıfa ait olduğu belirlenmek istenen gözlemin diğer bütün gözlemlere olan uzaklıkları bütün özellikler için hesaplanır. Eğitim veri seti ile belirlenen “ k ” hiper parametre değerine göre en yakın “ k ” tane komşu gözlem seçilir. Hangi sınıfa ait olduğu araştırılan gözlem çoğunluğu oluşturan en yakın komşu gözlemlerin ait olduğu sınıfa atanır (Lantz, 2015).

“ k ” en yakın komşu gözlem arasındaki uzaklıkları hesaplamak için farklı fonksiyonlar geliştirilmesine rağmen k-EYK algoritması eğitimi sırasında en çok tercih edilen ve R gibi programlarda varsayılan olarak kullanılan Öklid yaklaşımıdır. Öklid mesafesi iki nokta arasındaki en kısa mesafe, iki noktayı cetvelle birleştiren doğrunun uzunluğu veya kuş uçuşu uzaklık olarak tanımlanabilir (Bulut, 2019).

Öklid, en sık kullanılan uzaklık ölçümlerinden biridir. Veri setindeki özellik sayısı yüksek yani boyut sayısı çok fazla sayıda olduğu durumda, Öklid yönteminin yerine Manhattan ve Minkowski gibi farklı yöntemlerin kullanılması daha yararlı bulunmaktadır (Kawsar, 2023). Şekil 1.2.'de en sık kullanılan uzaklık ölçme yöntemleri görülmektedir.



Şekil 1.2. En sık kullanılan uzaklık ölçme yöntemleri (Grootendorst, 2021)

K-EYK büyük veri setlerinde diğer algoritmalara oranla çok daha kısa sürelerde anın fotoğrafını ortaya koyup çözümler üretmemize yardım eder. Kısa sürede doğruluk ve güvenilirlik oranı yüksek hızlı kararlar alınabilmesi için k-EYK gibi basit kolay yorumlanabilir algoritmalara ihtiyaç duyulmaktadır. Veri seti içerisindeki tahmin edilmek istenen hedef ya da bağımlı değişkenin iki ve üzeri, yani çok sınıflı (multiclass) kategorik değişken üzerinde tanımlı sınıflara veriyi en doğru bir biçimde dağıtmayı amaçlamaktadır (Arat, 2019).

Sınıflandırma problemlerinin çözümü için geliştirilen çoğu algoritma yalnızca iki sınıflı problemler için çözüm üretebilmektedir. K-EYK algoritmasının diğer bütün algoritmalara üstün olan tarafı çok sınıflı sınıflandırma problemlerinde kullanılabilir olmasıdır (Kulkarni, 2020). Çok sınıflı problemlerin çözümünde belirli bir varsayıma gerek duymaz. Her çeşit veri türünün sınıflandırmasında kullanılabilir. Veri setindeki örneklem sayısı 50.000'den az olması durumunda çok sınıflı sınıflandırma problemlerinde k-EYK algoritmasının kullanılması tavsiye edilmektedir (Arat, 2019).

1.10.2. C5.0 Karar Ağacı (KA) Algoritması

KA veri madenciliği ve makine öğreniminde regresyon ve sınıflandırma problemlerinin çözümünde kullanılan en yaygın yöntemlerdendir. En çok kullanılan güçlü KA yöntemlerinden birisi olan C5.0, Quinlan (1996) tarafından, ID3 ve C4.5 algoritmalarının geliştirilmesi ile oluşturulmuştur. Doğrusal ve doğrusal olmayan ilişkilerin modellenmesinde kullanılan ve oldukça iyi performans gösteren algoritmalar. Daha az veri hazırlığı gerektiren, anlaşılması ve yorumlanması kolay bir algoritmadır. Kategorik ve sayısal veriler kolayca işlenebilir. Büyük verilerin analizinde çok kısa süre içinde etkili ve iyi performans sağlar (Berry ve Linoff, 1997; Pang ve Gong, 2009; Tripathi vd., 2021).

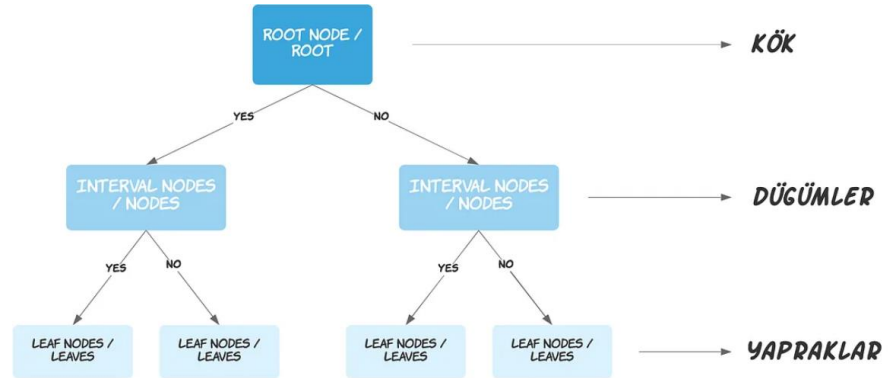
C5.0 algoritması ile oluşturulan karar ağaçları, C4.5 algoritması ile oluşturulan ağaçlar gibi tutarlıdır ve bu ağaçların tüm işlevlerini içerir. C5.0, C4.5 algoritmasından farklı olarak, artırıcı (boosted) teknoloji ve maliyet (cost) matrisi yapım teknolojisi kullanılmaktadır. C5.0 artırıcı teknolojilerin kullanılması vasıtasıyla çok daha hızlı çalışır. Yeni geliştirilen maliyet matrisi yapım teknolojisi ile daha az bellek kullanımı sağlanır ve daha az ağaç oluşturulur (Hao vd., 2022; Myint ve Tin, 2021).

KA, dallardan ve düğümlerden oluşan ters çevrilmiş ağaca benzer. En üstteki düğüme kök düğüm adı verilir. İç düğüm (parent), alt iki dala ayrılırken, yaprak (terminal) düğüm en

son noktadır ve hiç alt düğümü yoktur. Bir KA'nın kökü ve her bir iç düğümü bir soruyla etiketlenmiş karar noktalarıdır. Bölünmeye tüm gözlemleri barındıran kök (root) düğümünden başlanır. “Yukarıdan-Aşağıya” tekniği ile genellikle açgözlü veya iştahlı (greedy) yaklaşımla yapılır (Akman, 2010; Krishnan, 2021; Larose ve Larose, 2015).

İç düğüm, bir testi veya niteliği temsil eder ve tekrar tekrar bölündüğü her bir dal, bir testin sonucunu temsil eder. Her terminal düğüm veya yaprak bir sınıfı tutar. C5.0, her düğümdeki en iyi özelliği seçen yukarıdan aşağıya, yinelemeli bir yöntem kullanarak veri kümesini böler. Terminal veya yaprak düğüm bir sınıf değeri taşır ve bu düğüme gelen gözlem artık o sınıfın bir elemanıdır (Krishnan, 2021; Larose ve Larose, 2015).

Şekil 1.3'te görüldüğü gibi KA, akış şeması benzeri bir yapıdır. Ağaçtaki her bir iç düğüm her bir öznitelik ile ilgili bir “testi” temsil eder. Test sonuçları dallar ile, sınıf etiketleri ise düğümler ile temsil edilmektedir. Kök ile son düğüm arasındaki yollar sınıflandırma kurallarını temsil etmektedir (Saner vd., 2019).



Şekil 1.3. Kökten son düğüme kadar giden yolları gösteren KA yapısı (Akca, 2020)

Kurulma biçimleri KA'yı birbirinden ayırır. Soldan sağa ve yukarıdan aşağıya doğru kurulabilmektedir (Krishnan, 2021). KA yaprak odaklı (leaf-wise) ve seviye odaklı (level-wise) olmak üzere eğitim için iki strateji kullanmaktadır (Muratlar, 2020b).

C5.0 avantajları; ölçeklenebilir ve verimlidir, hem sürekli hem de kategorik özellikleri ele alır, gürültülü verilere karşı nispeten dayanıklıdır, anlaşılması ve yorumlanması genellikle basittir. Sürekli özellikleri daha iyi yönetir, bellek tüketimini en aza indirmek için verimli bellek tüketimi yapar, gelişmiş budama yöntemlerini kullanır ve olasılık tahminleri yapabilir. Dezavantajları ise; dikkatli parametre seçimi gerektirir, verilerdeki eksik

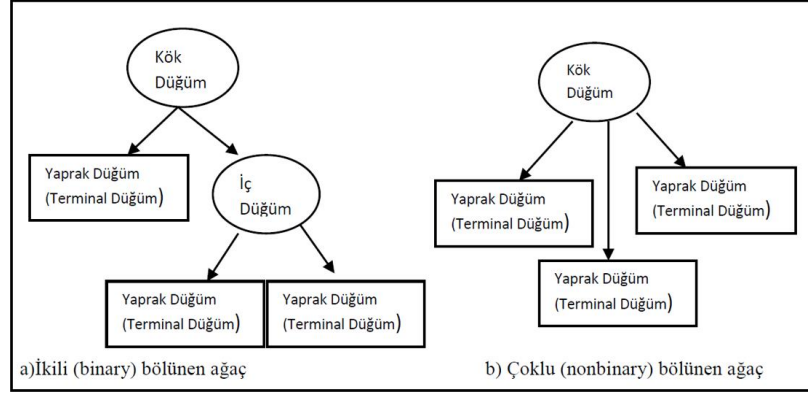
değerlerden etkilenebilir, karmaşık doğrusal olmayan ilişkiler için iyi modeller geliştiremez (Pang ve Gong, 2009).

1.10.2.1. En Uygun Bölünme Nasıl Seçilir?

En uygun (optimum) bölünmenin seçilme aşaması, sınıflandırma ve tahmin sonuçlarını etkilediği için C5.0 algoritmasının işleyişinde çok önemlidir. C5.0 modeli bilgi teorisine dayanarak oluşturur. Bu yaklaşımla aşağıya doğru dallara ayırma yaparken en iyi bölünmeyi sağlayacak değişkeni seçmek için “Entropi” ve “Bilgi kazancı” kriterleri kullanılmaktadır (Özgür, 2020).

C5.0 algoritmasında, bir veri setinin belirsizliği veya öngörülemezliği entropi ile ölçülür. Entropi, veri kümelerindeki düzensizlik veya safsızlık ölçüsü olarak kullanılır. Entropi değeri ne kadar yüksek ise herhangi bir sonuca ulaşmak da o kadar zor olur. Entropi büyük ise veri kümesi çok karışıktır ve bölme işlemi gerekir. Her bölünmede veriler daha homojen hale gelir ve bu da entropiyi azaltır. Ağaçlar yaprak düğümüne (terminale) ulaştığında maksimum saflık elde edilmiş olur. Alt düğümler çoğaldıkça homojenlik artar yani düğümlerin entropisi azalır (Pang ve Gong, 2009).

Bilgi kazancı, verideki safsızlığı belirten entropideki beklenen azalmayı ölçen bir ölçüttür. En düşük safsızlığı ve en yüksek bilgi kazanımını sağlayan özellik, kök düğüm olarak kabul edilir. Bilgi kazancı, ağacın oluşum aşamasının her adımında aşağıya doğru bölünmenin hangi özellikten başlayacağı kararını vermek için kullanılır. Daha fazla bilgi kazancı sağlayan bir öznitelik daha bilgilendiricidir ve veri belirsizliğini başarılı bir şekilde azaltabilir. Bilgi kazancı şu şekilde hesaplanır; veri seti alt bölümlere ayrılmadan önce entropisi hesaplanır hesaplanan bu değer ile sonrasında her bir alt bölüm için bulunan entropi değeri arasındaki fark hangi öznitelik için en yüksek ise en iyi dallara ayırma kriteri olarak seçilir. C5.0 algoritması, her özellik için tüm potansiyel bölünmeleri değerlendirdikten sonra bilgi kazanımını optimize eden bölünmeyi belirler (Pang ve Gong, 2009). İki ve çok sınıflı KA bölünmesini gösteren gösterim şekli, Şekil 1.4.’te verilmiştir.



Şekil 1.4. İki ve çok sınıflı KA bölünmesini gösteren şekil (Akman, 2010)

KA en fazla bilgi kazancını sağlayan alana göre bölünür. Yinelemeli olarak bu yöntem, en yüksek bilgi kazanımını sağlayan alana bağlı olarak ilk bölmeyle belirlenen her alt örneği böler. Her bir düğümdeki bölünme için gerekli olan en az veri sayısı ile ağaç derinliğinin miktarı gibi belirli durdurma kuralı veya hiper parametreler bölünmenin ne kadar devam edeceğini belirler. Bu işlem durdurma gereksinimi karşılanıncaya kadar tekrarlanır. Aşırı uyumu önlemek ve yeni verilere genellemeyi geliştirmek için budama mekanizmaları C5.0'a dahil edilmiştir (Akman, 2010; Pang ve Gong, 2009).

1.10.2.2. Budama (Pruning)

Budama, alt ağaçların kaldırılması ve gereksiz karşılaştırmaların çıkartılması işlemidir. Önbudama (pre-pruning) ve sonradan yapılan budama (post-pruning) olmak üzere iki yöntem ile budama işlemi yapılabilmektedir. İlk yöntem olan budama öncesi yaklaşımda, ağacın bölünmesi belirli bir düğümde durdurulur. İkinci yöntem ise budama sonrası yaklaşım olarak bilinir oluşturulan bütün ağaçtan alt ağaçlar çıkarılır. Bir düğümde bir alt ağaç budanır. Bir düğümdeki dallar çıkarılarak yerine bir yaprak düğüm yerleştirilebilir. Çoğu algoritma genellikle sonradan budama yöntemini kullanmaktadır (Akman, 2010). Budama, karar ağaçlarının sınıflandırma tahmin performansını artırır. Anlaşılması kolay, daha az karışık ve daha küçük ağaçlar elde edilir.

C5.0 algoritması maliyet karmaşıklığı budama stratejisini kullanarak KA'ya ait hata oranı ve karmaşıklığı arasında bir uzlaşa sağlar. Kullanıcı tarafından tanımlanan bir güven faktörüyle bir şubenin maliyet karmaşıklığını karşılaştırıp onu budayıp budamama kararını verir. Maliyet-karışıklık oranı güven faktöründen düşükse dal kesilir. Bu durumun istatistiksel olarak anlamlı olup olmadığı kararını verebilmek için C5.0 algoritması ayrıca

bir istatistiksel test kullanır. Hata oranlarında istatistiksel olarak anlamlı bir fark yoksa dallanma kesilir (Pang ve Gong, 2009).

Ağacın büyümesi tamamlandığında, gürültülü veya aykırı değerlerin varlığı karar ağaçlarında anormalliklere yol açabilir ve bunun sonucunda verilere aşırı uyum gösterebilir. Aşırı öğrenme durumunda, gürültülü veriler veya aşırı değerler (outliers) modelin hatalı sonuçlar üretmesine sebep olmaktadır. Bunu ortadan kaldırmak için ayıklama adı verilen bir yöntem kullanılır (Pang ve Gong, 2009).

1.10.3. Topluluk (Ensemble) Öğrenme Algoritmaları

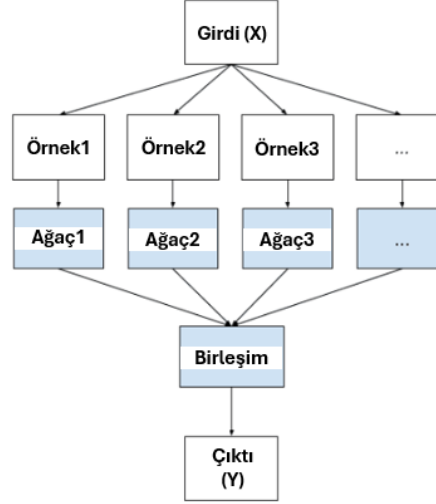
Topluluk öğrenme yöntemleri, farklı birleştirme teknikleri kullanarak tek tek yapılan doğruluk oranları düşük tahminlerin birleştirilmesi ile meta-modeller oluşturan öğrenme algoritmalarıdır (Burkov, 2019). Torbalama (Bagging), Artırma (Boosting) ve Yığılma (Stacking) olmak üzere üç farklı birleştirme metodolojisi bulunmaktadır. Bunların her biri hem bir MÖ algoritması iken hem de günümüzde kullanımı popüler hale gelmiş RO ve GAM ve diğer algoritmaların temelini oluştururlar.

Topluluk öğrenmesi, zayıf öğrencilerin bir araya getirilmesi ile oluşturulan güçlü öğrenciler vasıtasıyla yapılır. Tek bir KA'nın yaptığı sınıflandırma tahminlerinin doğruluğunu artırmak için tek karar vericilerin bir araya getirilmesi ile oluşur. Tahmin varyansı ve yanlılığı azaltma amacı ile kullanılan yaklaşımlardır. Kaggle türü yarış platformlarında bir dizi sınıflandırma ve tahmin probleminin çözümünde farklı algoritmalar yarışdırılmaktadır. Bu yarışlarda topluluk algoritmaları hızından ve doğruluk oranının yüksekliğinden dolayı en çok tercih edilen algoritmalar olmuştur (Ceylan, 2018).

1.10.3.1. Torbalama Algoritması (Bagging)

Torbalama (Bagging) yöntemi Breiman (1996) tarafından geliştirilen ismini “bootstrap aggregation” ifadesinin kısaltılmasından alan topluluk öğrenmesi yapan bir MÖ algoritmasıdır. Basit bir algortmadan daha çok kendisinden sonra gelecek algoritmalar için güçlü bir metodoloji ortaya koyan bir yöntemdir. Algoritma eğitimi sırasında önyükleme (Bootstrap) örneklem seçimi ile veri setinden her seferinde yerine koyarak farklı örnekler seçilir. Yeni veri setleri ile birbirinden bağımsız sınıflandırma tahminleri yapan bağımsız karar

ağaçlarının birleştirilmesi ile oluşturulur. Torbalama algoritması eğitim aşamasına ait şekil Şekil 1.5.'te verilmiştir.



Şekil 1.5. Torbalama algoritması eğitim aşaması görsel gösterim şekli (Brownlee, 2021)

Torbalama algoritmalarının asıl amacı tek bir öğrenici algoritmadan (genellikle KA) elde edilen modelin performansını artırmaktır. Bu yaklaşım ile aşırı öğrenme sorunlarına çözüm bulunabilmektedir. Yüksek eğitim varyansına sahip tek bir öğrenici algoritmanın yani KA'nın diğer karar ağaçları ile torbalanarak biraraya getirilmesi sonucunda eğitim varyansı düşer ve nihai model dengelenir. Bunun sonucunda tahmin doğruluğu (accuracy) artar. Sınıflandırma problemlerinde sınıf tahmini yapılan yeni örnek en çok oyu alan yani her bir bağımsız KA tarafından tahmin edilen çoğunluk sınıfa atanır (Keskin, 2019).

1.10.3.2. Rastgele Orman (RO) (Random Forest, RF)

Breiman (2001) tarafından geliştirilen Rastgele Orman (RO) algoritması, tekil karar ağaçlarının tek tek verdikleri kararları bir araya getirerek sınıflandırma ve regresyon problemlerinin çözümünde kullanılan topluluk öğrenme yöntemidir. RO algoritmasının tekil ağaçları oluşturma şekli Breiman vd. (1984) tarafından geliştirilen Sınıflandırma ve Regresyon Karar Ağaçları (SRKA) (Classification and Regression Trees, CART) ile benzerlik göstermektedir. RO algoritması iki bileşenden oluşur; torbalama ve rastgele alt uzay (RAU) (Random Subspace, RS) yöntemleri. Algoritmanın eğitimi sırasında kullanılan torbalama (bagging) metodolojisine öznitelik torbalama olarak da adlandırılan RAU yöntemi ile rastgele öznitelik seçimi yöntemi de eklenmiştir.

Torbalama metodolojisi zayıf öğrencilerin birbirinden bağımsız olarak aldıkları kararların bir araya getirilmesi ile oluşan güçlü öğrencilerin oluşturduğu topluluk öğrenme yöntemidir. Birden çok tekil KA'dan elde edilen tahminler kullanıldığı için RO algoritması sınıflandırma tahminine ait doğruluk oranları daha yüksek ve başarılıdır. Böylece aşırı uyum (overfitting) engellenmiş olur. Sınıflandırma sırasında, kuralı durdurma ya da budama işlemleri yapılmaz (Akman, 2010).

SRKA algoritmasının KA oluşturma metodolojisi kullanılarak önyükleme (bootstrap) örneklem seçimi yöntemi ile elde edilen alt örneklerden karar ormanındaki her bir KA oluşturulur. Her bir ağaç muhtemel en geniş oranda geliştirilir. Ağaç oluşumu sırasında dallara ayrılma kriterini belirleyen ve en iyi bölünmeyi sağlayacak olan öznitelige bilgi kazancı değeri ile karar verilir. Uygun sınıflandırma yapabilmek için dallanmayı sağlayacak özniteligin kesim (cut-off) noktası Gini katsayısı ile belirlenir (Akman, 2010; Keskin, 2019).

RO yönteminde her bir KA oluşturulurken Bradley Efron (1979) tarafından geliştirilen önyükleme örneklem seçimi yöntemi kullanılmıştır. RAU yöntemi ile az sayıda öznitelik rastgele seçilip ağacın her bir düğümünde en iyi dallara ayıran öznitelik ile birbirinden bağımsız ağaçlar ile karar ormanı oluşturulur (Akman, 2010).

RO algoritması ile sınıflandırma tahmininde kullanılan özniteliklerin önem derecesi de hesaplanmaktadır. KA oluşturulurken, bütün ağaçlardaki çeşitli hesaplamalar yapılarak her seviyedeki öznitelik belirlenir. Sonra diğer ağaçlardaki öznitelikler birleştirilir. En çok kullanılan öznitelik için önem derecesi hesaplanır. Önem derecesi yüksek olan öznitelikler belirlenip bunlar ile yeni modelin kurulması halinde daha etkin tahminler elde edilebilir (Akman, 2010).

MÖ algoritmaları ile sınıflandırma tahmin modelleri kurulurken, algoritmanın eğitimi ve elde edilen modelin doğruluğunun testi için genellikle veri seti eğitim ve test verisi olarak ikiye bölünür. RO algoritmasında ise veri seti eğitim ve test verisi olarak ikiye bölünmeden de sınıflandırma modelinden elde edilen sınıf tahminlerinin doğruluk oranlarının testi yapılabilmektedir.

Orijinal veri seti eğitim ve test olarak ikiye ayrılmadığında RO ile sınıflandırma ağaçları oluşturma şekli aşağıda açıklanmıştır.

1. İlk olarak veri seti içerisinde önyükleme (bootstrap) örneklem seçimi yöntemi ile n adet alt örnekler elde edilir.
2. Her alt örneklemin $2/3$ 'ü torba içi (Tİ) (In-Bag, inBag) yani eğitim verisi olarak kullanılarak karar ormanındaki her bir KA oluşturulur.
3. Tİ veri setinden de önyükleme (bootstrap) örneklem seçimi yöntemi ile alt örneklem seçilip sınıflandırma ağaçları budanmadan şu şekilde oluşturulur;
 - a. Her düğümde p (toplam öznitelik sayısı) kadar öznitelik yerine rastgele seçilen m tane öznitelik kullanılır. “ m ” değerinin $\sqrt{p}$ olarak alınması tavsiye edilir. En iyi bölünme ve en çok bilgi kazancını sağlayacak öznitelik ile dallara ayırma işlemi başlar.
 - b. En iyi dallanma kriterini bulabilmek için sınıflandırma tahmin kararını verecek değişkenin alt dallara ayıracak Gini indeks değeri hesaplanır.
 - c. Madde a ve b'deki işlemler yaprak düğüme ulaşıncaya kadar her düğümde yinelenir.
4. Eğitim veri seti ile oluşturulan KA üzerinde yukarıdan aşağıya doğru yöntemi ile test veri setindeki bütün veriler tek tek yerleştirilir. Her bir veri var olan sınıflardan birine atanır bu işlem bütün veriler için tekrarlanır.
5. Bütün KA tahminleri birleştirilip yeni tahminde bulunulur. En çok oyu alan tahmin sınıfı sonuç tahmini olur (Akman, 2010).

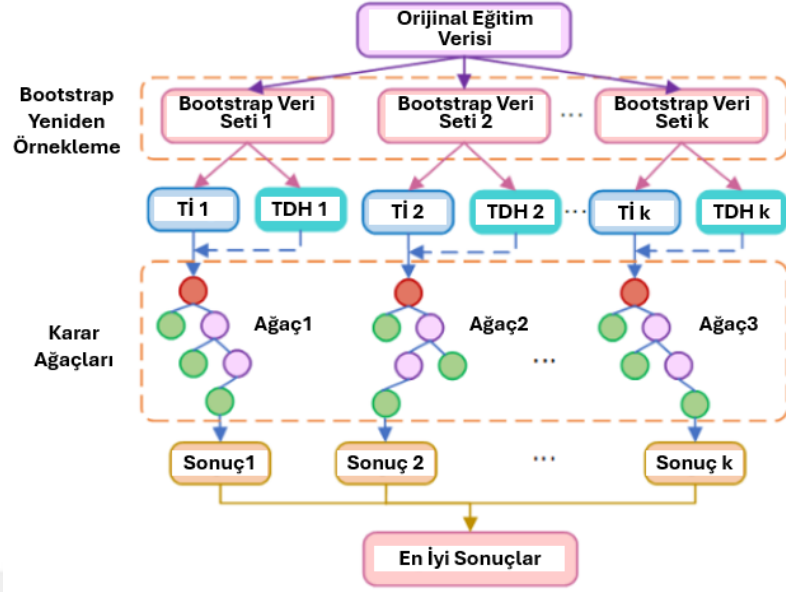
Orijinal veri setinin eğitim ve test veri seti olarak ikiye ayrılması durumunda RO algoritması sınıflandırma ağaçları şu şekilde oluşturulur;

1. İlk olarak %75-%25 ve %80-%20 gibi oranlarda eğitim ve test veri seti şeklinde orijinal veri seti bölünür.
2. Eğitim veri setinden sınıf dağılımı bozulmadan önyükleme örneklem seçimi yöntemi ile n adet alt örnek elde edilir. Her alt örneklemin $2/3$ 'ü eğitim verisi (Tİ) olarak kullanılır.

3. Yukardaki 3. maddenin a, b ve c uygulamaları tekrarlanarak karar ormanındaki her bir KA oluşturulur.
4. Geriye kalan 1/3'lük kısım ile modelin iç hata oranı olan torbadan çıkma tahmini olarak da adlandırılan torba dışı hata (TDH) (Out of Bag, OOB) hesaplanır. TDH oranları yardımıyla modelin düzgün kurulması denetlenir. Her ağaca verilecek ağırlık değeri belirlenir.
5. Önceden hesaplanmış TDH oranları her bir KA'ya verilecek ağırlık değerini belirler. En düşük TDH oranlı ağaca yüksek ağırlık, TDH oranı yüksek ağaca düşük ağırlık verilir.
6. Ağaçlar ağırlıklarına göre yaptıkları sınıf tahminleri için bir oy verme işlemine tabi tutulur. Ağaçların oyu oluştuktan sonra RO algoritmasında bu ağırlıklı oylar toplanır. Oluşturulan sınıflandırma ağaçlarından ağırlıklı olarak en çok oyu alan sınıf nihai sınıf tahmini olarak belirlenir (Akman, 2010).

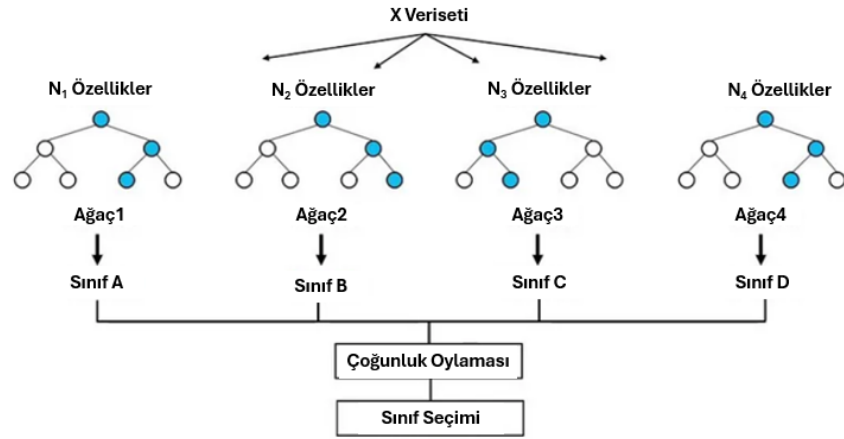
Kurulan karar ağaçlarının her biri için ayrılan TDH verisi ile oluşturulan tahmin modelinin testi yapıp tahmini hata oranı belirlenir. Karar ormanını oluşturan bütün karar ağaçları için elde edilen TDH oranları ortalaması alınıp karar ormanına ait TDH oranı kestirimi hesaplanır. TDH verisinin kullanıldığı test modelin iç testi olarak bilinmektedir.

Test veri seti ile model için elde edilen tahmini hata oranı TDH oranına benzer değer almaktadır. Eğer yeteri kadar KA oluşturulmuşsa TDH oranının sapmasız olarak doğru hesaplanabileceği ortaya konmuştur (Akman, 2010). Şekil 1.6.'da karar ormanı oluştururken kullanılan orijinal veri setinin "Tİ" ve "TDH" olarak bölünme şekli gösterilmiştir.



Şekil 1.6. RO sınıflandırma algoritmasında orijinal veri setinin “Tİ” ve “TDH” olarak bölünmesi ile oluşan çalışma şekli gösterimi (Wang vd., 2023)

Uygulaması oldukça kolay olan bu algoritma, oluşturulan karar ağaçları üzerinde budama işlemi uygulamaz. Algoritmanın eğitimi sırasında belirlenmesi gereken iki hiper parametre bulunur, bunlar “ntree” isimli ağaç sayısı ile “mtry” isimli rastgele seçilecek öz nitelik sayısıdır. Seçilen bu parametrelerin sonuca etkisi genellikle azdır (Akman, 2010). Şekil 1.7.’de RO sınıflandırma algoritmasının çalışma şekli görülmektedir.



Şekil 1.7. RO sınıflandırma algoritmasının çalışma şekli gösterimi (Khushvaktov, 2023)

Algoritma eğitime başlamadan önce “mtry” ismi ile bilinen hiper parametre değeri belirlenir ve bu sayı kadar öz nitelik rastgele seçilir. Bu sayede, ağaçlar gereğinden fazla büyümmez, aşırı öğrenme (overfitting) sorunu olmaz ve budamaya ihtiyaç duyulmaz.

- Her düğümde, dallara ayrılmaya en yüksek bilgi kazancına sahip öznitelikten başlar. Dallara ayrılma kriterini belirleyen öznitelik ile her düğümden aşağıya doğru hedef değişkenin sınıf sayısı kadar yani iki veya daha fazla dal oluşturulur. Gini indeksi ile özniteliğin hangi değerinin dallanmayı yapacağına karar verilir. Bu işlem her düğüm için oluşturulacak yeni bir dal kalmayıncaya kadar tekrar eder (Akman, 2010).
- Karar ormanı ile yeni bir verinin sınıflandırması yapılırken, her ağaç ile elde edilen sınıflardan biri için ağaç ağırlıklı olarak oy verir. Tüm ağaçların yaptığı tahminler bir araya getirilir ve ağırlıklı olarak en çok oyu alan sınıf seçilerek rastgele ormanın veri setindeki her veri için sınıf tahmini yapılmış olur (Akman, 2010).

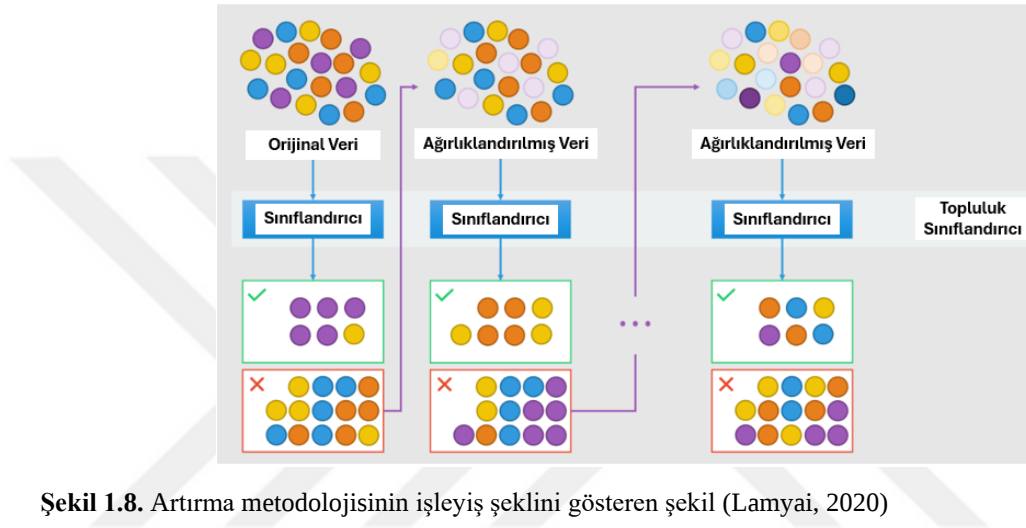
RO, çok boyutlu ve karmaşık veri setlerinde oldukça iyi çözümler vermektedir. Bu yöntem; değişken önemliliği, genelleştirilmiş hata ve gözlemler arası ilişkiyi gösteren uzaklık (proximity) matrisini hesaplamaktadır. Bu uzaklık matrisini kullanarak geliştirdiği algoritma ile veri setinde yer alan kayıp değerli gözlemlere atama yapmaktadır (Yılmaz, 2014).

Değişken önem derecesi ile model oluşumunda katkısı fazla olmayan öznitelikler modelden çıkartılabilir ve model indirgemesi yapılabilir. Karar ormanını oluşturan her bir ağaç benzersizdir yani birbirinden farklıdır. Karar ormanı içindeki ağaçların çeşitliliği yani ağaçlar arasındaki korelasyonun derecesi RO algoritmasının performansını etkiler. Korelasyon arttıkça toplam performans düşer. Ağaçlar arasındaki korelasyonu, rastgele seçilmiş özniteliklerden her düğümde farklı özniteliğin ağaç oluşumunda kullanılması ile düşürmek mümkündür. Bunun sonucunda ağaçlar daha güçlü ve sınıflandırma performansı daha yüksek olur. Çeşitliliği ve farklılığı daha çok ağaçlar oluşturularak, karar ormanı sınıflandırma tahminlerinde daha etkili kararlar alınması sağlanabilir (Akman, 2010).

Tekil ağaçlar oluşturulduktan sonra, sınıflandırma tahmini yapılacak veriler ile TDH ve Tİ değerleri yukarıdan aşağıya doğru ağaçta yerleştirilir. Aynı yaprak düğümde (terminal düğüm) sonlanan gözlemlerin sayıları kaydedilip yakınlık derecelerini gösteren uzaklık (proximity) matrisi oluşturulur. Uzaklık değeri, aşırı değer (outlier) tespitinde ve eksik değerlerin tahmininde de kullanılabilir (Akman, 2010).

1.10.3.3. Artırma (Boosting) Algoritması

İlk kanıtlanabilir ve etkili artırma algoritması Robert E. Schapire (1990) ve Yoav Freund (1995) tarafından sunulmuştur. Topluluktaki zayıf sınıflandırıcıların veya tahmin edicilerin bir araya getirilmesi ile güçlü bir sınıflandırıcı ya da tahmin edici oluşturmayı amaçlayan zayıf öğrenme varsayımına dayanan aşamalı bir yaklaşımdır (Polikar, 2012). Şekil 1.8.'de artırma metodolojisinin işleyiş şekli verilmiştir.



Şekil 1.8. Artırma metodolojisinin işleyiş şeklini gösteren şekil (Lamyai, 2020)

Artırma algoritması eğitim veri seti ile zayıf öğrencileri kullanarak aşamalı bir şekilde çoklu tahmin modelleri oluşturan topluluk öğrenme metodudur (Burkov, 2019).

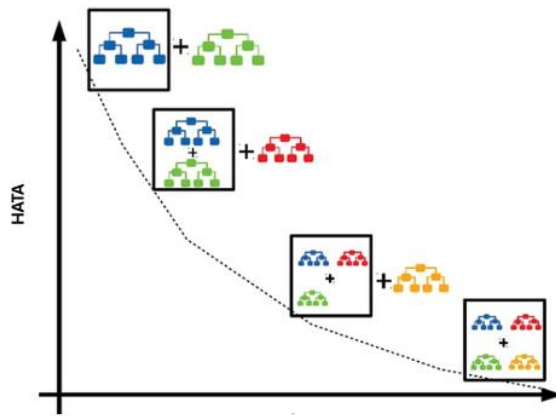
Bu algoritma, hata en iyilemesine (optimizasyonu) dayalı performans artırıcı MÖ yöntemidir. Oluşturulan her bir model kendinden önce oluşturulan modelin tahmin hatalarını en iyilemek için kurulur (Keskin, 2019). Ağırlıklı oylama ile karar ağaçlarının sonuçlarına göre son sınıf tahmini yapılır.

En yaygın kullanılan artırma yöntemleri; Adaboost, GAM, Aşırı Gradyan Artırma Makinesi (AGAM) (Extreme Gradient Boosting Machine, XGBoost), Hafif Gradyan Artırma Makinesi (HGAM) (Light Gradient Boosted Machine, LightGBM) ve Kategorik Gradyan Artırma Makinesi'dir (KGAM) (Category Gradient Boosting Machine, CatBoost).

1.10.3.4. Gradyan Artırıcı Makine (GAM) (Gradient Boosting Machine, GBM) Algoritması

Friedman (2001) tarafından geliştirilen bu algoritma sıklıkla kullanılan popüler bir makine öğrenme tekniğidir. Topluluk öğrenmesi yapan GAM topluluktaki karar ağaçlarını zayıf öğreniciler olarak kullanarak artırma yöntemi ile güçlü öğreniciler ya da tahmin ediciler oluşturan algoritmadır. Topluluk öğrenme yöntemleri ailesi içinde sınıflandırılan bu algoritma, zayıf öğrenme varsayımına dayanan aşamalı bir yaklaşımdır. Ağaçlar birbirine bağlı olduğu için seri bir şekilde eğitilir. Seri içerisindeki her bir zayıf öğrenici ağaç modeli serideki bir önceki modelin tahmin veya sınıflandırma hatasını en aza indirerek birleştiren ve güçlü bir tahmin modeli oluşturan topluluk öğrenme yöntemidir (Muratlar, 2020a).

Bu algoritma, hata en iyilemesine dayanan performans artırıcı MÖ yöntemidir (Polikar, 2012). GAM algoritması, uluslararası düzeyde yapılan çeşitli makine öğrenimi yarışmalarında (Kaggle gibi) ve yapılan araştırmalarda çokça kullanılarak popülerlik kazanmış güçlü bir topluluk öğrenme tekniğidir. Yüksek tahmin doğruluğu sağlayan ve özellikler arasındaki karmaşık ilişkileri hızlı ve doğru bir şekilde yöneten, esnek ve aşırı uyuma karşı dayanıklı bir algoritmadır (Kılıç, 2023). GAM, sınıflandırma ve regresyon görevlerinde kullanılabilen makine öğrenme tekniğidir (Currie, 2024). Bu algoritmadaki “Gradyan” kelimesi hataları en aza indirmek ve genel modelin performansını artırmak için kullanılan meyilli veya gradyan inişi (Mİ, gradient descent) optimizasyon tekniğinden gelmektedir. Şekil 1.9.’da meyilli (Gradyan) inişin gösterim şekli görülmektedir (Günay, 2023).



Şekil 1.9. Meyilli (Gradyan) inişin gösterim şekli (Günay, 2023)

GAM algoritması özellikle Mİ kavramını temel alır. Mİ optimizasyon problemlerinde fonksiyonlar için minimum noktayı bulmayı sağlayan iteratif bir yöntemdir. GAM algoritması

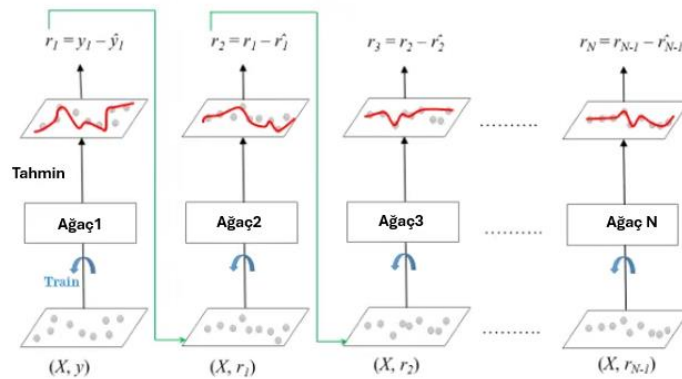
da Mİ yönteminin çalışmasını temel olarak eğitim sürecini tamamlar. Bir fonksiyonun minimumunu elde etmek için kullanılan bu teknik ile GAM algoritmasının işleyişinde hata teriminin minimumu aranır ve bulunan yeni tahminciler öncekilerin hatalarını düzeltmek için uğraşır (Öztemel, 2023).

MÖ ve veri analizi alanlarında oldukça yaygın olarak kullanılan bir yöntemdir. Mİ, her bir algoritma için farklı kayıp ya da maliyet fonksiyonunun en küçüklenmesi için kullanılır. Bu fonksiyonun ortaya koyduğu hataları ya da artıkları (tahminlerin gerçek değerlerden sapması) en küçük hale getirmek için Mİ tekniği devreye girer (Öztemel, 2023).

Bu yöntemin işleyiş şeklini basitçe açıklamak gerekirse; ilk olarak başlangıç noktası belirlenip fonksiyonun bu noktadaki eğimi (gradyanı) hesaplanır. Fonksiyonun en hızlı arttığı yönü gösteren bu eğim sayesinde atacağımız ileri adımın yönü belirlenir ve eğimin tersi yönde bir adım atılır. Bunun sonucunda, hata fonksiyonunun değeri azalır. Bu süreç her bir noktada tekrarlanır eğim hesaplanıp tekrar adım atılır. Belirli bir durma kriteri örneğin, belirlenen bir hata eşiğine ulaşıncaya kadar bu adımlar tekrarlanır (Öztemel, 2023).

1.10.3.4.1. GAM Nasıl Çalışır?

GAM, zayıf öğrenenleri (çoğunlukla karar ağaçlarını) tek bir güçlü öğrenicide birleştirip yeni modelleri önceki modellerin artıklarına (hatalarına) yinelemeli olarak uydurur ve böylece genel tahmin hatasını azaltmış olur. GAM algoritması çalışma adımları Şekil 1.10.'da gösterilmiştir.



Şekil 1.10. GAM algoritması çalışma adımları (Kılıç, 2023)

GAM algoritması çalışma adımları şu şekildedir;

1. **İlk Modelin Oluşumu:** GAM eğitime KA gibi ilk zayıf öğrenici ya da temel modelin oluşturulmasıyla başlar. Bu ilk model, doğru sınıflandırma tahminlerinde bulunmak için eğitim veri seti üzerinde eğitilir. İlk GAM modeli yani ilk KA yaprağı genellikle hedef değişkenin ortalaması olan sabit bir değerdir ve tüm veri noktaları için ilk tahmin görevini görmektedir (Kılıç, 2023).
2. **Hata veya Artık Hesabı:** Gerçek değerler ile başlangıç modelinden elde edilen tahminler arasındaki fark hata veya artık olarak hesaplanır. Her yinelemede mevcut model artıkları (gerçek ve tahmin değerleri arası fark) hesaplanır. Oluşturulan ilk ağacın ilk yaprağı hataları tahmin eden bir model olarak kurulacaktır. Hataları hesaplamak için en iyilenecek (Optimize) bir “Kayıp Fonksiyonu” kullanılır. Kayıp fonksiyonu türevlenebilir bir fonksiyondur. Sınıflandırma problemlerinin çözümünde GAM algoritmasında çoğunlukla kayıp fonksiyonu olarak logaritmik fonksiyon kullanılır.
3. **Sonraki Modellerin Oluşumu:** Bundan sonraki süreçte ikinci zayıf öğrenici bu hataları tahmin etmek için oluşturulur. Artıkları tahmin etmek için eğitilmiş olan yeni bir KA buraya sığdırılır. Bu model, ilk modelin yaptığı sınıflandırma hatalarını düzeltmek için tasarlanmıştır. İlk modelde görülen kötü performans alanlarına çözümler üretir. Her ağaç eklendiğinde oluşabilecek aşırı öğrenme sorununu çözmek için değer aralığı 0 ile 1 olan öğrenme oranı (learning rate) hiper parametresi eklenir.
4. **Ağırlıklı Kombinasyon:** Oluşturulan her bir zayıf öğrenici KA modelinden elde edilen tahminler birleştirilir. Kayıp fonksiyonu ile elde edilen kaybı en aza indireyecek önceki tahmincilerin hatalarını düzeltmeye çalışan tahmin edici ağaçları eklemek için “Katkı Modeli” olarak Mİ yaklaşımı kullanılır. Hataların minimumunu bularak tahmincilerin hatalarını düzeltmeye çalışıp güçlü bir tahmin edici oluşturmaya çalışır (Öztemel, 2023). Her model tahmini, performansına uygun bir biçimde ağırlıklandırılarak birleştirilir. Hataların azaltılmasında daha iyi performans gösteren modeller daha yüksek ağırlık ile daha kötü performans gösteren modeller ise daha düşük ağırlık ile ağırlıklandırılır. Yeni model, önceki modelin tahminlerine göre kayıp fonksiyonunun negatif gradyanına uyarlanmıştır. Bu durum aslında yeni modelin topluluk tarafından şu ana kadar yapılan hataları düzeltmeyi öğrendiği anlamına gelmektedir.

5. **Tekrarlanan Süreç:** Yukarıdaki 2-4 arasındaki işlemler önceden tanımlanmış bir performans eşiğine ulaşılan kadar belirli sayıda tekrarlanır. Öğrenme oranı ile GAM'ın her yinelemede modeli güncellediği adım boyutu belirlenir. Her yineleme ile önceki modellerden oluşan topluluğun sınıflandırma hatalarını tahmin eden yeni bir zayıf öğrenici model artıklarını tahmin etmek için eğitilir (Currie, 2024). Her yineleme ile hatayı en aza indirmek için ağırlıklar güncellenir. Yeni ağaç üretimi algoritma eğitimi sırasında ayarlanan ağaç sayısına ulaşıncaya veya ağaç eklenmesi anlamlı olmayıncaya kadar devam eder.

6. **Sonuç Tahmini:** Topluluktaki bütün modellere ait tahminler toplanır ve son tahmin yapılır. Her birinin genel performansa katkısı ağırlıklandırma oranlarını belirler. Oluşturulan tahmin modeli ile test veri setindeki yeni örneklerin sınıflandırması yapıp tahmin modelinin performansı ölçülür.

GAM algoritması eğitim veri setindeki model desenini öğrenirken oluşturduğu ağaç topluluklarının yaptığı hataları düzeltmeye odaklanır. Karar ağaçları yinelemeli olarak seri bir şekilde eklenerek sınıflandırma tahmin doğruluğu artırılır. Verinin farklı yönlerinde uzmanlaşmış iki veya daha fazla KA'nın birleşimi GAM algoritmasının verilerdeki karmaşık modelleri ve ilişkileri yakalamasına olanak sağlar. Bunun sonucu olarak regresyon ve sınıflandırma görevlerinde kullanılabilen güçlü bir makine öğrenimi tekniği elde edilmiş olur (Currie, 2024; Kılıç, 2023).

GAM algoritmasında topluluk içindeki ağaçlar açgözlü (greedy) yaklaşımla aşağıya doğru dallanarak büyür. Birçok ağaç tabanlı algoritma entropi, bilgi kazancı ve gini katsayısı gibi kriterler ile en iyi bölünme noktalarını seçer. Mevcut düğümdeki safsızlığı en aza indiren ya da bilgi kazancını üst düzeye çıkaran bölünmeler seçilir. Maksimum sayıda katman, düğüm, bölme veya yaprak düğümü sayılarının belirlenmesi ile zayıf öğrenicilerin zayıf kalması sağlanır.

GAM'ın bazı önemli hiper parametreleri ve hakkında kısa bilgiler şu şekildedir:

- **Öğrenme Oranı (learning rate):** GAM algoritmasının eğitim süreci boyunca her yinelemede modeli güncellediği adım boyutunu belirleyen çok önemli bir hiper parametredir. Küçük öğrenme oranları modelin daha yavaş yakınsamasına neden olup daha iyi genelleme yapmaya, büyük öğrenme oranları ise daha hızlı

yakınsamaya ve daha zayıf genelleme yapmaya yol açabilir. Varsayılan değer olarak 0,1 alınır. 0,01 ile 0,3 aralığında yer alır ancak en iyi değer çapraz doğrulama (k-fold cross validation) ile yapılan hiper parametre en iyilemesi yoluyla elde edilir.

- **Maksimum derinlik (max depth):** algoritma eğitimi sırasında kurulan her bir KA'nın (zayıf öğrenciler) derinliğini kontrol eder. Maksimum derinliğin büyük olması ağaçların daha karmaşık desenleri yakalaması için ideal bir ortam oluştururken aşırı uyum riskini de artırabilmektedir. Maksimum derinliğin küçük olması ise yetersiz sayıda ve daha basit ağaçların oluşmasına neden olur. Veri kümesinin büyüklüğü ve sınıflandırma probleminin karmaşıklığına bağlı olarak genellikle 3 ile 10 aralığında değerler alır. Varsayılan değeri 3'tür.
- **Tahmin Edici Sayısı (n estimators):** tahmincilerin sayısı ile topluluğa kaç adet zayıf öğrencinin dahil olacağı belirlenir. Bu hiper parametre yineleme sayısına benzetilebilir. Tahmincilerin sayısı arttırılarak genellikle model performansı arttırılabilir, ancak performans ile eğitim süresi arasında bir denge bulunur. Genellikle 50 ila 500 arasında değerler kullanılabilir fakat bu sayı veri kümesi boyutu ve karmaşıklığı ile değişebilir. Varsayılan değeri 100'dür.
- **Alt örnek:** her zayıf öğrencinin eğitimi sırasında kullanılacak eğitim verilerinin fraksiyonunu kontrol etmek için kullanılır. 1,0'dan küçük değerler eğitim sürecine rastgelelik katar ve aşırı uyumu engeller "stokastik gradyan artırma" olarak bilinir. 0,5 ile 1,0 arasındaki değerler yaygın olarak seçilir. 0,8 veya 0,5 gibi daha küçük değerler kullanılarak rastgelelik elde edilir. Varsayılan değeri 1'dir (Kılıç, 2023).

1.11. Çok Sınıflı (Multi-class) Sınıflandırma Yöntemleri

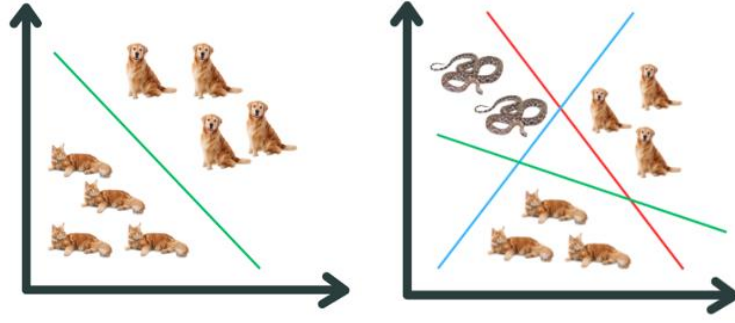
Sınıf sayısı iki olduğunda ikili sınıflandırma (0 veya 1) görevi olarak adlandırılır ve veri setindeki bütün veri noktaları sadece iki sınıftan birinde yer alır. Sınıf sayısı ikiden fazla olduğunda ise bu durum çok sınıflı sınıflandırma görevi olarak bilinir. İkili sınıflandırmada iki sınıf genellikle birbirinin tersi olduğu halde çok sınıflı sınıflandırmada bu durum MÖ için mutlaka doğru değildir. MÖ'de çok sınıflı sınıflandırma her bir veri noktasının p adet sınıftan yalnızca birine ait olacağı temel varsayımına dayanır. Veri içerisinde gözlenen model desenlerine göre çok sınıflı sınıflandırıcı algoritmalar, her bir veriyi p sayıda farklı sınıfa

ayırmak için tasarlanmıştır. Her sınıfa ait yeterli örneğin olması durumunda doğru sınıflandırma başarı oranı artar (Del Moral vd., 2022).

En az üç sınıftan birine sınıflandırma işlemi çok sınıflı sınıflandırma görevi olarak bilinir. Gerçek hayatta çok sınıflı sınıflandırma uygulamaları farklı bilim alanlarında oldukça fazla yer tutmaktadır. Çok sınıflı problemlerin en sık görüldüğü durumlar; veterinerlik alanında hayvan davranışlarının ve görüntülerin sınıflandırması, metin sınıflandırması, biyoinformatik alanında mikrodizi ve doku sınıflandırması, tıp alanında kanser ve radyoloji görüntü sınıflandırması, nesne tanıma ve bilgisayarlı görme alanları ile biyometri alanındaki yüzlerce sınıflandırma gibi her alanda her yerde bulunabilir (Del Moral vd., 2022; Galar vd., 2011; Krawczyk vd., 2015).

İkili sınıflandırma problemleriyle karşılaştırıldığında, çok sınıflı sınıflandırma problemleri her zaman daha karmaşıktır (Ding vd., 2019). Böl ve yönet stratejisi, karmaşık ve çok sınıflı veri kümelerinin çözümünde kullanılabilir. Bu strateji yardımıyla, orijinal görev her birinin çözümü daha kolay olan bir dizi alt sorunlara bölünür. Bu strateji, ayrıştırma-yeniden yapılandırma yöntemi olarakta adlandırılıp çok sınıflı sınıflandırma MÖ algoritmalarında da çok sık kullanılmaktadır (Hsu ve Lin, 2002).

Orijinal veri setinin ikili problemlere ayrıştırması en çok tercih edilen yaklaşımdır. Bu yöntem, çok sınıflı sınıflandırma problemlerini bir dizi ikili sınıflandırma problemine dönüştürür sonrasında ikili sınıflandırıcıları birleştirerek çok sınıflı sınıflandırmayı çalıştırmış olur. Bu ikili sınıflandırma problemlerine ait çözümlerin birleştirilmesi aşaması çok sınıflı problemin orijinal halinin yeniden inşası gibi çok önemlidir. Verileri ikili gruplara bölerek sorunu çözmek nispeten basit olsa da her birinden elde edilen çözümleri birleştirip yeniden oluşturmak için buna uygun bir birleştirme yöntemi bulmak zordur. En iyi çok sınıflı sınıflandırıcıyı bulmak, en iyi ayrıştırmayı bulmanın yanı sıra en iyi ikili sınıflandırıcıyı da bulmak anlamına gelir. Bu nedenle bu problemlerin çözümünde doğru ayrıştırma ve birleştirme yöntemleri kullanmak doğru tahmin başarısı yakalayabilmek için çok önemlidir (Del Moral vd., 2022; Krawczyk vd., 2015). Şekil 1.11.'de iki ve üç sınıflı sınıflandırma çözüm şekli gösterilmiştir.



Şekil 1.11. İki ve üç sınıflı sınıflandırma çözüm şekli (Trevisan, 2022)

Genellikle sınıf sayısındaki artışlar, tahmin edilen karar kurallarının karmaşıklığını da artırır ve bunun sonucunda problemlerin çözümü daha zorlu hale gelir. Bu durum, aşırı uyum olasılığını ve ihtiyaç duyulan hesaplama gücünü artırır (Krawczyk vd., 2015). Sınıf sayısının artışı ile yeni bir örneğin hangi sınıfa dahil olabileceği kararını doğru verebilmek az sınıf ile verilen karardan nispeten daha zor olmaktadır. Bunun nedenlerinden birisi, iki sınıflı sınıflandırma problemlerinde öğrenme algoritması bir karar sınırına göre sınıflandırma yapacakken üç ve üzeri sınıflı sınıflandırma problemlerinde aynı öğrenme algoritması eş zamanlı olarak iki ve ikiden fazla karar sınırı için sınıflandırma yapmaktadır (Del Moral vd., 2022).

Çok sınıflı sınıflandırma probleminin bütün zorluğunu yalnızca karar sınırlarının geometrik karmaşıklığı oluşturmaz, karar sınırlarının etkileşimi de ek zorluklar getirir. Karar sınırlarının heterojenliği kavramına dayanılarak farklı açıdan yeni çözümler getirilebilir. Yeni geliştirilen ayrıştırma teknikleri ile sınıflar ayrıştırılırken karar sınırlarının heterojenliği bilgisi doğru bir şekilde kodlanırsa daha iyi sonuçlar elde edilebilir. Farklı karar sınırlarının heterojenliğini ölçüp bu durumu hesaba katarak çok sınıflı problemi ayrıştırmak sınıflandırma performansında bir artış gösterebilir ya da modellerin eğitimi ve değerlendirilmesi aşamaları daha kısa sürede yapılabilir (Del Moral vd., 2022).

Çok sınıflı sınıflandırma problemlerinin çözümü için dört farklı yaklaşım mevcuttur. Bu yaklaşımlardan ilki, hiçbir ayrıştırma uygulamadan sorunu doğrudan çözen yaklaşımdır. Bu yaklaşımı uygulayan k-EYK ve SRKA algoritmalarıdır.

İkincisi ve en çok tercih edilen yaklaşım sınıf ikileştirme (binarization) yaklaşımıdır. Çok sınıflı bir sınıflandırma problemini ikili sınıflandırıcıları kullanarak ayrıştırmadan birden fazla yolu vardır (Del Moral, 2022). İyi bilinen üç türü şunlardır; bire karşı bir (BKB) (One-

vs-One, OvO), bire karşı hepsi (BKH) (One-vs-All, OvA) ve hata düzeltme çıkış kodlarıdır (Error-Correcting Output Codes, HDÇK). En çok bilinen ve kullanılan bu teknikler birçok uygulamaya başarılı bir şekilde uygulanmıştır (Lan vd., 2022). Çok sınıflı sınıflandırma tahmininde kullanılan MÖ algoritmaları ikili ayrıştırma yöntemlerini çoğunlukla kullanmakta olup çok sınıflı problemlerin çoğunda çok iyi performans gösterdiği görülmüştür (Krawczyk vd., 2015).

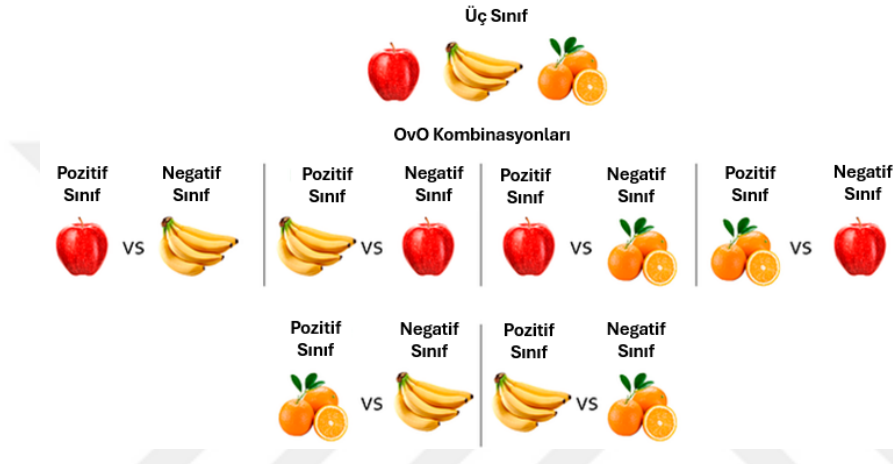
Üçüncüsü, algoritmaların çok sınıflı sorunların çözümü için tekrar ayarlanması yaklaşımıdır. Çoğu sınıflandırma algoritması ikili sınıflandırma için tasarlandığı için çok sınıflı sınıflandırma problemlerinin çözümünde kullanırken ek stratejilere ihtiyaç duyulur (Ding vd., 2019). Freund ve Schapire (1997) tarafından geliştirilen Adaboost.M1 ve Adaboost.M2, Zhu vd. (2009) tarafından Adaboost için geliştirilen SAMME (Stagewise Additive Modeling using Multi-class Exponential loss function), He vd. (2012) tarafından Destek Vektör Makineleri (Support Vector Machine, DVM) için geliştirilen (Simplified Multi-class DVM, SimMSVM) ve Tripathi vd. (2021) tarafından RO için geliştirilen (Multi-Class Random Forest, MCRF) algoritmalar örnek olarak verilebilir.

Dördüncüsü, hibrid yaklaşımdır. Polat ve Güneş (2009)'in C4.5 karar ağacı temelli BKH yaklaşımını kullanarak geliştirdiği yeni hibrid modeli, buna örnek olarak verilebilir.

Günümüzde kullanılan en popüler ikilileştirme yöntemleri BKH ve BKB'dir. BKB yöntemi olası tüm ikili sınıf kombinasyonlarını yaratırken, BKH yöntemi ise bir sınıfı pozitif sınıf olarak seçip geri kalanların hepsini negatif sınıf olarak kullanır. BKB ve BKH stratejilerinin basit ama çok güçlü yöntemler olduğu görülmüştür (Galar vd., 2011). Her iki ikili ayrıştırma yönteminde en basit toplama stratejisi olan oylama yöntemi kullanılmaktadır. BKB yönteminde her sınıf çifti için bir ikili sınıflandırıcı eğitilir ve tüm ikili sınıflandırıcıların tahmin çıktıları alınarak değerlendirilir. Nihai tahmine karar vermek için oylama yolu kullanılıp en yüksek oyu alan sınıf kazanır. Sınıflandırıcı sonuç tahminleri olasılıksal ise çıktılar farklı metotlar ile birleştirilebilir. Şekil 1.12.'de üç sınıflı sınıflandırma probleminde BKH yöntemi Şekil 1.13.'te ise BKB yöntemi çözüm şekli gösterilmiştir.



Şekil 1.12. Üç sınıflı sınıflandırma probleminde BKH yönteminin çözüm şekli (Trevisan, 2022)



Şekil 1.13. Üç sınıflı sınıflandırma probleminde BKB yönteminin çözüm şekli (Trevisan, 2022)

Son yıllarda araştırmacıların kullandığı en önemli ikilileştirme yöntemi BKH yaklaşımıdır. BKH yöntemi bir sınıfı pozitif sınıf olarak seçip sınıfların geri kalanını negatif sınıf olarak kullanarak ikili sınıflandırıcı algoritmaları eğitir. Tüm ikili sınıflandırıcılara ait tahminlerin oylamaları değerlendirilerek nihai tahmin elde edilir. Sınıflandırıcılar olasılıklı çıktılar verirse en yüksek olasılığı veren sınıflandırıcının sonucu nihai sonuç olarak alınır. Sonuçta elde edilen tahminler kazanan hepsini alır (Winners Takes All, WTA) stratejisine uygundur.

BKB ve BKH tahminlerini birleştirmek için kullanılan basit oylama yöntemlerinin kısıtları görüldüğü için ayrıştırma görevini yapacak daha karmaşık yöntemler önerilmiştir. Sınıflara ait karar sınırlarının heterojenliğinin artması BKH haricinde diğer tüm yaklaşımların performansını azaltmaktadır. Dengesiz dağılım gösteren sınıflandırma problemlerinde BKH'nın BKB'ya kıyasla daha düşük performans gösterdiği raporlanmıştır. Yapılan bir araştırmada BKH yöntemini kullanan RO algoritması çok sınıflı sınıflandırıcıdan daha iyi performans göstermiştir (Del Moral vd., 2022; Krawczyk vd., 2015).

Günümüzde çok sınıflı sınıflandırma problemlerinin çözümü için kullanılan popüler ayrıştırma yöntemlerinden bir diğeri ise HDÇK yöntemi olup, her biri bir sınıf grubunu diğlerinden ayıran bir dizi ikili sınıflandırıcıyı eğitir (Del Moral vd., 2022). HDÇK, çok sınıflı kategorizasyon problemiyle başa çıkmak için ikili sınıflandırıcıların kararlarının yeniden yapılanması yoluyla oluşturulmuş basit ve etkili bir yöntemdir. Her ikili sınıflandırıcı için benzersiz bir kod oluşturur. Çok sınıflı sınıflandırmaya yönelik bir sınıf ikilileştirme yöntemi olan HDÇK, iletişim teorisinde yer alan hata düzeltme kod iletim tekniklerinden ilham alınarak oluşturulmuştur (Krawczyk vd., 2015).

HDÇK matrisindeki her sütun bir ikili sınıflandırıcıyı eğitmek için kullanılır. HDÇK matrisi satırında benzersiz bir kod sözcüğü olması gerekirken sütunların aynı ya da tamamlayıcı olması gerekmez. Sınıflandırıcılar tarafından tahmin edilen kod sözcüğünü HDÇK kod sözcükleriyle eşleştirmek için Hamming mesafesi kullanılır. En küçük Hamming mesafesine sahip sınıf tahmin edilen sınıf olarak kabul edilir.

Bu yöntemde her sınıf için bir kod sözcüğü oluşturulur. Her sınıfın üyelik bilgisi bu kod sözcüklerinde kodlanır. Kod sözcüklerinin satırlarını oluşturduğu bir kodlama matrisi elde edilir. Bu ikili problemlerin her biri sınıf kümesini iki bölüme ayırır. Kod çözme aşamasında, n adet eğitilmiş ikili sınıflandırıcı algoritmanın uygulanması ile test veri setindeki her bir veri için bir kod elde edilir. Kodlama matrisinde tanımlanmış her bir sınıfın temel kod sözcükleriyle bu kod karşılaştırılır ve en yakın koda sahip sınıfa atanır (Krawczyk vd., 2015).

BKB ve BKH yöntemlerinde ikili sınıflandırma sayıları tam olarak belli iken HDÇK yönteminde ise her sınıf esnek sayıda ikili sınıflandırmayla kodlanabilmektedir. Daha büyük boyutlu HDÇK, hataları düzeltmek için daha fazla yer (bit) sağlarken çok daha fazla hesaplamaya mal olur (Lan vd., 2022).

HDÇK yedeklilik ve çeşitlilik üzerine kurulmuş olup birden çok sınıflandırıcı her karar sınırını öğrenir. Bunun sonucu olarak, sınıflandırıcılardan biri başarısız olsa da diğerleri doğrusya hatayı düzelteceklerdir. Bu durum, sınıflandırıcıların hataları birbiriyle ilişkili değilse ortaya çıkar. HDÇK'da her ikili sınıflandırıcı, orijinal problemin tüm sınıflarına iki etiketten birini atayarak benzersiz bir kodla tanımlanır (Del Moral vd., 2022).

Çoğu öğrenme algoritması (Adaboost ve DVM) yalnızca ikili çıktılar için çözüm önerirken SRKA, GAM ve k-EYK hem ikili hem de çok sınıflı sınıflandırma problemleri için

doğal çözümler sunabilir. K-EYK ve KA çok sınıflı sınıflandırma problemlerinin çözümünde en yaygın kullanımı olan algoritmalarıdır (Ahuja ve Yadav, 2012; Del Moral vd., 2022; Tanha vd., 2020).

1.12. Sınıf Dengesizliği Tanımı ve Çözüm Yöntemleri

Doğruluk oranları yüksek sınıflandırma tahmin modellerinin oluşumu sırasında karşılaşılan en önemli zorluklardan birisi de dengesiz veri problemidir. Dengesiz sınıf dağılımı, bir kategorik değişkenin sınıflarında bulunan gözlem sayılarının birbirlerinden çok farklı oranlarda yer alması bir sınıfın diğer sınıftan daha az yada daha fazla örneğe sahip olması durumunu ifade eder ve gerçek veri setlerinde yaygın olarak görülmektedir. Bu durumun sonucunda tahmin modellerinin performansı olumsuz yönde etkilenir, sonuçların yanıltıcı olmasına neden olur ve modelin genelleme yeteneği azalır (Demirsöz, 2024; Gong ve Kim, 2017).

Kategorik değişkenlerin sahip olduğu sınıflar arasındaki dağılımın homojen olmaması gözlem sayısının az olduğu azınlık ve gözlem sayısının çok olduğu çoğunluk sınıfları arasındaki oranların farklılığı veri setini dengesiz olarak kabul etmek için yeterlidir (Bölükbaşı, 2023; Pir, 2022).

Veri setindeki sınıflar arasındaki dengesizlik derecesi dengesizlik oranı olarak ifade edilir. Bu oran veri setinde azınlık ve çoğunluk sınıfları arasındaki dağılım farkını belirtmek için kullanılır. Dengesizlik oranı bir ise veri setleri sınıf dağılımı yönünden dengelidir ve bu durum ideal olarak kabul edilir. Bu oranın artması ile veri setinin dengesizlik derecesi o oranda artar (Demirsöz, 2024; Zhu vd., 2020).

Çoğunluk sınıfındaki gözlem sayısının azınlık sınıfındakilere bölümü ile elde edilen dengesizlik oranının sınıflar arasında ne kadar olması gerektiği hakkında birçok değer önerilmiş fakat açık ve kabul edilebilir bir kural ortaya konamamıştır (Özen vd., 2024). Önerilen oranlardan birisine örnek olarak Fernández vd. (2013) tarafından bir veri setinin dengesiz veri olarak tanımlanması için gerekli oran olan 1,50 değeri verilebilir.

Algoritmaların veri setindeki tahmin modelini doğru öğrenebilmesi için eğitim veri seti olarak ayrılan kısım ile eğitilmesi aşamasının sağlıklı bir şekilde ilerlemesi tahmin edilmeye çalışılan kategorik değişkendeki sınıf denge durumu ile yakından ilişkilidir (Talu, 2023).

Sınıf dengesizliği oranı yüksek dengesiz veri setleri kullanıldığında sınıflandırma algoritmaları ile elde edilen tahmin modellerinin eğitimi ve doğru sınıflandırma performansı olumsuz yönde etkilenir. Bu olumsuzlukların etkileri çeşitli boyutlarda görülmektedir. Bu etkilerden ilki, az sayıda örnek içeren sınıfı, sınıflandırma tahmin modeliyle doğru bir şekilde tanımlama yeteneği sınırlanır sonuçlar yanıltıcı hale gelir. Diğer olumsuz etkiler, model genelleme yeteneği düşer, yeni verilere yanıt verme kabiliyeti azalır, modelin dengeli bir şekilde eğitilmesi zorlaşır ve azınlık sınıfların önemi göz ardı edilir. Sınıf dağılımındaki homojenlik tahmin doğruluğunu arttırır (Demirsöz, 2024; He ve Garcia, 2009).

Dengesiz verinin neden olduğu olumsuz sonuçlardan uzak durabilmek, dengesiz sınıf dağılımıyla başa çıkmak ve böylece azınlık sınıfını doğru bir şekilde ayırt etmek için çok sayıda özel teknik ve strateji geliştirilmiştir (Aygün Çakıroğlu, 2024; Demirsöz, 2024; He ve Garcia, 2009). Bunlar şu şekilde sıralanabilir;

- Algoritma düzeyinde yaklaşımlar iç yaklaşımlar olarak bilinip dengesizlik problemini yeni algoritmalar oluşturarak veya mevcut algoritmaları değiştirerek aşar (Barandela vd., 2003).
- Veri düzeyindeki yaklaşımlar dış yaklaşımlar olarak bilinip dengesizlik problemini verileri önceden işleyerek problemin etkisini azaltmaya çalışır (Aygün Çakıroğlu, 2024).
- Hem iç hem de dış yaklaşımları birleştiren bir diğer yaklaşım olan maliyete duyarlı öğrenme, öğrenme aşamasında her sınıf için farklı yanlış sınıflandırma maliyetleri kullanır (Aygün Çakıroğlu, 2024).
- Topluluk öğrenmesine dayalı yöntemler, zayıf performans sergileyen bireysel modellerin birleştirilmesi yoluyla elde edilen ve yüksek performans sergileyen birleşik modeller oluşturur (Galar vd., 2012)

Literatürde sınıf dengesizliği probleminin olumsuz etkisini azaltmak için veri düzeyindeki yaklaşımlardan olan yeniden örnekleme yöntemleri sıklıkla kullanılmaktadır (Bölükbaşı, 2023). Bu yöntemler; örneklem azaltma (aşağı örnekleme), artırma (yukarı örnekleme) ve her ikisinin birlikte kullanıldığı hibrit yöntemler olmak üzere üç ana grup

altında toplanmaktadır (Krawczyk, 2016). Hangi yöntemin en iyi sonucu vereceği, veri seti özellikleri ve problemin gereksinimleri ile doğrudan ilişkilidir (Demirsöz, 2024).

Bu yöntemlerden birisi olan örneklem azaltma (aşağı örnekleme/undersampling), çoğunluk sınıfındaki fazla örneklerin çıkartılması veya alt örnekleme yöntemlerinden birisinin kullanılmasıyla gerçekleştirilir. Az örnek içeren sınıflar tahmin modellerinde daha az temsil edilir bu yüzden sınıflandırma performansı düşer bu teknik ile dengesizlik sorunu çözülerek performansın artırılması sağlanır. Örneklem azaltma yöntemlerinden bazıları; Küme Merkezleri, Tekrarlı Düzenlenen En Yakın Komşu, Tomek Linkleri ve Rastgele Örneklem Azaltma (RÖA) (Random Under Sampling, RUS) yöntemidir (Başarır, 2022).

Bunlardan, RÖA metodu en eski ve en fazla tercih edilen örneklem azaltma yöntemlerinden birisidir (Liu vd., 2008).

RÖA yöntemi modelin karmaşıklığını azaltır, veri setini dengeler, sınıflandırma sonuçlarını ve genelleme yeteneğini iyileştirir ve bu sayede azınlık sınıfının doğru sınıf tahmin performansının artmasını sağlar (Fernández vd., 2018).

RÖA tekniği ile çoğunluk sınıf gözlemlerinden rastgele bir alt örneklem seçilip azınlık sınıfındaki gözlemlerle orantılı büyüklükte seçilerek bir alt örneklem oluşturulur. Çoğunluk sınıfından seçilen yeni örneklem ile azınlık sınıfındaki gözlemler birleştirilip dengelenmiş bir veri seti elde edilir. Bu veri seti, sınıflandırma algoritmalarının model eğitiminde eğitim veri seti olarak kullanılır (Liu vd., 2008).

RÖA, basit ve hızlı şekilde uygulanabilen, azınlık sınıfındaki gözlemlerin kaybına neden olmadan sınıf dengesizliği problemini çözebilen, azınlık sınıfın tahmin performansının artmasını sağlayan ve büyük veri setlerinde etkili bir şekilde kullanılabilen etkili bir yaklaşımdır. Üç ve üzeri sınıflı hedef değişkenlerde sınıflar farklı oranlarda olduğunda doğru alt örnekleme oranı seçilirse veri setinin temsil ediciliği korunmuş olur (Fernández vd., 2018).

RÖA yöntemi dengesiz veri setlerinde azınlık sınıfı yada sınıfları dengelemek için etkili bir araç olmasına rağmen bu yöntemin bazı dezavantajları da bulunmaktadır. Rastgele seçilerek oluşturulan çoğunluk sınıfını temsil eden yeni örneklem tahmin modelinde bilgi kaybına yol açabileceği ve modelin genelleme yeteneğini azaltabileceği öngörülmektedir. Örneklem azaltılması ile veri dağılımı tam olarak dengeleyemeyebilir ve temsil ediciliği azalır

modelin performansını olumsuz etkilenebilir. Bu nedenle, uygun bir alt örnekleme oranı seçmek ve veri setinin temsil ediciliğini korumak önemlidir (Demirsöz, 2024).

Aşırı örnekleme (oversampling) yada örneklem artırma (yukarı örnekleme) yöntemi, sınıf dağılımındaki ciddi düzeydeki çarpıklığı düzeltmek, azınlık ve çoğunluk sınıfları arasındaki dengelenmeyi sağlamak için azınlık sınıflarındaki örnekleri farklı yollar ile veya sentetik örnekler üreterek çoğaltır. Tahmin modellerinde yer alan azınlık sınıfların daha fazla temsilini sağlamak amacı ile uygulanır. Oluşturulan dengeli eğitim veri setleri ile tahmin modelini öğrenen algoritmaların sınıflandırma performansı artar. Örneklem artırma yöntemlerinden bazıları; Rastgele Aşırı Örneklem (RAÖ) (Random Over Sampling, ROS), Sentetik Azınlık Örneklem Arttırma Tekniği (SAÖAT) (Synthetic Minority Oversampling Technique, SMOTE), Adaptif Sentetik Örneklem ve Rastgele Sentetik Azınlık Aşırı Örneklemedir (Chawla vd., 2002).

RAÖ yöntemi, dengesiz bir veri kümesi için çarpık sınıf dağılımını yeniden dengelemek için basit bir teknik sağlar. Azınlık sınıfındaki örnekleri rastgele üretilen yeni örnekler ile çoğaltmak için kullanılan kolay ve etkili bir yöntemdir (Zheng vd., 2015).

RAÖ tekniği ile azınlık sınıfına ait gözlemlerden rastgele bir alt örneklem seçilip azınlık sınıfındaki yada sınıflarındaki gözlemlere eklenerek çoğunluk sınıfındaki gözlem sayısına eşitlenir. Sınıf oranları dengeli hale getirilen dengelenmiş eğitim veri seti sınıflandırma algoritmalarının model eğitiminde kullanılır (Liu vd., 2008).

RAÖ, çoğunluk sınıfın tahmin modelindeki etki oranını azaltıp azınlık sınıfların modele katkısını artırarak model eğitimi sırasında algoritmaların daha dengeli ve gerçekçi tahmin modellerini ortaya çıkartmasını sağlar. Bu sayede elde edilen modellerin genelleme yeteneği artabilir ve modellerden elde edilen sınıflandırma performansları yükselebilir. Dengesiz veri setini dengelemek için kullanılan RAÖ yöntemi modelin performansını iyileştirmek amacıyla azınlık sınıfı örneklerini artırır ancak bu durum azınlık sınıfının veri setinde aşırı temsil edilmesine ve modelin bu sınıf yönünden aşırı uyum göstermesine neden olabilir. Bunun yanında yanlış sentetik veri üretimi ile tahmin model performansı olumsuz etkilenir ve sonucunda orijinal veri setinin temsil ediciliği korunamayabilir (Zheng vd., 2015).

SAÖAT veri kümelerini dengelemek için tercih edilen yaklaşımlardan birisidir. (Bölükbaşı, 2023). Bu yöntem ile azınlık sınıflarındaki veriler rastgele seçilmez veya

doğrudan kopyalanmaz bunun yerine gözlemlere ait en yakın komşular temel alınarak sentetik veriler üretilir (Chawla vd., 2002; Fernández vd., 2018).

Bu yöntem, veri setindeki azınlık sınıfını güçlendirir ve doğru temsil edilmesini sağlar. Azınlık sınıfın modele katkısı artar, modelin nadir durumları daha iyi öğrenmesi sağlanır ve modelin genelleme yeteneği yükselir.

SAÖAT ile üretilen sentetik örnekler, sınıflar arasındaki ayrımı belirsizleştirip sınıflandırma modelinin performansını olumsuz etkileyebilir. Sınıf içi varyasyon azalıp modelin belirli alt grupları doğru şekilde öğrenme yeteneği sınırlanabilir. Azınlık sınıfındaki yeni üretilen örneklerin diğer örnekler ile benzer derecesi yüksek olabilir ve bu durum da model aşırı uyum gösterebilir. Bunların sonucunda modelin genelleme yeteneği azalabilir (Demirsöz, 2024).

Hızlı ve etkili sonuçlar elde edilmesine olanak sağlayan SAÖAT büyük veri setlerinde uygulaması kolay bir yöntemdir. Ancak büyük veri setlerinde sentetik örnekler çok daha uzun sürede oluşturulur. Bunun yanında, veri işleme süreci ve işlem gücü artar, kaynak gereksinimi çoğalır. Eğer gürültü ve hatalı veriler çoğaltılır ise modelin doğruluğu ve güvenilirliği olumsuz etkilenir (Demirsöz, 2024).

Her yöntemin kendine göre avantajları ve dezavantajları mevcuttur. Çoğunluk sınıftan örnek azaltılarak azınlık sınıfına eşitlenmesi yolu ile bilgi kaybına yol açılması örneklem azaltma yönteminin dezavantajı, azınlık sınıfa örnek eklenerek çoğunluk sınıfına eşitlenmesi yoluyla elde edilen modelde aşırı uyum probleminin görülmesi aşırı örnekleme yönteminin dezavantajıdır. Örneklem azaltma yönteminde veri sayısı azaldığı için modelin eğitim süresi düşer ve bellekte kapladığı alan azalırken aşırı örnekleme yönteminde ise tam tersi durum gözlenir (Bölükbaşı, 2023).

Model performans karşılaştırmasında yaygın olarak kullanılan, hesabı ve yorumlanması kolay olan doğruluk ve hata oranının dengesiz veri setlerinden elde edilen modellerin değerlendirilmesinde bazı dezavantajlara sahip olduğu görülmüştür. Bu dezavantajlardan ilki, dengesizlik oranları yüksek veri setlerinden düşük hata oranı ve yüksek doğruluk oranı elde etmek kolaydır. İkincisi, hatalar eşit maliyetli olarak varsayılır. Son olarak, azınlık sınıflarını yanlış sınıflandırmak, çoğunluk sınıflarını yanlış olarak sınıflandırmaktan çok daha maliyetlidir (Provost ve Fawcett, 2001).

Bu nedenle sınıflar arasında bir dengesizlik olduğunda, bir sınıflandırma algoritmasının performansını ve kalitesini değerlendirmek için yaygın olarak kullanılan ölçütler şunlardır; kesinlik ve anma (recall/duyarlık) metriklerini tek bir değerde birleştirerek modelin etkinliğine ilişkin kapsamlı bir değerlendirme yapılmasını sağlayan F1 skoru, iki sınıf için duyarlılık ve seçiciliğin çarpımının karekökü alınarak hesaplanan üç ve üzeri sınıflar için ise anmanın küp kökü veya kuvvet kökleri alınarak bulunan G-ortalama değeri, sınıflandırma modelinin gerçek pozitif oranına karşı yanlış pozitif oranını gösteren bir grafik olan Alıcı İşlem Karakteristik (AİK) (Receiver Operating Characteristic, ROC) Eğrisi Altında Kalan Alan (EAKA) (Area Under Curve, AUC) ve karışıklık matrisine ait tüm bilgileri kullanan ölçüm metriği olan Matthews Korelasyon Katsayısı (MKK) (Matthews Correlation Coefficient, MCC) ve Çok Sınıflı Matthews Korelasyon Katsayısı (ÇSMKK) (Multiclass Matthews Correlation Coefficient, MMCC) değerleridir (Aygün Çakıroğlu, 2024).

Yukarıdaki bölümlerde çalışma ile ilgili genel bilgiler verilmiş olup tez çalışmasının amaçları, bilime katkı ve yöntem açısından ne gibi yenilik hedeflediği aşağıda cümlelerde belirtilmiştir.

Yapılan çalışma ile, günümüzde büyük ve karmaşık veri analizlerinde çokça tercih edilen MÖ sınıflandırma algoritmaları ile veteriner hekimliği alanında ve özellikle atlarda ilk kez bu teknikler kullanılmış, Türkiye'deki safkan İngiliz yarış atlarının yarış hayatı boyunca elde ettiği kazanca çeşitli çevresel faktörlerin etkisinin değerlendirilmesi için sınıflandırma tahmin modellerinin oluşturulması hedeflenmiştir. Yarış atlarının performans ve yarış kabiliyetinin göstergelerinden birisi olan kazanç değişkeni, genetik ve çevresel faktörlerden doğrudan ve dolaylı olarak etkilenmektedir. Yarış atlarında kazancın doğru tahmin edilmesi, yarış endüstrisinde başarıyı artırma, yetiştiricilikte genetik üstünlük sağlama ve seleksiyon süreçlerinde daha doğru kararlar alma açısından büyük bir öneme sahiptir.

Ayrıca, yapılan bu çalışma ile, iki sınıflı kazanç değişkenini içeren veri seti ile üç sınıflı kazanç değişkenini içeren RAÖ yöntemi ile dengelenmiş ve dengelenmemiş veri setlerinden öğrenilen modellerin performans ölçütlerinin karşılaştırılması ve en iyi tahmin modelinin elde edilmesi amaçlanmıştır.

2. GEREÇ VE YÖNTEM

2.1. Veri Ön İşleme Süreci

Araştırmanın materyalini, 2007-2020 yılları arasında doğan 30 Haziran 2022 tarihi itibarıyla yarış hayatı sonlanmış İngiliz yarış atlarına ait Türkiye Jokey Kulübü'nün düzenli olarak tutmuş olduğu kamuya açık kayıtlar oluşturmaktadır. Çalışmada her bir at bağımsız bir birey olarak kabul edilmiştir.

2007-2020 yılları arasında doğup düzenlenen yarışlara bir defa da olsa katılan 14.233 adet safkan İngiliz yarış atının dahil olduğu tez çalışmasında kullanılan öznitelikler, özniteliklerin adı ve açıklamaları, özniteliklerin tipleri ve tanımlayıcı bilgileri (değer aralıkları) Çizelge 2.1.'de gösterilmiştir.

Çizelge 2.1.'de atların yarışa başlama yaşı isimli YBY ile kodlanmış koşan atların yarış hayatlarına kaç yaşında (ay cinsinden) başladığını gösteren sürekli tipteki değişkenin ortalaması $30,64 \pm 4,39$ ay, koşu sayısı isimli KS ile kodlanmış koşan atların kaç yarışta start aldığını gösteren sürekli tipteki değişkenin ortalaması $17,31 \pm 18,79$ adet olarak bulunmuştur. Analize dahil edilen atların çoğunluğunun dişilerden oluştuğu (7.387, %51,90), sahiplerinin tek kişi olduğu (12.586, %88,40) ve yarış hayatı boyunca farklı tipte koşu parkurunda (KP) (10.720, %75,30) yarıştığı görülmüştür.

Çizelge 2.1. Öznitelik adları ve açıklamaları ile öznitelik tipleri ve tanımlayıcı bilgiler

Öznitelik Adı	Kod	Öznitelik Açıklaması	Tipi	Değer Aralıkları <i>A.Ort ± SS (Min.-Maks.)</i>
Yarışa Başlama Yaşı	YBY	Koşan atların yarış hayatlarına kaç yaşında (ay cinsinden) başladığı.	Nicel	$30,64 \pm 4,39$ (22-84)
Yarış Terketme Yaşı	YTY	Koşan atların yarış hayatlarının kaç yaşında (ay cinsinden) sonlandığı.	Nicel	$46,04 \pm 13,50$ (23-153)
Kariyer Süresi	KSÜ	Atların yarış hayatlarının kaç ay sürdüğü.	Nicel	$15,03 \pm 13,62$ (1-121)
Koşu Sayısı	KS	Koşan atların kaç yarışta start aldığı.	Nicel	$17,31 \pm 18,79$ (1-309)

Çizelge 2.1. Devam öznitelik adları ve açıklamaları ile öznitelik tipleri ve tanımlayıcı bilgiler

Anne Kazanç Durumu	AKD	Atların annelerinin yarış hayatları boyunca kazanma durumu.	Nitel	Kazanamadı: 1.620 (%11,40) Kazandı: 8.369 (%58,80) Yarışlara Katılmadı: 4.244 (%29,80)
Baba Kazanç Durumu	BKD	Atların babalarının yarış hayatları boyunca kazanma durumu.	Nitel	Kazanamadı: 241 (%1,70) Kazandı: 6.665 (%46,80) Yarışlara Katılmadı: 7.327 (%51,50)
Koşu Parkuru	KP	Yarış hayatı boyunca koştuğu parkur türü.	Nitel	Değişken Parkur:10.720 (%75,30) Her Daim Çim: 1.138 (%8,00) Her Daim Kum: 2.021 (%14,20) Her Daim Sentetik:354 (%2,50)
Sahiplik Durumu	SD	Koşan atların tek ya da birden fazla sahibinin olma durumu.	Nitel	Tek Sahip:12.586 (%88,40) Hisseli: 1.647 (%11,60)
Cinsiyet		Atların cinsiyeti.	Nitel	Erkek :6.846 (%48,10) Dişi :7.387 (%51,90)
Kazanç (Üç Kategorili)		Atların yarış hayatı boyunca ikramiye kazanma durumu üç sınıflı hali.	Nitel	Kazanç Yok: 8.132 (%57,10) Az Kazanç (<=5740\$): 3.052 (%21,46) Çok Kazanç (>5740\$): 3.049 (%21,44)
Kazanç (İki Kategorili)		Atların yarış hayatı boyunca ikramiye kazanma durumu iki sınıflı hali.	Nitel	Kazanç Yok: 8.132 (%57,10) Kazanç Var: 6.101 (%42,90)

Anne kazanç durumu (AKD) ve baba kazanç durumu (BKD) incelendiğinde analize dahil edilen atların çoğunluğunun annesi yarış kazanmış (8.369, %58,80) babası ise yarışlara dahil olmamıştır (7.327, %51,50).

Analize dahil edilmeyen, yarış terketme yaşı isimli YTY ile kodlanmış koşan atların yarış hayatlarının kaç yaşında (ay cinsinden) sonlandığını belirten sürekli tipteki değişkenin ortalaması 46,04±13,50 ay, en küçük değer 23 ay, en büyük değer 153 ay ve ortanca değeri 43 aydır. Kariyer süresi isimli KSÜ ile kodlanmış atların yarış hayatlarının kaç ay sürdüğünü belirten sürekli tipteki değişkenin ortalaması 15,03±13,62 ay, en küçük değer 1 ay, en büyük değer 121 ay ve ortanca değeri 12 aydır.

Çizelge 2.1.'de hedef değişken kazancın hem iki sınıflı hemde üç sınıflı haline ait tanımlayıcı bilgiler sunulmuştur. Yarış hayatı boyunca hiç kazanç elde edemeyen safkan at sayısı 8.132 (%57,10) kazananların sayısı ise toplam 6.101 (%42,90)'dir. Üç sınıflı kazanç değişkeni elde edebilmek için kazanan sınıfının iki kategoriye ayrılması gerekir. Bunun için kazanç değişkenine ait ortalama ve ortanca gibi tanımlayıcı istatistikler incelenmiştir.

Elde edilen kazancın TL bazında ortalaması $49.007,41 \pm 145.423,14$ TL, en küçük değeri 1.013 TL, en büyük değeri 6.695.450,00 TL ve ortanca değeri 12.000,00 TL'dir. Araştırmaya dahil edilen 2007-2020 arasında doğan İngiliz yarış atlarının 30 Haziran 2022 tarihine kadar yarış hayatı devam etmiş bu zaman aralığında Türk lirasındaki enflasyona bağlı değer kaybı hesaba katılarak doğru ve tutarlı sonuçlar elde etmek için kazanılan ikramiyeler Amerikan dolar (US\$) kuruna çevrilmiştir.

Merkez Bankasının yıllık ortalama kurları dikkate alınarak 2009-2022 aralığında koşu hayatı devam etmiş ve sonlanmış atların kazandığı ikramiyeler dolar kuruna çevrilmiştir. Elde edilen kazanç dolar kuruna çevrildiğinde atların kazanç ortalaması $14.380,35 \pm 37.699,85$ dolar en küçük değer 280,55 dolar en büyük değer 1.837.390,23 dolar olarak bulunmuştur. Kazanç değişkeni standart sapma değeri çok yüksek, yani dağılım çarpık olduğu için kazanamayan grup hariç diğer kazanan grup ortanca değerine (5.749,98 dolar ~5750 dolar) göre iki kategoriye ayrılmıştır. Bu durumda hedef değişken olan kazanç; hiç kazanamayan (0, $n=8.132$, %57,10), 5.750 dolar ve aşağısı (1=Az kazanç, $n=3052$, %21,46) ile 5.750 dolardan yüksek (2=Çok kazanç, $n=3049$, %21,44) olmak üzere 3 kategoriye ayrılmıştır.

Verilerin önceden işlenmesi iyi bir sınıflandırma ve tahmin performansı elde etmek için önemli bir adımdır. Veri setindeki eksik veriler tespit edilip TJK'nın düzenli olarak tuttuğu açık kayıtlardan eksikler tamamlanmıştır. Aykırı değerlerin olmadığı görülmüştür.

YBY ve KS isimleri ile kodlanmış sürekli tipteki özelliklere ait değerlerin ortak bir aralığa dönüşümünü sağlayarak algoritma öğrenme sürecine eşit katkıda bulunması için veri standardizasyonu işlemi gerçekleştirilmiştir (Öztürk, 2020).

Çalışmada kullanılan 14.233 gözlemlili ve üç sınıflı kazanç değişkenini içeren veri setindeki kazanç değişkeni çoğunluk sınıfı ile azınlık sınıfları arasında dengesizlik oranı 2,66 olarak bulunmuş olup Fernandez vd. (2013) tarafından belirtilen dengesiz veri olarak tanımlanma oranı 1,50'den yüksek bulunduğu için kazanç yönünden dengesiz dağılım problemi olduğu görülmüştür.

Çalışmada, (1) üç sınıflı kazanç değişkenini içeren dengeleme yapılmamış orjinal veri seti, (2) üç sınıflı kazanç değişkenini içeren ve rastgele aşırı örnekleme (RAÖ) yöntemi uygulanarak dengelenen veri seti ile (3) iki sınıflı kazanç değişkenini içeren veri setindeki

sınıflandırma tahmin desenlerinin ortaya çıkarılması için, k-EYK ve ağaç tabanlı algoritmaları (C5.0, RO, GAM) uygulanmıştır.

Üç sınıflı kazanç değişkenini içeren dengeleme yapılmamış veri seti ve RAÖ yöntemi ile dengelenmiş veri seti kullanılarak MÖ sınıflandırma algoritmaları ile elde edilen tahmin modellerinin performanslarının kıyaslanması için, doğruluk (accuracy), dengeli (balanced) doğruluk, seçicilik (specificity) ve negatif tahmin (Negative Predictive) değerlerinin ortalaması, kesinlik (precision), anma (recall) ve F skorunun (F-Measure, F-Ölçütü) makro ortalamaları, G-ortalama (G-Mean) ile Matthews korelasyon katsayısının (MKK) (Matthews Correlation Coefficient, MCC) üç ve üzeri sınıflar için geliştirilmiş özel halleri ve EAKA gibi metrikler ile karışıklık matrisi kullanılmıştır.

İki sınıflı kazanç değişkenini içeren veri seti kullanılarak MÖ sınıflandırma algoritmaları ile elde edilen tahmin modellerinin performanslarının kıyaslanması için, doğruluk, dengeli doğruluk, seçicilik, negatif tahmin değeri, kesinlik, anma, F1 skoru, MKK ve EAKA gibi metrikler ile karışıklık matrisi kullanılmıştır.

Sınıflandırma performansı en yüksek model belirlendikten sonra bu modelde yer alan tahmincilerin modele sundukları katkılar (önem düzeyleri) üzerine çıkarımlar yapılmıştır.

2.2. Model Performans Değerlendirme

Model performans değerlendirme yöntemleri mevcut veri setinin eğitim ve test veri seti olarak nasıl bölüneceğini, örneklerin nasıl elde edileceğini gösterir (Arat, 2018). Modellerin performansını; algoritma seçimi, sınıflardaki veri sayısı ve dağılımı etkilerken bunun yanında hangi oranlarda eğitim ve test veri seti seçildiği ile çapraz doğrulama yapıp yapılmaması da etkilemektedir (Alan, 2020).

Çalışmada, iki sınıflı ve üç sınıflı kazanç değişkenini içeren veri setleri dışarda tutma (hold-out), k-katlı çapraz doğrulama (ÇD) isimli iki farklı veri seti bölme tekniği ile iki parçaya bölünmüştür. Veri setleri dışarda tutma yüzdesel bölme yöntemi ile %70-%30, %75-%25 ve %80-%20 olmak üzere üç farklı oranda rastgele bölünmüş, k-EYK ve ağaç tabanlı (C5.0, RO, GAM) sınıflandırma algoritması ile eğitilip elde edilen sınıflandırma modeli test edilerek değerlendirilmiştir. Algoritmalarından elde edilen sınıflandırma modellerinin doğru sınıflandırma performanslarını ölçmek için k-katlı (k-fold) ÇD tekniği kullanılmıştır.

ÇD, MÖ algoritmalarının eğitim veri setinden öğrendiği modeli test etmek için kullanılan yaygın model performans değerlendirme yöntemi ve bir model doğrulama tekniğidir. Veri seti rastgele olarak k adet parçaya bölünür. “ k ” veri setinin rastgele olarak bölünme sayısını hem de model performansının sınanma sayını belirtir. En çok tercih edilen “ k ” sayıları 3, 5 ve 10’dur. K adet birbirine benzer homojen dağılım gösteren alt kümeden “ $k-1$ ” tanesi eğitim geriye kalanı test veri seti olarak kullanılır. “ $k-1$ ” adet eğitim veri setinden algoritma bir model öğrenir her bir model test veri seti ile sınanır (Karşlı, 2019; Özgür, 2020). Bu çalışmada, 10 katlı ÇD 10 tekrarlı olarak kullanılmıştır.

2.3. Hiper Parametre En İyilemesi (Tuning)

Hiper parametreler algoritmanın performansını ayarlayabilen parametrelerdir. Model eğitime başlamadan önce kullanıcı tarafından karar verilen değerlerdir. Hiper parametre en iyilemesi (tuning) modellemenin en önemli aşamalarından biridir. Hiper parametre değerlerinde yapılan en ufak değişiklikler model performansında çok büyük etkiler oluşturur (Çarkacı, 2018).

Bu çalışmada, modellerin tahmin performansını artırmak ve en iyi hiper parametrelere ulaşabilmek için eğitim verilerinde 10 kat ÇD 10 tekrarlı olarak kullanılmıştır. Izgara araştırma (grid search) yöntemi kullanılarak hiper parametre değerleri tekrar ve tekrar değiştirilip model başarısı gözlenmiş ve en iyi performansı gösteren hiper parametreler seçilmiştir.

2.4. Öznitelik Önem Dereceleri

K-EYK hariç C5.0, RO ve GAM algoritmalarının eğitimi sırasında özniteliklerin önem derecelerine ilişkin üretilen bilgiler grafik halinde sunulmuştur. RStudio programında “caret” paketinde “summary” fonksiyonu kullanılarak bu üç algoritma eğitim veri setlerinden öğrenilen sınıflandırma tahmin modelleri içerisinde yer alan özniteliklerin modele katkısı önem derecelerine göre elde edilmiştir.

2.5. Model Karşılaştırması ve Performans Değerlendirme Metrikleri

Bir sınıflandırma sürecinin sonucunda sınıflandırıcı performansını değerlendirilmek ve farklı modelleri karşılaştırmak için kullanılacak değerlendirme ölçütlerinin seçimi çok önemlidir. Farklı modelleri ve algoritmaları etkili bir şekilde karşılaştırmak için çözülmeye

çalışılan soruna ve hedeflere bağlı olarak uygun ölçüm metrikleri seçilmelidir (Chicco vd., 2021; Shah, 2023).

Aşağıda, bu çalışmada iki sınıflı sınıflandırma probleminin çözümünde kullanılan algoritmalarından elde edilen performans ölçütlerinden bu araştırmada modellerin değerlendirilmesi için tercih edilen ölçütler ve bunların üç ve üzeri sınıflar (çok sınıflı) için genişletilmiş halleri hakkında kısa bilgiler verilecektir.

Bu ölçütlerden ilki, karışıklık (hata) matrisi (confusion matrix) olup MÖ sınıflandırma modellerinin başarılı olup olmadığını değerlendirmek için performans değerlendirme kriteri olarak sıklıkla kullanılır. Matris içindeki bilgiler ile sınıflandırıcıların performanslarını değerlendirmek için sıklıkla tercih edilen; doğruluk, dengeli doğruluk, kesinlik, anma, F1 skoru, seçicilik, negatif tahmin değeri, MKK, G-ortalama ve EAKA gibi ölçüm metrikleri elde edilir. Modelin performansına ilişkin kapsamlı bir görünüm sunan karışıklık matrisi her sınıfın doğru ve yanlış tahminlerini analiz ederek modelin güçlü ve zayıf yönlerini teşhis edip performansını artırmak için kullanılabilir (Shah, 2023).

Performans metriklerinin hesaplanmasında kullanılan karışıklık matrisinin dört temel bileşeni olup bunlar; doğru pozitif (DP), yanlış pozitif (YP), yanlış negatif (YN) ve doğru negatiftir (DN). DP, gerçekte pozitif ve tahmin pozitif ise, YP gerçekte negatif fakat tahmin pozitif ise, YN gerçekte pozitif fakat tahmin negatif ise, DN gerçekte negatif iken negatif tahmin yapıldı ise oluşmaktadır. Matriste her satır tahmini sınıfı, her sütun gerçek sınıfı temsil eder. Verilerin sayısını, gerçek sınıf ve tahmin edilen sınıf oranları olarak özetleyen bir tablodur (Alan, 2020; Kundu, 2022).

Doğruluk, dengeli doğruluk, kesinlik, anma, F1 skoru, seçicilik, negatif tahmin değeri, MKK, G-ortalama ve EAKA gibi metrikler sınıflandırma algoritmaları ile elde edilen modelin performansını değerlendirirken sıkça kullanılır. İki sınıflı karışıklık matrisi içinde yalnızca DP, YP, YN ve DN bulunurken bu araştırmada üç sınıflı sınıflandırma probleminin çözümünde kullanılan BKH ikilileştirmesinde, b sınıfının pozitif diğer iki sınıfın (a ve c) negatif olarak kabul edildiği çok sınıflı bir sınıflandırıcıdan elde edilen sonuçların gösterildiği karışıklık matrisi Çizelge 2.2.'de görülmektedir.

Çizelge 2.2. Üç sınıflı karışıklık matrisi

Tahmin Değerleri	Gerçek Değerler			
	Sınıflar	a	b	c
	a	DN	YN	DN
	b	YP	DP	YP
	c	DN	YN	DN
		Toplam		
		a		
		b		
		c		

Doğruluk (Accuracy), doğru tahminlerin tüm tahminlere oranı olarak hesaplanır. İki ve çok sınıflı sınıflandırma problemlerinde çok tercih edilir. 0 ile 1 arasında değer alır. 1'e yakın olması doğruluk oranını artırmaktadır. Veri kümesi dengeliyse, yanlış pozitif ve negatiflerin maliyetleri eşitse doğruluk daha uygun bir değerlendirme ölçütüdür. İki ve çok sınıflı hallerine ait formül aşağıdaki gibi gösterilir.

$$\text{Doğruluk} = \frac{DP+DN}{DP+YP+DN+YN}$$

Dengeli doğruluk (Balanced Accuracy), iki ve çok sınıflı sınıflandırma problemlerinde çok tercih edilen ve en çok bilinen ölçüm metriklerinden birisidir. Her sınıf için elde edilen anma değerlerinin ortalaması alınarak bulunabilir. Sınıf dengesine ve bunların göreceli önemine bağlı olarak ağırlıklı ortalama kullanılabilir. Eğer sınıflar dengeli ise doğruluk değeri ile yakın sonuçlar elde edilir (Grandini vd., 2020). Dengeli doğruluk metriğinin iki ve çok sınıflı hallerine ait formül aşağıda sırasıyla gösterilmiştir.

$$\text{Dengeli doğruluk} = \frac{\frac{DP}{\text{Toplam}_a} + \frac{DN}{\text{Toplam}_b}}{2}$$

$$\text{Dengeli doğruluk} = \frac{\frac{DP}{\text{Toplam}_b} + \frac{DN}{\text{Toplam}_a} + \frac{DN}{\text{Toplam}_c}}{3}$$

Seçicilik (Specificity), model tarafından üretilen doğru negatif tahminlerin gerçekte negatif olanlara oranıdır. BKH ile ikileştirilen b sınıfının pozitif diğer iki sınıfın (a ve c) negatif olarak kabul edildiği çok sınıflı bir sınıflandırıcıdan elde edilen sonuçları gösteren karışıklık matrisine göre gerçekte doğru negatiflerin gerçekte doğru ve yanlış negatiflerin toplamının oranı seçiciliği verir. Aşağıda seçicilik metriği formülünün k adet sınıf için genişletilmiş hali görülmektedir.

$$\text{Seicilik}_k = \frac{DN_k}{\text{Gerekte}(DN_k + YN_k)}$$

Negatif tahmin (Negative Predictive) deęeri, doęru negatif tahminlerin sınıflandırma modeli tarafından tahmin edilen negatiflere oranını gösteren bir ölçüttür. Aşaęıda negatif tahmin deęeri formülünün k adet sınıf için genişletilmiş hali görölmektedir.

$$\text{Negatif tahmin deęeri}_k = \frac{DN_k}{\text{Tahmin}(DN_k + YN_k)}$$

Kesinlik ve anma, makine öğreniminde yanlış pozitifler ve yanlış negatifler arasındaki dengeyi anlamak için kullanılan bir dięer temel deęerlendirme ölçümleridir.

Kesinlik (Precision), gerçek doęru pozitif sonuçların, model tarafından üretilen tüm olumlu tahminlere yani doęru ve yanlış tüm pozitif sonuçlara oranıdır. Pozitif tahmin (positive predictive) deęeri olarakta bilinir. Olumlu tahminlerin ne kadar doęru olduęunun yüzdesini belirten bir ölçüdür. Sınıflandırmanın yanlış pozitifleri eleme kabiliyetini gösterir. Aşaęıda kesinlik metrięi formülünün k adet sınıf için genişletilmiş hali görölmektedir. (Shah, 2023).

$$\text{Kesinlik}_k = \frac{DP_k}{DP_k + YP_k}$$

Anma (Recall), duyarlılık (sensitivity) veya gerçek pozitif oran olarak bilinen anma gerçek pozitiflerin, model tahmini sonucu elde edilen doęru pozitif ve yanlış negatiflere oranıdır. Aşaęıda anma metrięi formülünün k adet sınıf için genişletilmiş hali görölmektedir.

$$\text{Anma}_k = \frac{DP_k}{DP_k + YN_k}$$

Tüm olumlu tahminler arasındaki gerçek olumlu tahminleri kesinlik, tüm gerçek olumlu örnekler arasındaki gerçek olumlu tahminleri ise anma deęeri ölçer. Yüksek kesinlik, daha az yanlış pozitive sahip olunduęu yüksek anma ise daha az hatalı negatife sahip olunduęu anlamına gelir.

Kesinlik ve anma bir denge sunar; kesinlięin yüksek olması durumunda, veri kümesindeki gerçek pozitif örneklerden bile şüphe edilir ve böylece anma puanı düşer. Anma puanı yükseldike aslında negatif olan veriler pozitifmiş gibi deęerlendirilerek pozitif olarak

sınıflandırılır bu durum kesinliği azaltır. İyi bir sınıflandırıcı hem kesinlik hem de anma yüzdeleri yüksek olan modellerdir. Kesinlik ve anma, sınıf dağılımı dengesiz olduğunda faydalı bilgiler sunar (Kundu, 2022).

2.5.1. F1 Skoru (F-Measure, F-ölçütü)

F1 skoru, modelin kesinlik ve anma değerlerini birleştirip modelin doğruluğunu ölçen bir MÖ değerlendirme ölçümüdür. Kesinlik ve anma performans ölçütlerini tek başına karşılaştırmak anlamlı olmadığı için her iki ölçütü beraber değerlendirmemize olanak sağlayan F-ölçütü tanımlanmıştır. F1 skoru, kesinlik ve anma değerlerinin harmonik ortalaması olup yanlış pozitifleri ve yanlış negatifleri dikkate alan bir ölçüm sağlar. Hem kesinlik hem de anmanın önemli olduğu durumlarda harmonik ortalamalarını kullanarak kesinlik ve anmayı birleştirir ve modelin performansına ilişkin uyumlu bir değerlendirme sağlar. F-ölçütü genel olarak, F1, F0,5 ve F2 olarak kullanılmaktadır (Chicco ve Jurman, 2020).

$$F1 \text{ Skoru} = 2 * \left(\frac{Kesinlik * Anma}{Kesinlik + Anma} \right)$$

Kesinlik ve anma değerleri eşit ağırlıklı olarak hesaplanırsa F-ölçütü genel olarak F1 ile gösterilir. F-ölçütünün aldığı değerler 0-1 aralığında yer alırken tahmin doğruluğu yüksek olan modeller için F1 skoru bire yakın değer alır. Sınıflar arasında dengesizlik oranları arttıkça modele ait doğruluk değerleri ile performansı ölçmek yanlış olabilmekte bu yüzden de yanlış pozitifler ve yanlış negatifler arasındaki dengeyi dikkate alan F1 değerini kullanmak daha isabetli olmaktadır (Alan, 2020; Kundu, 2022; Shah, 2023).

Çok sınıflı bir veri setindeki her bir sınıfa ait F-ölçüt puanlarını hesaplamak amacıyla bire-karşı-hepsi tekniği kullanılarak sınıf bazında kesinlik ve anma değerleri hesaplanır ve harmonik ortalaması alınır. Farklı ortalama teknikleri kullanılarak da genel F1 puanı hesaplanabilir (Kundu, 2022).

2.5.2. G-ortalama (G-Mean)

Dengesiz veri kümelerinde kullanılan bir başka performansını değerlendirme aracı G-ortalama'dır. İkili sınıflandırma problemlerinde duyarlılık ve seçicilik çarpımının karekökü alınarak bulunur. Olumlu örnekleri doğru şekilde tanımlayan duyarlılık ile olumsuz örnekleri

doğru şekilde tanımlayan seçiciliği dikkate alır. G-ortalama, duyarlılık ve seçicilik açısından mükemmel performansı gösteren 1 ile kötü performansı temsil eden 0 arasında değer alır (Aygün Çakıroğlu, 2024).

Çoklu sınıflarda G-ortalama değeri, anma değerlerinin çarpımının sınıf sayısı kadar kökü bulunarak elde edilir. Örneğin; 3 sınıf var ise üç anma değerinin çarpımının küp köküdür. Aşağıdaki ilk formül iki sınıflı ikinci formül çok sınıflı yaklaşım için geliştirilmiştir.

$$G\text{-ortalama}=(\prod_{i=1}^2Se\check{c}icilik * Duyarlilik)^{\frac{1}{c}}$$

$$G\text{-ortalama}=(\prod_{i=1}^CAnma_i)^{\frac{1}{c}}$$

2.5.3. Eğri Altında Kalan Alan (EAKA) (Area Under Curve, AUC)

Alıcı İşlem Karakteristik (AİK-ROC) eğrisi analizi, sadece bir duyarlılık ve seçicilik değeri kullanarak tanı koymanın getirdiği sakıncaları ortadan kaldırmak için geliştirilmiş istatistik değerlendirme yöntemidir. AİK eğrisi, farklı eşik değerleri için hesaplanan, dikey eksen üzerinde doğru pozitiflik (duyarlılık) ve yatay eksen üzerinde yanlış pozitiflik (1-seçicilik) oranlarının yer aldığı bir grafikdir (Tomak ve Bek, 2009). AİK eğrileri yüksek duyarlılık ve seçicilik değerine sahip bir eşik değerinin olup olmadığına karar vermek için kullanılır.

AİK eğrisi, tanı testlerinin geçerliliğinin denetlenmesi ve performanslarının kıyaslanmasına olanak sağlayan bir yöntem olan AİK eğrisi altında kalan alan tanı testlerinin doğru klinik tanı koymadaki başarısının karşılaştırılmasına olanak sağlayan bir ölçüttür (Kılıç, 2013). Sınıflandırma tahmin performansını, tanısal yeterliliğini veya test verilerini doğru sınıflara ayırabilme kapasitesini ölçmek için tanımlanan bir ölçüttür (Tomak ve Bek, 2009).

EAKA ile gösterilen AİK eğri alanı 0 ile 1 aralığında değerler almaktadır. Bu değer 1'e yaklaşması ile sınıflandırıcının tahmin doğruluğu artmakta, 0 değerine yakın değerler alması ile sınıflandırıcının tahmin doğruluğu azalmakta ve 0,5 değerini alması durumunda ise sınıflandırıcının rastgele tahminlerde bulundu sonucunu göstermektedir (Alan, 2020). EAKA ne kadar büyük ise sınıflandırma modelinin gözlemlerin sınıflarını doğru ayırt etme olasılığı artmaktadır (Tomak ve Bek, 2009).

Başarı karşılaştırma yöntemi olarak MÖ sınıflandırma algoritması tahminlerinde ikili ve çok sınıflı sınıflandırma modellerinin performansını değerlendirmek için kullanışlı ve yaygın olarak kullanılan EAKA metriği sınıflandırıcıların performansını test etmede sıklıkla başvurulan bir ölçüttür (Van Calster vd., 2008).

EAKA'nın, sınıflandırıcı performansını değerlendirmek ve sınıflandırma eğitimi sırasında en uygun çözümü ayırt etmede doğruluk metriğinden daha iyi sonuçlar ürettiği teorik ve deneysel olarak kanıtlanmıştır. EAKA performansı değerlendirme ve ayırt etme süreçlerinde mükemmel olmasına rağmen özellikle çok sınıflı problemler için sınıf sayısı arttıkça üretilen çözümlerin hesaplama maliyetinin çok fazla arttığı görülmüştür (Hossin ve Sulaiman, 2015).

2.5.4. Matthews Korelasyon Katsayısı (MKK) (Matthews Correlation Coefficient, MCC)

İlk olarak 1975 yılında Brian W. Mattheus tarafından kimyasal yapıların karşılaştırılması için geliştirilen korelasyon katsayısı, Karl Pearson tarafından geliştirilen phi (ϕ) katsayısının özel bir durumu olup sınıf dengesizliği sorununun üstesinden de gelen etkili bir çözüm sunmaktadır (Chicco ve Jurman, 2020; Chicco vd., 2021). MKK katsayısı Baldi vd. (2000) tarafından çok sınıflı sınıflandırma MÖ için standart bir performans ölçütü olarak önerilmiştir. Karışıklık matrisi ile ilgili bütün bilgileri içerdiği için çok sınıflı ve dengesiz sınıflardaki tahminlerde kullanılması tavsiye edilir. MKK $[-1, +1]$ aralığında değer alır. Sonuç $+1$ 'e yakın ise sınıflandırıcının doğru tahmin oranı yüksek, 0 'a yakınsa ise sınıflandırıcının yaptığı tahminler rastgeledir (Alan, 2020; Chicco ve Jurman, 2020; Grandini vd., 2020). Aşağıdaki ilk formül iki sınıflılar için ikinci formül ise k adet sınıflar için genişletilmiş halidir (Özen vd., 2024).

$$MKK = \frac{DP * DN - YP * YN}{\sqrt{(DP + YP)(DP + YN)(DN + YP)(DN + YN)}}$$

$$\text{ÇSMKK} = \frac{2}{c(c-1)} \sum_{i < j} MKK(i, j)$$

2.6. Makro ve Mikro Ortalamalar

Çok sınıflı sınıflandırma performans ölçüleri arasında genel performansın daha kısa bir özetini veren makro veya mikro ortalamaları kullanmak daha iyi bir seçenek olabilir. Her iki metrikte, Her iki metrikte, bir karışıklık matrisinin sınıf ikilileştirme (binarization) yaklaşımı ile çözülmesiyle hesaplanır. Makro ortalama, her sınıfa eşit derecede ağırlık vererek sınıflar arasındaki ortalama performansı gösterir (Grandini vd., 2020).

Kesinlik, anma ve F1 skoru çok sınıflı durumlar için mikro veya makro ortalama yöntemleriyle genişletilebilir (Tanha vd., 2020). Kesinlik, anma ve F1 skoruna ait makro ortalama, sınıf bazında hiçbir ağırlıklandırma yapmadan tüm sınıflar için elde edilen kesinlik, anma ve F1 puanlarının basit bir ortalamasıdır. Matematiksel olarak şu şekilde ifade edilir (“k” adet sınıf bulunan veri seti için):

$$\text{Makro Ortalama Kesinlik} = \frac{\sum_{k=1}^K \text{Kesinlik}_k}{K}$$

$$\text{Makro Ortalama Anma} = \frac{\sum_{k=1}^K \text{Anma}_k}{K}$$

$$\text{Makro Ortalama F1 Skoru} = 2 * \left(\frac{\text{Makro Ortalama Kesinlik} * \text{Makro Ortalama Anma}}{\text{Makro Ortalama Kesinlik} + \text{Makro Ortalama Anma}} \right)$$

Makro ortalamalı F1 puanı yalnızca kullanılan veri setinin her sınıfına ait oranlar birbirine yakın ise kullanışlıdır. F1 makro ortalama değeri algoritmanın sınıflara ayırmadaki başarısının yüksekliğini gösterir. Farklı kategorilerde farklı miktarda veri bulunması yani sınıflar açısından dengesizlik olması durumunda basit bir ortalama, yanıltıcı bir performans ölçütü olabilir (Grandini vd., 2020; Kundu, 2022).

Mikro ortalama aynı şekilde her örneğe eşit ağırlık verir ve tüm tahminlerde ortalama performansı gösterir. Mikro ortalama, tüm sınıflardaki doğru pozitiflerin, yanlış pozitiflerin ve yanlış negatiflerin sayımlarını toplar ve ardından toplam sayılara dayalı olarak performans ölçümünü hesaplar. Sınıf etiketine ve sınıftaki örnek sayısına bakılmadan her örneğe eşit ağırlık verir. Çok sınıflı sınıflandırma durumunda mikro ortalamalı kesinlik, anma ve F1 skoru ile doğruluk değerleri birbirinden farklılık gösterir (Grandini vd., 2020; Kundu, 2022).

Mikro ortalama kesinlik, anma ve F1 ölçütü için formüller aşağıdadır:

$$\text{Mikro Ortalama Kesinlik} = \frac{\sum_{k=1}^K DP_k}{\sum_{k=1}^K \text{Sütun Toplam}_k}$$

$$\text{Mikro Ortalama Anma} = \frac{\sum_{k=1}^K DP_k}{\sum_{k=1}^K \text{Satır Toplam}_k}$$

$$\text{Mikro Ortalama F1 Skoru} = \frac{\sum_{k=1}^K DP_k}{\left(DP + \frac{1}{2}(YP + YN)\right)}$$

Makro ortalamanın mikro ortalamaya göre bir üstünlüğü olduğu belirtilmiştir (Gupta, 2020).

2.7. Dengesiz Veri İçin Rastgele Aşırı Örnekleme (RAÖ) Yöntemi

Çalışmada kullanılan 14.233 gözlemlili ve üç sınıflı kazanç değişkenli veri setindeki kazanç değişkeni sınıfları arasında dengesizlik oranı 2,66 olarak bulunmuş olup Fernández vd. (2013) tarafından belirtilen dengesiz veri olarak tanımlanma oranı olan 1,50'den yüksek bulunduğu için kazanç yönünden dengesiz bir veri seti olarak tanımlanabilir (Özen vd., 2024). Bu yüzden dengesiz verilerde sınıflandırma probleminin çözümü için literatürde sıklıkla kullanılan yeniden örnekleme yöntemlerinden biri olan basit ama etkili bir yaklaşım olarak bilinen rastgele aşırı örnekleme (RAÖ) yöntemi kullanılmıştır (Bölükbaşı, 2023).

Batista vd. (2004) tarafından yapılan araştırmada dengesiz veri dengeleme yöntemleri karşılaştırılmış RAÖ daha karmaşık (SAÖAT ve türevleri) yöntemlerle rekabetçi sonuçlar sağlamış, diğer yöntemlerin üreteceğinden daha az maliyetli anlamlı sonuçlar üretmiştir. Çalışmadan elde edilen bulgulara göre Batista vd. (2004) tarafından büyük veri setleri için RAÖ yöntemi önerilmiştir.

Dengesiz veri setinin RAÖ yöntemi ile dengelenmesi için öncelikli olarak veri seti %70-%30, %75-%25 ve %80-%20 oranlarında eğitim ve test veri seti olarak ikiye bölünmüş, eğitim veri setleri RAÖ yöntemi ile yeniden oluşturulmuştur. Bu eğitim veri setlerine ÇD tekniği 10 katlı ve 10 tekrarlı olarak uygulanmış k-EYK, C5.0, RO ve GAM sınıflandırma algoritmaları model eğitime tabi tutulmuş ve performans ölçütleri ile en iyi hiper parametre değerleri elde

edilmiştir. Elde edilen tahmin modellerinin performansını ölçmek için %30, %25 ve %20 oranlarında test veri setlerine k-EYK, C5.0, RO ve GAM algoritmaların uygulanması sonucunda elde edilen karşılaştırmalı performans ölçütleri (doğruluk, dengeli doğruluk, kesinlik, anma ve F1 skoru makro ortalaması, seçicilik, negatif tahmin değerlerinin ortalaması, G-ortalama, ÇSMKK ve EAKA) çizelgelerde, sınıflandırma tahmin modellerinde yer alan değişkenlerin önem dereceleri ise grafiklerde gösterilmiştir. Her bir eğitim ve test veri setinde en iyi doğru tahmin değerine sahip algoritmanın test veri setinden elde ettiği doğru pozitif (DP), yanlış pozitif (YP), yanlış negatif (YN) ve doğru negatif (DN) değerleri karışıklık matrisi ile çizelgede sunulmuştur.

2.8. Veri Analizi İçin Kullanılan Program ve Paketler

Bu tez çalışmasında, MÖ sınıflandırma algoritmalarını uygulamak ve görselleştirmek için RStudio (sürüm 2023.09.1) istatistik yazılımı kullanılmıştır (RStudio Team, 2020). C5.0, RO, k-EYK ve GAM eğitim modelleri için “caret” (sürüm 6.0-94) paketi kullanılmıştır (Kuhn, 2008). MKK ve ÇSMKK değerlerini elde etmek için “mltools” (sürüm 0.3.5), çok sınıflı EAKA değerini elde etmek için ise “MultiClassROC” (sürüm 0.1.0) ve ‘pROC’ (sürüm 1.18.5) paketleri ile işlem yapılmıştır (Hand ve Till, 2001; Robin vd., 2011). Her bir sınıfın performans metrikleri “multiClassSummary” fonksiyonu ile elde edilmiş, çok sınıflı EAKA değerini elde etmek için ise tip olarak “prob” fonksiyonu seçilmiştir. Dengesiz dağılımı dengelemek için kullanılan RAÖ yöntemi eğitim veri setine “UBL” (sürüm 0.0.9) paketi ile uygulanmıştır (Branco vd., 2016). Tanımlayıcı istatistik olarak kategorik değişkenler için sıklık ve yüzde, sürekli değişkenler için aritmetik ortalama (A.Ort.), standart sapma (SS), ortanca, en küçük (Min.) ve en büyük değerler (Maks.) verilmiştir.

3. BULGULAR

Bu bölümde, ilk olarak kazanç hedef değişkeninin dengesiz dağılıma sahip üç kategorili hali ve iki kategorili hali için 7 açıklayıcı özniteliğin her bir sınıftaki tanımlayıcı istatistikleri incelenmiş ve Çizelge 3.1.'de gösterilmiştir.

İkinci olarak, Türkiye'de safkan İngiliz yarış atlarının kazanç durumu tahmini için yedi çevresel faktör kullanılarak k-EYK, C5.0, RO ve GAM sınıflandırma algoritmaları ile %70-%30, %75-%25 ve %80-%20 oranlarındaki eğitim ve test veri setlerinden sınıflandırma tahmin modelleri oluşturulmuştur. %70, %75 ve %80 oranlarındaki (1) dengelenmemiş-orjinal, (2) RAÖ yöntemi ile dengelenmiş ve (3) iki sınıflı eğitim veri setlerinden; k-EYK, C5.0, RO ve GAM sınıflandırma algoritmalarının öğrendiği tahmin modellerine ait, (a) hiper parametre değerleri ve tahmin modellerinin performans ölçütleri çizelgeler ile, (b) modellerde yer alan açıklayıcı özniteliklerin önem dereceleri ise grafikler ile gösterilmiştir. Eğitim veri setleri ile ulaşılan tahmin modelleri, %30, %25 ve %20 oranlarındaki test veri setleri ile sınanmış ve elde edilen performans karşılaştırma ölçütleri çizelgelerde gösterilmiştir.

3.1. Üç Sınıflı Dengesiz Veri Seti Bulguları

3.1.1. Tanımlayıcı İstatistikler

Yedi açıklayıcı özniteliğin üç sınıflı kazanç değişkenini içeren dengeleme yapılmamış orijinal veri setindeki dağılımları Çizelge 3.1.'de verilmiştir.

Çizelge 3.1. Üç sınıflı kazanç veri seti açıklayıcı öznitelik dağılımları (n, %)

		Kazanç Yok (n=8.132)	Az Kazanç (n=3.052)	Çok Kazanç (n=3.049)
Cinsiyet	Erkek	3.877 (%47,7)	1.443 (%47,3)	1.526 (%50,0)
	Dişi	4.255 (%52,3)	1.609 (%52,7)	1.123 (%50,0)
KP	Değişken Parkur	5.978 (%73,5)	2.430 (%79,6)	2.312 (%75,8)
	Her Daim Çim	634 (%7,8)	210 (%6,9)	294 (%9,6)
	Her Daim Kum	1.286 (%15,8)	366 (%12,0)	369 (%12,1)
	Her Daim Sentetik	234 (%2,9)	46 (%1,5)	74 (%2,4)
	Kazanamadı	1.036 (%12,7)	307 (%10,1)	277 (%9,1)
AKD	Kazandı	4.770 (%58,7)	1.808 (%59,2)	1.791 (%58,7)
	Yarışlara Katılmamış	2.326 (%28,6)	937 (%30,7)	981 (%32,2)

Çizelge 3.1. Devam Üç sınıflı kazanç veri seti açıklayıcı öznitelik dağılımları (n, %)

BKD	Kazanamadı	143 (%1,8)	62 (%2,0)	36 (%1,2)
	Kazandı	3.978 (%48,9)	1.398 (%45,8)	1.289 (%42,3)
	Yarışlara Katılmamış	4.011 (%49,3)	1.592 (%52,2)	1.724 (%56,5)
SD	Hisseli/Ortak	920 (%11,3)	362 (%11,9)	365 (%12,0)
	Tek Sahip	7.212 (%88,7)	2.690 (%88,1)	2.684 (%88,0)
YBY (A.Ort±SS) (Ay)		31,50±4,78	29,57±3,59	29,39±3,35
Ortanca (Min.-Maks.)		31 (22-84)	29 (22-57)	29 (22-55)
KS (A.Ort±SS) (Adet)		15,19±18,30	19,70±20,00	20,57±18,03
Ortanca (Min.-Maks.)		9 (1-265)	14 (1-309)	16 (1-282)

Çizelge 3.1.'de atların cinsiyet dağılımlarının her üç grupta da dengeli olduğu görülmüştür. Koşu parkuru (KP), sahiplik durumu (SD), anne kazanç durumu (AKD) ve baba kazanç durumuna (BKD) ait üç gruptaki dağılım homojen bulunmuştur. Yarışa başlama yaşı (YBY) kazancı olmayan grup için 31,50±4,78 ay, az kazancı olan için 29,57±3,59 ay ve çok kazancı olan için ise 29,39±3,35 ay olarak bulunmuştur. Koşu sayısı (KS) üç grup için sırasıyla; 15,19±18,30, 19,70±20,00 ve 20,57±18,03 olarak hesaplanmıştır.

3.1.2. MÖ Sınıflandırma Algoritma Sonuçları

Aşağıdaki bölümde k-EYK, C5.0, RO ve GAM sınıflandırma algoritmalarının %70, %75 ve %80 oranlarındaki üç sınıflı kazanç değişkenini içeren dengeleme yapılmamış eğitim veri setine ve %30, %25 ve %20 oranlarındaki test veri setlerine uygulanması sonucunda elde edilen performans ölçütleri, hiper parametre değerleri ve karşılaştırmalı performans ölçütleri çizelgelerde, değişkenlerin önem dereceleri grafiklerde gösterilmiş ve en iyi doğruluk oranına sahip algoritmanın karışıklık matrisi ise çizelgede sunulmuştur.

3.1.3. Eğitim (%70) ve Test (%30) Veri Seti Sonuçları

Veri seti ilk olarak, %70 eğitim ve %30 test verisi olarak bölünmüş, eğitim veri setini 9.976 adet örnek, test veri setini ise 4.257 adet örnek oluşturmuştur. Dengeleme yapılmamış eğitim veri setine performans değerlendirme yöntemi olan ÇD tekniği 10 katlı ve 10 tekrarlı olarak uygulanmış, eğitim veri setinden dört ayrı sınıflandırma algoritması ile sınıflandırma tahmin modeli öğrenilmiş, performans ölçütleri elde edilmiş ve en iyi hiper parametre değerleri için ayarlama yapılmıştır. Bu hiper parametre değerleri ile oluşturulan tahmin modelleri test veri setleri ile sınanmış ve karşılaştırmalı performans ölçütleri elde edilmiştir. Elde edilen bulgular sırasıyla aşağıda sunulmuştur.

3.1.4. Eğitim (%70) Veri Seti Performans Sonuçları

Eğitim veri setindeki 9.976 adet örnek ile ÇD model performans değerlendirme yöntemi 10 katlı ve 10 tekrarlı uygulanarak dört algoritma eğitilmiştir. Hiper parametre en iyileme yöntemi olarak ızgara araştırması (grid search) kullanılarak her bir algoritma için Çizelge 3.2.’de görülen hiper parametre değerleri elde edilmiştir.

Çizelge 3.2. Üç sınıflı %70 eğitim veri seti hiper parametre değerleri

Algoritmalar	Hiper parametreler	Değer
k-EYK	“k” en yakın komşu sayısı	9
RO	“mtry” köklerin dallara ayrılmasında kullanılan rastgele seçilmiş özellik sayısı	2
GAM	“n.trees” kurulacak KA sayısı	50
	“interaction.depth” her bir ağacın en büyük derinlik sayısı	3
	“shrinkage” öğrenme oranı	0,1
	“n.minobsinnode” terminal düğümlerdeki minimum gözlem sayısı	10

Hiper parametre en iyilemesi sonucunda, k-EYK için “k” en yakın komşuluk değeri 9, RO için “mtry” köklerin dallara ayrılmasında kullanılan rastgele seçilmiş özellik sayısı 2 olarak bulunmuştur. GAM algoritması için “n.trees” kurulacak karar ağaç sayısı 50, “interaction.depth” her bir ağacın en büyük derinlik sayısı 3, “shrinkage” öğrenme oranı 0,1 ve “n.minobsinnode” terminal düğümlerdeki minimum gözlem sayısı 10 olarak bulunmuştur (Çizelge 3.2.).

Yukarıda önerilen hiper parametre değerleri kullanılarak sınıflandırma algoritmalarının üç sınıflı kazanç değişkenini içeren dengeleme yapılmamış eğitim veri setinden öğrendikleri tahmin modellerine ait performans ölçüt sonuçları Çizelge 3.3.’te gösterilmiştir.

Çizelge 3.3. Üç sınıflı %70 eğitim veri seti performans metrikleri

	C5.0	k-EYK	RO	GAM
Doğruluk	0,581	0,539	0,576	0,583
Dengeli Doğruluk Ortalaması	0,521	0,535	0,504	0,524
Ortalama Kesinlik (Pozitif tahmin değeri)	0,444	0,397	-	0,472
Ortalama Anma (Duvarlılık)	0,359	0,377	0,338	0,363
Ortalama F1 skor	0,309	0,365	-	0,310
Ortalama Seçicilik	0,682	0,694	0,669	0,684
Ortalama Negatif tahmin değeri	0,761	0,713	0,811	0,760

Çizelge 3.3. incelendiğinde, %70 eğitim veri setinde GAM ve C5.0 algoritmalarının en yüksek doğruluk oranlarına (%58,30 ve %58,10) sahip olduğu görülmüştür. Olumlu

tahminlerin ne kadar doğru olduğunun yüzdesini belirten bir ölçüt olan kesinliğin ortalaması GAM için %47,20 ve ortalama negatif tahmin değeri RO için %81,10 olarak bulunmuştur.

K-EYK algoritmasının yüksek performans gösterdiği metrikler şu şekildedir; dengeli doğruluk ortalaması %53,50, ortalama seçicilik %69,40, anma metriği ortalaması %37,70, kesinlik ve anma performans ölçütlerini beraber değerlendiren F1 skoru ortalaması %36,50'dir.

3.1.5. Test (%30) Veri Seti Performans Sonuçları

Test veri seti olarak 4.257 adet örnek ile eğitim veri setinden öğrenilen model deseni k-EYK, C5.0, RO ve GAM sınıflandırma algoritmaları ve belirlenen en iyi hiper parametreler kullanılarak test edilmiştir. Sonuçta, doğruluk, dengeli doğruluk ve ÇSMKK ile kesinlik, anma, F-ölçütlerinin makro ortalamaları, seçicilik ve negatif tahmin değeri ortalaması ile G-ortalama ve EAKA değeri gibi ulaşılan performans ölçütleri Çizelge 3.4.'te gösterilmiştir.

Çizelge 3.4. Üç sınıflı %30 test veri seti performans metrikleri

	C5.0	k-EYK	RO	GAM
Doğruluk	0,570 %95GA (0,555-0,585)	0,521 %95GA (0,506-0,536)	0,567 %95GA (0,552-0,582)	0,575 %95GA (0,559-0,589)
Dengeli Doğruluk	0,516	0,526	0,503	0,519
Makro Ortalama Kesinlik (Pozitif tahmin değeri)	0,326	0,370	0,424	0,515
Makro Ortalama Anma (Duyarlılık)	0,353	0,363	0,337	0,356
Makro Ortalama F1 skor	0,286	0,348	0,250	0,291
Ortalama Seçicilik	0,679	0,689	0,669	0,681
Ortalama Negatif tahmin değeri	0,765	0,705	0,816	0,784
ÇSMKK	0,104	0,071	0,055	0,109
G-ortalama	0	0,239	0	0,063
EAKA	0,602	0,571	0,575	0,638

Çizelge 3.4. incelendiğinde, %30 test veri setinde GAM algoritmasının en yüksek doğruluk oranına (%57,50, %95GA; %55,90-%58,90) sahip olduğu görülmüştür. GAM algoritması için olumlu tahminlerin ne kadar doğru olduğunun yüzdesini belirten kesinliğin makro ortalaması %51,50 ve karışıklık matrisine ait tüm bilgileri kullanan ölçüm metriği olan ÇSMKK değeri %10,90 ve EAKA değeri %63,80 olarak elde edilmiş bunun yanında negatif tahmin değeri ortalaması %81,60 RO için en yüksek olarak bulunmuştur.

K-EYK algoritmasının yüksek performans gösterdiği metrikler şu şekildedir; dengeli doğruluk ortalaması %52,60, G-ortalama değeri %23,90, seçicilik ortalaması %68,90, anma metriği makro ortalaması %36,30 ile kesinlik ve anma performans ölçütlerini beraber değerlendiren F1 skoru makro ortalaması %34,80'dir.

Test veri seti sonuçlarına göre tahmin doğruluk oranı en yüksek olan GAM algoritmasına ait karışıklık matrisi Çizelge 3.5.'te görülmektedir.

Çizelge 3.5. GAM karışıklık matrisi

	Gerçek Değer			
	Sınıflar	a	b	c
	Tahmin Değeri			
	a	2369	880	843
	b	2	3	1
	c	35	50	74

a:Kazanç yok, b:Az kazanç, c:Çok kazanç

Çizelge 3.5.'teki karışıklık matrisinde kazanç yok sınıfına ait 2.406 verinin 2.369'u, az kazanç sınıfına ait 933 verinin 3'ü ve çok kazanç sınıfına ait 918 verinin 74'ü GAM algoritması ile doğru olarak sınıflandırılmıştır.

Tahmin doğruluk oranı en yüksek GAM algoritmasındaki tüm kazanç sınıflarına ait performans metrikleri Çizelge 3.6.'da görülmektedir.

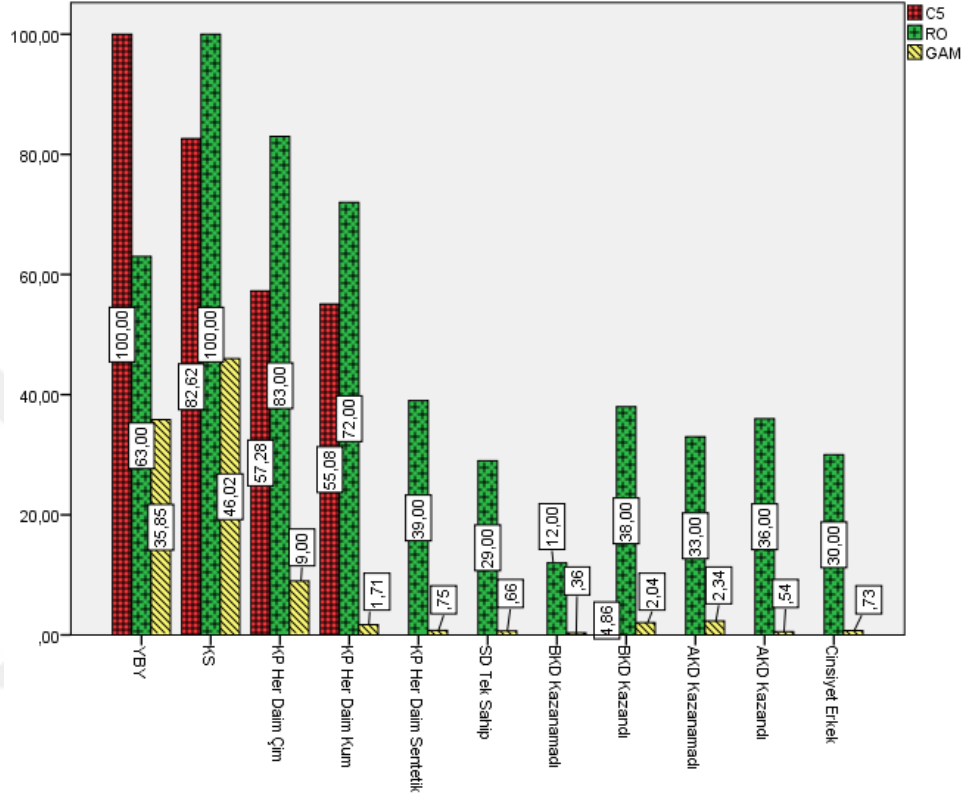
Çizelge 3.6. %30 test veri seti GAM tüm sınıflara ait performans metrikleri

	GAM		
	a	b	c
Doğruluk	0,575		
Dengeli doğruluk	0,527	0,501	0,528
Kesinlik (Pozitif tahmin değeri)	0,579	0,500	0,465
Anma (Duyarlılık)	0,985	0,003	0,081
F1 skor	0,729	0,006	0,137
Seçicilik	0,070	0,999	0,975
Negatif tahmin değeri	0,776	0,781	0,794

a:Kazanç yok, b:Az kazanç, c:Çok kazanç

Çizelge 3.6.'da GAM algoritma sonuçlarına göre, kazanç yok sınıfında kesinlik değeri %57,90, anma metriği %98,50 ve F1 skoru %72,90 diğer sınıflardan yüksek bulunmuştur. Seçicilik değeri %99,90 ve negatif tahmin değeri %79,40 az kazanç sınıfında yüksek bulunurken dengeli doğruluk oranı %52,80 ise çok kazanç sınıfında yüksek bulunmuştur.

Şekil 3.1’de, %70 oranındaki üç sınıflı kazanç değişkenini içeren dengeleme yapılmamış eğitim veri seti kullanılarak k-EYK hariç diğer algoritmalar ile elde edilen sınıflandırma tahmin modellerindeki özneliliklerin modele anlamlı katkılarını ve verdiği katkının önem derecesini gösteren grafik görülmektedir.



Şekil 3.1. Üç sınıflı %70 eğitim veri seti tahmin modeli öznelilik önem dereceleri

Şekil 3.1.’de %70 eğitim veri seti için değişken önem dereceleri incelendiğinde analize dahil olan tüm özneliliklerin C5.0, GAM ve RO algoritmalarında farklı önem dereceleri de olsa tahmin modeline katkı sağladığı görülmüştür. C5.0 ile kurulan sınıflandırma modelinde en etkili öznelilikler sırasıyla; koşu sayısı (KS), koşu hayatı boyunca çim pistte koşma (KPHerDaimÇim), koşu hayatı boyunca kum pistte koşma (KPHerDaimKum), baba kazanç durumu kazandı olanlar (BKDKazandı) ve yarışa başlama yaşıdır (YBY). RO ile kurulan sınıflandırma modelinde yer alan en etkili öznelilikler sırasıyla; KS, KPHerDaimÇim, KPHerDaimKum, YBY, koşu hayatı boyunca sentetik pistte koşma (KPHerDaimSentetik), BKDKazandı ve AKDKazandı. GAM ile kurulan sınıflandırma modelinde en etkili öznelilikler sırasıyla; KS, YBY, KPHerDaimÇim, KPHerDaimKum ve BKDKazandı.

3.1.6. Eğitim (%75) ve Test (%25) Veri Seti Sonuçları

Veri seti %75-%25 oranında eğitim ve test verisi olarak bölündüğünde eğitim veri setindeki örnek sayısı 10.671, test veri setinde ise 3.562 olarak bulunmuştur. Performans değerlendirme yöntemi olan ÇD tekniği 10 katlı ve 10 tekrarlı olarak uygulanmış, eğitim veri seti kullanılarak dört ayrı sınıflandırma algoritması ile sınıflandırma tahmin modeli öğrenilmiş, performans ölçütleri elde edilmiş ve en iyi hiper parametre değerleri için ayarlama yapılmıştır. Bu hiper parametre değerleri ile oluşturulan tahmin modeli test veri seti ile sınanmış ve karşılaştırmalı performans ölçütleri elde edilmiştir. Elde edilen bulgular sırasıyla aşağıda sunulmuştur.

3.1.7. Eğitim (%75) Veri Seti Performans Sonuçları

Eğitim veri setindeki 10.671 adet örnek ile ÇD model performans değerlendirme yöntemi 10 katlı ve 10 tekrarlı uygulanarak dört algoritma eğitilmiştir. Hiper parametre en iyileme yöntemi olarak ızgara araştırması kullanılarak her bir algoritma için Çizelge 3.7.' de görülen hiper parametre değerleri elde edilmiştir.

Çizelge 3.7. Üç sınıflı %75 eğitim veri seti hiper parametre değerleri

Algoritmalar	Hiper parametreler	Değer
k-EYK	“k” en yakın komşu sayısı	9
RO	“mtry” köklerin dallara ayrılmasında kullanılan rastgele seçilmiş özellik sayısı	2
GAM	“n.trees” kurulacak KA sayısı	50
	“interaction.depth” her bir ağacın en büyük derinlik sayısı	2
	“shrinkage” öğrenme oranı	0,1
	“n.minobsinnode” terminal düğümlerdeki minimum gözlem sayısı	10

Hiper parametre en iyilemesi sonucunda, k-EYK için “k” en yakın komşuluk değeri 9, RO algoritması için “mtry” köklerin dallara ayrılmasında kullanılan rastgele seçilmiş özellik sayısı 2 olarak bulunmuştur. GAM algoritması için “n.trees” kurulacak karar ağaç sayısı 50, “interaction.depth” her bir ağacın en büyük derinlik sayısı 2, “shrinkage” öğrenme oranı 0,1 ve “n.minobsinnode” terminal düğümlerdeki minimum gözlem sayısı 10 olarak elde edilmiştir (Çizelge 3.7.).

Yukarıda önerilen hiper parametre değerleri kullanılarak sınıflandırma algoritmalarının eğitim veri setinden öğrendikleri tahmin modellerine ait performans ölçüt sonuçları Çizelge 3.8.'de gösterilmiştir.

Çizelge 3.8. Üç sınıflı %75 eğitim veri seti performans metrikleri

	C5.0	k-EYK	RO	GAM
Doğruluk	0,570	0,526	0,569	0,583
Dengeli Doğruluk Ortalaması	0,521	0,532	0,504	0,520
Ortalama Kesinlik (Pozitif tahmin değeri)	0,415	0,388	0,406	0,465
Ortalama Anma (Duyarlılık)	0,360	0,372	0,338	0,359
Ortalama F1 skor	0,308	0,361	-	0,299
Ortalama Seçicilik	0,683	0,692	0,669	0,681
Ortalama Negatif tahmin değeri	0,749	0,708	0,813	0,770

Çizelge 3.8. incelendiğinde GAM algoritmasının en yüksek doğruluk oranına (%58,30) sahip olduğu görülmüştür. Olumlu tahminlerin ne kadar doğru olduğunun yüzdesini belirten bir ölçüt olan kesinliğin ortalaması GAM için %46,50 ve ortalama negatif tahmin değeri RO için %81,30 olarak bulunmuştur.

K-EYK algoritmasının yüksek performans gösterdiği metrikler şu şekildedir; dengeli doğruluk ortalaması %53,20, ortalama seçicilik %69,20 anma metriği ortalaması %37,20, F1 skoru ortalaması %36,10'dur.

3.1.8. Test (%25) Veri Seti Performans Sonuçları

Test veri seti olarak 3.562 adet örnek ile eğitim veri setinden öğrenilen model deseni k-EYK, C5.0, RO ve GAM sınıflandırma algoritmaları ve belirlenen en iyi hiper parametreler kullanılarak test edilmiştir. Sonuçta, doğruluk, dengeli doğruluk ve ÇSMKK ile kesinlik, anma, F1 skoru makro ortalamaları, seçicilik ve negatif tahmin değeri ortalaması ile G-ortalama ve EAKA değeri gibi ulaşılan performans ölçütleri Çizelge 3.9.'da gösterilmiştir.

Çizelge 3.9. Üç sınıflı %25 test veri seti performans metrikleri

	C5.0	k-EYK	RO	GAM
Doğruluk	0,591 %95GA (0,575-0,607)	0,539 %95GA (0,522-0,555)	0,589 %95GA (0,572-0,605)	0,597 %95GA (0,580-0,613)
Dengeli Doğruluk	0,521	0,534	0,504	0,521
Makro Ortalama Kesinlik (Pozitif tahmin değeri)	0,377	0,380	0,446	0,501
Makro Ortalama Anma (Duyarlılık)	0,359	0,371	0,338	0,360
Makro Ortalama F1 skor	0,305	0,362	0,257	0,305
Ortalama Seçicilik	0,682	0,696	0,669	0,682
Ortalama Negatif tahmin değeri	0,754	0,713	0,800	0,774
ÇSMKK	0,104	0,091	0,059	0,120
G-ortalama	0,053	0,257	0	0,080
EAKA	0,587	0,692	0,575	0,629

Çizelge 3.9. incelendiğinde GAM algoritmasının %59,70 (%95GA; %58,00-%61,30) ile en yüksek doğruluk oranına sahip olduğu görülmüştür. GAM algoritması için kesinliğin makro ortalaması %50,10 ve ÇSMKK değeri %12,00 olarak, negatif tahmin değeri ortalaması RO için %80,00 olarak bulunmuştur.

K-EYK algoritmasının yüksek performans gösterdiği metrikler şu şekildedir; dengeli doğruluk ortalaması %53,40, EAKA değeri %69,20, G-ortalama değeri %25,70, seçicilik ortalaması %69,60, anma metriği makro ortalaması %37,10 ile F1 skoru makro ortalaması %36,20'dir.

Test veri seti sonuçlarına göre tahmin doğruluk oranı en yüksek GAM algoritmasına ait karışıklık matrisi Çizelge 3.10.'da görülmektedir.

Çizelge 3.10. GAM karışıklık matrisi

		Gerçek Değer		
Tahmin Değeri	Sınıflar	a	b	c
	a	2050	685	691
	b	6	4	2
	c	32	21	71

a:Kazanç yok, b:Az kazanç, c:Çok kazanç

Çizelge 3.10.'daki karışıklık matrisinde kazanç yok sınıfına ait 2.088 verinin 2.050'si, az kazanç sınıfına ait 710 verinin 4'ü ve çok kazanç sınıfına ait 764 verinin 71'i GAM algoritması ile doğru olarak sınıflandırılmıştır.

Test veri seti sonuçlarına göre tahmin doğruluk oranı en yüksek GAM algoritmasında tüm kazanç sınıflarına ait performans metrikleri Çizelge 3.11.'de görülmektedir.

Çizelge 3.11. %25 test veri seti GAM tüm sınıflara ait performans metrikleri

	GAM		
	a	b	c
Doğruluk	0,597		
Dengeli Doğruluk	0,524	0,501	0,537
Kesinlik (Pozitif tahmin değeri)	0,598	0,333	0,573
Anma (Duyarlılık)	0,982	0,006	0,093
F1 skor	0,744	0,011	0,160
Seçicilik	0,066	0,997	0,981
Negatif tahmin değeri	0,721	0,801	0,798

a:Kazanç yok, b:Az kazanç, c:Çok kazanç

3.1.9. Eğitim (%80) ve Test (%20) Veri Seti Sonuçları

Veri seti ilk olarak, %80 eğitim ve %20 test verisi olarak bölünmüş, 11.342 adet örnek eğitim, 2.891 adet örnek ise test verisi olarak analiz edilmiştir. Performans değerlendirme yöntemi olan ÇD tekniği 10 katlı ve 10 tekrarlı olarak uygulanmış, eğitim veri seti kullanılarak dört ayrı sınıflandırma algoritması ile sınıflandırma tahmin modeli öğrenilmiş, performans ölçütleri elde edilmiş ve en iyi hiper parametre değerleri için ayarlama yapılmıştır. Bu hiper parametre değerleri ile oluşturulan tahmin modeli test veri seti ile sınanmış ve karşılaştırmalı performans ölçütleri elde edilmiştir. Elde edilen bulgular sırasıyla aşağıda sunulmuştur.

3.1.10. Test (%20) Veri Seti Performans Sonuçları

Eğitim veri setindeki 11.342 adet örnek ile ÇD model performans değerlendirme yöntemi 10 katlı ve 10 tekrarlı uygulanarak dört algoritma eğitilmiştir. Hiper parametre en iyileme yöntemi olarak ızgara araştırması kullanılarak her bir algoritma için Çizelge 3.12.'de görülen hiper parametre değerleri elde edilmiştir.

Çizelge 3.12. Üç sınıflı %80 eğitim veri seti hiper parametre değerleri

Algoritmalar	Hiper parametreler	Değer
k-EYK	“k” en yakın komşu sayısı	9
RO	“mtry” köklerin dallara ayrılmasında kullanılan rastgele seçilmiş özellik sayısı	2
GAM	“n.trees” kurulacak KA sayısı	50
	“interaction.depth” her bir ağacın en büyük derinlik sayısı	3
	“shrinkage” öğrenme oranı	0,1
	“n.minobsinnode” terminal düğümlerdeki minimum gözlem sayısı	10

Hiper parametre en iyilemesi sonucunda, k-EYK için “k” en yakın komşuluk değeri 9, RO algoritması için “mtry” köklerin dallara ayrılmasında kullanılan rastgele seçilmiş özellik sayısı 2 olarak bulunmuştur. GAM algoritması için “n.trees” kurulacak karar ağacı sayısı 50, “interaction.depth” her bir ağacın en büyük derinlik sayısı 3, “shrinkage” öğrenme oranı 0,1 ve “n.minobsinnode” terminal düğümlerdeki minimum gözlem sayısı 10 olarak bulunmuştur (Çizelge 3.12.).

Yukarıda önerilen hiper parametre değerleri kullanılarak sınıflandırma algoritmalarının eğitim veri setinden öğrendikleri tahmin modellerine ait performans ölçüt sonuçları Çizelge 3.13.'te gösterilmiştir.

Çizelge 3.13. Üç sınıflı %80 eğitim veri seti performans metrikleri

	C5.0	k-EYK	RO	GAM
Doğruluk	0,578	0,532	0,569	0,580
Dengeli Doğruluk Ortalaması	0,521	0,531	0,503	0,522
Ortalama Kesinlik (Pozitif tahmin	0,455	0,386	-	0,458
Ortalama Anma (Duyarlılık)	0,360	0,370	0,337	0,361
Ortalama F1 skor	0,307	0,358	-	0,304
Ortalama Seçicilik	0,682	0,692	0,669	0,683
Ortalama Negatif tahmin değeri	0,759	0,709	0,827	0,765

Çizelge 3.13. incelendiğinde GAM algoritmasının en yüksek doğruluk oranına %58,00 sahip olduğu görülmüştür. Olumlu tahminlerin ne kadar doğru olduğunun yüzdesini belirten bir ölçüt olan kesinliğin ortalaması GAM için %45,80 ve ortalama negatif tahmin değeri RO için %82,70 olarak bulunmuştur.

K-EYK algoritmasının yüksek performans gösterdiği metrikler şu şekildedir; dengeli doğruluk ortalaması %53,10, ortalama seçicilik %69,20, anma metriği ortalaması %37,00, F1 skoru ortalaması %35,80'dir.

3.1.11. Test (%20) Veri Seti Performans Sonuçları

Test veri seti olarak 2.891 adet örnek ile eğitim veri setinden öğrenilen model deseni k-EYK, C5.0, RO ve GAM sınıflandırma algoritmaları ve belirlenen en iyi hiper parametreler kullanılarak test edilmiştir. Sonuçta, doğruluk, dengeli doğruluk ve ÇSMKK ile kesinlik, anma, F-ölçütlerinin makro ortalamaları, seçicilik ve negatif tahmin değeri ortalaması ile EAKA değeri gibi ulaşılan performans ölçütleri Çizelge 3.14.'te gösterilmiştir.

Çizelge 3.14. Üç sınıflı %20 test veri seti performans metrikleri

	C5.0	k-EYK	RO	GAM
Doğruluk	0,580 %95GA (0,562-0,599)	0,537 %95GA (0,518-0,555)	0,576 %95GA (0,558-0,594)	0,583 %95GA (0,565-0,601)
Dengeli Doğruluk	0,522	0,535	0,506	0,524
Makro Ortalama Kesinlik (Pozitif tahmin değeri)	0,426	0,390	0,466	0,487
Makro Ortalama Anma (Duyarlılık)	0,362	0,387	0,341	0,364
Makro Ortalama F1 skor	0,304	0,375	0,258	0,309
Ortalama Seçicilik	0,682	0,695	0,671	0,683
Ortalama Negatif tahmin değeri	0,761	0,715	0,839	0,768
ÇSMKK	0,113	0,094	0,080	0,122
G-Mean	0,055	0,256	0	0,088
EAKA	0,601	0,704	0,678	0,636

Çizelge 3.14. incelendiğinde GAM algoritmasının doğruluk oranı (%58,30, %95GA; %56,50-%60,10) diğer algoritmalarından yüksek bulunmuştur. GAM algoritması için kesinliğin makro ortalaması %48,70 ve ÇSMKK değeri %12,20 olarak en yüksek, RO için negatif tahmin değeri ortalaması %83,90 ve olarak en yüksek bulunmuştur.

K-EYK algoritmasının yüksek performans gösterdiği metrikler şu şekildedir; dengeli doğruluk ortalaması %53,50, G-ortalama değeri %25,60, EAKA değeri %67,80, seçicilik ortalaması %69,50, anma metriği makro ortalaması %38,70 ile F1 skoru makro ortalaması %37,50'dir.

Test veri seti sonuçlarına göre performans ölçütlerinden tahmin doğruluğu oranı en yüksek GAM algoritmasına ait karışıklık matrisi Çizelge 3.15.'te görülmektedir.

Çizelge 3.15. GAM karışıklık matrisi

		Gerçek Değer		
Tahmin Değeri	Sınıflar	a	b	c
	a	1615	604	538
	b	5	4	3
	c	33	23	66

a:Kazanç yok, b:Az kazanç, c:Çok kazanç

Çizelge 3.15.'teki karışıklık matrisinde kazanç yok sınıfına ait 1.653 verinin 1.615'i, az kazanç sınıfına ait 631 verinin 4'ü ve çok kazanç sınıfına ait 607 verinin 66'sı GAM algoritması ile doğru olarak sınıflandırılmıştır.

Test veri seti sonuçlarına göre tahmin doğruluk oranı en yüksek GAM algoritmasında tüm kazanç sınıflarına ait performans metrikleri Çizelge 3.16.'da görülmektedir.

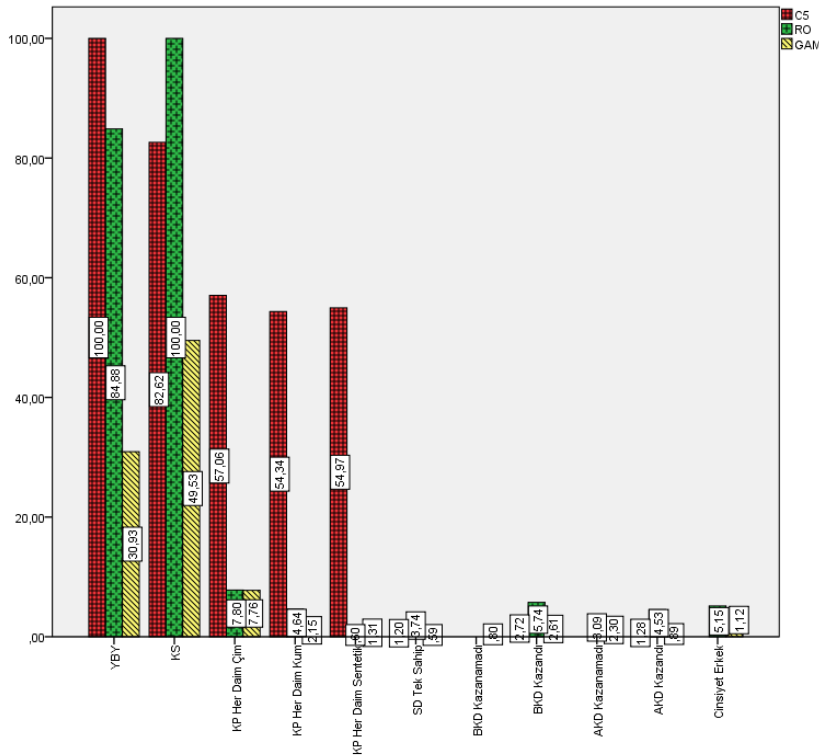
Çizelge 3.16. %20 test veri seti GAM tüm sınıflara ait performans metrikleri

	GAM		
	a	b	c
Doğruluk	0,583		
Dengeli Doğruluk	0,528	0,501	0,542
Kesinlik (Pozitif tahmin değeri)	0,586	0,333	0,541
Anma (Duyarlılık)	0,977	0,006	0,109
F1 skor	0,732	0,012	0,181
Seçicilik	0,077	0,996	0,975
Negatif tahmin değeri	0,716	0,782	0,805

a:Kazanç yok, b:Az kazanç, c:Çok kazanç

Çizelge 3.16. incelendiğinde kazanç yok sınıfına ait dengeli doğruluk oranı %52,80, anma metriği %97,70 ve kesinlik değeri %58,60 ile F1 skoru %73,20 olarak bulunmuştur. Negatif tahmin değeri %80,50 olarak çok kazanç sınıfında seçicilik değeri ise %99,60 az kazanç sınıfında diğer sınıflardan yüksek bulunmuştur.

Aşağıdaki Şekil 3.3. 'te, %80 oranındaki üç sınıflı kazanç değişkenini içeren dengeleme yapılmamış eğitim veri seti kullanılarak k-EYK hariç diğer algoritmalar ile elde edilen sınıflandırma tahmin modellerindeki özniteliklerin modele anlamlı katkılarını ve verdiği katkının önem derecesini gösteren grafik görülmektedir.



Şekil 3.3. Üç sınıflı %80 eğitim veri seti tahmin modeli öznitelik önem dereceleri

Şekil 3.3. 'te %80 eğitim veri seti için değişken önem dereceleri incelendiğinde analize dahil olan tüm özniteliklerin C5.0, GAM ve RO algoritmalarında farklı önem derecelerinde de olsa tahmin modeline katkı sağladığı görülmüştür. C5.0 ile kurulan sınıflandırma modelinde en etkili öznitelikler sırasıyla; YBY, KS, KPHerDaimÇim, KPHerDaimKum ve KPHerDaimSentetiktir. RO ile kurulan sınıflandırma modelinde yer alan en etkili öznitelikler sırasıyla; KS, YBY, KPHerDaimÇim, KPHerDaimKum ve BKDKazandıdır. GAM ile kurulan sınıflandırma modelinde en etkili öznitelikler sırasıyla; KS, YBY, KPHerDaimÇim, KPHerDaimKum, BKDKazandı ve AKDKazanamadıdır.

3.2. RAÖ ile Dengelenmiş Veri Seti Bulguları

Üç sınıflı kazanç değişkeninin olduğu orijinal veri seti %70-%30, %75-%25 ve %80-%20 eğitim ve test veri setlerine bölünmüş ve eğitim veri setlerindeki sınıf dengesizliği probleminin çözümünde RAÖ yöntemi kullanılarak yeni eğitim veri setleri oluşturulmuş ve bunlar ile k-EYK, C5.0, RO ve GAM sınıflandırma MÖ algoritmaları eğitilmiş ve test veri setleri ile sınanarak elde edilen bulgular aşağıdaki bölümde incelenmiştir.

3.2.1. Tanımlayıcı İstatistikler

Veri seti, %70 eğitim ve %30 test verisi olarak bölünmüş, eğitim veri setindeki iki azınlık sınıf RAÖ ile yeniden örnekleme yöntemi ile en yüksek sınıf sayısına (5.693 adet) eşitlenmiş ve toplam 17.079 adet gözlem oluşmuştur. Test veri seti ise 4.268 adet gözlemden oluşmaktadır.

Veri seti, %75 eğitim ve %25 test verisi olarak bölünmüş, eğitim veri setindeki iki azınlık sınıf RAÖ ile yeniden örnekleme yöntemi ile en yüksek sınıf sayısına (6.099 adet) eşitlenmiş ve toplam 18.297 adet gözlem oluşmuştur. Test veri seti ise 3.558 adet gözlemden oluşmaktadır.

Veri seti, %80 eğitim ve %20 test verisi olarak bölünmüş, eğitim veri setindeki iki azınlık sınıf RAÖ ile yeniden örnekleme yöntemi ile en yüksek sınıf sayısına (6.506 adet) eşitlenmiş ve toplam 19.518 adet gözlem oluşmuştur. Test veri seti ise 2.890 adet gözlemden oluşmaktadır.

3.2.2. MÖ Sınıflandırma Algoritma Sonuçları

Aşağıdaki bölümde k-EYK, C5.0, RO ve GAM sınıflandırma algoritmalarının %70, %75 ve %80 oranlarındaki dengelenmiş yeni eğitim veri setine ve %30, %25 ve %20 oranlarındaki orijinal test veri setlerine uygulanması sonucunda elde edilen performans ölçütleri, hiper parametre değerleri ve karşılaştırmalı performans ölçütleri çizelgelerde, değişkenlerin önem dereceleri grafiklerde gösterilmiş ve en iyi doğruluk oranına sahip algoritmanın karışıklık matrisi ise çizelgede sunulmuştur.

3.2.3. Eğitim (%70) ve Test (%30) Veri Seti Sonuçları

Veri seti, %70 eğitim ve %30 test verisi olarak bölünmüş, eğitim veri setindeki iki azınlık sınıf RAÖ ile yeniden örnekleme yöntemi ile en yüksek sınıf sayısına (5.693 adet) eşitlenmiş ve toplam 17.079 adet gözlem oluşmuştur. Test veri seti ise 4.268 adet gözlemden oluşmaktadır.

Eğitim veri seti kullanılarak dört ayrı sınıflandırma algoritması ile sınıflandırma tahmin modeli öğrenilmiş, performans ölçütleri, en iyi hiper parametre değerleri ve test veri setinden elde edilen karşılaştırmalı performans ölçütleri sırasıyla aşağıda sunulmuştur.

3.2.4. Eğitim (%70) Veri Seti Performans Sonuçları

Eğitim veri setindeki 17.079 adet gözlem ile ÇD model performans değerlendirme yöntemi 10 katlı ve 10 tekrarlı uygulanarak dört algoritma eğitilmiştir. Hiper parametre en iyileme yöntemi olarak ızgara araştırması kullanılarak her bir algoritma için Çizelge 3.17.'de görülen hiper parametre değerleri elde edilmiştir.

Çizelge 3.17. RAÖ'lü %70 eğitim veri seti hiper parametre değerleri

Algoritmalar	Hiper parametreler	Değer
k-EYK	“k” en yakın komşu sayısı	5
RO	“mtry” köklerin dallara ayrılmasında kullanılan rastgele seçilmiş özellik sayısı	11
GAM	“n.trees” kurulacak KA sayısı	150
	“interaction.depth” her bir ağacın en büyük derinlik sayısı	3
	“shrinkage” öğrenme oranı	0,1
	“n.minobsinnode” terminal düğümlerdeki minimum gözlem sayısı	10

Hiper parametre en iyilemesi sonucunda, k-EYK için “k” en yakın komşuluk değeri 5, RO için “mtry” köklerin dallara ayrılmasında kullanılan rastgele seçilmiş özellik sayısı 11 olarak bulunmuştur. GAM algoritması için “n.trees” kurulacak karar ağaç sayısı 150, “interaction.depth” her bir ağacın en büyük derinlik sayısı 3, “shrinkage” öğrenme oranı 0,1 ve “n.minobsinnode” terminal düğümlerdeki minimum gözlem sayısı 10 olarak bulunmuştur (Çizelge 3.17.).

Yukarıda önerilen hiper parametre değerleri kullanılarak sınıflandırma algoritmalarının eğitim veri setinden öğrendikleri tahmin modellerine ait performans ölçüt sonuçları Çizelge 3.18.'de gösterilmiştir.

Çizelge 3.18. RAÖ'lü %70 eğitim veri seti performans metrikleri

	C5.0	k-EYK	RO	GAM
Doğruluk	0,551	0,515	0,672	0,476
Dengeli Doğruluk Ortalaması	0,663	0,636	0,754	0,607
Ortalama Kesinlik (Pozitif Tahmin	0,554	0,518	0,670	0,477
Ortalama Anma (Duyarlılık)	0,551	0,515	0,672	0,476
Ortalama F1 skor	0,547	0,512	0,669	0,473
Ortalama Seçicilik	0,775	0,757	0,836	0,738
Ortalama Negatif Tahmin değeri	0,778	0,759	0,838	0,739

Çizelge 3.18. incelendiğinde RO algoritmasının en yüksek doğruluk oranına %67,20 sahip olduğu görülmüştür. RO için dengeli doğruluk ortalaması %75,40, kesinliğin ortalaması %67,00, anma metriği ortalaması %67,20, kesinlik ve anma performans ölçütlerini beraber değerlendiren F1 skor ortalaması %66,90, ortalama seçicilik %83,60 ve ortalama negatif tahmin değeri %83,80 olarak bulunmuştur.

3.2.5. Test (%30) Veri Seti Performans Sonuçları

Test veri seti olarak 4.268 adet örnek ile eğitim veri setinden öğrenilen model deseni k-EYK, C5.0, RO ve GAM sınıflandırma algoritmaları ve belirlenen en iyi hiper parametreler kullanılarak test edilmiştir. Sonuçta, doğruluk, dengeli doğruluk ve ÇSMKK ile kesinlik, anma, F-ölçütlerinin makro ortalamaları, seçicilik ve negatif tahmin değeri ortalaması ile G-ortalama ve EAKA değeri gibi ulaşılan performans ölçütleri Çizelge 3.19.'da gösterilmiştir.

Çizelge 3.19. RAÖ'lü %30 test veri seti performans metrikleri

	C5.0	k-EYK	RO	GAM
Doğruluk	0,426 %95GA (0,411-0,441)	0,383 %95GA (0,368-0,398)	0,440 %95GA (0,425-0,455)	0,455 %95GA (0,440-0,470)
Dengeli Doğruluk Ortalaması	0,555	0,547	0,537	0,594
Makro Ortalama Kesinlik (Pozitif tahmin değeri)	0,398	0,395	0,376	0,429
Makro Ortalama Anma (Duyarlılık)	0,400	0,390	0,378	0,433
Makro Ortalama F1 skor	0,387	0,364	0,375	0,412
Ortalama Seçicilik	0,710	0,705	0,698	0,733
Ortalama Negatif tahmin değeri	0,703	0,696	0,695	0,721

Çizelge 3.19. Devam RAÖ'lü %30 test veri seti performans metrikleri

ÇSMKK	0,111	0,095	0,168	0,168
G-ortalama	0,395	0,389	0,359	0,417
EAKA	0,562	0,558	0,575	0,621

Çizelge 3.19.'da GAM algoritmasının doğruluk oranı (%45,50, %95GA; %44,00-%47,00) diğer algoritmalarından daha yüksek bulunmuştur. GAM için dengeli doğruluk ortalaması %59,40, kesinliğin makro ortalaması %42,90, anma metriği makro ortalaması %43,30, F1 skoru makro ortalaması %41,20, ortalama seçicilik %73,30 ve ortalama negatif tahmin değeri %72,10 olarak bulunmuştur. Bunun yanında GAM algoritması için ÇSMKK değeri %16,80, EAKA değeri %62,10 ve G-ortalama değeri %41,70 olarak elde edilmiştir.

Test veri seti sonuçlarına göre tahmin doğruluk oranı en yüksek GAM algoritmasına ait karışıklık matrisi Çizelge 3.20.'de görülmektedir.

Çizelge 3.20. GAM Karışıklık matrisi

Tahmin Değeri	Gerçek Değer			
	Sınıflar	a	b	c
	a	1210	245	184
	b	504	256	253
	c	725	414	477

a:Kazanç yok, b:Az kazanç, c:Çok kazanç

Çizelge 3.20.'deki karışıklık matrisinde kazanç yok sınıfına ait 2.439 verinin 1.210'u, az kazanç sınıfına ait 915 verinin 256'sı ve çok kazanç sınıfına ait 914 verinin 477'si GAM algoritması ile doğru olarak sınıflandırılmıştır.

Test veri seti sonuçlarına göre performans ölçütlerinden tahmin doğruluğu oranı en yüksek GAM algoritmasında tüm kazanç sınıflarına ait performans metrikleri Çizelge 3.21.'de görülmektedir.

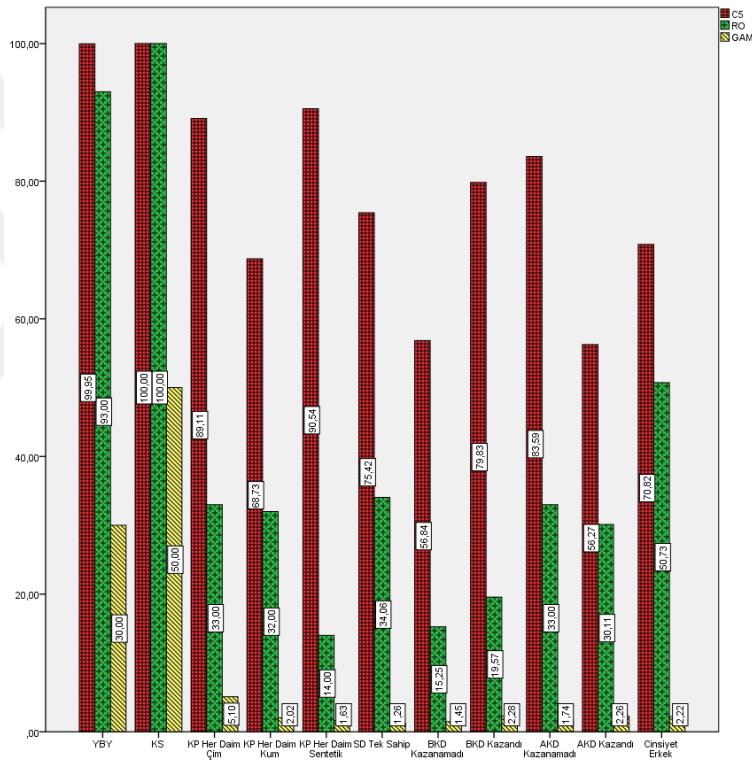
Çizelge 3.21. RAÖ'lü %30 test veri seti GAM performans metrikleri

	GAM		
	a	b	c
Doğruluk	0,455		
Dengeli Doğruluk Ortalaması	0,631	0,527	0,591
Kesinlik (Pozitif tahmin değeri)	0,738	0,253	0,295
Anma (Duyarlılık)	0,496	0,280	0,522
F1 skor	0,593	0,266	0,377
Seçicilik	0,765	0,774	0,660
Negatif tahmin değeri	0,533	0,798	0,835

a:Kazanç yok, b:Az kazanç, c:Çok kazanç

Çizelge 3.21.'deki GAM algoritma sonuçlarına göre, kazanç yok sınıfında dengeli doğruluk oranı %63,10, kesinlik değeri %73,80 ve F1 skoru %59,30 diğer sınıflardan yüksek bulunmuştur. Anma metriği %96,80 ve negatif tahmin değeri %83,50 çok kazanç sınıfında seçicilik değeri %77,40 ise az kazanç sınıfında yüksek bulunmuştur.

Aşağıdaki Şekil 3.4.'te, %70 oranlı eğitim veri setine RAÖ yöntemi uygulanarak elde edilen yeni eğitim veri seti kullanılarak k-EYK hariç diğer algoritmalar ile elde edilen sınıflandırma tahmin modellerine dahil olan öz niteliklerin modele anlamlı katkılarını ve verdiği katkının önem derecesini gösteren grafik görülmektedir.



Şekil 3.4. RAÖ ile dengelenen üç sınıflı %70 eğitim veri seti tahmin modeli öz nitelik önem dereceleri

Şekil 3.4.'te %70 eğitim veri seti için değişken önem dereceleri incelendiğinde analize dahil olan tüm öz niteliklerin C5.0, GAM ve RO algoritmalarında farklı önem derecelerinde de olsa tahmin modeline katkı sağladığı görülmüştür. C5.0 ile kurulan sınıflandırma modelinde en etkili öz nitelikler sırasıyla; YBY, KS, KPHerDaimSentetik, KPHerDaimÇim, AKDKazanamadı ve BKDKazandıdır. RO ile kurulan sınıflandırma modelinde yer alan en etkili öz nitelikler sırasıyla; KS, YBY, CinsiyetErkek, SDTekSahip, KPHerDaimÇim, KPHerDaimKum, AKDKazanamadı ve AKDKazandıdır. GAM ile kurulan sınıflandırma

modelinde en etkili öznitelikler sırasıyla; KS, YBY, KPHerDaimÇim, CinsiyetErkek, BKDKazandı ve AKDKazandıdır.

3.2.6. Eğitim (%75) ve Test (%25) Veri Seti Sonuçları

Veri seti, %75 eğitim ve %25 test verisi olarak bölünmüş, eğitim veri setindeki iki azınlık sınıf RAÖ ile yeniden örnekleme yöntemi ile en yüksek sınıf sayısına (6.099 adet) eşitlenmiş ve toplam 18.297 adet gözlem oluşmuştur. Test veri seti ise 3.558 adet gözlemden oluşmaktadır.

Eğitim veri seti kullanılarak dört ayrı sınıflandırma algoritması ile sınıflandırma tahmin modeli öğrenilmiş, performans ölçütleri, en iyi hiper parametre değerleri ve test veri setinden elde edilen karşılaştırmalı performans ölçütleri sırasıyla aşağıda sunulmuştur.

3.2.7. Eğitim (%75) Veri Seti Performans Sonuçları

Eğitim veri setindeki 18.297 adet örnek ile ÇD model performans değerlendirme yöntemi 10 katlı ve 10 tekrarlı uygulanarak dört algoritma eğitilmiştir. Hiper parametre en iyileme yöntemi olarak ızgara araştırması kullanılarak her bir algoritma için Çizelge 3.22.'de görülen hiper parametre değerleri elde edilmiştir.

Çizelge 3.22. RAÖ'lü %75 eğitim veri seti hiper parametre değerleri

Algoritmalar	Hiper parametreler	Değer
k-EYK	“k” en yakın komşu sayısı	5
RO	“mtry” köklerin dallara ayrılmasında kullanılan rastgele seçilmiş özellik sayısı	11
GAM	“n.trees” kurulacak KA sayısı	150
	“interaction.depth” her bir ağacın en büyük derinlik sayısı	3
	“shrinkage” öğrenme oranı	0,1
	“n.minobsinnode” terminal düğümlerdeki minimum gözlem sayısı	10

Hiper parametre en iyilemesi sonucunda, k-EYK için “k” en yakın komşuluk değeri 5, RO algoritması için “mtry” köklerin dallara ayrılmasında kullanılan rastgele seçilmiş özellik sayısı 11 olarak bulunmuştur. GAM algoritması için “n.trees” kurulacak karar ağaç sayısı 150, “interaction.depth” her bir ağacın en büyük derinlik sayısı 3, “shrinkage” öğrenme oranı 0,1 ve “n.minobsinnode” terminal düğümlerdeki minimum gözlem sayısı 10 olarak elde edilmiştir (Çizelge 3.22.).

Yukarıda önerilen hiper parametre değerleri kullanılarak sınıflandırma algoritmalarının eğitim veri setinden öğrendikleri tahmin modellerine ait performans ölçüt sonuçları Çizelge 3.23.'te gösterilmiştir.

Çizelge 3.23. RAÖ'lü %75 eğitim veri seti performans metrikleri

	C5.0	k-EYK	RO	GAM
Doğruluk	0,543	0,510	0,662	0,473
Dengeli Doğruluk Ortalaması	0,657	0,632	0,747	0,605
Ortalama Kesinlik (Pozitif tahmin)	0,545	0,514	0,661	0,475
Ortalama Anma (Duyarlılık)	0,543	0,510	0,662	0,473
Ortalama F1 skor	0,541	0,507	0,659	0,470
Ortalama Seçicilik	0,771	0,755	0,831	0,737
Ortalama Negatif tahmin değeri	0,772	0,756	0,833	0,738

Çizelge 3.23. incelendiğinde RO algoritmasının en yüksek doğruluk oranına %66,20 sahip olduğu görülmüştür. RO için dengeli doğruluk ortalaması %74,70, kesinliğin ortalaması %66,10, anma metriği ortalaması %66,20, F1 skor ortalaması %65,90, ortalama seçicilik %83,10 ve ortalama negatif tahmin değeri %83,30 olarak bulunmuştur.

3.2.8. Test (%25) Veri Seti Performans Sonuçları

Test veri seti olarak 3.558 adet örnek ile eğitim veri setinden öğrenilen model deseni k-EYK, C5.0, RO ve GAM sınıflandırma algoritmaları ve belirlenen en iyi hiper parametreler kullanılarak test edilmiştir. Sonuçta, doğruluk, dengeli doğruluk ve ÇSMKK ile kesinlik, anma, F skoru makro ortalamaları, seçicilik ve negatif tahmin değeri ortalaması ile G-ortalama ve EAKA değeri gibi ulaşılan performans ölçütleri Çizelge 3.24.'te gösterilmiştir.

Çizelge 3.24. RAÖ'lü %25 test veri seti performans metrikleri

	C5.0	k-EYK	RO	GAM
Doğruluk	0,430 %95GA (0,413-0,446)	0,402 %95GA (0,386-0,418)	0,453 %95GA (0,437-0,470)	0,454 %95GA (0,438-0,471)
Dengeli Doğruluk Ortalaması	0,557	0,557	0,553	0,586
Makro Ortalama Kesinlik (Pozitif tahmin değeri)	0,400	0,403	0,395	0,432
Makro Ortalama Anma (Duyarlılık)	0,404	0,404	0,399	0,439
Makro Ortalama F1 skor	0,391	0,380	0,394	0,414
Ortalama Seçicilik	0,710	0,709	0,706	0,733
Ortalama Negatif tahmin değeri	0,703	0,701	0,702	0,723
ÇSMKK	0,112	0,111	0,105	0,173
G-ortalama	0,401	0,404	0,385	0,422
EAKA	0,574	0,573	0,613	0,624

Çizelge 3.24. incelendiğinde GAM ve RO algoritmalarının doğruluk oranları (%45,40, %95GA; %43,80-%47,10 ve %45,30, %95GA; %43,70-%47,00) diğer algoritmalarından daha yüksek bulunmuştur. GAM için dengeli doğruluk ortalaması %58,60, kesinliğin makro ortalaması %43,20, anma metriği makro ortalaması %43,90, F1 skoru makro ortalaması %41,40, ortalama seçicilik %73,30 ve ortalama negatif tahmin değeri %72,30 olarak bulunmuştur. Bunun yanında GAM algoritması için ÇSMKK değeri %17,30, EAKA değeri %62,40 ve G-ortalama değeri %42,20 olarak elde edilmiştir.

Test veri seti sonuçlarına göre performans ölçütlerinden tahmin doğruluğu oranı en yüksek GAM ve RO algoritmalarına ait karışıklık matrisleri Çizelge 3.25. ve Çizelge 3.26.'da görülmektedir.

Çizelge 3.25. GAM karışıklık matrisi

Tahmin Değeri	Gerçek Değer			
	Sınıflar	a	b	c
	a	978	215	137
	b	419	214	201
	c	636	334	424

a:Kazanç yok, b:Az kazanç, c:Çok kazanç

Çizelge 3.26. RO karışıklık matrisi

Tahmin Değeri	Gerçek Değer			
	Sınıflar	a	b	c
	a	1122	306	323
	b	476	236	184
	c	435	221	255

a:Kazanç yok, b:Az kazanç, c:Çok kazanç

Çizelge 3.25.'teki karışıklık matrisindeki kazanç yok sınıfına ait 2.033 verinin 978'i, az kazanç sınıfına ait 763 verinin 214'ü ve çok kazanç sınıfına ait 762 verinin 424'ü GAM algoritması ile doğru olarak sınıflandırılmıştır.

Çizelge 3.26.'daki karışıklık matrisindeki kazanç yok sınıfına ait 2.033 verinin 1.122'si, az kazanç sınıfına ait 763 verinin 236'sı ve çok kazanç sınıfına ait 762 verinin 255'i RO algoritması ile doğru olarak sınıflandırılmıştır.

Test veri seti sonuçlarına göre tahmin doğruluk oranı en yüksek RO ve GAM algoritmalarında tüm kazanç sınıflarına ait performans metrikleri Çizelge 3.27.'de görülmektedir.

Çizelge 3.27. RAÖ'lü %25 test veri seti GAM ve RO performans metrikleri

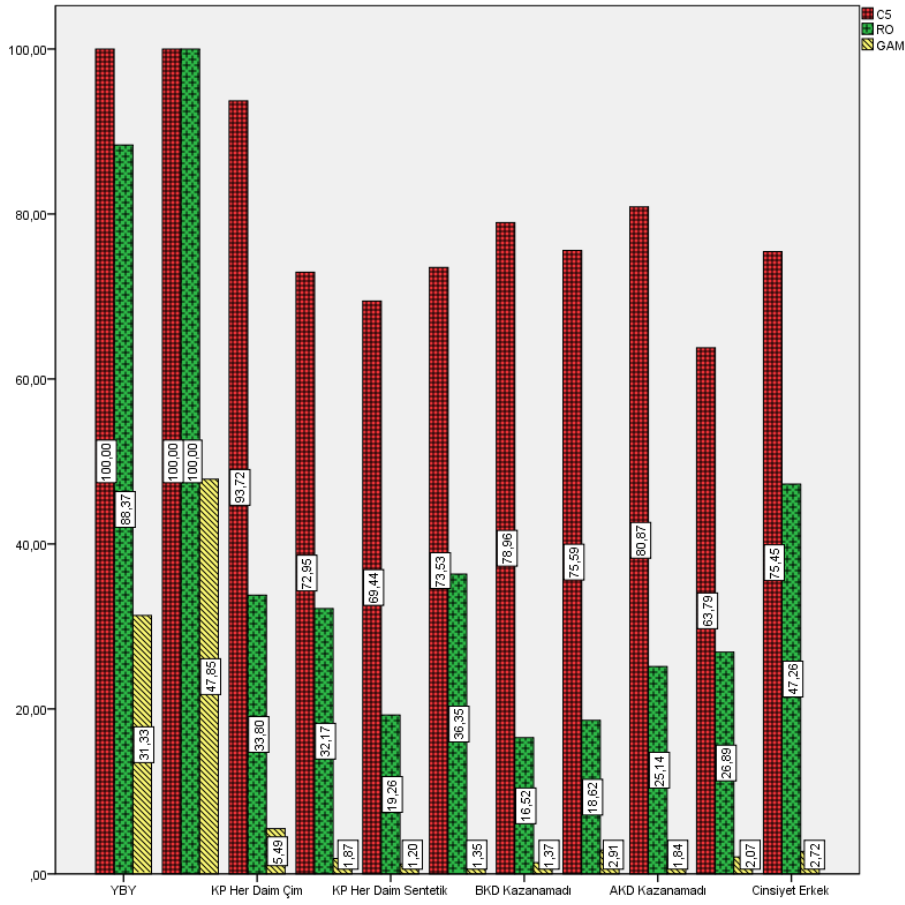
	GAM			RO		
	a	b	c	a	b	c
Doğruluk	0,454			0,453		
Dengeli Doğruluk	0,625	0,529	0,605	0,570	0,537	0,550
Kesinlik (Pozitif tahmin değeri)	0,735	0,257	0,304	0,641	0,263	0,280
Anma (Duyarlılık)	0,481	0,280	0,556	0,552	0,309	0,335
F1 skor	0,582	0,268	0,393	0,593	0,285	0,305
Seçicilik	0,769	0,778	0,653	0,587	0,764	0,765
Negatif tahmin değeri	0,527	0,798	0,844	0,496	0,802	0,808

a:Kazanç yok, b:Az kazanç, c:Çok kazanç

Çizelge 3.27.'deki GAM algoritma sonuçlarına göre, kazanç yok sınıfında dengeli doğruluk oranı %62,50, kesinlik değeri %73,50 ve F1 skoru %58,20 diğer sınıflardan yüksek bulunmuştur. Anma metriği %55,60 ve negatif tahmin değeri %84,40 çok kazanç sınıfında seçicilik değeri %77,80 ise az kazanç sınıfında yüksek bulunmuştur.

Çizelge 3.27.'deki RO algoritma sonuçlarına göre, kazanç yok sınıfında dengeli doğruluk oranı %57,00, kesinlik değeri %64,10, anma metriği %55,20 ve F1 skoru %59,30 diğer sınıflardan yüksek bulunmuştur. Seçicilik değeri %76,50 ve negatif tahmin değeri %80,80 çok kazanç sınıfında yüksek bulunmuştur.

Aşağıdaki Şekil 3.5.'te, %75 üç sınıflı kazanç değişkenini içeren RAÖ yöntemi uygulanarak dengelenmiş eğitim veri seti kullanılarak k-EYK hariç diğer algoritmalar ile elde edilen sınıflandırma tahmin modellerine dahil olan özneliliklerin modele anlamlı katkılarını ve verdiği katkının önem derecesini gösteren grafik görülmektedir.



Şekil 3.5. RAÖ ile dengelenen üç sınıflı %75 eğitim veri seti tahmin modeli öz nitelik önem dereceleri

Şekil 3.5.'te %75 eğitim veri seti için değişken önem dereceleri incelendiğinde analize dahil olan tüm öz niteliklerin C5.0, GAM ve RO algoritmalarında farklı önem dereceleri de olsa tahmin modeline katkı sağladığı görülmüştür. C5.0 ile kurulan sınıflandırma modelinde en etkili öz nitelikler sırasıyla; YBY, KS, KPHerDaimÇim, AKDKazanamadı, BKDKazanamadı, BKDKazandı ve CinsiyetErkektir. RO ile kurulan sınıflandırma modelinde yer alan en etkili öz nitelikler sırasıyla; KS, YBY, CinsiyetErkek, SDTekSahip, KPHerDaimÇim, KPHerDaimKum, AKDKazandı ve AKDKazanamadıdır. GAM ile kurulan sınıflandırma modelinde en etkili öz nitelikler sırasıyla; KS, YBY, KPHerDaimÇim, CinsiyetErkek, BKDKazandı ve AKDKazandıdır.

3.2.9. Eğitim (%80) ve Test (%20) Veri Seti Sonuçları

Veri seti, %80 eğitim ve %20 test verisi olarak bölünmüş, eğitim veri setindeki iki azınlık sınıf RAÖ ile yeniden örnekleme yöntemi ile en yüksek sınıf sayısına (6.506 adet) eşitlenmiş

ve toplam 19.518 adet gözlem oluşmuştur. Test veri seti ise 2.890 adet gözlemden oluşmaktadır.

Eğitim veri seti kullanılarak dört ayrı sınıflandırma algoritması ile sınıflandırma tahmin modeli öğrenilmiş, performans ölçütleri, en iyi hiper parametre değerleri ve test veri setinden elde edilen karşılaştırmalı performans ölçütleri sırasıyla aşağıda sunulmuştur.

3.2.10. Eğitim (%20) Veri Seti Performans Sonuçları

Eğitim veri setindeki 19.518 adet örnek ile ÇD model performans değerlendirme yöntemi 10 katlı ve 10 tekrarlı uygulanarak dört algoritma eğitilmiştir. Hiper parametre en iyileme yöntemi olarak ızgara araştırması kullanılarak her bir algoritma için Çizelge 3.28.'de görülen hiper parametre değerleri elde edilmiştir.

Çizelge 3.28. RAÖ'lü %80 eğitim veri seti hiper parametre değerleri

Algoritmalar	Hiper parametreler	Değer
k-EYK	“k” en yakın komşu sayısı	5
RO	“mtry” köklerin dallara ayrılmasında kullanılan rastgele seçilmiş özellik sayısı	11
GAM	“n.trees” kurulacak KA sayısı	150
	“interaction.depth” her bir ağacın en büyük derinlik sayısı	3
	“shrinkage” öğrenme oranı	0,1
	“n.minobsinnode” terminal düğümlerdeki minimum gözlem sayısı	10

Hiper parametre en iyilemesi sonucunda, k-EYK için “k” en yakın komşuluk değeri 5, RO algoritması için “mtry” köklerin dallara ayrılmasında kullanılan rastgele seçilmiş özellik sayısı 11 olarak bulunmuştur. GAM algoritması için “n.trees” kurulacak karar ağaç sayısı 150, “interaction.depth” her bir ağacın en büyük derinlik sayısı 3, “shrinkage” öğrenme oranı 0,1 ve “n.minobsinnode” terminal düğümlerdeki minimum gözlem sayısı 10 olarak bulunmuştur (Çizelge 3.28.).

Yukarıda önerilen hiper parametre değerleri kullanılarak sınıflandırma algoritmalarının eğitim veri setinden öğrendikleri tahmin modellerine ait performans ölçüt sonuçları Çizelge 3.29.'de gösterilmiştir.

Çizelge 3.29. RAÖ'lü %80 eğitim veri seti performans metrikleri

	C5.0	k-EYK	RO	GAM
Doğruluk	0,542	0,517	0,663	0,469
Dengeli Doğruluk Ortalaması	0,656	0,638	0,747	0,602
Ortalama Kesinlik (Pozitif tahmin değeri)	0,544	0,521	0,661	0,470
Ortalama Anma (Duyarlılık)	0,542	0,517	0,663	0,469
Ortalama F1 skor	0,539	0,514	0,659	0,467
Ortalama Seçicilik	0,771	0,758	0,831	0,735
Ortalama Negatif tahmin değeri	0,772	0,760	0,834	0,736

Çizelge 3.29. incelendiğinde RO algoritmasının en yüksek doğruluk oranına %66,30 sahip olduğu görülmüştür. RO için dengeli doğruluk ortalaması %74,70, kesinliğin ortalaması %66,10, anma metriği ortalaması %66,30, F1 skor ortalaması %65,90, ortalama seçicilik %83,10 ve ortalama negatif tahmin değeri %83,40 olarak bulunmuştur.

3.2.11. Test (%20) Veri Seti Performans Sonuçları

Test veri seti olarak 2.890 adet örnek ile eğitim veri setinden öğrenilen model deseni k-EYK, C5.0, RO ve GAM sınıflandırma algoritmaları ve belirlenen en iyi hiper parametreler kullanılarak test edilmiştir. Sonuçta, doğruluk, dengeli doğruluk ve ÇSMKK ile kesinlik, anma, F skoru makro ortalamaları, seçicilik ve negatif tahmin değeri ortalaması ile EAKA değeri gibi ulaşılan performans ölçütleri Çizelge 3.30.'da gösterilmiştir.

Çizelge 3.30. RAÖ'lü %20 test veri seti performans metrikleri

	C5.0	k-EYK	RO	GAM
Doğruluk	0,419 %95GA (0,401-0,438)	0,388 %95GA (0,370-0,407)	0,462 %95GA (0,443-0,480)	0,461 %95GA (0,442-0,479)
Dengeli Doğruluk	0,552	0,548	0,553	0,589
Makro Ortalama Kesinlik (Pozitif tahmin değeri)	0,394	0,394	0,396	0,438
Makro Ortalama Anma (Duyarlılık)	0,396	0,390	0,399	0,440
Makro Ortalama F1	0,383	0,367	0,396	0,421
Ortalama Seçicilik	0,707	0,705	0,708	0,738
Ortalama Negatif tahmin değeri	0,700	0,696	0,705	0,725
ÇSMKK	0,103	0,096	0,110	0,180
G-ortalama	0,394	0,390	0,381	0,432
EAKA	0,569	0,554	0,606	0,622

Çizelge 3.30. incelendiğinde GAM ve RO algoritmalarının doğruluk oranları (%46,10, %95GA; %44,20-%47,90 ve %46,20, %95GA; %44,30-%48,00) diğer algoritmalarından daha

yüksek bulunmuştur. GAM için dengeli doğruluk ortalaması %58,90, kesinliğin makro ortalaması %43,80, anma metriği makro ortalaması %44,00, F1 skor makro ortalaması %42,10, ortalama seçicilik %73,80 ve ortalama negatif tahmin değeri %72,25 olarak bulunmuştur. Bunun yanında GAM algoritması için ÇSMKK değeri %18,00, EAKA değeri %62,20 ve G-ortalama değeri %43,20 olarak elde edilmiştir.

Test veri seti sonuçlarına göre performans ölçütlerinden tahmin doğruluğu oranı en yüksek RO ve GAM algoritmalarına ait karışıklık matrisleri Çizelge 3.31. ve Çizelge 3.32.'de görülmektedir.

Çizelge 3.31. RO karışıklık matrisi

		Gerçek Değer		
Tahmin Değeri	Sınıflar	a	b	c
	a	934	270	241
	b	376	179	168
	c	316	161	200

a:Kazanç yok, b:Az kazanç, c:Çok kazanç

Çizelge 3.32. GAM karışıklık matrisi

Gerçek Değer				
Tahmin Değeri	Sınıflar	a	b	c
	a	809	159	110
	b	349	196	193
	c	468	255	306

a:Kazanç yok, b:Az kazanç, c:Çok kazanç

Çizelge 3.31.'de karışıklık matrisindeki kazanç yok sınıfına ait 1.626 verinin 934'ü, az kazanç sınıfına ait 610 verinin 179'u ve çok kazanç sınıfına ait 609 verinin 200'ü RO algoritması ile doğru olarak sınıflandırılmıştır. Çizelge 3.32.'de karışıklık matrisindeki kazanç yok sınıfına ait 1.626 verinin 809'u, az kazanç sınıfına ait 610 verinin 196'sı ve çok kazanç sınıfına ait 609 verinin 306'sı GAM algoritması ile doğru olarak sınıflandırılmıştır.

Test veri seti sonuçlarına göre performans ölçütlerinden tahmin doğruluğu oranı en yüksek RO ve GAM algoritmalarındaki tüm kazanç sınıflarına ait performans metrikleri Çizelge 3.33.'te görülmektedir.

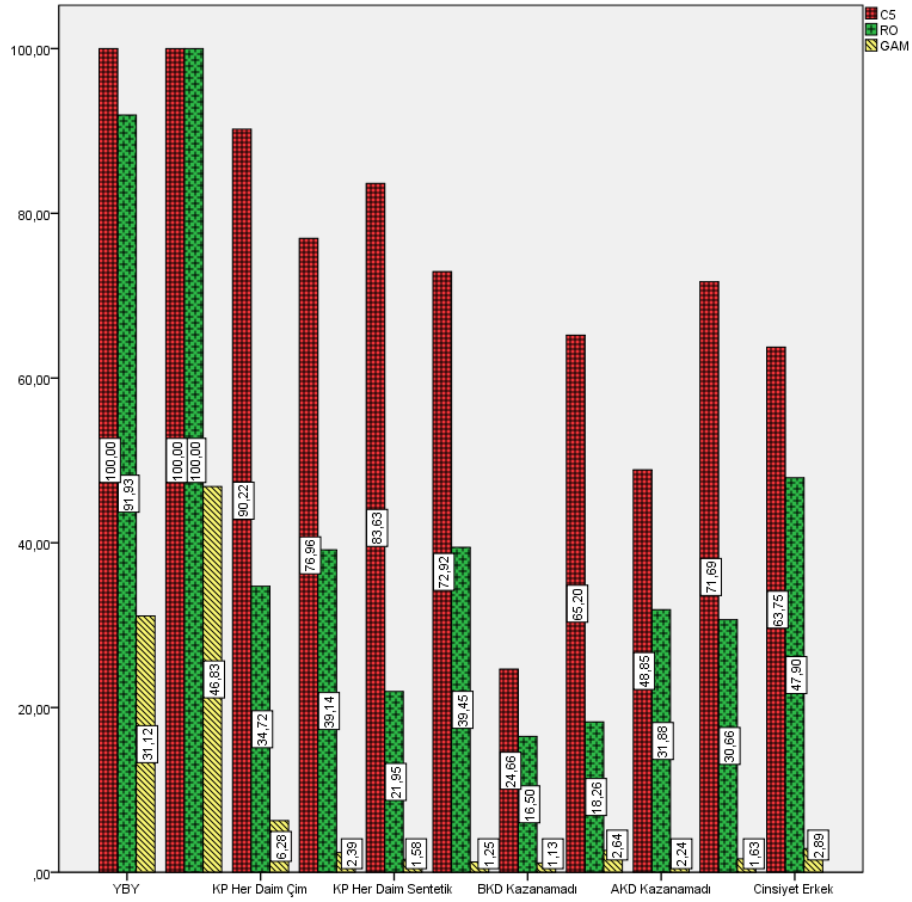
Çizelge 3.33. RAÖ'lü %20 test veri seti GAM ve RO performans metrikleri

	RO			GAM		
	a	b	c	a	b	c
Doğruluk	0,462			0,461		
Dengeli Doğruluk	0,578	0,525	0,558	0,638	0,539	0,590
Kesinlik (Pozitif tahmin değeri)	0,646	0,248	0,295	0,751	0,266	0,297
Anma (Duyarlılık)	0,574	0,293	0,328	0,498	0,321	0,503
F1 skor	0,608	0,268	0,311	0,598	0,291	0,373
Seçicilik	0,581	0,757	0,787	0,779	0,757	0,677
Negatif tahmin değeri	0,506	0,797	0,811	0,538	0,804	0,833

a:Kazanç yok, b:Az kazanç, c:Çok kazanç

Çizelge 3.33. incelendiğinde RO algoritması için kazanç yok sınıfına ait dengeli doğruluk oranı %57,80, kesinlik değeri %64,60, anma metriği %57,40 ve F1 skoru %60,80 diğer sınıflardan yüksek bulunmuştur. Çok kazanç sınıfında da seçicilik değeri %78,70 ve negatif tahmin değeri %81,10 olarak bulunmuştur. GAM algoritması için kazanç yok sınıfına ait dengeli doğruluk oranı %63,80, kesinlik değeri %75,10, F1 skoru %59,80 ve seçicilik değeri %77,90 diğer sınıflardan yüksek bulunmuştur. Çok kazanç sınıfında ise anma metriği %50,30 ve negatif tahmin değeri %83,30 olarak bulunmuştur.

Aşağıdaki Şekil 3.6.'da, %80 eğitim veri setine RAÖ yöntemi uygulanarak elde edilen yeni eğitim veri seti kullanılarak k-EYK hariç diğer algoritmalar ile elde edilen sınıflandırma tahmin modellerine dahil olan özneliliklerin modele anlamlı katkılarını ve verdiği katkının önem derecesini gösteren grafik görülmektedir.



Şekil 3.6. RAÖ ile dengelenen üç sınıflı %80 eğitim veri seti tahmin modeli öznitelik önem dereceleri

Şekil 3.6. incelendiğinde analize dahil olan tüm özniteliklerin C5.0, GAM ve RO algoritmalarında farklı önem dereceleri de olsa tahmin modeline katkı sağladığı görülmüştür. C5.0 ile kurulan sınıflandırma modelinde en etkili öznitelikler sırasıyla; YBY, KS, KPHerDaimÇim, KPHerDaimSentetik, KPHerDaimKum, SDTekSahip, AKDKazandı, BKDKazandı ve CinsiyetErkektir. RO ile kurulan sınıflandırma modelinde yer alan en etkili öznitelikler sırasıyla; KS, YBY, CinsiyetErkek, SDTekSahip, KPHerDaimKum, KPHerDaimÇim, AKDKazanamadı ve AKDKazandıdır. GAM ile kurulan sınıflandırma modelinde en etkili öznitelikler sırasıyla; KS, YBY, KPHerDaimÇim, BKDKazandı, CinsiyetErkek ve AKDKazanamadıdır.

3.3. İki Sınıflı Veri Seti Bulguları

İki sınıflı kazanç değişkenini içeren eğitim veri seti kullanılarak k-EYK, C5.0, RO ve GAM sınıflandırma MÖ algoritmaları eğitilmiş ve oluşturulan tahmin modeline ait eğitim ve test aşamasındaki bütün bulgular çizelge ve grafikler vasıtasıyla aşağıda gösterilmiştir.

3.3.1. Tanımlayıcı İstatistikler

Aşağıda ilk olarak yedi açıklayıcı özniteliğin iki sınıflı kazanç değişkenini içeren veri seti içerisindeki dağılımları Çizelge 3.34.'te görülmektedir.

Çizelge 3.34. İki sınıflı kazanç veri seti açıklayıcı öznitelik dağılımları (n, %)

		Kazanç Yok (n=8.132)	Kazanç Var (n=6.101)
Cinsiyet	Erkek	3.877 (%47,7)	2.969 (%48,7)
	Dişi	4.255 (%52,3)	3.132 (%51,3)
Koşu Parkuru (KP)	Değişken Parkur	5.978 (%73,5)	4.742 (%77,7)
	Her Daim Çim	634 (%7,8)	504 (%8,3)
	Her Daim Kum	1.286 (%15,8)	735 (%12,0)
	Her Daim Sentetik	234 (%2,9)	120 (%2,0)
Anne Kazanç Durumu (AKD)	Kazanamadı	1.035 (%12,7)	585 (%9,6)
	Kazandı	4.771 (%58,7)	3.598 (%59,0)
	Yarışlara Katılmamış	2.326 (%28,6)	1.918 (%31,4)
Baba Kazanç Durumu (BKD)	Kazanamadı	143 (%1,8)	98 (%1,6)
	Kazandı	3.978 (%48,9)	2.687 (%44,0)
	Yarışlara Katılmamış	4.011 (%49,3)	3.316 (%54,4)
Sahiplik Durumu (SD)	Hisseli/Ortak	920 (%11,3)	727 (%11,9)
	Tek Sahip	7.212 (%88,7)	5.374 (%88,1)
Yarışa Başlama Yaşı (YBY) (A.Ort±SS) (Ay)		31,50±4,78	29,48±3,47
Ortanca (Min.-Maks.)		31 (22-84)	29 (22-57)
Koşu Sayısı (KS) (A.Ort±SS) (Adet)		15,19±18,30	20,14±19,05
Ortanca (Min.-Maks.)		9 (1-265)	15 (1-309)

Atların cinsiyet dağılımları iki sınıfta da dengeli görülmüştür. KP, SD, AKD ve BKD değişkenleri iki sınıfta da homojen dağılmıştır. YBY ortalaması (ort±ss) kazancı olmayan sınıfta 31,50±4,78 ay ve kazancı olanda ise 29,48±3,47 ay KS ortalaması kazancı olmayan sınıfta 15,19±18,30 adet ve kazancı olanda ise 20,14±19,05 adet olarak bulunmuştur (Çizelge 3.34.).

3.3.2. MÖ Sınıflandırma Algoritma Sonuçları

Aşağıdaki bölümde k-EYK, C5.0, RO ve GAM sınıflandırma algoritmalarının %70, %75 ve %80 oranlarındaki iki sınıflı kazanç değişkenini içeren eğitim veri setine ve %30, %25 ve %20 oranlarındaki test veri setlerine uygulanması sonucunda elde edilen performans ölçütleri, hiper parametre değerleri ve karşılaştırmalı performans ölçütleri çizelgelerde, değişkenlerin önem dereceleri grafiklerde gösterilmiş ve en iyi doğruluk oranına sahip algoritmanın karışıklık matrisi ise çizelgede sunulmuştur.

3.3.3. Eğitim (%70) ve Test (%30) Veri Seti Sonuçları

Veri seti ilk olarak, %70 eğitim ve %30 test verisi olarak bölünmüş, eğitim veri setini 9.976 adet örnek, test veri setini ise 4.257 adet örnek oluşturmuştur. Performans değerlendirme yöntemi olan ÇD tekniği 10 katlı ve 10 tekrarlı olarak uygulanmış, eğitim veri seti kullanılarak dört ayrı sınıflandırma algoritması ile sınıflandırma tahmin modeli öğrenilmiş, performans ölçütleri elde edilmiş ve en iyi hiper parametre değerleri için ayarlama yapılmıştır. Bu hiper parametre değerleri ile oluşturulan tahmin modeli test veri seti ile sınanmış ve karşılaştırmalı performans ölçütleri elde edilmiştir. Elde edilen bulgular sırasıyla aşağıda sunulmuştur.

3.3.4. Eğitim (%70) Veri Seti Performans Sonuçları

Eğitim veri setindeki 9.976 adet örnek ile ÇD model performans değerlendirme yöntemi 10 katlı ve 10 tekrarlı uygulanarak dört algoritma eğitilmiştir. Hiper parametre en iyileme yöntemi olarak ızgara araştırması kullanılarak her bir algoritma için Çizelge 3.35.'te görülen hiper parametre değerleri elde edilmiştir.

Çizelge 3.35. İki sınıflı %70 eğitim veri seti hiper parametre değerleri

Algoritmalar	Hiper parametreler	Değer
k-EYK	“k” en yakın komşu sayısı	9
RO	“mtry” köklerin dallara ayrılmasında kullanılan rastgele seçilmiş özellik sayısı	2
GAM	“n.trees” kurulacak KA sayısı	100
	“interaction.depth” her bir ağacın en büyük derinlik sayısı	3
	“shrinkage” öğrenme oranı	0,1
	“n.minobsinnode” terminal düğümlerdeki minimum gözlem sayısı	10

Hiper parametre en iyilemesi sonucunda, k-EYK için “k” en yakın komşuluk değeri 9, RO için “mtry” köklerin dallara ayrılmasında kullanılan rastgele seçilmiş özellik sayısı 2 olarak bulunmuştur. GAM algoritması için “n.trees” kurulacak karar ağaç sayısı 100, “interaction.depth” her bir ağacın en büyük derinlik sayısı 3, “shrinkage” öğrenme oranı 0,1 ve “n.minobsinnode” terminal düğümlerdeki minimum gözlem sayısı 10 olarak bulunmuştur (Çizelge 3.35.).

Yukarıda önerilen hiper parametre değerleri kullanılarak sınıflandırma algoritmalarının eğitim veri setinden öğrendikleri tahmin modellerine ait performans ölçüt sonuçları Çizelge 3.36.'da gösterilmiştir.

Çizelge 3.36. İki sınıflı %70 eğitim veri seti performans metrikleri

	C5.0	k-EYK	RO	GAM
Doğruluk	0,613	0,594	0,618	0,624
Dengeli Doğruluk	0,596	0,576	0,589	0,610
Kesinlik (Pozitif tahmin değeri)	0,649	0,633	0,636	0,663
Anma (Duyarlılık)	0,715	0,693	0,782	0,700
F1 skor	0,679	0,662	0,701	0,681
Seçicilik	0,476	0,460	0,396	0,521
Negatif tahmin değeri	0,556	0,527	0,576	0,563

Çizelge 3.36.'da %70 eğitim veri seti incelendiğinde, GAM algoritmasının en yüksek doğruluk oranına %62,40 sahip olduğu görülmüştür. Algoritmanın yüksek performans gösterdiği metrikler şu şekildedir; dengeli doğruluk oranı %61,00, kesinlik %66,30 ve seçicilik %52,10 olarak bulunmuştur. RO algoritmasının yüksek performans gösterdiği metrikler şu şekildedir; anma metriği %78,20, F1 skoru %70,10 ve negatif tahmin değeri %57,60'dır.

3.3.5. Test (%30) Veri Seti Performans Sonuçları

Test veri seti olarak 4.257 adet örnek ile eğitim veri setinden öğrenilen model deseni k-EYK, C5.0, RO ve GAM sınıflandırma algoritmaları ve belirlenen en iyi hiper parametreler kullanılarak test edilmiştir. Sonuçta doğruluk, dengeli doğruluk, kesinlik, anma, F1 skoru, seçicilik, negatif tahmin değeri, MKK ve EAKA değeri gibi ulaşılan performans ölçütleri Çizelge 3.37.'de gösterilmiştir.

Çizelge 3.37. İki sınıflı %30 test veri seti performans metrikleri

	C5.0	k-EYK	RO	GAM
Doğruluk	0,632 %95GA (0,617-0,646)	0,574 %95GA (0,559-0,589)	0,621 %95GA (0,606-0,636)	0,640 %95GA (0,625-0,654)
Dengeli Doğruluk	0,617	0,559	0,595	0,627
Kesinlik (Pozitif tahmin değeri)	0,657	0,612	0,631	0,667
Anma (Duyarlılık)	0,730	0,674	0,795	0,725
F1 skor	0,692	0,641	0,703	0,695
Seçicilik	0,505	0,445	0,395	0,529
Negatif tahmin değeri	0,590	0,512	0,597	0,597
MKK	0,240	0,121	0,208	0,259
G-ortalama	0,607	0,548	0,560	0,619
EAKA	0,669	0,614	0,685	0,696

Çizelge 3.37.'de GAM algoritması doğruluk oranı (%64,00, %95GA; %62,50-%65,40) diğerlerinden yüksek bulunmuştur. Algoritmanın yüksek performans gösterdiği metrikler şu şekildedir; dengeli doğruluk oranı %62,70, kesinlik %66,70, seçicilik %52,90, negatif tahmin

değeri %59,70, MKK değeri %25,90, G-ortalama %61,90 ve EAKA değeri %69,60 olarak bulunmuştur. RO algoritmasının yüksek performans gösterdiği metrikler şu şekildedir; anma metriği %79,50 ve F1 skoru %70,30'dır.

Test veri seti sonuçlarına göre tahmin doğruluk oranı en yüksek GAM algoritmasına ait karışıklık matrisi Çizelge 3.38.'de görülmektedir.

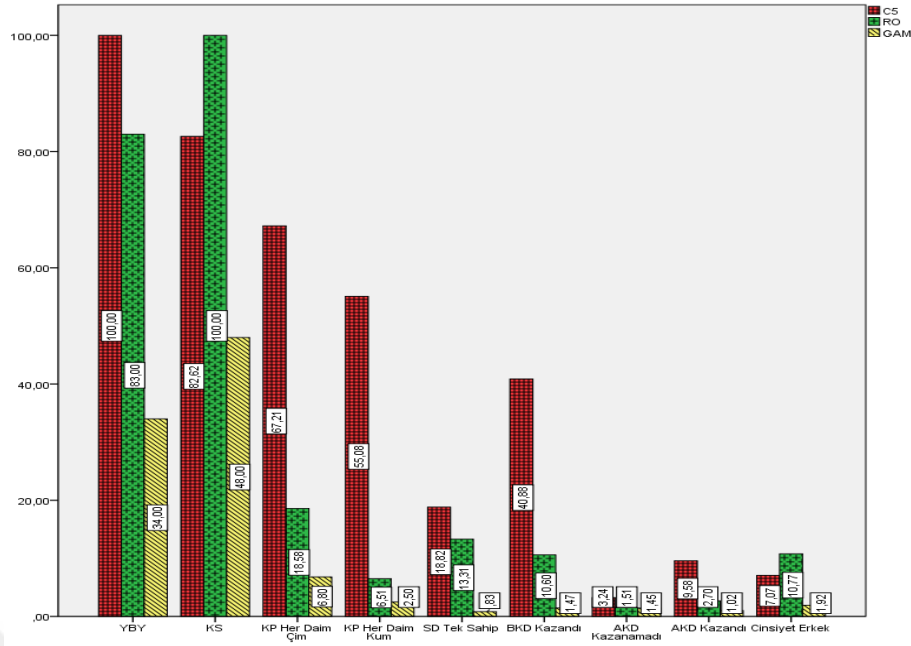
Çizelge 3.38. GAM karışıklık matrisi

Tahmin Değeri	Gerçek Değer		
	Sınıflar	a	b
	a	1745	872
	b	661	979

a:Kazanç yok, b:Kazanç var

Çizelge 3.38.'deki karışıklık matrisinde kazanç yok sınıfına ait 2.406 verinin 1.745'i ve kazanç var sınıfına ait 1.851 verinin 979'u GAM algoritması ile doğru olarak sınıflandırılmıştır.

Aşağıdaki Şekil 3.7.'de, %70 iki sınıflı kazanç değişkenini içeren eğitim veri seti kullanılarak k-EYK hariç diğer algoritmalar ile elde edilen sınıflandırma tahmin modellerine dahil olan özneliliklerin modele anlamlı katkılarını ve verdiği katkının önem derecesini gösteren grafik görülmektedir.



Şekil 3.7. İki sınıflı %70 eğitim veri seti tahmin modeli öz nitelik önem dereceleri

Şekil 3.7.'de %70 eğitim veri seti için değişken önem dereceleri incelendiğinde analize dahil olan tüm öz niteliklerin C5.0, GAM ve RO algoritmalarında farklı önem dereceleri de olsa tahmin modeline katkı sağladığı görülmüştür. C5.0 ile kurulan sınıflandırma modelinde en etkili öz nitelikler sırasıyla; YBY, KS, KPHerDaimÇim, KPHerDaimKum, BKDKazandı, SDTekSahip ve AKDKazandıdır. RO ile kurulan sınıflandırma modelinde yer alan en etkili öz nitelikler sırasıyla; KS, YBY, KPHerDaimÇim, SDTekSahip, BKDKazandı, CinsiyetErkek, BKDKazanamadıdır. GAM ile kurulan sınıflandırma modelinde en etkili öz nitelikler sırasıyla; KS, YBY, KPHerDaimÇim, KPHerDaimKum, BKDKazandı ve CinsiyetErkektir.

3.3.6. Eğitim (%75) ve Test (%25) Veri Seti Sonuçları

Veri seti %75-%25 oranında eğitim ve test verisi olarak bölündüğünde eğitim veri setindeki örnek sayısı 10.671, test veri setinde ise 3.562 olarak bulunmuştur. Performans değerlendirme yöntemi olan ÇD tekniği 10 katlı ve 10 tekrarlı olarak uygulanmış, eğitim veri seti kullanılarak dört ayrı sınıflandırma algoritması ile sınıflandırma tahmin modeli öğrenilmiş, performans ölçütleri elde edilmiş ve en iyi hiper parametre değerleri için ayarlama yapılmıştır. Bu hiper parametre değerleri ile oluşturulan tahmin modeli test veri seti ile sınanmış ve karşılaştırmalı performans ölçütleri elde edilmiştir. Elde edilen bulgular sırasıyla aşağıda sunulmuştur.

3.3.7. Eğitim (%75) Veri Seti Performans Sonuçları

Eğitim veri setindeki 10.671 adet örnek ile ÇD model performans değerlendirme yöntemi 10 katlı ve 10 tekrarlı uygulanarak dört algoritma eğitilmiştir. Hiper parametre en iyileme yöntemi olarak ızgara araştırması kullanılarak her bir algoritma için Çizelge 3.39.'da görülen hiper parametre değerleri elde edilmiştir.

Çizelge 3.39. İki sınıflı %75 eğitim veri seti hiper parametre değerleri

Algoritmalar	Hiper parametreler	Değer
k-EYK	“k” en yakın komşu sayısı	9
RO	“mtry” köklerin dallara ayrılmasında kullanılan rastgele seçilmiş	2
GAM	“n.trees” kurulacak KA sayısı	100
	“interaction.depth” her bir ağacın en büyük derinlik sayısı	3
	“shrinkage” öğrenme oranı	0,1
	“n.minobsinnode” terminal düğümlerdeki minimum gözlem sayısı	10

Hiper parametre en iyilemesi sonucunda, k-EYK için “k” en yakın komşuluk değeri 9, RO algoritması için “mtry” köklerin dallara ayrılmasında kullanılan rastgele seçilmiş özellik sayısı 2 olarak bulunmuştur. GAM algoritması için “n.trees” kurulacak karar ağaç sayısı 100, “interaction.depth” her bir ağacın en büyük derinlik sayısı 3, “shrinkage” öğrenme oranı 0,1 ve “n.minobsinnode” terminal düğümlerdeki minimum gözlem sayısı 10 olarak elde edilmiştir (Çizelge 3.39.).

Yukarıda önerilen hiper parametre değerleri kullanılarak sınıflandırma algoritmalarının eğitim veri setinden öğrendikleri tahmin modellerine ait performans ölçüt sonuçları Çizelge 3.40.'da gösterilmiştir.

Çizelge 3.40. İki sınıflı %75 eğitim veri seti performans metrikleri

	C5.0	k-EYK	RO	GAM
Doğruluk	0,627	0,591	0,624	0,624
Dengeli Doğruluk	0,624	0,578	0,609	0,611
Kesinlik (Pozitif	0,680	0,629	0,653	0,663
Anma (Duyarlılık)	0,647	0,677	0,718	0,700
F1 skor	0,662	0,643	0,683	0,681
Seçicilik	0,601	0,479	0,501	0,521
Negatif tahmin	0,566	0,532	0,577	0,563

Çizelge 3.40. incelendiğinde, C5.0 algoritmasının en yüksek doğruluk oranına %62,70 sahip olduğu görülmüştür. C5.0 algoritmasının yüksek performans gösterdiği metrikler şu şekildedir; dengeli doğruluk oranı %62,40, kesinlik %68,00 ve seçicilik %60,10 olarak bulunmuştur. RO algoritmasının yüksek performans gösterdiği metrikler şu şekildedir; anma metriği %71,80, F1 skoru %68,30 ve negatif tahmin değeri %57,70'dir.

3.3.8. Test (%25) Veri Seti Performans Sonuçları

Test veri seti olarak 3.562 adet örnek ile eğitim veri setinden öğrenilen model deseni k-EYK, C5.0, RO ve GAM sınıflandırma algoritmaları ve belirlenen en iyi hiper parametreler kullanılarak test edilmiştir. Sonuçta doğruluk, dengeli doğruluk, kesinlik, anma, F1 skoru, seçicilik, negatif tahmin değeri, MKK ve EAKA değeri gibi ulaşılan performans ölçütleri Çizelge 3.41.'de gösterilmiştir.

Çizelge 3.41. İki sınıflı %25 test veri seti performans metrikleri

	C5.0	k-EYK	RO	GAM
Doğruluk	0,611 %95GA (0,595-0,627)	0,599 %95GA (0,582-0,615)	0,615 %95GA (0,599-0,631)	0,621 %95GA (0,605-0,637)
Dengeli Doğruluk	0,590	0,585	0,600	0,606
Kesinlik (Pozitif tahmin değeri)	0,655	0,656	0,665	0,671
Anma (Duyarlılık)	0,712	0,661	0,692	0,695
F1 skor	0,682	0,659	0,678	0,683
Seçicilik	0,467	0,509	0,507	0,517
Negatif Tahmin değeri	0,534	0,515	0,537	0,545
MKK	0,184	0,171	0,201	0,214
G-ortalama	0,577	0,580	0,592	0,599
EAKA	0,632	0,707	0,658	0,675

Çizelge 3.41. incelendiğinde GAM algoritmasının (%62,10, %95GA; %60,50-%63,70) ile en yüksek doğruluk oranına sahip olduğu görülmüştür. Dengeli doğruluk oranı %60,60, kesinlik %67,10, F1 skoru %68,30, seçicilik %51,70, negatif tahmin değeri %54,50, MKK değeri %21,40 ve G-ortalama %59,90 GAM algoritmasında yüksek bulunmuştur. Anma metriği değeri %71,20 C5.0 algoritmasında ve EAKA değeri %70,70, k-EYK algoritmasında yüksek bulunmuştur.

Test veri seti sonuçlarına göre performans ölçütlerinden tahmin doğruluğu oranı en yüksek GAM algoritmasına ait karışıklık matrisi Çizelge 3.42.'de görülmektedir.

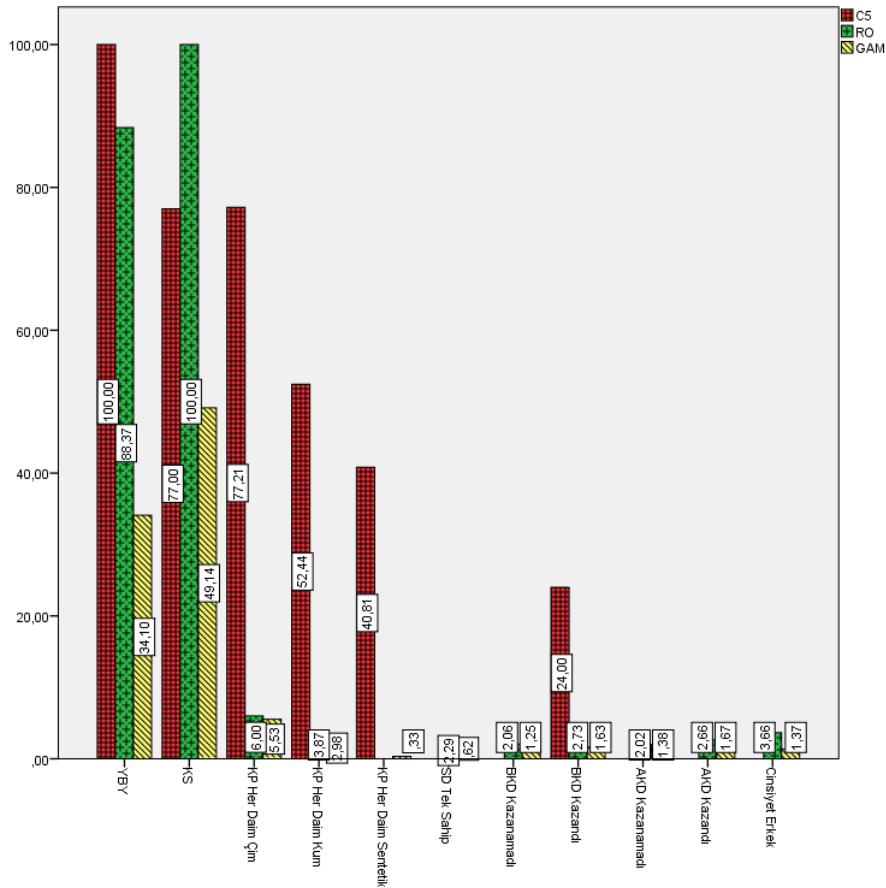
Çizelge 3.42. GAM karışıklık matrisi

Tahmin Değeri	Gerçek Değer		
	Sınıflar	a	b
	a	1451	712
	b	637	762

a:Kazanç yok, b:Kazanç var

Çizelge 3.42.'deki karışıklık matrisinde kazanç yok sınıfına ait 2.088 verinin 1.451'i ve kazanç var sınıfına ait 1.474 verinin 762'si GAM algoritması ile doğru olarak sınıflandırılmıştır.

Aşağıdaki Şekil 3.8.'de, %75 iki sınıflı kazanç değişkenini içeren eğitim veri seti kullanılarak k-EYK hariç diğer algoritmalar ile elde edilen sınıflandırma tahmin modellerine dahil olan öz niteliklerin modele anlamlı katkılarını ve verdiği katkının önem derecesini gösteren grafik görülmektedir.



Şekil 3.8. İki sınıflı %75 eğitim veri seti tahmin modeli öz nitelik önem dereceleri

Şekil 3.8.'de %75 eğitim veri seti için değişken önem dereceleri incelendiğinde analize dahil olan tüm özniteliklerin C5.0, GAM ve RO algoritmalarında farklı önem derecelerinde de olsa tahmin modeline katkı sağladığı görülmüştür. C5.0 ile kurulan sınıflandırma modelinde en etkili öznitelikler sırasıyla; YBY, KS, KPHerDaimÇim, KPHerDaimKum, KPHerDaimSentetik ve BKDKazandıdır. RO ile kurulan sınıflandırma modelinde yer alan en etkili öznitelikler sırasıyla; KS, YBY, KPHerDaimÇim, KPHerDaimKum, CinsiyetErkek, BKDKazandı ve AKDKazandıdır. GAM ile kurulan sınıflandırma modelinde en etkili öznitelikler sırasıyla; KS, YBY, KPHerDaimÇim, KPHerDaimKum, BKDKazandı, CinsiyetErkek ve AKDKazandıdır.

3.3.9. Eğitim (%80) ve Test (%20) Veri Seti Sonuçları

Veri seti ilk olarak, %80 eğitim ve %20 test verisi olarak bölünmüş, 11.342 adet örnek eğitim, 2.891 adet örnek ise test verisi olarak analiz edilmiştir. Performans değerlendirme yöntemi olan ÇD tekniği 10 katlı ve 10 tekrarlı olarak uygulanmış, eğitim veri seti kullanılarak dört ayrı sınıflandırma algoritması ile sınıflandırma tahmin modeli öğrenilmiş, performans ölçütleri elde edilmiş ve en iyi hiper parametre değerleri için ayarlama yapılmıştır. Bu hiper parametre değerleri ile oluşturulan tahmin modeli test veri seti ile sınanmış ve karşılaştırmalı performans ölçütleri elde edilmiştir. Elde edilen bulgular sırasıyla aşağıda sunulmuştur.

3.3.10. Eğitim (%20) Veri Seti Performans Sonuçları

Eğitim veri setindeki 11.342 adet örnek ile ÇD model performans değerlendirme yöntemi 10 katlı ve 10 tekrarlı uygulanarak dört algoritma eğitilmiştir. Hiper parametre en iyileme yöntemi olarak ızgara araştırması kullanılarak her bir algoritma için Çizelge 3.43.'te görülen hiper parametre değerleri elde edilmiştir.

Çizelge 3.43. İki sınıflı %80 eğitim veri seti hiper parametre değerleri

Algoritmalar	Hiper parametreler	Değer
k-EYK	"k" en yakın komşu sayısı	7
RO	"mtry" köklerin dallara ayrılmasında kullanılan rastgele seçilmiş özellik sayısı	2
GAM	"n.trees" kurulacak KA sayısı	100
	"interaction.depth" her bir ağacın en büyük derinlik sayısı	3
	"shrinkage" öğrenme oranı	0,1
	"n.minobsinnode" terminal düğümlerdeki minimum gözlem sayısı	10

Hiper parametre en iyilemesi sonucunda, k-EYK için “k” en yakın komşuluk değeri 7, RO algoritması için “mtry” köklerin dallara ayrılmasında kullanılan rastgele seçilmiş özellik sayısı 2 olarak bulunmuştur. GAM algoritması için “n.trees” kurulacak karar ağaç sayısı 100, “interaction.depth” her bir ağacın en büyük derinlik sayısı 3, “shrinkage” öğrenme oranı 0,1 ve “n.minobsinnode” terminal düğümlerdeki minimum gözlem sayısı 10 olarak bulunmuştur (Çizelge 3.43.).

Yukarıda önerilen hiper parametre değerleri kullanılarak sınıflandırma algoritmalarının eğitim veri setinden öğrendikleri tahmin modellerine ait performans ölçüt sonuçları Çizelge 3.44.’te gösterilmiştir.

Çizelge 3.44. İki sınıflı %80 eğitim veri seti performans metrikleri

	C5.0	k-EYK	RO	GAM
Doğruluk	0,621	0,588	0,626	0,628
Dengeli Doğruluk	0,616	0,574	0,612	0,615
Kesinlik (Pozitif tahmin değeri)	0,677	0,630	0,656	0,668
Anma (Duyarlılık)	0,649	0,677	0,717	0,677
F1 skor	0,661	0,653	0,684	0,680
Seçicilik	0,583	0,471	0,507	0,540
Negatif tahmin değeri	0,556	0,523	0,579	0,562

Çizelge 3.44. incelendiğinde, GAM algoritmasının en yüksek doğruluk oranına %62,80 sahip olduğu görülmüştür. C5.0 algoritmasının yüksek performans gösterdiği metrikler şu şekildedir; dengeli doğruluk oranı %61,60, kesinlik %67,70 ve seçicilik %58,30 olarak bulunmuştur. RO algoritmasının yüksek performans gösterdiği metrikler şu şekildedir; anma metriği %71,70, F1 skoru %68,40 ve negatif tahmin değeri %57,90’dır.

3.3.11. Test (%20) Veri Seti Performans Sonuçları

Test veri seti olarak 2.891 adet örnek ile eğitim veri setinden öğrenilen model deseni k-EYK, C5.0, RO ve GAM sınıflandırma algoritmaları ve belirlenen en iyi hiper parametreler kullanılarak test edilmiştir. Sonuçta doğruluk, dengeli doğruluk, kesinlik, anma, F1 skoru, seçicilik, negatif tahmin değeri, MKK ve EAKA değeri gibi ulaşılan performans ölçütleri Çizelge 3.45.’te gösterilmiştir.

Çizelge 3.45. İki sınıflı %20 test veri seti performans metrikleri

	C5.0	k-EYK	RO	GAM
Doğruluk	0,630 %95GA (0,612-0,647)	0,595 %95GA (0,577-0,613)	0,652 %95GA (0,634-0,669)	0,615 %95GA (0,599-0,631)
Dengeli Doğruluk	0,612	0,582	0,638	0,600
Kesinlik	0,658	0,639	0,683	0,665
(Pozitif tahmin değeri)				
Anma (Duyarlılık)	0,733	0,671	0,731	0,692
F1 skor	0,694	0,655	0,706	0,678
Seçicilik	0,492	0,494	0,546	0,507
Negatif tahmin değeri	0,579	0,528	0,603	0,537
MKK	0,231	0,166	0,281	0,201
G-ortalama	0,600	0,576	0,632	0,592
EAKA	0,664	0,718	0,709	0,690

Çizelge 3.45. incelendiğinde RO algoritmasının en yüksek doğruluk oranına %65,20 (%95GA; %63,40-%66,90) sahip olduğu görülmüştür. RO algoritmasının yüksek performans gösterdiği metrikler şu şekildedir; dengeli doğruluk oranı %63,80, kesinlik %68,30, F1 skoru %70,60, seçicilik %54,60, negatif tahmin değeri %60,30 ve MKK değeri %28,10 olarak bulunmuştur. Anma metriği %73,30 ile C5.0 algoritmasında, G-ortalama %63,20 ile RO algoritmasında ve EAKA %71,80 ile k-EYK algoritmasında yüksek performans göstermiştir.

Test veri seti sonuçlarına göre tahmin doğruluk oranı en yüksek RO algoritmasına ait karışıklık matrisi Çizelge 3.46.'de görülmektedir.

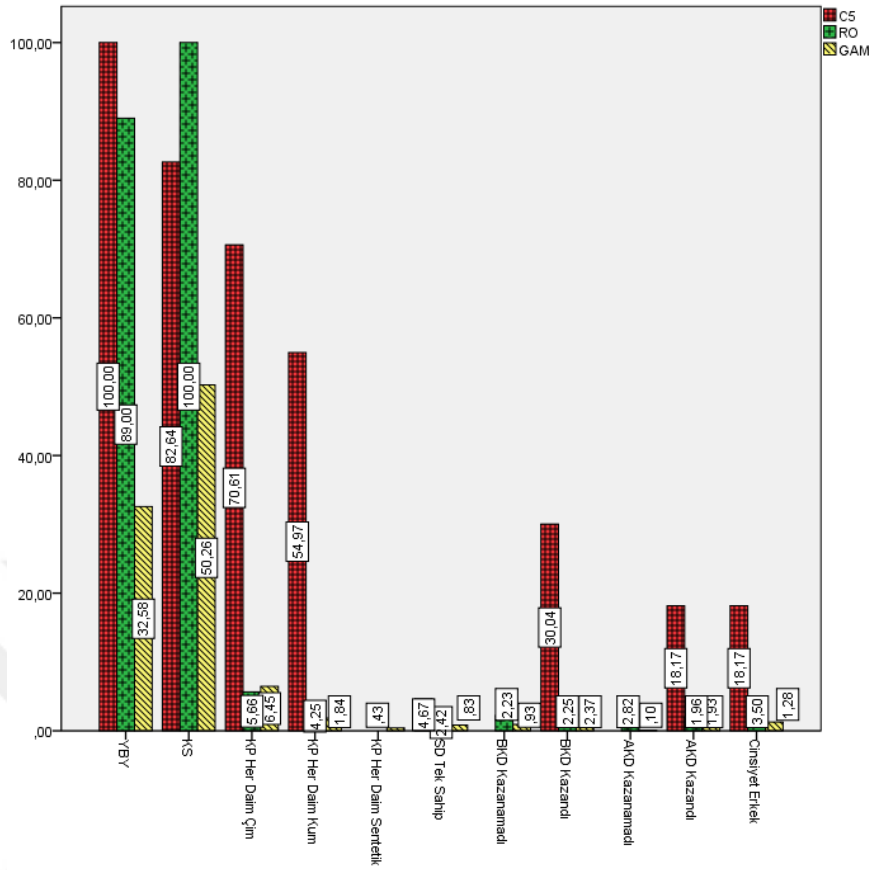
Çizelge 3.46. RO karışıklık matrisi

Gerçek Değer			
Tahmin Değeri	Sınıflar	a	b
	a	1209	562
	b	445	675

a:Kazanç yok, b:Kazanç var

Çizelge 3.46.'daki karışıklık matrisinde kazanç yok sınıfına ait 1.654 verinin 1.209'i ve kazanç var sınıfına ait 1.237 verinin 675'i RO algoritması ile doğru olarak sınıflandırılmıştır.

Aşağıdaki Şekil 3.9.'da, iki sınıflı kazanç değişkenini içeren %80 oranındaki eğitim veri seti kullanılarak k-EYK hariç diğer algoritmalar ile elde edilen sınıflandırma tahmin modellerine dahil olan öz niteliklerin modele anlamlı katkılarını ve verdiği katkının önem derecesini gösteren grafik görülmektedir.



Şekil 3.9. İki sınıflı %80 eğitim veri seti tahmin modeli öznitelik önem dereceleri

Şekil 3.9.'da %80 eğitim veri seti için değişken önem dereceleri incelendiğinde analize dahil olan tüm özniteliklerin C5.0, GAM ve RO algoritmalarında farklı önem derecelerinde de olsa tahmin modeline katkı sağladığı görülmüştür. C5.0 ile kurulan sınıflandırma modelinde en etkili öznitelikler sırasıyla; YBY, KS, KPHerDaimÇim, KPHerDaimKum, BKDKazandı, AKDKazandı ve CinsiyetErkek'tir. RO ile kurulan sınıflandırma modelinde yer alan en etkili öznitelikler sırasıyla; KS, YBY, KPHerDaimÇim, KPHerDaimKum ve CinsiyetErkek'tir. GAM ile kurulan sınıflandırma modelinde en etkili öznitelikler sırasıyla; KS, YBY, KPHerDaimÇim, KPHerDaimKum, BKDKazandı ve AKDKazandı'dır.

3.4. Tüm Test Veri Seti Performans Ölçüm Sonuçları

3.4.1. Test (%30) Veri Seti Performans Ölçüm Sonuçları

Aşağıdaki çizelge 3.47.'de çalışmada kullanılan sınıflandırma algoritmalarına ait %30 test veri setleri için elde edilen tüm performans metrikleri gösterilmiştir.

Çizelge 3.47. %30 test veri setleri tüm performans metrikleri

	Üç sınıflı veri seti sonuçları				Üç sınıflı RAÖ veri seti sonuçları				İki sınıflı veri seti sonuçları			
	C5.0	k-EYK	RO	GAM	C5.0	k-EYK	RO	GAM	C5.0	k-EYK	RO	GAM
Doğruluk	0,570	0,521	0,567	0,575	0,426	0,383	0,440	0,455	0,632	0,574	0,621	0,640
	%95GA	%95GA	%95GA	%95GA	%95GA	%95GA	%95GA	%95GA	%95GA	%95GA	%95GA	%95GA
	(0,555- 0,585)	(0,506- 0,536)	(0,552- 0,582)	(0,559- 0,589)	(0,411- 0,441)	(0,368- 0,398)	(0,425- 0,455)	(0,440- 0,470)	(0,617- 0,646)	(0,559- 0,589)	(0,606- 0,636)	(0,625- 0,654)
Dengeli Doğruluk Ortalaması	0,516	0,526	0,503	0,519	0,555	0,547	0,537	0,594	0,617	0,559	0,595	0,627
Makro Ortalama Kesinlik	0,326	0,370	0,424	0,515	0,398	0,395	0,376	0,429	0,657	0,612	0,631	0,667
Makro Ortalama Anma (Duyarlılık)	0,353	0,363	0,337	0,356	0,400	0,390	0,378	0,433	0,730	0,674	0,795	0,725
Makro Ortalama F1	0,286	0,348	0,250	0,291	0,387	0,364	0,375	0,412	0,692	0,641	0,703	0,695
Ortalama Seçicilik	0,679	0,689	0,669	0,681	0,710	0,705	0,698	0,733	0,505	0,445	0,395	0,529
Ortalama Negatif tahmin değeri	0,765	0,705	0,816	0,784	0,703	0,696	0,695	0,721	0,590	0,512	0,597	0,597
MKK/ CSMKK	0,104	0,071	0,055	0,109	0,111	0,095	0,168	0,168	0,240	0,121	0,208	0,259
G-ortalama	0	0,239	0	0,063	0,395	0,389	0,359	0,417	0,607	0,548	0,560	0,619
EAKA	0,602	0,571	0,575	0,638	0,562	0,558	0,557	0,621	0,669	0,614	0,685	0,696

3.4.2. Test (%25) Veri Seti Performans Ölçüm Sonuçları

Aşağıdaki çizelge 3.48.'de çalışmada kullanılan sınıflandırma algoritmalarına ait %25 test veri setleri için elde edilen tüm performans metrikleri gösterilmiştir.

Çizelge 3.48. %25 test veri setleri tüm performans metrikleri

	Üç sınıflı veri seti sonuçları				Üç sınıflı RAÖ veri seti sonuçları				İki sınıflı veri seti sonuçları			
	C5.0	k-EYK	RO	GAM	C5.0	k-EYK	RO	GAM	C5.0	k-EYK	RO	GAM
Doğruluk	0,591	0,539	0,589	0,597	0,430	0,402	0,453	0,454	0,611	0,599	0,615	0,621
	%95GA	%95GA	%95GA	%95GA	%95GA	%95GA	%95GA	%95GA	%95GA	%95GA	%95GA	%95GA
	(0,575- 0,607)	(0,522- 0,555)	(0,572- 0,605)	(0,580- 0,613)	(0,413- 0,446)	(0,386- 0,418)	(0,437- 0,470)	(0,438- 0,471)	(0,595- 0,627)	(0,582- 0,615)	(0,599- 0,631)	(0,605- 0,637)
Dengeli Doğruluk Ortalaması	0,521	0,534	0,504	0,521	0,557	0,557	0,553	0,586	0,590	0,585	0,600	0,606
Makro Ortalama Kesinlik	0,377	0,380	0,446	0,501	0,400	0,403	0,395	0,432	0,655	0,656	0,665	0,671
Makro Ortalama Anma (Duyarlılık)	0,359	0,371	0,338	0,360	0,404	0,404	0,399	0,439	0,712	0,661	0,692	0,695
Makro Ortalama F1	0,305	0,362	0,257	0,305	0,391	0,380	0,394	0,414	0,682	0,659	0,678	0,683
Ortalama Seçicilik	0,682	0,696	0,669	0,682	0,710	0,709	0,706	0,733	0,467	0,509	0,507	0,517
Ortalama Negatif tahmin değeri	0,754	0,713	0,800	0,774	0,703	0,701	0,702	0,723	0,534	0,515	0,537	0,545
MKK/ÇSMKK	0,104	0,091	0,059	0,120	0,112	0,111	0,105	0,173	0,184	0,171	0,201	0,214
G-ortalama	0,053	0,257	0	0,080	0,401	0,404	0,385	0,422	0,577	0,580	0,592	0,599
EAKA	0,587	0,692	0,575	0,629	0,574	0,573	0,613	0,624	0,632	0,707	0,658	0,675

3.4.3. Test (%20) Veri Seti Performans Ölçüm Sonuçları

Aşağıdaki çizelge 3.49.'da çalışmada kullanılan sınıflandırma algoritmalarına ait %20 test veri setleri için elde edilen tüm performans metrikleri gösterilmiştir.

Çizelge 3.49. %20 test veri setleri tüm performans metrikleri

	Üç sınıflı veri seti sonuçları				Üç sınıflı RAÖ veri seti sonuçları				İki sınıflı veri seti sonuçları			
	C5.0	k-EYK	RO	GAM	C5.0	k-EYK	RO	GAM	C5.0	k-EYK	RO	GAM
Doğruluk	0,580	0,537	0,576	0,583	0,419	0,388	0,462	0,461	0,630	0,595	0,652	0,615
	%95GA	%95GA	%95GA	%95GA	%95GA	%95GA	%95GA	%95GA	%95GA	%95GA	%95GA	%95GA
	(0,562- 0,599)	(0,518- 0,555)	(0,558- 0,594)	(0,565- 0,601)	(0,401- 0,438)	(0,370- 0,407)	(0,443- 0,480)	(0,442- 0,479)	(0,612- 0,647)	(0,577- 0,613)	(0,634- 0,669)	(0,599- 0,631)
Dengeli Doğruluk Ortalaması	0,522	0,535	0,506	0,524	0,552	0,548	0,553	0,589	0,612	0,582	0,638	0,600
Makro Ortalama Kesinlik	0,426	0,390	0,466	0,487	0,394	0,394	0,396	0,438	0,658	0,639	0,683	0,665
Makro Ortalama Anma (Duyarlılık)	0,362	0,387	0,341	0,364	0,396	0,390	0,399	0,440	0,733	0,671	0,731	0,692
Makro Ortalama F1	0,304	0,375	0,258	0,309	0,383	0,367	0,396	0,421	0,694	0,655	0,706	0,678
Ortalama Seçicilik	0,682	0,695	0,671	0,683	0,707	0,705	0,708	0,738	0,492	0,494	0,546	0,507
Ortalama Negatif tahmin değeri	0,761	0,715	0,839	0,768	0,700	0,696	0,705	0,725	0,579	0,528	0,603	0,537
MKK/CSMKK	0,113	0,094	0,080	0,122	0,103	0,096	0,110	0,180	0,231	0,166	0,281	0,201
G-ortalama	0,055	0,256	0	0,088	0,394	0,390	0,381	0,432	0,600	0,576	0,632	0,592
EAKA	0,601	0,704	0,678	0,636	0,569	0,554	0,606	0,622	0,664	0,718	0,709	0,690

4. TARTIŞMA

Bu araştırma kapsamında, Türkiye’deki safkan İngiliz yarış atlarının kazanç durumlarına etki eden bazı çevresel faktörler ilk kez MÖ sınıflandırma algoritmaları ile değerlendirilmiştir.

Günümüzde, artan veri sayısı ve büyük verilerin karmaşıklığı, MÖ ihtiyacını artırmıştır. MÖ, geleneksel veri işleme araçları ile analizinin yapılmasında zorlanan, yüksek boyut ve büyük miktardaki yapılandırılmamış veri setlerini kullanarak anlamlı ilişkiler bulan, model ve desenler geliştiren yöntemler ailesidir (Çağlayan Akay, 2018).

İstatistiksel modeller için ilk olarak model hipotezleri kurulur ve sonra varsayımları belirlenip çıkarımlara odaklanırken geleneksel yöntemlerin aksine, MÖ’nün hipotezsiz ve veri dağılımından bağımsız doğası onu büyük veri analizi için daha önemli hale getirmiştir (Gouda vd., 2022).

Çevresel özellikler düşük kalıtsallığa sahip olsa da değeri yüksek olan yarış performansının iyileştirilmesine önemli katkılarda bulunabileceği düşünülmektedir. Son yıllarda büyük veri analizlerinde çokça tercih edilen MÖ sınıflandırma algoritmaları ile veteriner hekimliği alanında ve özellikle atlarda ilk kez bu teknikler kullanılıp tahmin modelleri oluşturulmuş ve literatüre katkıda bulunulmuştur.

Bu bağlamda yapılan çalışmada elde edilen bulgular ile hem yöntemsel ve hem de hayvan yetiştiriciliğine ait iki adet tartışma konusu belirlenmiştir.

Yöntemsel karşılaştırmada ilk olarak kazanç için bir tahmin modeli geliştirilip geliştirilmediği sorgulanmış, literatür tarandığında toplam kazanç için tahmin modellerinin geliştirildiği ve bunun için doğrusal ve LR gibi geleneksel parametrik yöntemlerden yararlanıldığı görülmüştür. Bu çalışmalarda kazanç değişkeni, farklı farklı ölçek ve sınıflamalar ile değerlendirilmiştir. Cheetham vd. (2010) tarafından yapılan çalışmada doğrusal regresyon ile 2, 3, 4 ve 5 yaşlarındaki safkan İngiliz atlarının toplam kazançlarını tahmin eden model oluşturulmuştur. Langlois ve Hernu (2003) tarafından yapılan çalışmada ise LR ile 2, 3, 4 ve 5 yaşlarındaki safkan İngiliz atlarının toplam kazançları yıllık indekse çevrilip 4 kategoriye ayrılarak tahmin modeli geliştirilmiştir. Bu tez çalışması ile kazanç durumu, yöntemsel olarak literatürden farklı olarak makine öğrenmesi algoritmaları ile

değerlendirilmiş; kazanç durumu ise ölçeksel olarak ikili ve çok sınıflı bir yapıda değerlendirilmiştir.

Mevcut çalışmada %70, %75 ve %80 oranlarındaki dengelenmemiş, RAÖ yöntemi ile dengelenmiş ve iki sınıflı kazanç değişkenini içeren eğitim veri setlerinden k-EYK, C5.0, RO ve GAM sınıflandırma algoritmalarının öğrendiği tahmin modellerinin üç farklı test veri seti ile sınanması sonucu elde edilen tahmin performanslarının literatürdeki iki ve çok sınıflı sınıflandırıcı performansları ile benzerlik ve farklılıklarının değerlendirilmesi, diğer bir yönetsel tartışma konusu olarak belirlenmiştir. Algoritma eğitimi ve model testi sırasında kullanılan hiper parametre en iyileme tekniği, çok sınıflıları ayırıştırma, yani ikilileştirme yöntemi, performans değerlendirme yöntemi ile model tahminlerinin başarısının ölçüldüğü performans karşılaştırma ölçütlerinin literatür ile uyumlu olup olmadığı araştırılmıştır.

Çalışma bulgularına göre %30, %25 ve %20 test veri setleri ile k-EYK, C5.0, RO ve GAM sınıflandırma algoritmaları kullanılarak iki sınıflı kazanç değişkeni için elde edilen performans metrik değerleri, seçicilik ve negatif tahmin değerleri haricinde, üç sınıflı kazanç değişkenini içeren dengeli ve dengesiz veriler için elde edilen performans metriklerinden daha yüksek bulunmuştur. Araştırmada ulaşılan bu sonuç beklenen ve doğal bir durum olarak belirtilmiştir (Özen vd., 2024). Bunun en önemli nedenlerinden birisi olarak, ikili sınıflandırma problemleriyle karşılaştırıldığında çok sınıflı sınıflandırma problemlerinin karışıklık düzeyinin daha fazla olması olarak bildirilmiştir. Sınıf sayısı arttıkça farklı sınıfların dağılımları daha karmaşık hale gelmektedir (Ding vd., 2019).

İkili sınıflandırmalarda çok sınıflılara göre daha yüksek doğruluk değerlerinin elde edildiği çalışmalar incelenmiş, Viazzi vd. (2013) yaptıkları çalışmada sınıfların artmasıyla sınıflandırma görevinin daha zorlu hale gelebileceğini çok sınıflı sınıflandırma modellerinin performansının ikili sınıf modelleri kadar yüksek olmadığını göstermiştir. Araújo vd. (2017) tarafından meme dijital patoloji görüntüleri kullanılarak yapılan araştırmada elde edilen bulgular mevcut tez çalışması ile uyumlu bulunmuştur. Meme kanserli hastalar öncelikle karsinom var/yok şeklinde ikili sınıflandırılmış %83,30 doğruluk değeri elde edilmiş sonrasında normal doku, iyi huylu lezyon, yerinde karsinom ve invaziv karsinom olarak dört sınıflı olarak sınıflandırılmış doğruluk değeri %77,80 olarak bulunmuştur. Buna benzer şekilde yine meme kanseri olan hastalar üzerine Mi vd. (2021) tarafından yapılan araştırmada meme dijital patoloji görüntülerine ait veriler kullanılarak malign olmayan ile malign şeklinde ikili sınıflandırılmış %96,30 doğruluk değeri elde edilmiş sonrasında normal doku, iyi huylu

lezyon, yerinde karsinom ve invaziv karsinom olarak dört sınıflı olarak sınıflandırılmış doğruluk değeri %85,19 olarak bulunmuştur. Tummala vd. (2022) tarafından BreaKHis veri setindeki histopatolojik görüntüler ile meme kanserini bening ve yedi alt tipte malignite olarak sekiz sınıfa ayırdığı çalışmada elde edilen bulgulara göre, iki sınıflı sınıflandırma için %99,60 test doğruluğu, sekiz sınıf için %96,00'lık doğruluk elde edilmiştir. BreaKHis veri seti kullanılarak yapılan diğer çalışmalarda da benzer şekilde Alom vd. (2019), Bardou vd. (2018), Han vd. (2017) ve Jiang vd. (2019) tarafından iki sınıflı sınıflandırmalarda elde edilen doğruluk oranının sekiz sınıf için elde edilenden daha yüksek olduğu bulunmuştur. Biswas ve Gini (2024) tarafından yürütülen Alzheimer hastalığının hem var ve yok şeklinde ikili hem de normal, orta ve şiddetli olarak üçlü sınıflandırıldığı çalışmada beynin üç farklı bölgesine ait 3D MRI ve PET imajları kullanılmış bunlardan elde edilen iki farklı veri seti ile k-EYK, KA, RO ve GAM algoritmaları kullanılarak sınıflandırma yapılmıştır. Çalışmadan elde edilen bulguların ışığında çok sınıflı sınıflandırma çözümünün ikili sınıflandırmalara göre daha fazla zaman aldığı fakat buna karşın elde edilen çözümlerin maliyet etkililik düzeyinin düşük olduğu rapor edilmiştir. Imada vd. (2024) tarafından Kanada süt ineği tekrarlı verileri kullanılarak ağaç tabanlı (KA ve RO) sınıflandırma algoritmaları ile Johne hastalığının teşhisinde yararlanılan Johne test sonuçları tahmin edilmeye çalışılmıştır. Geçmişteki Johne testi sonuçlarının yanı sıra süt bileşeni testi sonuçlarından elde edilen veri setine RO algoritması uygulanmış ikili sınıflandırmada (pozitif ve negatif) Johne ELISA sonucu için performans metriği olarak 0,626'lık kappa ve 0,915'lik EAKA değeri üretmiş dört sınıflı (negatif, şüpheli, pozitif ve yüksek) sınıflandırma için ise daha düşük değerler olan 0,518'lik bir kappa ve 0,763'lük bir EAKA değeri elde edilmiştir. Infante vd. (2022) tarafından yapılan çalışmada trafik kazalarında görülen yaralanma şiddeti dört sınıflı olarak incelenmiş mevcut tez çalışmasında kullanılan RO, k-EYK ve C5.0 algoritmalarının yanında farklı MÖ yöntemleri (LR, Naïve Bayes (NB) ve DVM) kullanılmıştır. Ulaşılan bulgular mevcut tez çalışmasındaki üç sınıflı tahminlerden elde edilen düşük sonuçlar ile benzerlik taşımakta olup doğruluk oranları (0,610-0,656) ve diğer performans metriklerinin (örneğin; MKK 0,289-0,353) düşük olduğu görülmüştür. Deng vd. (2021) biyomedikal veri kümelerini kullanarak ölüm nedenlerini beş sınıflı olarak sınıflandırmış KA, RO, YSA, DVM ve çok terimli LR kullanarak yaptıkları araştırmada tahmin doğruluk oranlarını sırasıyla %69,21, %70,23, %70,16, %69,06 ve %68,12 olarak bulmuş olup bulunan bu düşük değerler mevcut çalışmadaki üç sınıflı tahminler için elde edilen sonuçlar ile benzerlik göstermektedir.

Mevcut çalışmada, kazanç değişkeni sınıfları arasında sınıf dengesizliği problemi görülmüş bunu gidermek için Batista vd. (2004) tarafından yapılan araştırmada, daha karmaşık

(SAÖAT ve türevleri) yöntemlere karşı rekabetçi sonuçlar sağlayan, daha az maliyetli anlamlı sonuçlar üreten, basit ve etkili bir yeniden örnekleme yöntemi olan ve büyük veri setleri için önerilen RAÖ yöntemi kullanılmıştır. Bu yöntem ile %70, %75 ve %80 oranlarındaki orijinal eğitim veri setleri yeniden oluşturulmuştur. Bu dengelenmiş eğitim veri setlerinden elde edilen tahmin modellerinin sınındığı test veri setlerine ait doğruluk oranları, seçicilik, negatif tahmin ortalamaları ve EAKA değerleri hariç bütün performans metrik değerleri dengelenmemiş üç sınıflı test veri setinden elde edilen değerlerden yüksek bulunmuştur. Sağlık alanında yöntemsel açıdan benzer nitelikteki çalışmalar incelendiğinde, Turhan vd. (2020) diyabet hastalığı tanısı almış hastalar üzerinde yaptıkları çalışmada, oluşturdukları dengesiz veri setini kullanarak diyabet tanısı üzerinden sınıflandırma performanslarını karşılaştırmayı amaçlamışlardır. Sınıf dengesizliği sorununu çözmek için RAÖ, SAÖAT ve RÖA yöntemleri kullanılmış, RAÖ ile yapılan sınıflandırmaların performansı için kullanılan doğruluk, Kappa istatistiği, duyarlılık ve seçicilik ölçütleri diğerlerinden çok daha başarılı tahmin gücüne sahip bulunmuştur.

Sınıflar arasında bir dengesizlik olduğunda, bir sınıflandırma algoritmasının performansını ve kalitesini değerlendirmek için yaygın olarak kullanılan ölçütler olan F1 skoru, G-ortalama, EAKA ve ÇSMKK değerleri yönünden mevcut tez sonuçları değerlendirildiğinde RAÖ ile dengelenmiş modellerden elde edilen performans ölçütleri EAKA haricinde yükseliş göstermiştir. Dengelenmiş veride F1 skoru, G-ortalama, EAKA ve ÇSMKK değerleri yönünden en iyi performansı GAM algoritması göstermiştir. Dengelenmemiş veride ise F1 skoru, G-ortalama ve EAKA değerleri yönünden en iyi performansı k-EYK algoritması ÇSMKK değerleri yönünden GAM algoritması göstermiştir.

Demirsöz (2024) tarafından dört farklı veri seti ve üç farklı MÖ sınıflandırma algoritması ile yürütülen tez çalışmasında F1 skoru, EAKA ve MKK performans metrikleri üzerinden değerlendirme yapılmış EAKA metriği haricinde F1 skoru ve MKK ölçütleri yönünden mevcut çalışmadan elde edilen bulgular ile örtüşen sonuçlara ulaşılmıştır. Dengesizlik oranı (10,74) yüksek öznitelik sayısının fazla olduğu “climate” veri seti için RO sınıflandırma algoritması kullanılarak dengelenmemiş veri seti ile F1 skoru 0,904 ve EAKA 0,884 olarak RAÖ yeniden örnekleme ile ise F1 skoru 0,969 ve EAKA 0,949 olarak bulunmuştur. Dengesizlik oranı (32,73) yüksek öz nitelik sayısının az olduğu “yeast” veri seti için RO sınıflandırma algoritması kullanılarak dengelenmemiş veri seti ile F1 skoru 0,969, EAKA 0,795 ve MKK 0,435 olarak RAÖ veri seti ile ise F1 skoru 0,978, EAKA 0,983 ve MKK 0,662 olarak bulunmuştur. Dengesizlik oranı (2,88) düşük öznitelik sayısının fazla

olduğu “vehicle” veri seti için RO sınıflandırma algoritması kullanılarak dengelenmemiş veri seti ile F1 skoru 0,976, EAKA 0,995 ve MKK 0,931 olarak RAÖ veri seti ile ise F1 skoru 0,988, EAKA 0,999 ve MKK 0,966 olarak bulunmuştur.

Elde edilen bütün performans metrikleri değerlendirildiğinde, RAÖ yöntemi ile dengelenmiş üç sınıflı veri setinde ve iki sınıflı veri setinde, %30 ve %25 oranlarında GAM sınıflandırma algoritmasına ait performans metrikleri en yüksek bulunurken %20 oranında ise RO sınıflandırma algoritmasından elde edilen performans metrikleri yüksek bulunmuştur. Dengelenmemiş üç sınıflı üç farklı orandaki test veri setinde ise doğruluk oranı ve kesinlik makro ortalaması yönünden en iyi performansı GAM algoritması göstermiştir. Mevcut çalışmada, GAM ve RO sınıflandırma algoritmalarının başarılı sonuçlar verdiği görülmüştür. Biswas ve Gini (2024) tarafından Alzheimer hastaları üzerine yapılan çalışmada da benzer algoritmaların başarılı olduğu görülmüştür. Alzheimer hastalığının ön tanısı için yapılan çalışmada beyin üç farklı bölgesine ait 3D MRI ve PET imajları kullanılmış bunlardan elde edilen iki farklı veri seti ile k-EYK, KA, RO ve GAM algoritmaları kullanılarak normal, orta ve şiddetli olarak sınıflandırılan hastalığın tahmini yapılmıştır. Yapılan çok sınıflı sınıflandırma çözümünde ilk veri setinde RO ve ikinci veri setinde ise GAM algoritmaları için yüksek doğruluk oranları bulunmuştur.

Hyde vd. (2020) tarafından süt sığırlarında otomatik mastitis teşhis aracı geliştirmek için yapılan iki sınıflı tahmin çalışmasında en iyi tahmin modelini belirlemek için RO, GAM ve radyal çekirdekli DVM algoritmaları kullanılmış RO tahmin modeli doğruluk, pozitif ve negatif tahmin değerleri yönünden en iyi model performansı göstermiştir. Imada vd. (2024) tarafından Johne hastalığının teşhisi için kullanılan Johne test sonuçlarını tahmin etmede algoritmaların yeteneğini araştırmak için Kanada süt ineğinden alınan tekrarlı süt testi verilerinin analizi için ağaç tabanlı (KA ve RO) algoritmalar uygulanmıştır. Dört sınıflı (negatif, şüpheli, pozitif ve yüksek) sınıflandırma için performans ölçütleri olarak kappa, EAKA, duyarlılık, seçicilik, pozitif tahmin ve negatif tahmin değerleri kullanılmış RO algoritmasının en yüksek değerlere sahip olduğu görülmüştür. Abnoosian vd. (2023) çalışmasında katılımcıları diyabetik olmayan, prediyabet ve diyabetik olmak üzere üç sınıfa ayırarak çok sınıflı sınıflandırma tahmin modeli geliştirmek için k-EYK, DVM, KA, RO, AdaBoost ve GAM algoritmaları kullanmış bu algoritmaların tahmin doğruluğunu değerlendirmek için ortalama doğruluk, kesinlik, anma, F1 skoru ve EAKA metrikleri kullanılmış en iyi performansı GAM ve RO göstermiştir. Dineva ve Atanasova (2023) tarafından süt inekleri üzerine yürütülen çalışmada süt ineklerinin sağlık durumunu sağlıklı,

şüpheli ve sağlıklı olmak üzere üç sınıfa ayrılmış bir tahmin modeli geliştirmek için kullandığı k-EYK, NB, KA, AGAM ve RO algoritmalarının tahmin doğruluğunu karşılaştırmak için ortalama doğruluk, kesinlik, anma ve F1 skoru kullanılmıştır. En yüksek doğruluk RO (%96) ve k-EYK (%82) ile elde edilmiş algoritma seçiminde en iyi performansı gösteren algoritmayı tespit etmek için belirleyici faktör olan anma metriği k-EYK ve RO (%73 ve %96) için en yüksek olarak elde edilmiş ve bu sonuçlara göre RO sınıflandırıcısı en iyi performansı göstermiştir. Deng vd. (2021) tarafından canlı, meme dışı kanser, meme kanseri, kardiyovasküler hastalık ve diğer nedenler olmak üzere beş kategorili sınıflandırma problemi çözümü için biyomedikal veri seti ile KA, RO, YSA ve DVM algoritmaları kullanılarak bir çalışma yürütülmüştür. Elde edilen tahmin modellerinin performanslarını değerlendirmek için doğruluk, seçicilik, kesinlik, anma, F1 skoru ve EAKA değerleri kullanılmış karışıklık matrislerine dayanarak dört MÖ modelinin de benzer performansa sahip olduğu görülmüştür. Ancak performans metriklerinin ayrıntılı karşılaştırılması sonucunda, mevcut çalışmaya benzer şekilde RO modeli sınıflandırılma tahmininde anma ve F1 skoru bazında daha iyi performans göstermiştir. Arif vd. (2023) kronik karaciğer yetmezliğini erken tespit etmek için iki sınıflı tahmin modeli geliştirmeyi amaçlamış ve k-EYK ile NB algoritmalarından elde edilen tahminlerin doğruluğunu karşılaştırmak için doğruluk, kesinlik, anma ve F1 skorları kullanılmış sonuçta k-EYK en iyi performansı göstermiştir. Vidyarthi vd. (2022) tarafından beyin tümörü hastaları ile ilgili yapılan çalışmada beş tip maligniteden oluşan MR görüntüleme veri seti kullanılmış altı sınıflı sınıflandırma tahmin problemi çözümünde k-EYK, çok sınıflı DVM ve sinir ağları (SA) (Neural Network, NN) kullanılmıştır. Tahmin performansları doğruluk, kesinlik, anma ve F1 skoru ile değerlendirilmiş ve doğruluk oranlarına göre mevcut tez çalışmasından farklı olarak en yüksek doğruluk SA için bulunmuştur. Ghosh ve Kole (2021) tarafından yapılan araştırmada MR görüntüleri kullanılarak tümörsüz bölgeler ile glioma, meningioma ve hipofiz tümörü içeren bölgeleri tespit etmek için DVM, k-EYK, RO ve AGAM olmak üzere dört MÖ algoritması kullanılmıştır. MÖ algoritmalarının performansları doğruluk, anma, kesinlik ve F1 skoruna göre karşılaştırılmış mevcut tez çalışmasından farklı olarak AGAM tüm performans değerlerinde diğer algoritmaları geride bırakmıştır. Abdullah vd. (2023) tarafından yürütülen çalışmada meme kanseri normal, erken tümör ve geç tümör olmak üzere üç sınıflı olarak sınıflandırılmış yüksek boyutlu mikrodizi verileriyle DVM, çok değişkenli LR, NB ve RO algoritmaları kullanılarak tahmin modeli elde edilmiştir. Algoritma performansları mevcut çalışmaya benzer şekilde makro ve mikro doğruluk, anma (duyarlılık) ve kesinlik değerleri ile değerlendirilmiştir. Sonuçta, mevcut çalışmadan farklı olarak en iyi performansı DVM göstermiştir (makro doğruluk=%86,7, makro anma=%77,8 ve makro kesinlik=%62,0).

Dengelenmemiş üç sınıflı üç farklı orandaki test veri setinden elde edilen sonuçlara göre k-EYK algoritması dengeli doğruluk ve seçicilik ortalaması, G-ortalama, EAKA, ÇSMKK, anma metriği ve F1 skoru makro ortalaması yönlerinden GAM ve RO gibi göreceli daha kompleks algoritmaların performansından daha iyi performans sergilemiştir. Bunun nedenlerinden birisinin kazanç grubuna ait dağılımın dengesizliği olduğu düşünülmektedir. Kullanılan farklı algoritmaların farklı performans ölçütlerinde daha başarılı oldukları mevcut araştırmada görülmüş olup buna benzer bir duruma Infante vd. (2022) tarafından trafik kazalarında görülen yaralanma şiddetini yaralanma yok, hafif, ciddi ve ölümcül yaralanma olarak dört sınıflı olarak inceledikleri çalışmada rastlanmıştır. RO, k-EYK ve C5.0 algoritmalarının yanında farklı MÖ yöntemleri (LR, NB ve DVM) performans metriği olarak doğruluk, duyarlılık, seçicilik, pozitif ve negatif tahmin değeri, F1 skoru, G-ortalama ve MKK kullanılmıştır. En yüksek doğruluk oranına NB sahip iken G-ortalaması, F1 skoru ve diğer performans ölçütleri değerlendirildiğinde C5.0 ve RO algoritmasının LR ile benzerlik gösterdiği ancak NB, DVM ve k-EYK'den daha yüksek performansa sahip olduğu görülmüştür. Abbad Ur Rehman vd. (2021) sağlam, hiper ve hipo tiroid hastaları olmak üzere üç sınıflı olarak sınıflandırdığı çalışmada k-EYK, LR, NB, KA ve DVM algoritmaları kullanmış performans ölçütü olarak bu çalışmaya benzer şekilde doğruluk, anma (duyarlılık), seçicilik, EAKA, F1 skoru ve MKK kullanmıştır. İki adet öznelik seçme yöntemi (L1 ve L2) kullanıp üç farklı veri seti elde edilmiş NB'in en yüksek doğruluk, EAKA, MKK, F1 skor, kesinlik ve anma değerlerine sahip olarak mevcut çalışmadan ayrıldığı görülmüştür. Ancak en düşük doğruluk değerine sahip KA'nın MKK değeri yönünden k-EYK, DVM ve LR'den daha yüksek olduğu görülmüştür. Ishfaq vd. (2022) 11 farklı veri seti dördü iki sınıflı diğerleri çok sınıflı sınıflandırma problemi çözümünde NB, KA, RO, GAM ve derin öğrenme tabanlı evrişimli sinir ağlarını (ESA) (Convolutional Neural Network, CNN) kullanmıştır. Elde edilen bulgular tez çalışması ile benzerlik göstermiş, doğruluk, kesinlik ve anma açısından GAM'ın performansının diğer algoritmalarından daha iyi olduğu görülmüştür. RO ve ESA için çoğu veri kümesinde benzer sonuçlara ulaşılmış NB ve KA'dan daha iyi performans gösterdikleri görülmüş ve NB'nin en düşük performanslı algoritma olduğu rapor edilmiştir.

Araştırmada iki sınıflı sınıflandırıcıların tahmin doğruluğunu değerlendirmek için kesinlik, anma ve F1 skoru üç sınıflılar için bunların makro ortalaması, iki sınıflı sınıflandırıcılar için doğruluk, dengeli doğruluk, seçicilik ve negatif tahmin değeri ile üç sınıflılar için bunların ortalamaları, iki sınıflı sınıflandırıcılar için EAKA ve MKK ile üç sınıflılar için ÇSMKK, EAKA ve G-ortalama performans ölçütleri ile karışıklık matrisi beraber kullanılmıştır. Tummala vd. (2022) yaptıkları çalışmada Swin görüntü transformatörü

(SwinT) tarafından histopatolojik görüntüleri ile göğüs kanserini dört iyi ve dört kötü huylu olarak sekiz sınıflı sınıflandırma tahmini probleminin çözümünde performans metriği olarak mevcut tez çalışmasına benzer şekilde doğruluk, dengeli doğruluk, EAKA, F1 skoru, MKK ve karışıklık matrisi kullanmıştır. Shekar ve Dagnev (2019) yaptıkları çalışmada üç sınıfa ayırarak sınıflandırdıkları Lösemi (kan kanseri) hastalarını mikrodizi verileri ile tahmin etmek için farklı MÖ algoritmaları kullanmış ve bu tez çalışmasında kullanılan performans ölçütleri gibi doğruluk, kesinlik, anma, F1 skor ve karışıklık matrisi kullanmış ve bunların yanına yanlış sınıflandırma oranını dahil etmişlerdir. Özen vd. (2024) dört sınıflı dengesiz verilerde yaptığı çalışmada RO, C5.0 ve AGAM algoritmalarını kullanmış, tahmin doğruluğunu kesinlik, anma, F1 skoru, MKK ve G-ortalama metrikleri ile değerlendirmiştir. Mevcut tez çalışmasına benzer şekilde kesinlik, anma, F1 skoru için makro ortalamalar tercih edilmiştir.

Doğruluk metriğinin sınıflandırıcıların performansını ölçmek için tek başına yeterli olup olmadığını belirlemek ve sınıflandırma algoritmalarının seçimini sağlarken hangi performans metriklerinin daha faydalı olacağına karar vermek için Alan (2020) tarafından tez çalışması düzenlenmiştir. Kaggle'dan elde edilen 32 farklı veri seti için dokuz farklı sınıflandırıcı algoritma ile iki ve çok sınıflı sınıflandırma problemi çözümleri için modeller oluşturulmuş tahmin performansları karışıklık matrisi yardımıyla elde edilen doğruluk, duyarlılık, anma, kesinlik, seçicilik, F1 skoru, MKK, EAKA ve AİK eğrisi metrikleri ile değerlendirilmiştir. Mevcut çalışmaya benzer şekilde her iki sınıflandırma tahmin çözümünde de doğruluğun yeterli olmadığı doğruluk değerleri düşük olan algoritmaların MKK/ÇSMKK, EAKA ve F1 skoru değerlerinin yüksek olup daha iyi tahminlerde bulunduğu rapor edilmiştir.

Çalışmaların değerlendirmesinde doğruluk oranının yeterli olmadığı Infante vd. (2022) tarafından trafik kazaları yaralanma şiddetleri üzerine yapılan çalışmada da görülmüş kullanılan sınıflandırma algoritmalarından en yüksek doğruluk oranına NB sahip iken G-ortalama, F1 skoru ve genel ölçüm göz önüne alındığında C5.0 ve RO algoritması LR modeliyle benzer şekilde NB, DVM ve k-EYK'den daha yüksek performans sergilemiştir.

Abbad Ur Rehman vd. (2021) tarafından tiroid hastaları üzerine yapılan üç sınıflı sınıflandırma çalışmasında k-EYK, LR, NB, KA ve DVM algoritmaları kullanılmış NB'in en yüksek doğruluk, EAKA, MKK, F1 skor, kesinlik ve anma değerlerine sahip olduğu görülmüş ancak en düşük doğruluk değerine sahip KA'nın MKK değeri yönünden k-EYK, DVM ve LR'den daha yüksek olduğu görülmüştür.

Araştırma bulgularına göre en yüksek doğruluk oranlarına sahip algoritmalar yani GAM ve RO sınıflandırma algoritmaları ile elde edilen karışıklık matrisleri incelendiğinde üç sınıflı dengelenmemiş veri setinden elde edilen karışıklık matrislerindeki gözlem sayısı düşük azınlık gruplardaki tahminlerin azlığı dağılımın homojen olmadığını ancak iki sınıflılarda ve RAÖ ile dengelenmiş üç sınıflılarda azınlık gruplardaki tahminlerin çokluğu ve dağılımların daha homojen olduğu görülmüştür. RAÖ yönteminin kullanımı ile azınlık gruptaki tahmin doğruluğu artmış, tahmin değerlerine ait dağılım daha homojen hale gelmiş ve bu durum F1 skor, ÇSMKK, G-ortalama ve EAKA değerlerinde yükselişe neden olmuştur.

Araştırmadan elde edilen bulgular ile Bölükbaşı (2023) tarafından yürütülen tez çalışması sonuçları örtüşmekte olup çalışmada dengesizlik oranı 3,64 olan iki sınıflı diyabet veri seti iki farklı bölme tekniği (yüzdesel bölme, k-kat çapraz doğrulama), üç farklı yeniden örnekleme tekniği (RAÖ, RÖA ve SAÖAT) ve altı farklı MÖ kullanılarak dengelenmiş ve dengelenmemiş veri setleri performansları karşılaştırılmıştır. Veri seti %70 eğitim-%30 test verisi olarak ayrıldıktan sonra, RAÖ ile dengelenmiş ve k-EYK algoritması ile elde edilmiş modellerde doğruluk oranı %82,96'dan %83,70'e, EAKA değeri 0,778'den 0,829'a, ortalama F1 skoru 0,644'ten 0,764'e çıkmıştır. C4.5 modellerinde doğruluk oranı değeri %95,56'dan %97,04'e, ortalama F1 skoru 0,942'den 0,955'e, RO modellerinde ise doğruluk oranı değeri %94,81'den %97,04'a ve ortalama F1 skoru 0,915'ten 0,944'a çıkmıştır.

Demirsöz (2024) tarafından yürütülen tez çalışmasında farklı dengesizlik oranlarına sahip dört farklı veri seti kullanılmış dengelenmiş ve dengelenmemiş veri setleri performansları karşılaştırıldığında “climate” veri seti için RAÖ ile dengelenmiş ve RO algoritması ile elde edilmiş modellerde EAKA değeri 0,884'ten 0,949'a, F1 skoru 0,904'ten 0,969'a çıkmıştır. Diğer bir veri seti olan “yeast” için F1 skoru 0,969'dan 0,978'e, EAKA 0,795'ten 0,983'e ve MKK 0,435'ten 0,662'e çıkmış “vehicle” veri seti için ise F1 skoru 0,976'dan 0,988'e, EAKA 0,995'ten 0,999'a ve MKK 0,931'den 0,966'ya çıkmış olup mevcut çalışma ile benzer sonuçlara ulaşılmıştır.

Mevcut tez bulgularına göre eğitim ve test veri setlerinden elde edilen performans metrikleri arasındaki benzerlik ve farklılıklar değerlendirilmiş üç sınıflı dengelenmemiş ve iki sınıflı veri setlerinden elde edilen bütün performans ölçüm değerlerinin birbirine yakın olduğu görülmüştür. Bunun yanında test veri setlerinden elde edilen doğruluk oranına ait %95 güven aralıklarının çoğu itibarıyla eğitim veri seti doğruluk oranlarını kapsadığı görülmüştür.

RAÖ yöntemi ile üç farklı orandaki eğitim veri setinden elde edilen performans ölçümlerin test veri setinden yüksek olduğu, üç farklı orandaki eğitim veri setinde yüksek performans sergileyen algoritmaların test veri setinde aynı performansları sergileyemediği görülmüştür. Bu farkın kaynağının eğitim veri setinin aşırı yeniden örnekleme ile oluşturulması fakat test veri setindeki orijinal sınıf oranlarının korunması olduğu düşünülmektedir.

Üç sınıflı kazanç değişkenini içeren dengesiz ve dengeli farklı oranlardaki eğitim veri setlerinde sınıflandırma algoritmalarından en kararlı tavrı k-EYK algoritması ürettiği performans ölçütleri ile göstermiştir. RO algoritmasından elde edilen G-ortalama değerleri farklı oranlardaki üç sınıflı kazanç değişkenini içeren dengesiz ve dengeli eğitim veri setlerinde kararsız davranış göstermiştir.

Mevcut tez çalışmasında performans değerlendirme yöntemi olarak Alpaydın (2010) ve Han vd. (2012) tarafından da kullanılması tavsiye edilen ÇD 10 katlı ve 10 tekrarlı olarak kullanılmıştır. Awaysheh vd. (2016), Hyde vd. (2020) ve Özen vd. (2024) tarafından yapılan çalışmalarda da benzer şekilde bu performans değerlendirme yaklaşımı tercih edilmiştir.

Makine öğreniminde k-kat çapraz doğrulama yönteminde k değeri, oluşturulan modelin performansını önemli ölçüde etkilemektedir. Kabul gören bir kural bulunmamasına rağmen yapılan çalışmalarda MÖ uygulayıcıları ve uzmanları k değerini genellikle beş ya da on olarak almaktadır. Bunun en önemli nedenlerinden birisinin bu iki değerlerin ortalama tahminler üretmesi olarak düşünülmektedir (Barut ve Altuntaş, 2023; Nti vd., 2021).

Algoritmanın eğitim etkinliğini artırmak için tercih edilen ÇD yönteminin 10 katlı hali literatürde en sık yararlanılan performans değerlendirme yöntemi olup mevcut çalışmada yararlanıldığı gibi Abonyi ve Szeifert (2003), Aruna ve Nandakishore (2011), Aryal ve Paudel (2020), Lavanya ve Usha Rani (2012), Salama vd. (2012) ve Saygılı (2018) tarafından Wisconsin göğüs kanseri (Wisconsin Breast cancer, WBC), Wisconsin göğüs kanseri tanısı (Wisconsin Diagnosis Breast Cancer, WDBC) ve Wisconsin göğüs kanseri ilerlemesi (Wisconsin Prognostic Breast Cancer, WPBC) veri setleri kullanılarak yapılan çalışmalarda da tercih edilmiştir.

İki sınıflı sınıflandırma probleminin çözümünde olduğu gibi çok sınıflı sınıflandırma problemlerinin çözümünde de performans değerlendirme yöntemi olarak çoğunlukla 10 kat

ÇD teknikleri tercih edilen yöntem olmuştur. Chaganti vd. (2022) tiroid hastalığı tahmin yaklaşımı sunmak için tasarladıkları çalışmada dört sınıflı sınıflandırma problemi çözümünde, Shekar ve Dagnev (2019) lösemi (kan kanseri) hastalarının mikrodizi verileri ile üç sınıflı sınıflandırma çalışmasında, Deng vd. (2021) biyomedikal veri kümeleri kullanarak ölüm nedenlerini beş kategorili sınıflandırmak için ve Ishfaq vd. (2022) yedi farklı veri setindeki çok sınıflı sınıflandırma problemini çözmek için bu yöntemi kullanmıştır.

Farklı modellerin eğitiminde farklı k değerlerinin kullanılmasına yönelik az sayıda çalışma yapıldığı görülmüştür. Teorik olarak genel kabul görse ve çoğunlukla kullanılması önerilse de k-kat ÇD yönteminde kat sayısı arttıkça eğitilen model sayısı artmakla birlikte maliyet ve sürede artmaktadır (Nematzadeh vd., 2015).

Bu bulguların aksine, Wong ve Yeh (2020) tarafından 20 farklı veri seti üzerinde yapılan çalışmada ÇD yöntemi farklı kat sayıları ile denenmiş doğruluk oranlarındaki artış ile kat sayılarındaki artış arasında ilişki olduğu görülmüştür. Büyük sayıda k kat sayısı ile küçük sayıda tekrarların sınıflandırma algoritmalarında yüksek performans göstermesi için uyarlanması gerektiği belirtilmiştir. Bu çalışmalar yapılan araştırmalarda yalnızca 10 kat değil farklı katlarda ve tekrarlar ÇD yönteminin denenmesi gerektiğini göstermiştir.

MÖ sınıflandırma algoritmaları ile en iyi tahmin performansına ulaşabilmek için hiper parametrelere ince ayar gerekmektedir bu nedenle mevcut tez çalışmasında iki sınıflı sınıflandırma probleminin çözümünde olduğu gibi çok sınıflı sınıflandırma problemlerinin çözümünde de literatürde sıklıkla tercih edilen ızgara araştırma yöntemi kullanılmıştır.

Anggoro ve Mukti (2021) tarafından kronik böbrek yetmezliği verileri kullanılarak yapılan araştırmada tercih edilen AGAM hiper parametrelerini en iyilemek için ızgara ve rastgele arama yöntemleri uygulanmış elde edilen bulgulara göre ızgara arama yöntemi ile ulaşılan doğruluk ve F1 skor değerlerinin daha yüksek olduğu bulunmuştur.

İki ve çok sınıflı sınıflandırma probleminin çözümünde en iyi performansı gösteren hiper parametreleri elde edebilmek için ızgara araştırma tekniği tercih edilmiştir. Shekar ve Dagnev (2019) lösemi (kan kanseri) hastalarının mikrodizi verileri ile üç sınıflı sınıflandırma çalışmasında ızgara araştırma tekniğini tercih etmiştir. Abnoosian vd. (2023) çalışmasında katılımcıları diyabetik olmayan, prediyabet ve diyabetik olmak üzere üç sınıfa ayırmıştır. Algoritma eğitimi sırasında k-katlı ÇD yöntemi ile performans optimizasyonu ve hiper

parametre ayarı için ızgara araması ve Bayesian optimizasyonu teknikleri kullanılmıştır. RO algoritması hiper parametrelerinin ayarında Bayesian optimizasyonu kullanılmış ve en yüksek performans elde edilmiştir. Sumathi (2020) tarafından metin sınıflandırması üzerine yapılan çalışmada RO algoritması kullanılmış hiper parametre ayarlanmadan önceki tahmin doğruluğu ile ızgara arama yöntemi kullandıktan sonraki tahmin doğruluğu performansları karşılaştırılmış ve ızgara arama yöntemi ile elde edilen sonuçların daha yüksek olduğu rapor edilmiştir. Rahman vd. (2021) tarafından yedi tip cilt lezyonunu tahmin etmeye yarayacak otomatik bir sistem geliştirmek amacıyla beş derin öğrenme yöntemi kullanarak bir çalışma yürütülmüştür. Sinir ağlarının performansında hayati bir rol oynayan hiper parametrelerin en iyi kombinasyonunu bulmak için ızgara arama yöntemi kullanılarak gelişmiş ayırt etme yeteneğine sahip bir tahmin modeli ortaya konmuştur.

Çok sınıflı bir sınıflandırma probleminin çözümü için farklı yöntemler olmasına karşın bu tez çalışmasında ikili sınıflandırıcılardan literatürde en çok kullanılan bire karşı hepsi (BKH) sınıf ikileştirme (binarization) yaklaşımı tercih edilmiştir. Tummala vd. (2022) tarafından sekiz sınıflı meme kanseri tahmininde ve Özen vd. (2024) tarafından dört farklı sınıftaki rahvan koşan at ırklarının ayrımı için düzenledikleri çalışmalarda çok sınıflı sınıflandırma problemi çözümü için mevcut çalışmaya benzer şekilde bire karşı hepsi (BKH) yaklaşımı kullanılmıştır. Varpa vd. (2011) otonörolojik veriler ile hastalık sınıflandırma problemi tahmini için yaptıkları çalışmada k-EYK ve DVM algoritmalarını kullanmış bu çalışmanın aksine BKB ikili ayrıştırma yönteminin BKH yönteminden daha iyi sonuçlar verdiğini ve özellikle k-EYK sonuçlarında BKB yaklaşımının en yüksek sınıflandırma doğruluğuna ulaştığını rapor etmişlerdir. Abnoosian vd. (2023) diyabet hastaları üzerine yaptıkları çalışmada üç sınıflı sınıflandırma tahmini için bu tez çalışmasından farklı olarak ikili ayrıştırma yöntemi olarak BKB yaklaşımını benimsemiştir.

İkinci tartışma konusunu, elde edilen kazanç tahmin modellerinde yer alan özniteliklerin at yetiştiriciliğindeki önemlerinin sorgulanması oluşturmaktadır. Yapılan araştırmada, geleneksel parametrik yöntemler ile elden edilen tahmin modellerindeki değişkenlere ait katsayıların aksine MÖ algoritmaları ile eğitim veri setlerinden elde edilen sınıflandırma tahmin modellerindeki özniteliklerin önem dereceleri yani modeldeki etki sıralamaları elde edilmiştir. Elde edilen bulguların ışığında at yetiştiriciliğine ait önem dereceleri belirlenen risk faktörleri aşağıda tartışılmıştır.

Araştırmanın sonuçları, %70, %75 ve %80 oranlarındaki üç sınıflı dengelenmemiş, RAÖ yöntemi ile dengelenmiş veri setleri ve iki sınıflı eğitim veri setleri ile eğitilen C5.0, RO ve GAM sınıflandırma algoritmaları ile elde edilen tahmin modellerinin hepsinde yer alan yarışa başlama yaşı (YBY) özniteliği önem derecesi en yüksek ilk iki öznitelik arasında yer almıştır.

Elde edilen bulgular %70 oranındaki üç sınıflı kazanç değişkenini içeren dengelenmemiş eğitim veri seti haricindeki diğer eğitim veri setlerinde C5.0 sınıflandırma algoritmasından elde edilen tahmin modellerinde YBY özniteliğinin önem derecesi ilk sırada yer alırken tahmin modeline katkısı RO ve GAM tahmin modellerine katkısından daha yüksek bulunmuştur.

Bu çalışmada, YBY diğer çalışmalardan farklı olarak yıl değil ay olarak alınmış olup elde edilen bulgular ile dünyanın farklı yerlerinde yapılan çalışmalardan elde edilenler bulgular örtüşmektedir.

Yarışa başlama yaşını yıl olarak araştırmaya dahil eden Özen ve Gürçan (2017) safkan Türk Arap yarış atları üzerinde yaptıkları çalışmada 3,5 yaşından önce yarış kariyerine başlayan atların sonra başlayanlara göre yüksek kazanç elde etme olasılıklarının 2,35 kat daha fazla olduğunu rapor etmiştir. Kazancı performans ölçütü olarak alan Langlois ve Hernu (2003) 2, 3, 4 ve 5 yaşlarındaki İngiliz atları üzerine yaptıkları çalışmada, yaş arttıkça kazancın artma olasılığının yükseldiğini göstermiştir. Cheetham vd. (2010) beş yaşındaki safkan İngiliz atlarının diğerlerine göre daha yüksek kazanç elde ettiğini rapor etmiştir. Ekiz vd. (2005b) yaptığı çalışmada, 2 ile 7 yaş üzeri safkan İngiliz atları en düşük kazanç performansını sergilerken 3, 4, 5 ve 6 yaş atların kazanç performansı daha yüksek ve birbirine benzer bulunmuştur. Mota vd. (2005) 1.000-1.600 m safkan İngiliz atlarının yaşlarını 3 yaş ve öncesi, 4, 5 ve 5 yaş üzeri olarak kategorize etmişler, 1.100 m haricindeki diğer mesafelerde dört yaşındakilerin daha yüksek performans gösterdiğini ve beş yaş üzeri atların ise daha uzun yarış sürelerine sahip olduklarını rapor etmişlerdir. Ekiz vd. (2005a) tarafından safkan Türk Arap yarış atlarında yaşı yarış süresine anlamlı etkisinin olduğu gösterilmiştir. Moritsu vd. (1994), safkan İngiliz atlarında yaptığı araştırmada yaşı 1.200 ve 1.800 m koşularında anlamlı etken olduğu sonucuna ulaşmıştır. Ekiz ve Koçak (2007) Türkiye'deki İngiliz atları üzerine yapılan çalışmada, performans ölçütü olarak yarış süresinin alındığı çalışmada yaşı 1.200-2.400 m aralığındaki bütün mesafelerde etkisini gözlemlemiştir. Paksoy (2016), İngiliz atlarında yaptığı tez çalışmasında, yaşla beraber yarış performansı (hız) artmış, iki yaşında en düşük ve

altı üzeri yaşlarda ise en yüksek hız elde edildiğini rapor etmiştir. Buxadera ve Mota (2008) ile Park vd. (2011) yaptıkları araştırmalarla İngiliz atlarında yarış performansına yaşın etkisinin önemli olduğunu bildirmişlerdir. Bu sonuçlardan farklı olarak Mota vd. (1998) safkan İngiliz atlarında ve Villela vd. (2002) Quarterlarda yaşın yarış süresine etkisinin anlamlı olmadığını rapor etmiştir.

Atlarda genelde yaş ilerledikçe vücut ölçüleri az da olsa artar fakat safkan İngiliz atları büyüme ve gelişmesini iki yaşında genel olarak tamamlamış olurlar (Özbeyaz ve Akçapınar, 2007; Paksoy, 2016). Daha erken yaşlarda yarış hayatına başlayan genç atların saha tecrübeleri ve yarış koşullarına uyum sağlaması zamanla arttığı için bu durumun yarış performansına yani kazanca olumlu etkide bulunabileceği değerlendirilmektedir. Yaş ile beraber artan tecrübe ve koşulan yarış sayısının artması atların performansını olumlu bir şekilde etkileyeceği düşünülmektedir. Sağlıklı olan atlar, yaş aldıkça daha tecrübeli ve güçlü bir şekilde yarış koşarak daha yüksek performanslara ulaşabilmektedirler (Paksoy, 2016).

Yapılan araştırmada elde edilen sonuçlara göre, %70, %75 ve %80 oranlarındaki iki ve üç sınıflı eğitim veri setleri ile eğitilen üç sınıflandırma algoritmasından elde edilen tahmin modellerinin hepsinde koşu sayısının (KS) modellere katkısının ilk ve ikinci sırada yer aldığı görülmüştür. Bu özneliliğin önem derecesinin yüksekliği literatürdeki sonuçlar ile benzerlik göstermiştir.

Elde edilen bulgular, %70, %75 ve %80 oranlarındaki tüm eğitim veri setlerinde RO ve GAM sınıflandırma algoritmaları ile elde edilen tahmin modellerinde modele katkısı en yüksek olan özneliliğin KS olduğunu göstermiştir.

Cheetham vd. (2010) yaptıkları çalışmada, safkan İngiliz atlarında koşu sayısının kazanca pozitif yönde etki ettiğini rapor etmişlerdir. Kazancı performans ölçütü olarak alan Özen ve Gürçan (2017) yaptıkları çalışmada, koşu sayısı altıdan fazla olan Türk safkan Arap yarış atlarının daha yüksek kazanç elde etme olasılığının altıdan az olanlara göre 9,5 kat daha fazla olduğunu belirtmişlerdir. Yapılan araştırmada elde edilen bulgular ile literatürdeki bulguların örtüştüğü görülmüştür.

Cinsiyet özneliliğinin tüm tahmin modellerine katkısı değerlendirildiğinde, farklı önem derecelerinde de olsa %70, %75 ve %80 oranlarındaki iki ve üç sınıflı eğitim veri setlerinden elde edilen tahmin modellerinin çoğunda yer aldığı görülmüştür. Dünyanın farklı ülkelerinde

yapılan çalışmalarda da cinsiyet değişkeni atların kazanç yönünün değerlendirildiği tahmin modellerinde yer almış ve sıklıkla benzer sonuçlar rapor edilmiştir. Araştırma sonuçları, dişilerin referans olarak alındığı çalışmada erkek atları temsil eden cinsiyet özniteliğinin önem derecesi RAÖ yöntemi ile dengelenmiş üç farklı orandaki eğitim veri setinden elde edilen tahmin modellerinde diğer eğitim setlerinden elde edilenlere göre daha yüksek bulunmuştur. Cinsiyet özniteliğinin RO algoritması ile üç farklı orandaki eğitim veri setlerinden elde edilen tahmin modellerinde önem derecesinin diğer algoritmalarından daha yüksek olduğu görülmüştür. RAÖ yönteminin kullanılması sonucunda dengelenmiş eğitim veri setinden elde edilen bütün tahmin modellerindeki cinsiyet özniteliğinin modele olan katkısı artmıştır.

Kazancı performans ölçütü olarak alan Langlois ve Hernu (2003)'nın İngiliz atları üzerinde yaptıkları çalışmada 3, 4 ve 5 yaşlarındaki erkek atların, Özen ve Gürcan'ın (2017) safkan Türk Arap yarış atları üzerinde yaptıkları çalışmada da benzer şekilde erkek atların daha fazla kazanç elde ettikleri görülmüştür. Ekiz vd. (2005b) performans ölçütü olarak kazanç, süre ve sıralamayı aldıkları çalışmalarında, erkek safkan İngiliz atlarının dişilere göre yüksek performans gösterdiklerini bildirmişlerdir.

Yarış zamanını ve en iyi yarış zamanını yarış performans ölçütü olarak alan çalışmalarda; Mota vd. (1998) safkan İngiliz atlarında ve Ekiz vd. (2005a) Türk safkan Arap yarış atlarında erkek atların dişilere göre daha yüksek performans gösterdiklerini, Mota vd. (2005) ise safkan İngiliz atlarında 1.000-1.600 m de erkek atların dişilere göre daha yüksek performans gösterdiklerini rapor etmiştir. Ekiz ve Koçak (2007) Türkiye'deki İngiliz atları üzerine yapılan çalışmada, performans ölçütü olarak yarış süresinin alındığı çalışmada cinsiyetin 1.200-2.400 m aralığındaki bütün mesafelerde etkisini gözlemlemiştir. Moritsu vd. (1994) safkan İngiliz atları üzerine yaptığı çalışmada cinsiyetin 1.800 m koşularda anlamlı etken iken 1.200 m koşu için anlamlı olmadığını rapor etmiştir.

Paksoy (2016) İngiliz atlarında yaptığı tez çalışmasında erkeklerin yarış performansının (hız) dişilerden daha yüksek olduğunu belirtmiş, Buxadera ve Mota (2008) ile Park vd. (2011) yaptıkları araştırmalarla buna benzer sonuçlar rapor etmişlerdir. Bu çalışmada kullanılan MÖ algoritması sonuçları ve yukarıda dünyanın farklı yerlerinde yapılan araştırmaların gösterdiği sonuçların aksine Bakhtiari ve Kashan (2009) tarafından İngiliz atları üzerinde yapılan bir çalışmada cinsiyetin performansa etkisi önemsiz bulunmuştur.

Erkek atların dişilerden daha yüksek performans göstermesi birçok nedene dayanabilir. Bunlardan ilki vücut ölçüleri ve fiziksel güçtür. Paksoy (2016) tarafından safkan İngiliz atları üzerinde yapılan araştırmada, cıdago yüksekliği, baş genişliği ve ön incik çevresi gibi vücut ölçülerinde erkeklerin dişilerden daha yüksek değerler elde ettiği görülmüştür. Erkek atların fiziksel olarak daha güçlü olmasının, dişilerden daha yüksek performans gösterme sebeplerinden birisi olabileceği düşünülmektedir.

Erkek atların dişilerden daha yüksek performans gösterme nedenlerinden bir diğeri yarış hayatı süresidir. Özen (2014) tarafından safkan Türk Arap atlarında yapılan çalışma erkeklerin yarış hayatı süresinin dişilere göre daha uzun olduğu bulunmuştur. Benzer bulgular, Sobczynska (2007) tarafından safkan yarış atı popülasyonu üzerinde yapılan çalışmada elde edilmiştir. Bu sonuçlar, yarış hayatlarının erkek atlarda daha uzun olması kazanç performanslarının dişilerden daha fazla olmasının nedenlerinden birisi olabileceğini düşündürmektedir. Diğer bir sebep yarış atlarının yarışlardaki psikolojileridir. Sobczynska (2007), bireysel olarak safkan yarış atlarının yarışlardaki psikolojilerinin performanslar üzerinde etkili olabileceğini, Harkins vd. (1992) ise erkeklerin yarış sırasındaki rekabet ve kazanma duygusunun dişi atlardan daha fazla olduğunu gözlemlemiştir. Bu durumun, dişilerdeki kötü yarış performansını açıklayıcı bir diğer faktör olabileceği düşünülmektedir.

Dişilerin performans düşüklüğü nedenlerinden bir diğeri de başarılı dişi atların sağlık problemleri, sakatlıklar ve idmanlarda yaşadığı aksaklıklar yüzünden yetiştiricilikteki değerleri düşmektedir. Bu nedenle sahipleri tarafından erkeklere kıyasla daha erken yaşlarda damızlığa ayrılma durumunda kalmaktadırlar. Bunun sonucunda daha az yarışa dahil olup daha az kazanç elde etmektedirler.

Araştırma sonuçları, safkan atların yarış hayatı boyunca farklı tip koşu parkurunda koşma durumunun referans olarak alındığı çalışmada yarış hayatı boyunca çim, kum ve sentetik gibi tek tip koşu parkurunda koşmasının kazanca etkisinin olduğunu göstermiştir. Literatürde dünyanın farklı yerlerinde yapılan çalışmalar, kum ve çim pistlerde atların yarış performanslarının farklı olduğunu rapor etmiştir.

KP özniteliğinin koşu hayatı boyunca çim pistte (KPHerDaimÇim), kum pistte (KPHerDaimKum) ve sentetik pistte (KPHerDaimSentetik) koşma kategorilerinin farklı önem derecelerinde de olsa üç farklı orandaki dengelenmemiş, RAÖ yöntemi ile dengelenmiş ve iki sınıflı eğitim veri setlerinden elde edilen tahmin modellerinin hepsinde yer aldığı görülmüştür.

KP özniteliğinin koşu hayatı boyunca çim pistte (KPHerDaimÇim) koşma kategorisinin üç farklı orandaki eğitim setlerinden elde edilen tahmin modellerine katkısının YBY ve KS özniteliklerinden sonra üçüncü önem derecesine sahip olduğu görülmüştür. Koşu hayatı boyunca çim pistte (KPHerDaimÇim) koşma kategorisinin bütün tahmin modellerine katkısının koşu hayatı boyunca kum pistte (KPHerDaimKum) ve sentetik pistte (KPHerDaimSentetik) koşma kategorilerinden daha fazla olduğu görülmüştür.

Yarışların çim veya kum pistte düzenlenmesi atların yarış performansını doğrudan etkilemektedir. İngiliz atların çim pistte kum piste göre daha iyi performans gösterdikleri görülmüştür (Paksoy ve Ünal, 2010).

Performans değişkenini hız olarak alan çalışmalarda, Mota vd. (1998) ve Moritsu vd. (1994) tarafından İngiliz yarış atları için, Ekiz vd. (2005a) tarafından Türk safkan Arap atları için yapılan çalışmalarda çim pistte koşan atların kum piste göre daha hızlı oldukları rapor edilmiştir. Köseman ve Özbeyaz (2009) safkan Türk Arap atları üzerine yaptıkları çalışmada, 616 erkek tayın verisini kullanmış kum piste göre, çim pistte kazanılan ikramiyenin daha yüksek ve koşu süresinin ise daha az olduğunu göstermiştir.

Paksoy (2016) İngiliz atları üzerine yaptığı tez çalışmasında çim pistte koşan atların kum pistte koşanlara göre daha iyi performans (hız) gösterdiklerini rapor etmiştir. Buxadera ve Mota (2008) ile Park vd. (2011) İngiliz atları üzerinde yaptıkları araştırmalarda, çim pistte performansın kum pistten daha yüksek olduğunu bildirmişlerdir. Buna benzer şekilde, Arpacık (1999) tarafından farklı mesafelerde (1.200-1.800 m arası) birinciliği elde eden üç yaşındaki İngiliz atlarının çim pistte yüksek performans gösterdiği belirtilmiştir. Oki vd. (1995) pist türünün yarış süresini önemli ölçüde etkilediğini bildirmiştir. Cheetham vd. (2010) İngiliz yarış atlarında çim ve kum pist türlerinin kazanç performansına etkisinin anlamlı olduğunu ancak diğer çalışmaların aksine kum pistte koşanların daha yüksek kazanç elde ettiğini rapor etmiştir. Özen ve Gürcan (2017)'in safkan Türk Arap yarış atlarında yaptıkları çalışmada mevcut tez çalışmasından farklı olarak hem kum hemde çim pistte koşan yarış atlarının tek tip pistte koşanlara göre daha yüksek ilk yıl kazancı elde ettiği görülmüştür. Özen (2014) tarafından safkan Türk Arap atları üzerinde yapılan tez çalışmasında, farklı tipteki pistlerde (kum/çim) yarış hayatını geçirmiş olanların ortalama yarış hayatı tek tip pistte geçirenlere göre daha uzun olarak hesaplanmıştır. Yarış hayatı sonlanma riski, tek tip pistte koşanların farklı tip pistte koşan atlara göre 1,44 kat daha fazla olduğu bulunmuştur. Kas ve iskelet sistemi

sakatlığının en önemli nedenlerinden birisi yarışların yapıldığı pist tipi ve koşulları olduğu için farklı tipteki pistlerde yaşanan performans farklılıkları bu sakatlıklarla ilişkilendirilebilir.

Elde edilen tüm tahmin modellerindeki özniteliklerden önem derecesi yüksek diğer bir öznitelik sahiplik durumudur (SD). Tek sahibi olan yarış atlarının performansını temsil eden SD özniteliğinin, farklı önem derecelerinde de olsa farklı oranlardaki bütün veri setlerinden elde edilen tahmin modellerinin hepsinde yer aldığı görülmüştür. RAÖ yöntemi ile dengelenmiş üç farklı orandaki eğitim veri seti ile eğitilen sınıflandırma algoritmalarının öğrendiği tahmin modellerine SD özniteliğinin diğer veri setleri ile bulunan tahmin modellerinden daha fazla katkıda bulunduğu görülmüştür. Her üç eğitim veri seti ile GAM sınıflandırma algoritması kullanılarak elde edilen tahmin modellerindeki SD özniteliğinin katkısının derecesi diğer sınıflandırma algoritmaları ile elde edilen tahmin modellerine göre daha yüksek bulunmuştur. Özen ve Gürçan'ın (2016) safkan Türk Arap yarış atları üzerine yaptıkları çalışmada SD'nin yarış hayatı süresine etkisinin anlamlı olmadığı rapor edilmiştir.

Araştırma sonucunda, farklı oranlardaki üç eğitim veri setinde de C5.0, RO ve GAM sınıflandırma algoritmaları ile elde edilen sınıflandırma modellerinde baba kazanç durumu (BKD) ve anne kazanç durumu (AKD) öznitelikleri farklı önem derecelerinde yer almıştır. Hayatı boyunca hiç yarışa katılmayan anne ve babanın referans kategori olarak alındığı BKD ve AKD öznitelikleri yarış hayatı boyunca yarış kazanan ve kazanamayan kategorilerinin de dahil olduğu toplam üç kategoriden oluşmaktadır.

Literatür tarandığında safkan İngiliz atlarının yarış performansına anne ve babanın yarışlara katılması veya yarışlardaki başarı durumunun etkisiyle ilgili hiçbir araştırmaya rastlanmamıştır. Baba kazanma durumunu temsil eden BKDKazandı özniteliğinin çoğu tahmin modeline AKD (AKDKazandı ve AKDKazanamadı), SD ve cinsiyet (CinsiyetErkek) özniteliklerinden daha fazla katkıda bulunduğu görülmüştür. Literatür incelendiğinde baba ve anne kazanç durumlarını gösteren değişkenlerin dahil edildiği tahmin modelleme çalışmalarına rastlanmamıştır.

AKD özniteliği çoğu tahmin modelinde önem derecesi olarak SD ve cinsiyet özniteliklerinden önde gelmektedir. RAÖ ile dengelenmiş farklı oranlardaki veri setlerinden elde edilen sınıflandırma tahmin modellerinde ise AKD özniteliğinin BKD, SD ve cinsiyet özniteliklerinden daha önemli olduğu sonucuna ulaşılmıştır. Atların kazanç performansına

babalarının yarış hayatları boyunca yarış kazanmasının annelerin yarış kazanmasından daha yüksek derecede etki ettiği görülmüştür.

Yapılan araştırmada analize dahil edilmeyen faktörler ve uygulanmayan yöntemler gibi kısıtlılıkların, performans ölçütlerinde düşük sonuçlara neden olduğu değerlendirilmektedir. Araştırmanın kısıtlılıkları arasında yer alan durumlar şu şekilde sıralanabilir;

- Atlarda yarış performansını genotip, vücut yapısı ve çevresel faktörler etkilemekte olup bu çalışma kapsamında yalnızca yedi adet çevresel faktör kullanılmış bakım-besleme, idman, handicap ağırlığı, binicinin durumu, ana yaşı ve hastalıklar gibi faktörler çalışmaya dahil edilememiştir.
- Sınıf dengesizliği durumunda dengesiz verilerde sınıflandırma probleminin çözümü veri düzeyinde, algoritma düzeyinde ve melez yöntemler ile yapılabilirken bu çalışmada yalnızca veri düzeyinde örnekleme yöntemlerinden birisi olan RAÖ uygulanmıştır.
- AGAM, HGAM ve KGAM gibi daha güncel ve derin öğrenme algoritmaları gibi daha kapsamlı ve detaylı sınıflandırma algoritmaları ile değerlendirilmemiştir.
- Çalışmada analizler yalnızca RStudio programı ile yapılmış Python programlama dili gibi zengin kütüphaneleri (TensorFlow ve Scikit-Learn gibi) olan diller kullanılmamıştır.
- Araştırma dahilinde çok sınıflı sınıflandırma probleminin çözümünde yalnızca bire karşı hepsi (BKH) yöntemi kullanılmış olup, bire karşı bir (BKB) ve hata düzeltme çıkış kodları (HDÇK) gibi farklı yaklaşımlar ve melez yöntemler kullanılmamıştır.

5. SONUÇ VE ÖNERİLER

Yarış atlarının performans ve yarış kabiliyetinin göstergelerinden birisi olan kazanç değişkeni, genetik ve çevresel faktörlerden doğrudan ve dolaylı olarak etkilenmektedir. Yapılan çalışma ile, günümüzde büyük ve karmaşık veri analizlerinde çokça tercih edilen MÖ algoritmalarının hipotezsiz ve veri dağılımından bağımsız özelliği kullanılarak ilk kez Türkiye’deki safkan İngiliz yarış atlarının yarış hayatı boyunca elde ettiği kazanç çevresel faktörlerin etkisinin değerlendirilmesi için sınıflandırma tahmin modelleri oluşturulmuştur.

Yapılan araştırmada, ilk olarak iki ve üç sınıflı kazanç değişkenini içeren orijinal veri setleri %70-%30, %75-%25 ve %80-%20 oranlarında eğitim ve test veri setlerine bölünmüştür. %70, %75 ve %80 oranlarındaki üç sınıflı dengelenmemiş eğitim veri setine, RAÖ yöntemi ile dengelenmiş yeni eğitim veri setine ve iki sınıflı eğitim veri setine 10 katlı ÇD yöntemi 10 tekrarlı olarak uygulanmış, ızgara araştırma yöntemi ile k-EYK, C5.0, RO ve GAM sınıflandırma algoritmaları için hiper parametre en iyilemesi yapılmış ve eğitim setlerindeki tahmin modelleri öğrenilmiştir. Sonrasında tahmin modellerinin performansı %30, %25 ve %20 oranlarındaki orijinal test veri setleri ile sınanmıştır.

Elde edilen sonuçlar, üç farklı orandaki tüm test veri setlerinde iki sınıflı sınıflandırıcılardan elde edilen performans ölçüm değerlerinin seçicilik ve negatif tahmin değerleri haricinde dengeli ve dengesiz üç sınıflı sınıflandırıcılardan elde edilenlerden daha yüksek olduğunu göstermiştir. Sınıf sayısı arttıkça farklı sınıfların dağılımları daha karmaşık hale gelmekte olup çok sınıflı sınıflandırma problemlerinin karışıklık düzeyinin daha fazla olmasının bu sonuca götürdüğü düşünülmektedir.

Mevcut çalışmada, kazanç değişkeni sınıfları arasında sınıf dengesizliği problemi görülmüş bunu gidermek için RAÖ yöntemi kullanılmış %70, %75 ve %80 oranlarında yeni eğitim veri setleri oluşturulmuştur. RAÖ yöntemi ile dengelenmiş ve dengelenmemiş veri setleri ile elde edilen performans metrikleri karşılaştırıldığında, eğitim veri setlerinden elde edilen bütün performans metriklerinde RAÖ yöntemi ölçümleri yüksek bulunmuş olup test veri setlerine ait performans ölçüm değerlerinde ise doğruluk oranları, negatif tahmin ortalamaları ve EAKA değerleri tamamen kesinlik değeri ise kısmen hariç diğer ölçümlerde yüksek bulunmuştur. Buradan fazla zaman almayan uygulaması kolay ve etkili bir yeniden örnekleme yöntemi olan RAÖ’nün performans ölçütleri üzerindeki olumlu etkisinin görüldüğü sonucu çıkartılabilir. Bu olumlu etkinin karışıklık matrislerinde azınlık grupların doğru tahmin

oranının artmasına, azınlık ve çoğunluk sınıfları arasındaki doğru tahmin dengesinin sağlanmasına ve bunun sonucunda performanstaki yükselişe neden olduğu değerlendirilebilir.

Araştırma sonuçları, tüm test veri setlerinden elde edilen sınıflandırma tahmin modelindeki çoğu performans ölçümleri dikkate alındığında en iyi tahmin performansını GAM ve RO algoritmaları göstermiştir.

Veri setinde sınıflar arasında bir dengesizlik problemi olduğunda sınıflandırma tahmin performansını ve kalitesini değerlendirmek için yaygın olarak kullanılan F1 skoru, G-ortalama ve MKK metriklerinde en yüksek değerler daha dengeli veri seti olduğu düşünülen iki sınıflı kazanç değişkenini içeren verilerden elde edilmiştir. En düşük ölçüm değerleri ise dengelenmemiş üç sınıflı veri setinden elde edilmiştir. Bu veri seti RAÖ ile dengelendiğinde bu metriklerde yükseliş olduğu görülmüştür. Hedef değişken için sınıf sayısı arttıkça veri setindeki dengesizliğin artma olasılığı yükselmekte olup bunun da F1 skoru, G-ortalama ve MKK değerlerinde düşme olasılıklarını arttırdığı düşünülmektedir.

Veri setinin RAÖ ile dengelenmesi sonucunda çoğu metrikte beklendiği şekilde yükseliş görülmüş ancak bütün algoritmalarından elde edilen sonuçlarda doğruluk oranı, negatif tahmin değeri ve EAKA değerlerinde düşüş gözlenirken kesinlik metriğinde ise yalnızca GAM ve RO'da düşüş gözlenmiştir. Mevcut çalışmadan elde edilen bu bulgular, tahmin modellerinin karşılaştırmasında başta doğruluk oranı olmak üzere diğer performans karşılaştırma metriklerinin kullanılmasında daha dikkatli olunması ve performans kalitesine karar verilirken diğer ölçümlerin de değerlendirilmesi gerektiği sonucuna ulaştırmıştır.

Elde edilen bulguların ışığında, dengelenmemiş üç sınıflı eğitim veri setleri ile eğitilmiş k-EYK algoritmasının %30, %25 ve %20 oranlarındaki test veri setleri ile sınanması sonucu elde edilmiş dengeli doğruluk ortalaması, G-ortalama, EAKA, seçicilik ortalaması, anma metriği ve F1 skoru makro ortalaması performansının çalışma boyunca en başarılı algoritmalar olan RO ve GAM'dan daha yüksek olduğu görülmüştür. Bu sonuçlara göre, çok sınıflı tahmin problemlerinin çözümünde hızlı ve etkili sonuçlar elde edebilmek için uygulanması kolay ve hızlı çözümler sunan k-EYK algoritmasının çalışmalarda kullanımının gerekliliği değerlendirilmelidir.

RAÖ yönteminin olumlu diğer bir etkisi karışıklık matrisinde görülmüş, azınlık ve çoğunluk sınıfları arasındaki oran dengesi sağlandığı için sınıflar arasında bir dengesizlik

olduğunda bir sınıflandırma algoritmasının performansını ve kalitesini değerlendirmek için yaygın olarak kullanılan F1 skoru, G-ortalama ve MKK değerlerinde yükseliş görülmüştür. Bu yöntemin kullanımı ile azınlık gruptaki tahmin doğruluğu artmış, tahmin değerlerine ait dağılım daha homojen hale gelmiş ve bu durum kesinlik ve anma performans ölçütlerini beraber değerlendiren F1 skoru, duyarlılık ve seçicilik değerlerini beraber değerlendiren G-ortalama ve karışıklık matrisine ait tüm bilgileri içeren MKK değerlerinde yükselişe neden olmuştur. Bu sonuçlar, çok sınıflı sınıflandırmalarda elde edilen karışıklık matrislerindeki azınlık grupların tahmin sayılarının ve dağılımın homojenliğinin dengesiz veri setini anlamak için incelenmesi gerekliliğini göstermiştir.

Üç sınıflı dengesiz veri setinde performans ölçüm başarısının homojen olmadığı yani farklı algoritmaların farklı performans metriklerinde başarı gösterdikleri RAÖ ile dengelenmiş veri seti performans ölçümlerinde ise doğruluk oranı en yüksek algoritmanın diğer metriklerinde de başarı gösterdiği görülmüştür. İki sınıflı veri setinde ise dengeli veri setlerinden beklenen davranış olan doğruluk oranı en yüksek GAM ve RO algoritmalarının diğer metriklerinde de başarı göstermesi beklenmiş fakat C5.0 algoritması anma (duyarlılık) değeri diğerlerinden yüksek bulunmuştur. Bu durumun iki sınıflı veri setinde belli bir oranda dengesizlik sorunu olma olasılığına işaret ettiği düşünülmektedir.

Ulaşılan bu sonuçlar, GAM ve RO algoritmalarının tahmin başarısının yüksekliğini, iki sınıflı sınıflandırıcıların üç sınıflılara göre daha iyi tahmin performansı sergilediklerini, algoritmaların tahmin doğruluğunu değerlendirmek için yalnızca doğruluk oranının yeterli olmadığını, RAÖ ile dengelenmiş verilerin dengelenmemişlerden çoğu performans ölçütü yönünden daha iyi performans sergilediğini göstermiştir. Dengeli ve dengesiz veri setlerinde çok sınıflı sınıflandırıcı algoritmaların tahmin doğruluğunu değerlendirirken doğruluk oranlarının yanında, F1 skoru, G-ortalama ve ÇSMKK değerleri ile karışıklık matrisinin de beraber düşünülmesi gerektiği, EAKA performans ölçütünün kullanımında daha dikkatli olunması gerektiği görülmüştür.

At yetiştiriciliğinde harcanan para, zaman ve emek tasarrufu açısından bir yarış atının kazanç performansına etki eden risk faktörlerinin belirlenmesi oldukça önemlidir. Araştırma kapsamında, MÖ algoritmaları ile elde edilen sınıflandırma tahmin modellerinde yer alan risk faktörlerinin hangisinin modele daha çok katkı verdiğini gösteren önemlilik dereceleri elde edilmiştir.

Tahmin modellerindeki özniteliklere ait önem dereceleri bulgularının ışığında, yarış performansının kazanç boyutuna etki eden risk faktörleri olan YBY, KS, KP (KPHerDaimÇim, KPHerDaimKum ve KPHerDaimSentetik), BKD (BKDKazandı ve BKDKazanamadı), AKD (AKDKazandı ve AKDKazanamadı), SD ve CinsiyetErkek özniteliklerinin sınıflandırma modellerine farklı önem derecelerine göre katkıda bulunduğu gözlenmiştir.

Araştırmanın sonuçları, YBY ve KS özniteliklerinin tüm sınıflandırma tahmin modellerine ilk ve ikinci sırada katkıda bulunduğunu göstermiştir. Sonraki önem derecesi yüksek özniteliğin KP olduğu yarış hayatı boyunca farklı tip koşu parkurunda koşma durumunun referans olarak alındığı çalışmada koşu hayatı boyunca çim ve kum parkurda yarışmanın kazanca etkisinin yüksek olduğu görülmüştür.

SD, cinsiyet (CinsiyetErkek), BKD (BKDKazandı ve BKDKazanamadı) ve AKD (AKDKazandı ve AKDKazanamadı) özniteliklerinin farklı tahmin modellerinde farklı önem dereceleri ile modellerde yer aldıkları görülmüştür. Bu özniteliklerden BKD'nın (BKDKazandı) çoğu tahmin modeline diğerlerinden daha önemli katkıda bulunduğu yani babanın kazanmasının yarış atlarının kazanmasına daha çok etki ettiği sonucuna ulaşılmıştır. AKD (AKDKazandı ve AKDKazanamadı) özniteliği çoğu tahmin modelinde önem derecesi olarak SD ve cinsiyet (CinsiyetErkek) özniteliklerinden önde gelmektedir. Son olarak cinsiyet (CinsiyetErkek) özniteliğinin çoğu tahmin modeline katkısı SD özniteliğinden yani tek sahibi olan yarış atlarının katkısından daha yüksek bulunmuştur.

Bu sonuçların ışığında, ileride yapılacak çalışmalar için öneriler iki başlık altında sunulacaktır. İlki gelecekte yapılacak araştırmalarda yapay zekanın farklı uygulamalarının bu sorunu çözmek için kullanılmasıdır. Bu uygulamalar şu şekilde sıralanabilir;

- Farklı denetimli öğrenme MÖ sınıflandırma algoritmalarından (DVM, AGAM, HGAM ve KGAM gibi) yararlanılabilir.
- Hedef değişken olan kazancın sürekli veri olması halinde MÖ regresyon algoritmaları kullanılabilir.
- Denetimsiz, yarı denetimli ve takviyeli (pekiştirmeli) MÖ yöntemleri kullanılabilir.

- Derin öğrenme algoritmaları kullanılarak daha kapsamlı ve detaylı modeller geliştirilebilir.
- Bu çalışmada kullanılan RStudio programı haricinde çok sınıflı sınıflandırma problemleri için farklı çözüm yolları öneren örneğin Python programlama dili gibi zengin kütüphaneleri (TensorFlow ve Scikit-Learn gibi) olan diller kullanılabilir.
- Çalışmada ikili ayrıştırma yöntemlerinden bire karşı hepsi (BKH) kullanılmış olup, ileride yapılacak çalışmalarda çok sınıflı sınıflandırma problemlerinin çözümü için farklı yaklaşımlardan olan bire karşı bir (BKB) ve hata düzeltme çıkış kodları (HDÇK) gibi yöntemler kullanılabilir.
- Çok sınıflı sınıflandırma probleminin çözümünde ikili ayrıştırmalara gitmeden tek parça halinde çözüm üreten yeni geliştirilen farklı yöntemler veya modifiye edilen algoritmalar denenebilir. Örneğin; Adaboost.M1, Adaboost.M2, SAMME, SimMSVM ve MCRF.

Araştırma kapsamında safkan İngiliz atlarının yarış hayatı boyunca elde ettiği kazanca etki eden risk faktörleri önem derecelerine göre belirlenmiş, ulaşılan bulgular doğrultusunda yarış atı yetiştiricilerine şu önerilerde bulunulmuştur;

1. İlk olarak, sağlığı yerinde, yeterli antreman seviyesine ulaşmış, yarış eğitim seviyesi yüksek ve fiziksel kapasitesi yeterli yarış atları için en erken ve en uygun yarışa başlama yaşı seçilip yarışlara dahil edilebilir.
2. Her atın kendine has özellikleri dikkate alınarak en uygun beslenme ve antreman programı uygulanıp düzenli sağlık kontrolleri yapılarak atların yarış hayatları uzatılabilir.
3. Atların en uygun yarışa başlama yaşını, en uygun beslenme ve antreman programlarını belirlemek için giyilebilir teknolojiler, sensörler ve internet bağlantılı akıllı ürünlerin (IoT) kullanılması düşünülebilir.
4. Atların sağlıkları ve yarışa hazırlık durumları dikkate alınarak yarışlara katılma sayısı yani kazanç performansına doğrudan etki eden koşu sayısı arttırılabilir.

5. Atlar hangi zeminde daha çok başarılı ise antrenmanlarını ve yarış hayatı boyunca yarışlarını da aynı zemin tiplerinde koşturmaları tavsiye edilebilir.
6. İyi bir yarış atı yetiştiriciliğinde atların anne ve babalarının yarış hayatlarının da önemli olduğu unutulmamalıdır.
7. Popülasyon odaklı yapılan bu araştırmada kullanılan çevresel parametrelere klinik yaklaşımın bireysel başarıları artırmak için kullandığı parametreler yani beslenme tipi, antrenman süresi, sakatlık ve hastalık gibi bireysel özelliklerde eklenerek daha kapsamlı ve detaylı yarış performansı modellemelerine ulaşılabilir.

Bu uygulamaların, veteriner hekimlik alanında da yeni fırsatlar sunabileceği ve daha önce üstesinden gelinemeyen sorunlar için yeni çözüm yolları üretebilme potansiyeline sahip olabileceği görülmüş olup her koşulda veteriner hekimlerin daha hızlı karar almasına yardımcı olabileceği düşünülmektedir (Türkmen ve Özen, 2023).

KAYNAKLAR

- Abbad Ur Rehman, H., Lin, C.Y., Mushtaq, Z. & Su, S.F. (2021). Performance Analysis of Machine Learning Algorithms for Thyroid Disease. *Arabian Journal for Science and Engineering*, 46, 9437–9449. <https://doi.org/10.1007/s13369-020-05206-x>
- Abdullah, M.N., Yap, B.W., Sapri, N.N.F.F. & Wan Yaacob, W.F. (2023). Multi-class Classification for Breast Cancer with High Dimensional Microarray Data Using Machine Learning Classifier. In: Wah, Y.B., Berry, M.W., Mohamed, A., Al-Jumeily, D. (Eds.), *Lecture Notes on Data Engineering and Communications Technologies: Vol. 165. Data Science and Emerging Technologies* (pp. 329–342). Singapore: Springer. https://doi.org/10.1007/978-981-99-0741-0_24
- Abnoosian, K., Farnoosh, R. & Behzadi, M.H. (2023). Prediction of diabetes disease using an ensemble of machine learning multi-classifier models. *BMC bioinformatics*, 24(1), 337. <https://doi.org/10.1186/s12859-023-05465-z>
- Abonyi, J. and Szeifert, F. (2003). Supervised fuzzy clustering for the identification of fuzzy classifiers. *Pattern Recognition Letters*, 24(14), 2195-2207.
- Ahuja, Y. and Yadav, S.K. (2012). Multiclass classification and support vector machine. *Global Journal of Computer Science and Technology Interdisciplinary*, 12(11), 14-20.
- Akca, M.F. (2020, 07 Eylül). Karar Ağaçları (Makine Öğrenmesi Serisi-3). *Medium*. <https://medium.com/deep-learning-türkiye/karar-ağaçları-makine-öğrenmesi-serisi-3-a03f3ff00ba5>
- Akçapınar, H. ve Özbeyaz, C. (1999). *Hayvan Yetiştiriciliği Temel Bilgileri (ss:13-14)*. Ankara: Kariyer Matbaacılık.
- Akman, M. (2010). *Veri Madenciliğine genel bakış ve Random Forest yönteminin incelenmesi: sağlık alanında bir uygulama* [Yüksek Lisans tezi]. Ankara Üniversitesi.
- Akson, S. (1971). *Sait Akson'un yarışçılık anıları*. Ankara: Türkiye Jokey Kulübü Yayınları.
- Akyüz, İ. (2012). Geçmişten Günümüze Şanlıurfa Hipodromu At Yarışları Üzerine Bir Araştırma. *Yüzüncü Yıl Üniversitesi Veteriner Fakültesi Dergisi*, 3(2), 20-34.
- Alan, A. (2020). *Makine öğrenmesi sınıflandırma yöntemlerinde performans metrikleri ile test tekniklerinin farklı veri setleri üzerinde değerlendirilmesi* [Yüksek Lisans tezi]. Fırat Üniversitesi.
- Alom, M.Z., Yakopcic, C., Nasrin, M.S., Taha, T.M. & Asari, V.K. (2019). Breast cancer classification from histopathological images with inception recurrent residual convolutional neural network. *Journal of digital imaging*, 32, 605-617.
- Alpaydin, E. (2020). *Introduction to machine learning* (2nd Edition). Massachusetts: MIT Press.

- Altuntaş, M. (2010, 20 Ekim). At Yetiştiriciliğimiz ve Türk Safkan Arap Atının Gelişimi. <https://mustafaaltuntas.wordpress.com/2010/10/20/at-yetistirciligimiz-ve-turk-safkan-arap-atinin-gelisimi/>
- Arat, M.M. (2019, 02 Ekim). Multiclass Classification-One-vs-Rest/One-vs-One. <https://mmuratarat.github.io/2019-10-02/multi-class-classification>
- Araújo, T., Aresta, G., Castro, E., Rouco, J., Aguiar, P., Eloy, C., Polónia, A. & Campilho, A. (2017). Classification of breast cancer histology images using convolutional neural networks. *PloS one*, 12(6), e0177544.
- Arif, M.S., Mukheimer, A., & Asif, D. (2023). Enhancing the early detection of chronic kidney disease: A robust machine learning model. *Big Data and Cognitive Computing*, 7(3), 144.
- Arpacık, R. (1999). *At yetiştiriciliği* (3. baskı). Ankara: Şahin Matbaası.
- Aruna, S., Rajagopalan, S.P. & Nandakishore, L.V. (2011). Knowledge based analysis of various statistical tools in detecting breast cancer. *Computer Science & Information Technology*, 2(2011), 37-45.
- Aryal, S., Paudel, B., Aryal, S. & Paudel, B. (2020). Supervised classification using gradient boosting machine: Wisconsin breast cancer dataset. *International Journal of Scientific Research and Engineering Trends*, 6(3), 1887-1892.
- Anggoro, D.A. and Mukti, S.S. (2021). Performance Comparison of Grid Search and Random Search Methods for Hyperparameter Tuning in Extreme Gradient Boosting Algorithm to Predict Chronic Kidney Failure. *International Journal of Intelligent Engineering & Systems*, 14(6).
- Anonim (2018). Türkiye’de Haralar. *Z Dergisi At: Tematik Mevsimlik Kültür, Sanat, Şehir Dergisi Sayı*, 3. <https://www.zdergisi.istanbul/makale/turkiyede-haralar-216>
- Awaysheh, A., Wilcke, J., Elvinger, F., Rees, L., Fan, W. & Zimmerman, K.L. (2016). Evaluation of supervised machine-learning algorithms to distinguish between inflammatory bowel disease and alimentary lymphoma in cats. *Journal of veterinary diagnostic investigation*, 28(6), 679-687. <https://doi.org/10.1177/1040638716657377>
- Aygün Çakıroğlu, M. (2024). *Dengesiz Veri Kümelerinde Sınıflandırma Problemleri İçin Bulanık Tip-2 Tabanlı Yeni Bir Algoritma*. [Doktora tezi]. Erciyes Üniversitesi.
- Bailey, E., Petersen, J.L. & Kalbfleisch, T.S. (2022). Genetics of Thoroughbred Racehorse Performance. *Annu. Rev. Anim. Biosci.*, 10, 131-150. <https://doi.org/10.1146/annurev-animal-020420-035235>
- Baldi, P., Brunak, S., Chauvin, Andersen, Y.C.A.F. & Nielsen, H. (2000). Assessing the accuracy of prediction algorithms for classification: An overview. *Bioinformatics*, 16(5), 412-424. <https://doi.org/10.1093/bioinformatics/16.5.412>
- Balta, A. (2018). *Makine öğrenmesi teknikleri ile ekolojik verilerin değerlendirilmesi*. [Yüksek Lisans tezi]. Fırat Üniversitesi.

- Barandela, R., Sánchez, J. S., García, V. & Rangel, E. (2003). Strategies for learning in class imbalance problems. *Pattern Recognition*, 36(3), 849-851.
- Bardou, D., Zhang, K. & Ahmad, S.M. (2018). Classification of breast cancer based on histology images using convolutional neural networks. *Ieee Access*, 6, 24680-24693.
- Barut, Z., & Altuntaş, V. (2023). Comparison Of Performance Of Different K Values With K-Fold Cross Validation In A Graph-Based Learning Model For Incrna-Disease Prediction. *Kırklareli Üniversitesi Mühendislik ve Fen Bilimleri Dergisi*, 9(1), 63-82. <https://doi.org/10.34186/klujes.1248062>
- Başarır, S. (2022). *Suç veri setini analiz etmek için makine öğreniminde örnekleme teknikleri ve uygulaması*. [Yüksek Lisans tezi]. Yıldız Teknik Üniversitesi.
- Batista, G.E., Prati, R.C. & Monard, M.C. (2004). A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD explorations newsletter*, 6(1), 20-29.
- Berry, M.J. and Linoff, G. (1997) *Data Mining techniques: for marketing, sales and customer support*. (1st ed.). New York: Wiley.
- Biswas, R. and Gini J.R. (2024). Multi-class classification of Alzheimer's disease detection from 3D MRI image using ML techniques and its performance analysis. *Multimedia Tools and Applications*, 83(11), 33527-33554.
- Bonnie, L.H. (1995). *International Encyclopedia of Horse Breeds* (ss. 172-175). Norman: University of Oklahoma Press. https://books.google.com.tr/books?id=rhNFW0sD3jMC&pg=PA52&source=gbs_selected_pages&cad=1#v=onepage&q&f=false
- Bower, M.A., McGivney, B.A., Campana, M.G., Gu, J., Andersson, L.S., Barrett, E., Davis, C.R., Mikko, S., Stock, F., Voronkova, V., Bradley, D.G., Fahey, A.G., Lindgren, G., MacHugh, D. E., Sulimova, G. & Hill, E.W. (2012). The genetic origin and history of speed in the Thoroughbred racehorse. *Nature communications*, (3), 643. <https://doi.org/10.1038/ncomms1644>
- Bölükbaşı, İ.B. (2023). *Dengesiz Bir Diyabet Veri Setinde Makine Öğrenmesi Yöntemlerini Kullanarak Diyabet Hastalığının Teşhisi*. [Yüksek Lisans Tezi]. Uludağ Üniversitesi.
- Branco, P., Ribeiro, R.P. & Torgo, L. (2016). UBL: an R package for Utility-based Learning. *ArXiv*, abs/1604.08079.
- Breiman, L., Friedman, J.H., Olshen, R.A. & Stone, C.J. (1984). *Classification and Regression Tree* (1st ed.). New York: Routledge.
- Breiman, L. (1996). Bagging predictors. *Machine Learning*, 24(1), 123-140. <https://doi.org/10.1007/BF00058655>

- Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1), 5-32. <https://doi.org/10.1023/A:1010933404324>
- Brownlee, J. (2021, April 27). A Gentle Introduction to Ensemble Learning Algorithms. *Machine Learning Mastery*. <https://machinelearningmastery.com/tour-of-ensemble-learning-algorithms/>
- Bulut, M.A. (2019). *Kredi analizinde makine öğrenmesi kullanımı: tarımsal kredilerde uygulama örneği*. [Doktora tezi]. Eskişehir Osmangazi Üniversitesi.
- Burkov, A. (2019). *The hundred-page machine learning book* (Vol. 1, pp. 83). Quebec City, QC, Canada: Andriy Burkov. <http://ema.cri-info.cm/wp-content/uploads/2019/07/2019BurkovTheHundred-pageMachineLearning.pdf>
- Buxadera, A.M. and Mota M.D.S. (2008). Variance component estimations for race performance of thoroughbred horses in Brazil by random regression model. *Livestock Science*, 117, 298-307.
- Ceylan, T. (2018). *Perakende sektöründe makine öğrenmesine dayalı yaklaşımlar*. [Yüksek Lisans tezi]. Yıldız Teknik Üniversitesi.
- Chaganti, R., Rustam, F., De La Torre Díez, I., Mazón, J.L.V., Rodríguez, C.L. & Ashraf, I. (2022). Thyroid Disease Prediction Using Selective Features and Machine Learning Techniques. *Cancers*, 14(16), 3914. <https://doi.org/10.3390/cancers14163914>
- Chawla, N.V., Bowyer, K.W., Hall, L.O. & Kegelmeyer, W.P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16, 321-357.
- Cheetham, J., Riordan, A.S., Mohammed, H.O., McIlwraith, C.W. & Fortier, L.A. (2010). Relationships between race earnings and horse age, sex, gait, track surface and number of race starts for Thoroughbred and Standardbred racehorses in North America. *Equine Vet J*, 42(4), 346-350. <https://doi.org/10.1111/j.2042-3306.2010.00032>
- Chicco, D. and Jurman, G. (2020). The Advantages of the Matthews Correlation Coefficient (MCC) over F1 Score and Accuracy in Binary Classification Evaluation. *BMC Genomics*, 21(6), 12864. <https://doi.org/10.1186/s12864-019-6413-7>
- Chicco, D., Starovoitov, V. & Jurman, G. (2021). The Benefits of the Matthews Correlation Coefficient (MCC) over the Diagnostic Odds Ratio (DOR) in Binary Classification Assessment. *IEEE Access*, 9, 47112-47124. <https://doi.org/10.1109/ACCESS.2021.3068614>
- Cochran, C. (2023, December 14). Thoroughbred Horse Breed: Characteristics, Health & Nutrition Guide. *MADBARN*. <https://madbarn.com/thoroughbred-horse-breed-profile/#characteristics>
- Cover, T. and Hart, P. (1967). Nearest Neighbor Pattern Classification. *IEEE Transactions on Information Theory*, 13, 21-27. <http://dx.doi.org/10.1109/TIT.1967.1053964>
- Cunningham, E.P., Dooley, J.J., Splan, R.K. & Bradley, D.G. (2001). Microsatellite diversity, pedigree relatedness and the contributions of founder lineages to thoroughbred horses. *Animal genetics*, 32(6), 360-364.

- Currie, D. (2024, February 9). How Gradient Boosting Works. *Medium*. <https://medium.com/@Currie32/how-gradient-boosting-works-76e3d7d6ac76>
- Çağlayan Akay, E. (2018). Ekonometride Yeni Bir Ufuk: Büyük Veri ve Makine Öğrenmesi. *Sosyal Bilimler Araştırma Dergisi*, 7(2), 41-53.
- Çarkacı, N. (2018, February 12). Derin Öğrenme Uygulamalarında Hiper Parametre Seçim Yöntemleri. *Medium*. <https://medium.com/deep-learning-turkiye/derin-ogrenme-uygulamalarinda-model-dogrulama-ve-hiper-parametre-secim-yontemleri-823812d95f3>
- Çetin, C. (2018). Menzil Beygirleri. *Z Dergisi At: Tematik Mevsimlik Kültür, Sanat, Şehir Dergisi Sayı: 3*. <https://www.zdergisi.istanbul/makale/menzil-beygirleri-178>
- Çetinkaya, R. (2018). *Sultansuyu Harası (1865-1909)*. [Yüksek Lisans tezi]. Kütahya Dumlupınar Üniversitesi.
- Del Moral, P., Nowaczyk, S. & Pashami, S. (2022). Why Is Multiclass Classification Hard? *IEEE Access*, 10(1), 1-1. <https://doi.org/10.1109/access.2022.3192514>
- Demirsöz, S. (2024). *Dengesiz Veri Setlerinde Sınıflandırma Performansını Etkileyen Yaklaşımların İncelenmesi*. [Yüksek Lisans tezi]. Selçuk Üniversitesi.
- Deng, F., Huang, J., Yuan, X., Cheng, C. & Zhang, L. (2021). Performance and efficiency of machine learning algorithms for analyzing rectangular biomedical data. *Laboratory Investigation*, 101(4), 430-441.
- Dineva, K. and Atanasova, T. (2023). Health Status Classification for Cows Using Machine Learning and Data Management on AWS Cloud. *Animals*, 13(20), 3254.
- Ding, S., Zhao, X., Zhang, J., Zhang, X. & Xue, Y. (2019). A review on multi-class TWSVM. *Artificial Intelligence Review*, 52(2), 775–801. <https://doi.org/10.1007/s10462-017-9586-y>
- Drummond, C. and Holte, R.C. (2003, August). C4. 5, class imbalance, and cost sensitivity: why under-sampling beats over-sampling. In Workshop on learning from imbalanced datasets II (Vol. 11, No. 1–8).
- Ekiz, B., Koçak, Ö. & Demir, H. (2005a) Estimates of genetic parameters for racing performances of Arabian horses. *Turkish Journal of Veterinary and Animal Science*, 29(2), 543 – 549.
- Ekiz, B., Koçak, Ö. & Yılmaz, A. (2005b). Phenotypic and genetic parameter estimates for racing traits of Thoroughbred horse in Turkey. *A Archiv fur Tierzucht*, 48(2), 121-129. <https://doi.org/10.5194/aab-48-121-2005>
- Ekiz, B. and Koçak, Ö. (2007). Estimates of Genetic Parameters for Racing Times of Thoroughbred Horses. *Turkish Journal of Veterinary and Animal Science*, 31(1), 1-5.

- Eliaçık, M. (2018). Baytarnâmeler. *Z Dergisi At: Tematik Mevsimlik Kültür, Sanat, Şehir Dergisi Sayı: 3*. <https://www.zdergisi.istanbul/makale/baytarnameler-269>
- Endicott, R. (2018). At Lisânı, (H.F. Dursun, Çev.). *Z Dergisi At: Tematik Mevsimlik Kültür, Sanat, Şehir Dergisi Sayı: 3*. <https://www.zdergisi.istanbul/makale/at-lisani-258>
- Fernández, A., García, S., Galar, M., Prati, R.C., Krawczyk, B. & Herrera, F. (2018). *Learning from imbalanced data sets*. (Vol. 10). Cham, Switzerland: Springer.
- Fernández, A., López, V., Galar, M., Del Jesus, M. J. & Herrera, F. (2013). Analysing the classification of imbalanced data-sets with multiple classes: Binarization techniques and ad-hoc approaches. *Knowledge-based systems*, 42, 97-110.
- Fix, E. and Hodges, J.L. (1951). *Discriminatory Analysis, Nonparametric Discrimination: Consistency Properties*. [Project]. Technical Report 4, USAF School of Aviation Medicine, Randolph Field.
- Freund, Y. and Schapire, R.E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55, 119-39.
- Friedman, J.H. (2001). Greedy Function Approximation: A Gradient Boosting Machine. *Annals of Statistics*, 29, 1189-1232. <https://doi.org/10.1214/aos/1013203451>
- Galar, M., Fernandez, A., Barrenechea, E., Bustince, H. & Herrera, F. (2012). A Review on Ensembles for the Class Imbalance Problem: Bagging-, Boosting-, and Hybrid-Based Approaches. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 42(4), 463-484.
- Galar, M., Fernández, A., Tartas, E.B., Bustince, H. & Herrera, F. (2011). An overview of ensemble methods for binary classifiers in multi-class problems: Experimental study on one-vs-one and one-vs-all schemes. *Pattern Recognit.*, 44(8), 1761-1776. <https://doi.org/10.1016/j.patcog.2011.01.017>.
- Gedik, E. and Kolay, A. (2023). Osmanlı Ordusu İçin At Yetiştirilmesi: Kütahya Hara-yı Hümâyûnu. *Vakanüvis - Uluslararası Tarih Araştırmaları Dergisi*, 8(Dr. Recep Yaşa'ya Armağan), 2548-2582. <https://doi.org/10.24186/vakanuvis.1344515>
- Ghosh, A. and Kole, A. (2023). A comparative study of enhanced machine learning algorithms for brain tumor detection and classification. *Authorea Preprints*. <https://doi.org/10.36227/techrxiv.16863136.v1>
- Grandini, M., Bagli, E. & Visani, G. (2020). Metrics for Multi-Class Classification: an Overview. <https://doi.org/10.48550/arXiv.2008.05756>.
- Grootendorst, M. (2021, February 1). Distance Measures in Data Science, The advantages and pitfalls of common distance measures. *Medium*. <https://towardsdatascience.com/9-distance-measures-in-data-science-918109d069fa>

- Gong, J. and Kim, H. (2017). RHSBoost: Improving classification performance in imbalance data. *Computational Statistics Data Analysis*, 111, 1-13.
- Gouda, H.F., Hassan, F.A.M., El-Araby, E.E. & Moawed, S.A. (2022). Comparison of machine learning models for bluetongue risk prediction: a seroprevalence study on small ruminants. *S BMC Vet Res*, 18, 394. <https://doi.org/10.1186/s12917-022-03486-z>
- Gupta, M. (2020, April 21). Calculating Precision & Recall for Multi-Class Classification. *Medium*. <https://medium.com/data-science-in-your-pocket/calculating-precision-recall-for-multi-class-classification-9055931ee229>
- Günay, D. (2023, 12 Eylül). Gradient Boosting. *Medium*. <https://medium.com/@denizgunay/gradient-boosting-6578f7f5c974>
- Halaçoğlu, Y. (1991). *At: İslami Devir*. <https://islamansiklopedisi.org.tr/at#2-islami-devir>
- Han, J., Kamber, M. & Pei, J. (2012). *Data mining: concepts and techniques* (3rd Edition). Waltham: Morgan Kaufmann Publishers.
- Han, Z., Wei, B., Zheng, Y., Yin, Y., Li, K. & Li, S. (2017). Breast cancer multi-classification from histopathological images with structured deep learning model. *Scientific reports*, 7(1), 4172.
- Hand, D.J. and Till, R.J. (2001). A Simple Generalisation of the Area Under the ROC Curve for Multiple Class Classification Problems. *Machine Learning*, 45, 171-186.
- Hao, W.-T., Lu, Y., Dong, R.-H., Yong, Y.-L. & Zhang, Q.-Y. (2022). Adaptive Intrusion Detection Model Based on CNN and C5.0 Classifier. *International Journal of Network Security*, 24(4), 648-660. [https://doi.org/10.6633/IJNS.20220724\(4\).07](https://doi.org/10.6633/IJNS.20220724(4).07)
- He, H. and Garcia, E.A. (2009). Learning from imbalanced data. *IEEE Transactions on knowledge data engineering*, 21, 1263-1284.
- Hintz, R.L. (1980). Genetics of performance in the horse. *J. Anim. Sci.*, 51, 582–594.
- Hossin, M. and Sulaiman, M.N. (2015). A review on evaluation metrics for data classification evaluations. *International journal of data mining & knowledge management process*, 5(2), 1.
- Hsu, C.W. and Lin, C.J. (2002). A Comparison of Methods for Multiclass Support Vector Machines. *IEEE Transactions on Neural Networks*, 13, 415-425. <http://dx.doi.org/10.1109/72.991427>
- Hyde, R.M., Down, P.M., Bradley, A.J., Breen, J.E., Hudson, C., Leach, K.A. & Green, M.J. (2020). Automated prediction of mastitis infection patterns in dairy herds using machine learning. *Scientific reports*, 10(1), 4289.
- Imada, J., Arango-Sabogal, J.C., Bauman, C., Roche, S. & Kelton, D. (2024). Comparison of Machine Learning Tree-Based Algorithms to Predict Future Paratuberculosis ELISA Results Using Repeat Milk Tests. *Animals*, 14(7), 1113.

- Infante, P., Jacinto, G., Afonso, A., Rego, L., Nogueira, V., Quaresma, P., Saias, J., Santos, D., Nogueira, P., Silva, M., Kosta, R.P., Gois, P. & Manuel, P.R. (2020). Comparison of Statistical and Machine-Learning Models on Road Traffic Accident Severity Classification. *Computers*, 11(5), 80. <https://doi.org/10.3390/computers11050080>.
- Ishfaq, U., Shabbir, D., Khan, J., Khan, H.U., Naseer, S., Irshad, A., Shafiq, M. & Hamam, H. (2022). Empirical analysis of machine learning algorithms for multiclass prediction. *Wireless Communications and Mobile Computing*, 2022(1), 7451152.
- Jiang, Y., Chen, L., Zhang, H. & Xiao, X. (2019). Breast cancer histopathological image classification using convolutional neural networks with small SE-ResNet module. *PloS one*, 14(3), e0214587.
- Kafesoğlu, İ. (1991). *At: İslam Öncesi*. <https://islamansiklopedisi.org.tr/at#1-islam-oncesi>
- Karslı, Ö.B. (2019). *Makine Öğrenme Yöntemleri ile Karaciğer Hastalığının Teşhisi*. [Yüksek Lisans tezi]. Ağrı İbrahim Çeçen Üniversitesi.
- Kawsar, S. (2023, Nov. 09). K-Nearest Neighbors (KNN): Machine Learning Interview Prep 11. *Medium*. <https://kawsar34.medium.com/machine-learning-interview-quiz-11-k-nearest-neighbors-knn-09eac050cbdf>
- Kesik, M. (2018). Selçuklularda At Eğitimi. *Z Dergisi At: Tematik Mevsimlik Kültür, Sanat, Şehir Dergisi Sayı: 3*. <https://www.zdergisi.istanbul/makale/selcuklularda-at-egitimi-281>
- Keskin, M.V. (2018, 05 Ağustos). Makine Öğrenmesi ile İstatistiksel Öğrenme Arasında Ne Fark Var? *Veri Bilimi Okulu*. <https://www.veribilimiokulu.com/makine-ogrenmesi-ile-istatistiksel-ogrenme-arasinda-ne-fark-var/>
- Keskin, M.V. (2019, 18 Şubat). Ağaca Dayalı Yöntemlerde Bagging ve Boosting Arasında Ne Fark Var? *Veri Bilimi Okulu*. <https://www.veribilimiokulu.com/agaca-dayali-yontemlerde-bagging-ve-boosting-arasinda-ne-fark-var/>
- Khushvaktov, F. (2023, 26 August). Introduction Random Forest Classification By Example. *Medium*. <https://medium.com/@mrmaster907/introduction-random-forest-classification-by-example-6983d95c7b91>
- Kılıç, İ. (2023, 23 Eylül). Python Örneği ile Gradyan Arttırma Makineleri (GBM). *Medium*. <https://medium.com/@ilyurek/gradient-boosting-machines-gbm-with-python-example-b65421d10f2a>
- Kılıç, S. (2013). Klinik karar vermede ROC analizi. *Journal of Mood Disorders*, 3(3), 135-40.
- Kıral Özkan, T. (2015). *Makine öğrenmesi yöntemleri ile Vakıf Üniversiteleri doluluk tahminlemesi*. [Doktora tezi]. Marmara Üniversitesi.
- Koçkar, M.T. (2018). Atlı Oyun ve Sporlar. *Z Dergisi At: Tematik Mevsimlik Kültür, Sanat, Şehir Dergisi Sayı: 3*. <https://www.zdergisi.istanbul/makale/atli-oyun-ve-sporlar-209>

- Köksal, O. (2009). Osmanlı Dönüşüm Sürecinde Bir Devlet Teşebbüsü Olarak Çifteler Hâra-Yı Hümayunu ve Türk Atçılığına Katkıları. *Eskişehir Osmangazi Üniversitesi Sosyal Bilimler Dergisi*. 10(2), 333-363.
- Köseman A. and Özbeyaz, C. (2009). Some phenotypic and genetic parameters of racing performance in Arabian horses. *Ankara Üniversitesi Veteriner Fakültesi Dergisi*, 56(3), 219–224. https://doi.org/10.1501/Vetfak_0000002221.
- Köstem, R. (2000). *Tarihsel sürecinde atçılığımızın yapısı ve yarışçılığımızın oluşumu*. İstanbul: Türkiye Jokey Kulübü Yayınları.
- Köstem, R. (2018a). At Yarışlarının Kısa Târihi. *Z Dergisi At: Tematik Mevsimlik Kültür, Sanat, Şehir Dergisi Sayı: 3*. <https://www.zdergisi.istanbul/makale/at-yarislarinin-kisa-tarihi-220>
- Köstem, R. (2018b). At Yarışları ve Müşterek Bahis. *Z Dergisi At: Tematik Mevsimlik Kültür, Sanat, Şehir Dergisi Sayı: 3*. <https://www.zdergisi.istanbul/makale/at-yarislari-ve-musterek-bahis-221>
- Krawczyk, B. (2016). Learning from imbalanced data: open challenges and future directions. *Progress in Artificial Intelligence*, 5, 221-232.
- Krawczyk, B., Wozniak, M. & Herrera, F. (2015). On the usefulness of one-class classifier ensembles for decomposition of multi-class problems. *Pattern Recognition*, 48(12), 3969-3982. <https://doi.org/10.1016/j.patcog.2015.06.001>.
- Krishnan, S. (2021, August 26). Decision Tree for Classification, Entropy, and Information Gain. *Medium*. <https://medium.com/codex/decision-tree-for-classification-entropy-and-information-gain-cd9f99a26e0d>
- Kuhn, M. (2008). Building Predictive Models in R Using the caret Package. *Journal of Statistical Software*, 28(5), 1-26. <https://doi.org/10.18637/jss.v028.i05>
- Kulkarni, R. (2020, May 23). Summary of KNN algorithm when used for classification. *Medium*. <https://medium.com/analytics-vidhya/summary-of-knn-algorithm-when-used-for-classification-4934a1040983>
- Kundu, R. (2022, December 16). F1 Score in Machine Learning: Intro & Calculation. V7. <https://www.v7labs.com/blog/f1-score-guide>
- Kurnaz Şahin, F. (2022). II. Meşrutiyet Döneminde At Neslinin Islahına Yönelik Bir Örgütlenme Örneği: Islah-ı Nesl-i Feres Cemiyeti. *Selçuk Üniversitesi Edebiyat Fakültesi Dergisi*, 47, 315-334. <https://doi.org/10.21497/sefad.1128835>
- Lamyai, S. (2020, January 14). Ensemble_Boosting. https://commons.wikimedia.org/wiki/File:Ensemble_Boosting.svg

- Lan, G., Gao, Z., Tong, L. & Liu, T. (2022). Class binarization to neuroevolution for multiclass classification. *Neural Comput & Applic*, 34, 19845–19862. <https://doi.org/10.1007/s00521-022-07525-6>
- Langlois, B. (1980). Heritability of racing ability in Thoroughbreds - a review. *Livest. Prod. Sci.*, 7, 591–605.
- Langlois, B. (1996). A consideration of genetic aspects of some current practices in thoroughbred horse breeding. *Ann. Zootech.* 45, 41–51.
- Langlois, B. and Hernu V. (2003). An attempt to predict the earning status of a thoroughbred in France by genealogical data. *Animal Research*, 52 (1), 79-85. <https://doi.org/10.1051/animres:2003006>
- Lantz, B. (2015). *Machine learning with R*. (2nd ed.). Birmingham: Packt Publishing Ltd.
- Larose, D.T. and Larose, C.D. (2015). *Data mining and predictive analytics* (2nd ed.). New York: Wiley.
- Lavanya, D. and Rani, K.U. (2012). Ensemble decision tree classifier for breast cancer data. *International Journal of Information Technology Convergence and Services*, 2(1), 17-24.
- Liu, X. Y., Wu, J. & Zhou, Z. H. (2008). Exploratory undersampling for class-imbalance learning. *IEEE Transactions on Systems, Man, Cybernetics, Part B*, 39, 539-550.
- Mi, W., Li, J., Guo, Y., Ren, X., Liang, Z., Zhang, T. & Zou, H. (2021). Deep Learning-Based Multi-Class Classification of Breast Digital Pathology Images. *Cancer management and research*, 13, 4605–4617. <https://doi.org/10.2147/CMAR.S312608>
- Moritsu, Y., Funakoski, H. & Ichikawa, S. (1994). Genetic evaluation of sires and environmental factors influencing best racing times of thoroughbreds horses in Japan. *J. Equine.*, 5, 53-58. <https://doi.org/10.1294/jes.5.53>
- Mota M.D.S. (2006). Genetic correlations between performance at different racing distances in Thoroughbreds. *Livestock Science*, 104, 227-232.
- Mota, M.D.S., Abrahão, A.R. & Oliveira, H.N. (2005). Genetic and environmental parameters for racing time at different in brazilian thoroughbreds. *J. Anim. Breed. Genet.*, 122, 393-399. <https://doi.org/10.1111/j.1439-0388.2005.00551.x>
- Mota, M.D.S., Oliveira, H.N. & Silva, R.G. (1998). Genetic and environmental factors that affect the best time of thoroughbred horses in Brazil. *J. Anim. Breed. Genet.*, 115, 123-129. <https://doi.org/10.1111/j.1439-0388.1998.tb00335.x>
- Muratlar, E.R. (2020a, 24 Ocak). Gradient Boosted Regresyon Ağaçları. *Veri Bilimi Okulu*. <https://www.veribilimiokulu.com/gradient-boosted-regresyon-agaclari/>

- Muratlar, E.R. (2020b, 29 Nisan). LightGBM. Veri Bilimi Okulu. <https://www.veribilimiokulu.com/lightgbm/>
- Myint, K.L. and Tin, H.H. (2021, February 27-28). *Analyzing the Comparison of C4.5, CART and C5.0 Algorithms on Heart Disease Dataset using Decision Tree Method* [Proceeding]. Proceedings of the 2nd International Conference on ICT for Digital, Smart, and Sustainable Development, ICIDSSD 2020, Jamia Hamdard, New Delhi, India.
- Nematzadeh, Z., Ibrahim, R. & Selamat, A. (2015, May 31-June 03). Comparative studies on breast cancer classifications with k-fold cross validations using machine learning techniques. In *2015 10th Asian control conference (ASCC)*, (pp. 1-6). IEEE. doi: 10.1109/ASCC.2015.7244400.
- Nti, I.K., Nyarko-Boateng, O. & Aning, J. (2021). Performance of machine learning algorithms with different K values in K-fold cross-validation. *International Journal of Information Technology and Computer Science*, 13(6), 61-71.
- Quinlan, J.R. (1996). *Bagging, Boosting and C4.5*. In 13th National Conference on Artificial Intelligence (pp. 725-730). Portland, OR: AAAI Press.
- Oki, H., Sasaki, Y. & Willham, R.L. (1995). Genetic parameter estimates for racing time by restricted maximum likelihood in the thoroughbreds horse of Japan. *J. Anim. Breed. Genet.*, 112, 146-150. <https://doi.org/10.1111/j.1439-0388.1995.tb00550.x>
- Özen, D. (2014). *Türk safran Arap atlarında yarış hayatına etki eden faktörlerin belirlenmesi*. [Doktora tezi]. Ankara Üniversitesi.
- Özen, D. and Gürcan, İ.S. (2017). Factors that affect whether Arabian horses have earnings during their first year of racing. *Turkish Journal of Veterinary & Animal Sciences*, 41(4), 460-463. <https://doi.org/10.3906/vet-1612-61>
- Özen, D. and Gürcan, İ.S. (2016). Determination of factors affecting the length of racing career of Arabian horses in Turkey. *Turkish Journal of Veterinary & Animal Sciences*, 63(3), 303-309. https://doi.org/10.1501/Vetfak_0000002744
- Özen, D., Kaya, U., Özen, H., Ambarcioğlu, P., Ünal, N. & Gürcan, İ.S. (2021). Investigation of Factors Influencing Thoroughbred Horses' Racing Career Length in Turkey. *Journal of Equine Veterinary Science*, 107, 103782. <https://doi.org/10.1016/j.jevs.2021.103782>
- Özen, H., Özen, D., Yüceer Özkul, B. & Özbeyaz, C. (2024). Comparison of Some Balancing Methods for Classification of Pacing Horses Using Tree-based Machine Learning Algorithms. *Kafkas Univ Vet Fak Derg*, 30 (1), 31-39. <https://doi.org/10.9775/kvfd.2023.30325>
- Özgür, E.G. (2020). *Rasch modeline uygun yanıt desenlerinde bilgisayar uyarlamalı test yöntemi ile makine öğrenmesi yöntemlerinin tahmin performanslarının benzetim çalışması ile karşılaştırılması*. [Doktora tezi]. Ankara Üniversitesi.
- Öztemel, H. (2023, 09 Ağustos). Gradient Boosting (Gradyan Artırma): Regresyon ve Sınıflandırmada Güçlü Bir Tahmin Metodu. *Medium*. <https://medium.com/@hilaloztemel/gradient-boosting-gradyan-artırma-regresyon-ve-sınıflandırmada-güçlü-bir-tahmin-metodu-fb12536d79ee>

- Paksoy, Y. (2016). *İngiliz atlarında vücut ölçülerinin yarış performansına etkileri*. [Doktora tezi]. Ankara Üniversitesi.
- Paksoy, Y. and Ünal, N. (2010). Atlarda Yarış Performansını Etkileyen Faktörler. *Lalahan Hay. Araşt. Enst. Derg.* 50(2), 91-101.
- Paksoy, Y., Ünal, N., Polat, M., Tekin, M. & Özbeyaz, C. (2018). Arap ve İngiliz Atlarında Tırnak Büyüklüğünün Yarış Performansına Etkisinin Araştırılması. *Lalahan Hay. Araşt. Enst. Derg.* 58(1), 21-33.
- Pang, S.-L. and Gong, J.-Z. (2009). C5.0 Classification Algorithm and Its Application on Individual Credit Score for Banks. *Systems Engineering-Theory & Practice*, 29, 94-104. [https://doi.org/10.1016/S1874-8651\(10\)60092-0](https://doi.org/10.1016/S1874-8651(10)60092-0).
- Park, J., Lee, J., OH, S., Lee, J.W., OH, H. & Kim, H. (2011). Principal components analysis applied to genetic evaluation of racing performance of Thoroughbred race horses in Korea. *Livestock Science*, 135, 293-299.
- Pir, M.Ş. (2022). *Dengesiz Veri Setlerinde Sınıflandırma Problemlerinin Çözümünde Melez Yöntem Uygulaması*. [Yüksek Lisans tezi]. Uludağ Üniversitesi.
- Polat, K. and Günes, S. (2009). A novel hybrid intelligent method based on C4.5 decision tree classifier and one-against-all approach for multi-class classification problems. *Expert Syst. Appl.*, 36, 1587-1592.
- Polikar, R., Zhang, C. (Ed.). & Ma, Y. (Ed.). (2012). *Ensemble Machine Learning: Methods and Applications* (pp. 1-34). Boston, MA, USA: Springer.
- Provost, F. and Fawcett, T. (2001). Robust Classification for Imprecise Environments. *Machine Learning*, 42(3), 203-231.
- Rahman, Z., Hossain, M.S., Islam, M.R., Hasan, M.M. & Hridhee, R.A. (2021). An approach for multiclass skin lesion classification based on ensemble learning. *Informatics in Medicine Unlocked*, 25, 100659.
- Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C. & Müller, M. (2011). pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics* 12, 77. <https://doi.org/10.1186/1471-2105-12-77>
- Rosenblatt, F. (1962). *Principles of neurodynamics: Perceptron and theory of brain mechanisms*. Washington D.C.: Spartan Books.
- RStudio Team (2020). RStudio: Integrated Development for R. RStudio, PBC, Boston, MA URL <http://www.rstudio.com/>
- Sachinsoni. (2023, June 11). K Nearest Neighbours — Introduction to Machine Learning Algorithms. *Medium*. <https://medium.com/@sachinsoni600517/k-nearest-neighbours-introduction-to-machine-learning-algorithms-9dbc9d9fb3b2>

- Salama, G.I., Abdelhalim, M.B. & Zeid, M.A.E. (2012, November). *Experimental comparison of classifiers for breast cancer diagnosis*. In 2012 Seventh International Conference on Computer Engineering & Systems (ICCES) (pp. 180-185). Ieee.
- Samuel, A.L. (1959). Some Studies in Machine Learning Using the Game of Checkers. *IBM Journal of Research and Development*, 3(3), 210-229.
- Saner, C.B., Kesici, M., Mahdı, M., Yaslan, Y. & Genç, V.M.İ. (2019). Güç Sistemlerinde Geçici Hal Kararsızlığının Arıza Öncesi Fazör Ölçümleri Kullanarak Karar Ağacı Tabanlı Kestirimi. *Süleyman Demirel Üniversitesi Fen Bilimleri Enstitüsü Dergisi*, 23(1), 6-14. <https://doi.org/10.19113/sdufenbed.474888>
- Saygılı, A. (2018). Classification and diagnostic prediction of breast cancers via different classifiers. *International Scientific and Vocational Studies Journal*, 2(2), 48-56.
- Sevim, M. (2022). *Yarış atlarında yumuşak damağın dorsale deplasmanının insidansının araştırılması*. [Yüksek Lisans tezi]. Fırat Üniversitesi.
- Shah, D. (2023, May 04). Top Performance Metrics in Machine Learning: A Comprehensive Guide. V7. <https://www.v7labs.com/blog/performance-metrics-in-machine-learning>
- Shekar, B.H. and Dagnev, G. (2019, February). *Grid search-based hyperparameter tuning and classification of microarray cancer data*. In 2019 second international conference on advanced computational and communication paradigms (ICACCP) (pp. 1-8). IEEE.
- Sivrikaya, Ö. (2017). At Kültürü ve At Yarışları-Düzce Örneği. *Karabük Üniversitesi Sosyal Bilimler Enstitüsü Dergisi*, 7(2), 564-573.
- Sumathi, B. (2020). Grid search tuning of hyperparameters in random forest classifier for customer feedback sentiment prediction. *International Journal of Advanced Computer Science and Applications*, 11(9).
- Talu, H.S. (2023). *Vadeli Mevduat Hesabı Aboneliği Tahmini Makine Öğrenmesi Yaklaşımı*. [Yüksek lisans tezi]. Yıldız Teknik Üniversitesi.
- Tan, S. (2015). *Osmanlı Devleti'nde at yetiştiriciliği (1842-1918)*. [Doktora tezi]. Akdeniz Üniversitesi.
- Tanha, J., Abdi, Y., Samadi, N. & Razzaghi N. (2020). Boosting methods for multi-class imbalanced data classification: an experimental review. *J Big Data*, 7(1), 70, 19845–19862. <https://doi.org/10.1186/s40537-020-00349-y>
- Taşagıl, A. (2018). İslâm Öncesi Türklerde Atlı Hayat. *Z Dergisi At: Tematik Mevsimlik Kültür, Sanat, Şehir Dergisi*, 3. <https://www.zdergisi.istanbul/makale/islam-oncesi-turklerde-atli-hayat-293>
- Tepeoğlu, Y. (2018). *Orta mesafe koşan yarış atlarında performansın değerlendirilmesi ve yarış performansı ile ilişkisinin araştırılması*. [Doktora tezi]. Balıkesir Üniversitesi.

- Thiruvankadan, A.K., Kandasamy, N., & Panneerselvam, S. (2009). Inheritance of racing performance of Thoroughbred horses. *Livestock Science*, 121(2), 308-326. <https://doi.org/10.1016/j.livsci.2008.07.009>
- TJK (2024). Türkiye Jokey Kulübü. TJK. <https://www.tjk.org/TR/YarisSever/Static/Page/AyrintiliIstatistikler>
- Trevisan, V. (2022, Feb. 12). Multiclass classification evaluation with ROC Curves and ROC AUC. *Towards Data Science*. <https://towardsdatascience.com/multiclass-classification-evaluation-with-roc-curves-and-roc-auc-294fd4617e3a>
- Tripathi, A., Goswami, T., Trivedi, S.K. & Sharma, R. (2021). A multi class random forest (MCRF) model for classification of small plant peptides. *Int. J. Inf. Manag. Data Insights*, 1(2), 100029. <https://doi.org/10.1016/j.jjimei.2021.100029>
- Tolley, E. A., Notter, D. R. & Marlowe, T. J. (1985). A review of the inheritance of racing performance in horses. *Anim. Breed. Abstr.*, 16-185.
- Tomak, L. and Bek, Y. (2009). İşlem karakteristik eğrisi analizi ve eğri altında kalan alanların karşılaştırılması. *Journal of Experimental and Clinical Medicine*, 27(2). <https://doi.org/10.5835/jecm.v27i2.1569>
- Tummala, S., Kim, J. & Kadry, S. (2022). BreaST-Net: Multi-Class Classification of Breast Cancer from Histopathological Images Using Ensemble of Swin Transformers. *Mathematics*, 10(21), 4109. <https://doi.org/10.3390/math10214109>
- Türkmen, V. and Özen, D. (2023). Makine Öğrenmesinde Regresyon Algoritmalarının Veteriner Hekimliği Alanında Uygulamaları. *Veteriner Farmakoloji ve Toksikoloji Derneği Bülteni*, 14(3), 118-132. <https://doi.org/10.38137/vftd.1307581>
- Ünlüoğlu, M. (2022). At Yarışı Endüstrisinin Türkiye Ekonomisine Katkıları Üzerine Bir Değerlendirme. *Turkish Studies Social Sciences*, 17(6), 1249-1263.
- Vapnik, V.N. (1998). *The nature of statistical learning theory*. New York: Springer Verlag.
- Varpa, K., Joutsijoki, H., Iltanen, K. & Juhola, M. (2011). Applying one-vs-one and one-vs-all classifiers in k-nearest neighbour method and support vector machines to an otoneurological multi-class problem. *Studies in health technology and informatics*, 169, 579–583.
- Van Calster, B., Van Belle, V., Condous, G., Bourne, T., Timmerman, D. & Van Huffel, S. (2008, June 1-8). Multi-class AUC metrics and weighted alternatives. In *2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence)* (pp. 1390-1396). IEEE.
- Viazzi, S., Bahr, C., Schlageter-Tello, A., Van Hertem, T., Romanini, C.E.B., Pluk, A., Halachmi, I., Lokhorst, C. & Berckmans, D. (2013). Analysis of individual classification of lameness using automatic measurement of back posture in dairy cattle. *Journal of Dairy Science*, 96(1), 257-266.

- Vidyarthi, A., Agarwal, R., Gupta, D., Sharma, R., Draheim, D. & Tiwari, P. (2022). Machine learning assisted methodology for multiclass classification of malignant brain tumors. *IEEE Access*, 10, 50624-50640.
- Villela, L.C.V., Mota, M.D.S. & Oliveria, H.N. (2002). Genetic parameters of racing performance traits of quarter horses in Brazil. *J. Anim. Breed. Genet.*, 119, 229-234. <https://doi.org/10.1046/j.1439-0388.2002.00338.x>
- Wang, H., Li, D., Li, D., Liu, C., Yang, X. & Zhu, G. (2023). Remaining Useful Life Prediction of Aircraft Turbofan Engine Based on Random Forest Feature Selection and Multi-Layer Perceptron. *Applied Sciences*, 13(12), 7186. <https://doi.org/10.3390/app13127186>
- Wong, T.T. and Yeh, P.Y. (2019). Reliable accuracy estimates from k-fold cross validation. *IEEE Transactions on Knowledge and Data Engineering*, 32(8), 1586-1594.
- Yılmaz, H. (2014). *Random Forests yönteminde kayıp veri probleminin incelenmesi ve sağlık alanında bir uygulama*. [Yüksek Lisans tezi]. Eskişehir Üniversitesi.
- Yılmaz, O. (2012). Türkiye yerli at ırkları ve bir koruma çalışması. *Yüzüncü Yıl Üniversitesi Tarım Bilimleri Dergisi*, 22(2), 117-133.
- Zheng, Z., Cai, Y. & Li, Y. (2015). Oversampling method for imbalanced classification. *Computing Informatics*, 34, 1017-1037.
- Zhu, J., Zou, H., Rosset, S. & Hastie, T. (2009). Multi-class AdaBoost. *Statistics and its interface*, 2, 349-360
- Zhu, R., Guo, Y. & Xue, J-H. (2020). Adjusting the imbalance ratio by the dimensionality of imbalanced data. *Pattern Recognition Letters*, 133, 217-223.