ANKARA ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

YÜKSEK LİSANS TEZİ

BULANIK REGRESYON FONKSİYONLARI YÖNTEMİ İLE KONUT KREDİSİ RİSKİ MODELLEMESİ

Elif Hande EDİNSEL

GAYRİMENKUL GELİŞTİRME VE YÖNETİMİ ANABİLİM DALI

ANKARA
2023

ÖZET

Yüksek Lisans Tezi

BULANIK REGRESYON FONKSİYONLARI YÖNTEMİ İLE KONUT KREDİSİ RİSKİ MODELLEMESİ

Elif Hande EDİNSEL

Ankara Üniversitesi
Fen Bilimleri Enstitüsü
Gayrimenkul Geliştirme ve Yönetimi Anabilim Dalı

Danışman: Doç. Dr. Furkan BAŞER

Kredi riskinin değerlendirilmesi son yıllarda yaşanan finansal krizler nedeniyle önemli bir konu halini almıştır. Kredi riskinin doğru tespit edilmesiyle finansal kuruluşlar, kredi başvurusu yapan kişi ve kurumlara kredi vermemeyi ya da ek teminatlar talep etmeyi tercih ederek; kredi kaynaklı risklerini mümkün olduğunca aza indirebilmektedir. Finansal kuruluşların veriye dayalı doğru kredi verme kararı alabilmeleri için kredi riski değerlendirme modelleri kullanılmaktadır. Ancak kredi riski verileri genelde riskli ve riskli olmayan bireylerin oranına göre dengesiz bir dağılım sergilemektedir. Sınıflandırma modellerinde bu tür dengesiz form, çoğunluk sınıfı lehine yanlı tahminler ile sonuçlanmakta ve sınıflandırma doğruluğu riskli bireylerin yer aldığı azınlık sınıfı için çok düşük kalmaktadır. Bununla birlikte, göreceli olarak yüksek riskinden ve ciddi finansal kayıplara yol açmasından dolayı bankalar ve finans kurumları, pratikte azınlıkta olan sınıf vakalarının tespitine daha fazla önem vermektedir.

Bu çalışmada konut kredisi riskinin değerlendirilmesi amacıyla Bulanık Regresyon Fonksiyonları (BRF) yaklaşımından faydalanılmıştır. Yeniden örnekleme yöntemi ile dengesiz veri setinin getirmiş olduğu dezavantajlardan kaçınmak ve azınlık sınıfına ait doğruluğu daha yüksek tahminler elde etmek amaçlanmış olup, BRF yaklaşımı ile geleneksel makine öğrenmesi yöntemlerinin performansında iyileştirme gerçekleştirilmesi hedeflenmiştir. Açık kaynaktan temin edilen gerçek konut kredisi veri seti kullanılarak elde edilen bulgular farklı performans ölçütleri ile değerlendirilmiştir. Dengesiz bir dağılım gösteren konut kredisi veri setinde yeniden örnekleme yöntemine göre gerçekleştirilen farklı uygulamalarda, BRF yaklaşımının klasik makine öğrenmesi yöntemlerine göre daha üstün sınıflandırma doğruluğuna ulaştığı belirlenmiştir.

Eylül 2023, 89 Sayfa

Anahtar Kelimeler: Kredi riski, konut kredisi, bulanık regresyon fonksiyonları, makine öğrenmesi algoritmaları

ABSTRACT

Master Thesis

HOME CREDIT RISK MODELLING WITH THE FUZZY REGRESSION FUNCTIONS METHOD

Elif Hande EDİNSEL

Ankara University
Graduate School of Natural and Applied Sciences
Department of Real Estate Development and Management

Supervisor: Doç. Dr. Furkan BAŞER

The assessment of credit risk has become an important issue due to the financial crises experienced in recent years. By determining the credit risk correctly, financial institutions can reduce their credit-related risks as much as possible by choosing not to give credit to individuals or institutions that apply for credit or to demand additional collateral. Credit risk assessment models are used to enable the financial institutions to make the right decision based on data. However, credit risk data generally show an unbalanced distribution with respect to the ratio of risky and non-risky individuals. In classification models, an unbalanced form results in biased predictions in favor of the majority class, and the classification accuracy remains very low for the minority class which includes risky individuals. Due to its relatively high risk and resulting serious financial losses, banks and financial institutions, in practice, attach greater importance to determining minority class cases.

In this study, Fuzzy Regression Function (BRF) approach was used to evalaute the home credit risk. It is aimed to avoid the disadvantages brought by the unbalanced data set and to obtain higher accuracy estimates of the minority class by using the resampling method, as well as to improve the performance of traditional machine learning methods with the BRF approach. The findings obtained using the open source home credit dataset were evaluated using different performance criteria. Different applications using the resampling method for the unbalanced home credit data set show that BRF approach achieves superior classification accuracy compared to classical machine laarning methods.

September 2023, 89 Pages

Key Words: Credit risk, home credit risk, fuzzy regression functions, artifical intelligence

ÖNSÖZ VE TEŞEKKÜR

Çalışmalarımda beni yönlendiren, bilgi ve birikimlerini esirgemeyerek tüm süreç boyunca yanımda olan, yalnızca akademik ortamda değil, zor zamanlarımda da desteğini eksik etmeyen danışman hocam sayın Doç. Dr. Furkan Başer'e teşekkür etmeyi bir borç bilirim. Ayrıca yüksek lisans eğitimim boyunca kıymetli bilgilerini bizlerle paylaşan değerli hocalarım Sayın Prof. Dr. Harun Tanrıvermiş, Sayın Doç. Dr. Yeşim Tanrıvermiş, başta olmak üzere tüm bölüm hocalarıma, bir yandan çalışırken bir yandan eğitimimi sürdürmemde her daim destek olan şirketimiz yönetim kurulu başkanı Sayın Dr. Ahmet Cem Gel'e en derin duygularımla teşekkür ederim. Eğitimimi her zaman gönülden destekleyen ve bu süreçte birçok fedakârlık gösteren kıymetli eşime ve aileme de destekleri için minnetlerimi sunarım.

Bu tezi, hayatta en değer verdiğim varlığım olan ve yakın zamanda kaybettiğim "annem"e adamak istiyorum. İçimde bıraktığı boşluğu doldurmanın bir yolunu ömrümce bulamayacağımı biliyorum. Umarım her zaman söylediği gibi şimdi de yanımdadır ve benimle gurur duyuyordur. Onu hasret ve sevgiyle anıyorum.

Elif Hande EDİNSEL

Ankara, Eylül 2023

İÇİNDEKİLER

ŞEKİLLER DİZİNİ

ÇİZELGELER DİZİNİ

1. GİRİŞ

Kredi riski, bir banka borçlusunun veya karşı tarafın, üzerine anlaşmaya varılan şartlara uygun şekilde yükümlülüklerini yerine getirmemesinden kaynaklanan muhtemel zarar olarak tanımlanmaktadır (Council of Europe Development Bank, 2023). Özellikle dünyada son yıllarda yaşanan finansal krizler nedeniyle kredi riskinin değerlendirilmesinin önemi artmıştır. Bununla bağlantılı olarak, finansal kuruluşların doğru kredi verme kararı alabilmeleri için kredi riski değerlendirme modelleri de önemli hale gelmiştir (Yu vd. 2011). Ayrıca kredi derecelendirmeleri; tahvil yatırımcıları, borç verenler ve hükümet yetkilileri tarafından şirketlerin ve tahvillerin riskliliğini temsil eden bir ölçü şeklinde yaygın olarak kullanılmaktadır. Öyle ki kredi riski; risk primlerinin ve dahası tahvillerin pazarlanabilirliğinin de önemli belirleyicilerindendir (Huang vd. 2003).

Birçok finansal kurumda kredi riski ölçülmesine karşın, özellikle bankalar için kredi riskinin ölçülmesi büyük önem taşımaktadır. Banka riski ölçebildiğinde; risk notu yüksek olan ve temerrüde düşme riski taşıyan bir müşterisine kredi vereceğinde, daha güçlü teminatlar almaya yönelirken, düşük riski olan bir müşteriye ise daha kolay şartlar ile kredi verebilmektedir. Tüm bu kararların alınmasında ise riskin ölçümlenmesinin önemi ortaya çıkmaktadır (İldaş 2021).

Geçmişten günümüze kredi riskinin ölçülmesine yönelik farklı yöntemler uygulanmıştır. Bu amaçla; bankalar bünyelerinde uzmanlar istihdam etmiş ve bu kişilerin görüşleri doğrultusunda kredi vermeye yoğunlaşmıştır. Bu durum ise piyasada iyi bir izlenim yaratan firma veya kişilerin kredilerinin onaylanmasına, diğer kişi ya da firmaların ise kimi zaman bilgi eksikliğinden kimi zaman ise subjektif duyumlar neticesinde kredi kullanamamasına neden olmuştur. Yıllar içerisinde kredi taleplerinin artması, bankaların kredi kullandırmak için daha objektif ve güvenilir yöntemlere ihtiyaç duyması gibi sebeplerle farklı yöntemler geliştirilmiştir (İldaş 2021).

Bankacılık sektörü; finansal piyasalar, düzenleyici müdahaleler, operasyonel çerçeve ve küresel gelişmeler ile iç içe yapısı nedeniyle finansal riski oldukça belirsiz ve öngörmesi zor olan riske açık durumdadır. Bunlar arasında; piyasa riski, ekonomik durgunluklar ve

diğer özelliklere göre ortaya çıkabilen temerrüt riski ise en yaygın risk türleri olarak kabul edilmektedir (Lessmann vd. 2015). 2008'de ABD'de yaşanan eşik-altı ipotekli konut piyasası krizi ile bankacılık sektöründeki kontrolsüz büyüme ve finansal küreselleşmenin sonucu olarak, başta ABD'de gerçekleşen konut kredisi işlemlerinden kaynaklanan finansal kriz tüm dünyaya yayılmış ve dünyada her seviyede büyük finansal felaketlere sebep olmuştur. Ev sahiplerinin kredi riski ise özellikle bu kriz sonrasında, bankalar için başka bir kaygı sebebi olmaya başlamıştır. Sonuç olarak, finansal krizden önce bankacılık sektöründe risk yönetiminde yaşanan zafiyet, öngörülemeyen kayıplara yol açmıştır. Basel III gibi düzenleyici çerçeve, bankaların uyumluluklarını, rekabet güçlerini, finansal krizlere dayanıklılıklarını ve karlılıklarını tespit etmek için risk yönetimini uygulamalarını gerekli kılmaktadır (Gatzert ve Wesker 2012). Bu nedenle, bankalarının kurumsal ve bireysel müşterileri bazında kredi skorlarını sayısallaştırarak, karşı karşıya kalabilecekleri riskleri öngörmeleri de önem kazanmaktadır.

Kredi verme riskinin ölçülmesi yalnızca uzman görüşüne (kredi analisti) ya da müşterinin mevcut ve geçmiş finansal gücüne dayanılarak gerçekleştirilebilmektedir (Thomas vd. 1997). Bu nedenle güvenilir bir risk ölçüsünü tahmin edebilmek için tahmine dayalı yöntemleri somutlaştırmak, risk yönetimi sisteminde önem kazanmaktadır. Kredi skorlaması; aralarında diskriminant analizi ve lojistik regresyon gibi anlaşılır ve kolay uygulanabilir olması sebebiyle oldukça popüler olan modellerin de bulunduğu bir dizi ana akım modelleme yaklaşımına dayanmaktadır. Olasılıksal ve ekonometrik birçok tahmine dayalı yaklaşımlar ile en uygun makine öğrenmesi sınıflandırma algoritmalarını kullanmak; sağlam, güvenilir ve modernize edilmiş bir risk değerlendirmesi sağlamaktadır (Anderson 2007).

Makine Öğrenmesi uygulamaları, özellikle büyük veri setlerinin işlenmesinde oldukça cazip bir hale gelmekte (Alpaydın 2016) ve tahminde yüksek performans sağlayarak modellemedeki gücü göstermektedir. Son on yılda yaşanan ilerlemelere bağlı olarak gelişen makine öğrenmesi metotları; özellikle yapısı karmaşık, çok boyutlu, diğer faktörlerle iç içe ve dışsal etkilere duyarlı finansal verilerde, metotların seçimindeki çeşitlilikten, veri işleme ve modellemedeki esneklikten faydalanmaya imkan sağlamaktadır (Lessmann vd. 2015). Literatürde; yapay sinir ağları (Zhao vd. 2015, Liang

ve Cai 2020), destek vektör makineleri (Harris 2015, Yu vd. 2011), karar ağaçları (Teles vd. 2020, Golbayani vd. 2020), en yakın komşuluk (Hand ve Henley 1997) ve Naïve Bayes sınıflandırıcısı (Marqués vd. 2012) kullanımını öneren kredi skorlaması üzerine yapılmış çok sayıda makine öğrenmesi çalışması sunulmaktadır. Bahsi geçen sınıflandırıcıların kullanılmasının yanı sıra, birçok araştırmacı tarafından verinin gerçek içeriklerini yeterince iyi şekilde ifade etmek için karmaşık kredi skorlama modelleri de incelenmiştir.

Makine öğrenmesi metotlarının uygulanması; tahmin başarısı yüksek orana sahip algoritma seçimine, özellik seçimine, veri dönüşümüne ve doğrulama yapısına bağlıdır (Selçuk vd. 2022). Bu metotların farklı kombinasyonları sayısız deneme oluşturur. Bu nedenle en ilgili olanı seçmek sadece tatmin edici bir performans ölçümü ile yapılmamakta olup, aynı zamanda makine öğrenmesi algoritmalarının teknik özelliklerini ve değişkenlerini de iyi anlamayı gerektirmektedir (Koç 2019). Kredi skorlama analizleri sıklıkla; çoğu belirsiz, tanımlaması zor, hatta birbiriyle ilişkili ve çelişen birçok faktörün dikkate alınmasını gerektirmektedir (Syau vd. 2001). Risk değerlendirmesinde ortaya çıkan bu belirsizlik, karar alma süreçlerine destek olmak için güvenilir ve tutarlı bir sistem modelleme yaklaşımı gereğini ortaya çıkarmaktadır. Bu sayede de bulanık mantık tekniği, belirsizliklerin, muğlaklıkların ve insan düşünme sürecinin farklı formlarını modellemede geniş bir kullanım alanına ulaşmıştır.

Bulanık tabanlı makine öğrenmesi yaklaşımları, her veri örneğini eşit şekilde eğitmekten kaynaklanan istenmeyen aşırı uyum sorununa bir çare olarak kullanılabilmektedir. Bununla birlikte, bulanık modelleri eğitirken aykırı değerlerin etkisi, kredi skorlama modelleri bağlamında ilginç bir sorun olarak ortaya çıkmaktadır (Shieh ve Yang 2008). Temerrüde düşmeyen gözlemlerin genellikle temerrüde düşenlerden fazla olması nedeniyle kredi verilerinin sınıf dağılımı açısından dengesiz olduğu kabul edilir. Sınıflandırma modellerinde bu tür dengesiz form, çoğunluk sınıfı (temerrüde düşmeyen) lehine yanlı tahminler ile sonuçlanır. Böylece sınıflandırma doğruluğu azınlık sınıfı (temerrüde düşen) için çok düşük kalmaktadır. Bununla birlikte, ciddi finansal kayıplara yol açan göreceli olarak yüksek riskinden dolayı bankalar ve finans kurumları, pratikte azınlıkta olan sınıf vakalarının (temerrüde düşen) tespitine daha fazla önem vermektedir

(Galar vd. 2011, Chang vd. 2020). Literatür çoğunlukla temerrüde düşme olasılığını tahmin etmek ya da yeni bir kredi başvurusunu onaylayıp onaylamamaya karar vermek için daha önce borçlananlar hakkındaki eski bilgilere dayanan tek bir kredi riski modeli sunmaktadır. Ancak, borçlulara bağlı sistem girdilerinin heterojen olması nedeniyle, tek bir sınıflandırma kuralı, en doğru davranışa sahip modeli geliştirmek için yeterli olmayabilir (Lim and Sohn 2007). Bu engel, ikili çıktılar arasındaki gözlemlere göreli ağırlıklar veren bir bulanık kümeleme yaklaşımı ile çözülebilmektedir (Çelikyılmaz ve Türkşen 2007, 2008a,b). Bu çerçevede, Çelikyılmaz ve Türkşen (2009), sınıflandırma problemlerinin bulanık yapısını tanımlamak için kümeleme tabanlı bir bulanık modelleme olan Bulanık Regresyon Fonksiyonlarını (BRF) önermiştir. Bu metodun en önemli karakteristiği; sistemin yapısını daha iyi ifade edebilmek için girdi değişkenleri arasına bulanık üyelik değerlerini de eklemektir. Bir girdinin kümelere olan üyeliklerini veren Bulanık k-Ortalama (BKO) kümeleme algoritması; BRF'nin temelini oluşturur. Her bir küme için bulanık regresyon fonksiyonu parametrelerini tahmin etmek üzere, farklı çekirdek fonksiyonlarını kullanan Lojistik Regresyon (LR) ve Destek Vektör Makineleri (DVM) uygulanmaktadır.

Bahsedildiği gibi, borçlulara bağlı sistemin heterojen olmaması, tespiti daha önemli olan azınlık vakaların tespitini zorlaştırmaktadır. Bu gibi dengesiz veri setleri için yeniden örnekleme yöntemlerine başvurulabileceği görülmektedir. Yeniden örneklemenin amacı, veri setindeki bu dengesiz yapının giderilmesini sağlayarak, kullanılacak yöntemlerden doğruluğu daha yüksek sonuçlar elde edebilmektedir. Bu çalışmada, konut kredisi riskinin tespiti amaçlanmakta olup, veri seti incelendiğinde asıl tespiti istenen ve temerrüde düşme riski taşıyan verilerin azınlıkta olduğu ve veri setinin dengesiz olduğu görülmektedir. Bu çalışmada Çelikyılmaz ve Türkşen (2009) tarafından önerilen ve Baser vd. (2023) tarafından geliştirilen bulanık sınıflandırma yöntemi, yeniden örnekleme ile entegre edilerek konut kredisi veri seti üzerinden elde edilen sonuçlar karşılaştırılacaktır. Kredi riski veri setlerinin bahsedilen özelliklerine ilişkin olarak; bu çalışmanın amacı (i) Başer vd. (2023) tarafından geliştirilen; kümeleme tabanlı bulanık sınıflandırma yöntemini bir yeniden örnekleme stratejisi ile entegre ederek konut kredisi model performansını arttırmak (ii) her bir kümedeki model çıktılarını bulanık üyelik değerleri

ile ağırlıklandırarak her bir girdi için tek bir temerrüt olasılığını (Probability of Default - PD) hesaplamak için bir algoritma geliştirmektir.

Çalışmanın Birinci Bölümünde, kredi riski tanımlanmış ve konu ile ilgili genel bilgiler verilmiş olup, çalışmanın İkinci Bölümünde literatür taramasına yer verilmiştir. Çalışmanın Üçüncü Bölümünde kredi riski, konut kredisi finansman yöntemleri ve Türkiye'de kredi istatistiklerine de yer verilerek detaylıca incelenmiştir. Çalışmanın Dördüncü Bölümünde, sınıflandırma problemlerinde makine öğrenmesi algoritmaları ele alınmıştır. Çalışmanın Beşinci Bölümünde bulanık regresyon fonksiyonları yaklaşımı ve algoritmalarına yer verilmiş olup, çalışmanın Altıncı Bölümünde ise ikili sınıflandırma problemlerinde dengesiz veri setleri için yeniden örnekleme stratejileri sunulmuştur. Çalışmanın Yedinci Bölümünde konut kredisi veri seti kullanılarak BRF yaklaşımının uygulamasına yer verilmiştir. Çalışmanın Sekizinci Bölümünde ise ulaşılan önemli sonuçlar özetlenmiştir.

2. ÖNCEKİ ARAŞTIRMALARIN İNCELENMESİ

Kredi skorlaması modellerinin oluşturulmasında Lineer Diskriminant Analizi (LDA) ve Lojistik Regresyon (LR) Analizi gibi bazı istatistiksel yöntemler geniş bir kullanım alanı bulmuştur. Buna karşın bu yöntemlerin varsayımlarının, sınırlı sayıdaki örnek için sağlanmasının zor olması bu yöntemlerin kullanımını zorlaştırmaktadır (Wang vd. 2011). Ayrıca bu istatistiksel modellerin kolay uygulanabilir olmasına karşın, göreceli olarak düşük bir tahmin performansına sahip olması özellikle geniş veri setlerinde kullanılmasını da kısıtlamaktadır. Kullanışlı ve verimli bir modelden bahsedebilmek için kredi skorlama modellerinin hem sınıflandırma performansında hem de yorumlanabilirlik açısından dengeli dağılım göstermesi gerekmektedir (Bao vd. 2019).

Bağımlı değişkenin ikili değişken olması halinde kullanılan bir regresyon analizi türü olan Lojistik Regresyon, modeldeki bağımsız değişkenler arasında olabilecek çoklu doğrusallık problemlerine karşı hassastır. Değişken sayısı arttıkça modelin yeterliliği düşebilmektedir. Geleneksel istatistiksel yöntemlerin yorumlanması kolay olmasına karşın varsayımlarının sağlanması ve parametre tahmini sonucunda modelin belirlenmesi kullanımını zorlaştıran etkenlerdendir. Makine öğrenmesi yöntemleri modelin yapısının verilerden öğrenilmesini sağlamaktadır (Huang vd. 2004). Bu çerçevede literatürde makine öğrenmesi yöntemlerinin sıklıkla tercih edildiği görülmektedir.

Khandani vd. (2010) çalışmalarında bir makine öğrenmesi tekniği kullanmış olup, müşteri verilerinin sınıflandırılmasında CART (Classification and Regression Tree) modelinden faydalanmışlardır. ROC eğrisi ile başarının ölçümlendiği bu modelde modelin %82,9 doğruluk yüzdesine ulaştığı görülmektedir. Budak ve Erpolat (2012) ise yine bir bankadan temin edilen müşteri verilerini kullanarak hem lojistik regresyon hem de bir makine öğrenmesi algoritması olan yapay sinir ağları algoritması ile analiz gerçekleştirmiştir. Çalışma sonucunda yapay sinir ağları yöntemi ile %70,3 ve lojistik regresyon yöntemi ile %65,1 oranında sınıflama doğruluğuna ulaşıldığı görülmüştür (Budak ve Erpolat 2012).

Tian vd. (2019), kredi riskinin modellenmesinde Gradient Boosting Karar Ağacı yöntemini kullanmışlardır. Bu yöntemin seçiminde diğer bazı makine öğrenmesi yöntemlerinin özelliklerinden de bahsetmişlerdir. Örneğin; literatürde çokça tercih edilen makine öğrenmesi yöntemlerinden destek vektör makineleri (DVM), ikili sınıflandırma problemlerinde tercih edilen geleneksel yöntemlerden biridir. İlk kez 1964 yılında tanıtılmış olan yöntem zaman içinde görece yetersiz kalmış ve gelişen diğer algoritmaların yanında performansının zayıfladığı görülmüştür. Karar ağaçları algoritması ise tümevarımsal bir akıl yürütme yöntemi olarak tanımlanmaktadır. Burada ise dayanıklı (robust) olmama durumu söz konusudur ve verilerdeki küçük bir değişiklik bile ağacı değiştirebilmektedir. Gradient Boosting Karar Ağacı yöntemi, etkili ve güçlü bir sınıflandırıcı elde etmek için birkaç zayıf sınıflandırıcıyı bir araya getiren bir yöntemdir. Algoritma ile zayıf olan sınıflandırıcılara daha fazla, güçlü sınıflandırıcılara ise daha düşük ağırlıklar verilerek iteratif olarak model tahminine ulaşılmaktadır. Bu şekilde daha iyi performansa sahip sınıflandırıcılar elde edilebilmektedir (Tian vd. 2019).

Son yıllarda, bulanık küme analizinin makine öğrenmesi teknikleriyle birleştirilmesiyle gerçekleştirilen kredi riski değerlendirmesinin model performansını geliştirmede daha etkili olduğu gösterilmiştir (Bai vd. 2019). Örneğin, Malhotra ve Malhotra (2002) tüketici kredi uygulamalarında nöro-bulanık modellemeyi önermiş ve nöro-bulanık sistemlerin geleneksel istatistiksel tekniklere göre avantajlarını göstermişlerdir. Ramkumar (2016) bulanık çıkarım sistemi ve değiştirilmiş bir analitik ağ süreci kullanarak üçünü-şahıs e-tedarik sistemleri için bir risk değerlendirmesi modeli kurmuştur. Sohn vd. (2016) temerrüde düşme olasılığını tahmin etmek için bulanık mantığı geleneksel lojistik regresyon modeline entegre etmiştir. Önerilen bulanık lojistik regresyon modelinin doğruluğunun geleneksel lojistik regresyondan daha yüksek olduğu tespit edilmiştir. Shi ve Xu (2016), kredi riski analizi için bulanık üyelik değerlerinin sınıflandırma hiperdüzleminin öğrenmesinin her girdi noktasının farklı katkısını temin ettiği bir bulanık DVM algoritması önermiştir. Sun vd (2022) kredi riskinin niteliksel göstergeleri için bulanık mantık kullanılmasına dayanan çok katmanlı bir modelleme yaklaşımı geliştirmiştir. Bu bulanık karar verme yöntemi küçük sanayi işletmelerinin kredi derecelerinin tahmin edilmesine uygulanmaktadır.

Aynı kümelerde yer alan kredi kullananlar benzer davranış örüntüleri sergileyebileceğinden, her küme için farklı sınıflandırma kurallarının geliştirilmesi risk sınıflandırmasının doğruluğunun artırılmasına katkıda bulunabilir (Correa vd. 2012, Scitovski ve Šarlija 2014, Ghanbari vd. 2014). Kredi kullanıcılarının benzerliklerine göre kümelendirildiği çok sayda çalışmada, kredi riski analizinde her bir küme için ayrı model geliştirmenin etkinliği gösterilmiştir. Harris (2015) geleneksel doğrusal olmayan destek vektör makinelerinin kısıtlamalarından bazılarını araştırmış ve kredi notu değerlendirmesi için kümeleme tabanlı destek vektör makine yaklaşımının kullanılmasını önermiştir. Zhang vd. (2018), sınıflandırma algoritmasının eğitim ve test aşamalarında bulanık kümelemeyi kesin kümeleme yaklaşımına entegre etmiş, bir genetik algoritma kullanarak kredi notlandırmasında en iyi sınıflandırma modelini tespit etmişlerdir. Bao vd. (2019) bireylerin kredi riskini hesaplamak için kümeleme ve sınıflandırma yöntemlerinin birleştirilmesine dayanan bir algoritma önermiş ve bu bağlamda k-ortalama ve öz-düzenlemeli (kendi kendini düzenleyen) kümeleme modelleri ile birlikte yedi denetimli sınıflandırma modelini tartışmıştır. Bougaci vd. (2021) finansal stres tahmin performansını iyileştirmek için rastgele orman topluluk modelinin tahmini öncesinde k-ortalama kümelemesini kullanmıştır.

Uzun zamandan beri üzerinde araştırmalar yapılmaya başlanmış olmasına karşın, belirsizlik durumunda modelleme sorunu henüz tam anlamı ile sonuca ulaştırılamamıştır (Çelikyılmaz ve Türkşen 2009). Modelde bulunan değişkenlerin önemli değerlerinin açıklanabilmesi, kredi riskinin tahmin edilmesinde modellerin oluşturulabilmesinin temelini oluşturmaktadır. Bu değerlerin elde edilmesi kolay olmamaktadır. Bunun sebebi özellikle hayatın içerisindeki problemlerin yapısı itibariyle belirsizlik içermesidir. Bu çalışmada amaç; elde bulunan bilginin korunması ile belirsizliği parametrelerde ve bulanık regresyon fonksiyonlarının bünyesinde tutan bir bulanık modelleme oluşturabilmektedir. Bu çerçevede, sistem parametrelerini en iyi hale getirmek suretiyle gizli yapıları açıklayabilen güçlü bulanık modellere değinilecektir.

3. KREDİ RİSKİ

Kredi riskinin belirlenmesinin finans kurumları açısından önemi büyüktür. Bu amaçla gerçekleştirilen kredi notlandırmasının amacı, krediye başvuranları iki tipe sınıflandırmaktır. Bunlardan biri iyi kredibilitesi olan başvuranlar, diğeri ise kötü kredibilitesi olan başvuranlar şeklindedir. İyi kredibilitesi olan müşterilerin finansal yükümlülüklerini yerine getirme olasılığı yüksekken, kötü olanların ise finansal yükümlülüklerini yerine getirme olasılığı düşüktür. Bu noktada kredi notlandırmasının doğruluğu, finansal kuruluşun karlılığı açısından kritiktir. Kredibilitesi kötü olan müşterilerin kredi notlandırmasında yapılacak %1'lik bir iyileştirme bile finansal kuruluşun kaybını oldukça düşürebilecektir (Wang vd. 2011).

Kredi riskinin ölçümünde iki temel ve önem arz eden kavram bulunmaktadır. Bunlardan bir tanesi kredi notlaması (scoring), diğeri ise kredi derecelendirmesi (rating) şeklindedir. Temel olarak bu iki kavram arasında bazı farklılıklar bulunmakta olup, Çizelge 3.1'de sunulmuştur (SPL Lisanslama Sınavları Çalışma Notları, 2023).

Çizelge 0.1 Kredi notlaması ve kredi derecelendirmesi karşılaştırılması

Kredi Notlaması (Scoring)	Kredi Derecelendirmesi (Rating)
Kredinin geri ödenmemesi yani temerrüt riskini ölçen bir matematiksel modeldir.	Bir borçlunun borcunu geri ödeme isteği ve kapasitesi hakkında bir bağımsız görüş şeklindedir.
Temerrüt olasılığını hesaplar.	Matematiksel ve istatistiksel bir başarısızlık olasılığını sağlamaz.
Matematiksel olasılıkları notlara dönüştürmek mümkündür.	Uzman kuruluşlar tarafından ve genelde A, B, C gibi harfler ve +,- gibi semboller kullanılmak suretiyle kamuyla paylaşılır.
Risk bazlı fiyatlamaya imkan vermesi ve kredi riskinin ölçümünü sağlaması itibariyle özellikle bankacılık sistemine şeffaflık sağlar.	Risk bazlı fiyatlamaya imkan vermesi ve kredi riskinin ölçümünü ve yönetimini sağlaması itibariyle özellikle piyasalara sistemine şeffaflık sağlar.

3.1 Konut Kredisi

3.1.1 Konut finansman sistemleri

Konut talebi her ülkenin kendisine ait sosyokültürel ve demografik yapısına göre şekillenmektedir. Bu da her ülkenin konut finansman sisteminin kendisine özgü yapılanmasını sağlamaktadır. Ancak genel olarak bir konut finansman sisteminde; bir tasarruf sistemi, bu sistemden faydalanmak isteyen bireyler ve sistemi organize eden bir kamu otoritesi bulunmaktadır (Kılıç 2007). Bu çerçevede, kredi tasarruf sistemleri ana hatlarıyla dört başlık altında toplanabilmektedir (Sezer 2011).

- Doğrudan Finansman Sistemi (The Direct Route)
- Sözleşme-Mukavele Sistemi (The Contractual Route)
- Mevduat Finansmanı Sistemi (The Deposite Financing Route)
- İpotek Bankası Sistemi – İpoteğe Dayalı Finansman Sistemi (The Mortgage Bank Route)

3.1.1.1 Doğrudan finansman sistemi

Doğrudan finansman sistemi, genellikle gelişmiş bir finansman sistemine sahip olmayan ve gelişmekte olan ülkelerde görülen bir sistemdir. Bu sistemde konut talebi olan kişilerin, finansal olarak fazlası olan kişilerle kişisel ilişkileri ya da iş ilişkileri sayesinde doğrudan iletişimde oldukları ve finansman ihtiyaçlarını bu şekilde karşıladıkları bir sistemdir (Sezer 2011).

3.1.1.2 Sözleşme-mukavele sistemi

Bu finansman yönteminde uzmanlaşan kurumlara tasarruf sahiplerince yatırılan paralar, yatırım dönemi süresince piyasa fiyatlarının daha altında bir getiri elde etmektedir. Sürenin sonunda ise bekleyen tasarrufun önceden belirlenmiş bir katı kadar kredi kullanan tasarruf sahipleri, bu krediye de piyasanın altında bir faiz ödemektedir. Sisteme

yeni katılanlar, krediye hak kazanana dek tasarruflarını sistemde bekletirken, kredi kullananların ise bu bekleyen tasarruflardan kredi ihtiyaçları karşılandığından, sisteme sürekli yeni katılımcıların gelmesi gerekmektedir. Bu sistemin de gelişmekte olan ülkelerde (Brezilya, Meksika ve Filipinler gibi) uygulama örneklerine rastlanmaktadır. Fransa ve Almanya gibi gelişmiş ülkelerde ise özel kurumlar aracılığı ile bu sistemin daha kurumsal ve gelişmiş şekilde işletildiği görülmektedir (Sezer 2011).

3.1.1.3 Mevduat finansmanı sistemi

Konut finansmanı sistemleri içerisinde en sık karşılaşılan bu yöntem, mevduat toplayan finans kuruluşlarının topladıkları mevduatların hepsini ya da bir kısmını konut finansmanına aktarması şeklinde çalışmaktadır. Bu sisteme dahil kuruluşlar genellikle ticari bankalar, tasarruf bankaları, yapı kooperatifleri ve kredi birlikleri gibi kuruluşlardır. Bu kuruluşların asli işi mevduat toplamak olup, daha sonrasında toplanan mevduatın ne şekilde kullanılacağına karar verilmektedir. Tasarruf ve kredi birlikleri ABD'de yasal zorunluluklar çerçevesinde verdikleri kredilerin %65'ini konut veya diğer tüketici kredileri şeklinde vermekle yükümlüdür. Bu durum da ABD'de bu kurumları konut piyasalarında yaşanan kredilere karşı daha savunmasız bırakmaktadır (Sezer 2011).

3.1.1.4 İpotek bankası sistemi – ipoteğe dayalı finansman sistemi

İpotek bankacılığında, ipotek sürecinin tamamı bu kuruluşlarca yürütülmektedir. Bu süreçte, kredilendirme, fonlama, fon satışları ve kredi hizmetleri gibi tüm aşamalar mevcuttur. İpotek ise borcun geri ödenmesinin teminatı olarak mülkün rehin alınması işlemidir. İpotek sisteminin amacı, uzun vadeli finansman sağlamaktır (Sezer 2011).

Bu sistemde iki aşamalı bir aracılık söz konusudur. Mevduat fazlası bulunan kişiler bu mevduatlarını tasarruf bankalarına ya da ticari bankalara yatırır. Bu bankalar ise yatırılan bu mevduatlar ile ipotek bankaları tarafından ihraç edilen tahvilleri satın almak için kullanır. İpotek bankaları da bu şekilde konut kredisi almak isteyenler için kullanılacak fonları oluşturmuş olur. İpotek bankalarınca ihraç edilen ipotek tahvilleri, hayat sigortası

şirketleri, emeklilik fonları gibi kurumsal yatırımcılar tarafından da alınabilmektedir. Bu kurumlar da aldıkları bu menkul kıymetler aracılığı ile verdikleri kredileri temin etmiş olur. Burada sistemin düzgün işlemesi için ikincil piyasalarda tahvil satılabilmesi önem arz etmektedir (Sezer 2011).

İpotek bankası sisteminin benzer işleyişlere sahip olmakla birlikte temel iki uygulaması mevcuttur. Bunlar Almanya ve Danimarka sistemleridir. Almanya sisteminde; yeni bir kredi verildiğinde tahvil ihraç edilmekte ve bu şekilde kaynak yaratılmaktadır. İhraç edilmiş krediler ipotek bankasının bilançosunda kalmaktadır. Bu da kredi riskinin ipotek bankasında kalması anlamına gelmektedir. Danimarka sisteminde ise ipotek bankaları tahvilin ihracatçısı değil aracısı olarak çalışmaktadır. Bu sistemde aracı rolü üstlenen bankalar riski tahvil yatırımcılarına satmaktadır (Sezer 2011).

Bu sistemlerin en gelişmiş hali ise Amerika'da yaygın şekilde kullanımı olan ipotek karşılığı menkul hale getirilen krediler sistemidir. Burada; gelecekte bir nakit girişi oluşturacak olan alacaklar, krediler vb. araçlar bir araya getirilir ve bir havuz oluşturulur. Bu havuzlara göre menkul kıymetler oluşturulur ve derecelendirilerek üçüncü parti yatırımcılara satılır (Sezer 2011).

Mortgage Sistemi ise; ipotek bankası sisteminin gelişmiş bir halidir. Bu sistemde ipotek bankası tarafından kaynak sağlamak yerine yatırımcılara konut kredileri satılmaktadır. Bu sistemde birincil piyasalarda gayrimenkul almak isteyenler ile bu gayrimenkule kredi sağlamak isteyen kurumlar bir araya gelir. İkincil piyasalarda ise mevcut mortgage kredileri çalışmaktadır. İkincil piyasalar birincil piyasalara fon sağlamaktadır (Sezer 2011).

3.1.2 Konut talebini etkileyen faktörler

Fiyat: Konut talebini etkileyen önemli bir faktör olarak konut fiyatı sayılabilmektedir. Hem ikamet ya da kiralama amaçlı, hem de yatırım amaçlı olarak konut edinmek istenmektedir. Konut kiralarında yaşanan artışın konut fiyatlarında da artışı getirdiği

görülmektedir. Yine benzer şekilde enflasyonun yüksek olduğu durumlarda, konut edinmek için gereken konut finansman olanakları da olumsuz etkilenmektedir. Bununla beraber enflasyon beklentisi de konut talebini arttırmakta, fiyat düşüşü beklentisi ise talebi azaltmaktadır. Bu çerçevede incelendiğinde konut talebi ile konut fiyatları arasında ters yönlü bir ilişki olduğundan bahsedilebilmektedir (Öztürk ve Fitöz 2009).

Gelir: Gelirdeki artışın evlilik oranlarını arttırdığı ve bunun da dolaylı olarak konut talebini arttırdığı değerlendirilmektedir. Konut talebinin analizinde reel gelir önemli bir araç şeklinde kullanılmaktadır. Yatırım veya tüketim amacıyla olması fark etmeksizin, hane halkının ömrü boyunca elde etmeyi planladıkları gelire göre konut talepleri şekillenmektedir. O halde konut talebi ile gelir arasında aynı yönde pozitif bir ilişkiden bahsedilebilmektedir (Öztürk ve Fitöz 2009).

Kredi Koşulları ve Faizleri: Özellikle orta gelir gruplarına yönelik olarak esnek ödeme kolaylığı olan krediler ile konut talebinin önemli ölçüde artabildiği görülmektedir. Günümüzde konut faiz oranlarının düşmesi ve mortgage kredileri gibi uzun vadeli kredilerin kullandırılmasının konut kredilerine artış yarattığı anlaşılmaktadır (Öztürk ve Fitöz 2009).

Sosyal Talepler: Özellikle gelişmekte olan ülkelerde konut alımı aynı zamanda yarını garanti altına almak için bir güvenlik aracı olarak görülmektedir. Bu toplumlarda etkin bir sosyal güvenlik sistemi oluşmadığından, konut bir güvence olarak görülmekte, gerekirse teminat amaçlı kullanılmakta ve gelecek nesillere bırakılmak için konut talep edilmektedir (Öztürk ve Fitöz 2009).

Demografik Faktörler: Köyden kente veya kentten kente göçler sonucunda konut taleplerinin artması ile konut üretimi de artmaktadır. Yine nüfus artışı da diğer tüm mal ve hizmetlerin talebini arttırdığı gibi konut talebini de arttıran bir unsur olarak öne çıkmaktadır (Öztürk ve Fitöz 2009).

3.1.3 Konut arzını etkileyen faktörler

Fiyat: Konut arzını etkileyen önemli faktörlerden birisi de yapım maliyetlerinde yaşanan değişikliklerdir. Arz miktarı sabitken kullanıcıların maliyetlerinin artması halinde öncelikle reel konut fiyatları artış göstermektedir. Akabinde konut arzının artması ile konut fiyatları düşerek bir dengeye oturmaktadır. Fiyatın artış hızı da inşaat sayısına ve bu inşaatların yapım maliyetleri ile ilişkilidir. Özetle, konut fiyatları ile konut arzı arasında aynı yönlü bir pozitif ilişkiden bahsedilebilmektedir (Öztürk ve Fitöz 2009).

Kentleşme Hızı: Kentleşme hızı arttıkça konut arzının da artması beklenmektedir. Planlı bir kentleşme; konut arzını arttırırken, çarpık yapılaşmayı ve aşırı kentleşme durumunu düzenleyebilmektedir. Türkiye'de de örnekleri görüldüğü üzere, kent merkezlerinin dışına inşa edilen ve ihtiyaçlara yönelik donanıma sahip siteler yapılması ile kent merkezlerine bağlılığın azaldığı gözlenmiştir (Öztürk ve Fitöz 2009).

Hükümet Politikaları: Talep yönlü politikalarda tüketicilere teşvikler sağlanmaktadır. Örneğin konut senetleri, konut vergilerinin vergilerden düşülmesi, kredi faizlerinden düşülme vb. teşvikler ile konut talebi arttırılır ve bu da konut fiyatlarının artmasına ve arzın da artmasına neden olur. Arz yönlü politikalarda ise konut arzını sağlayanlar tarafına teşvikler sağlanarak, toplu konut uygulamaların, arsa sahibi olanlara teşvikler vb. şeklinde olabilmektedir (Öztürk ve Fitöz 2009).

3.1.4 Amerika Birleşik Devletleri örneği ve eşik-altı (subprime) krizi

Amerika Birleşik Devletleri'nde yaşanan konut kredisi kaynaklı krizler ile özellikle konut piyasasının finansal sistemi istikrarsızlaştırmada ne kadar kritik bir role sahip olduğunu göstermiştir. 1980'li yıllar öncesinde ABD'de konut kredileri genel olarak tasarruf ve kredi kurumları ya da ticari bankalar gibi mevduat toplayan kuruluşlarca verilmekte idi. Bu sistemde, diğer adıyla birincil mortgage piyasalarında, kredinin verilmesinde riskin yönetimine dek tüm sorumluluk krediyi veren kurum ile krediyi kullanan arasında paylaşılmaktadır. Özellikle de faiz oranı, kredi riski gibi birçok riskin büyük kısmının

bankacılık sektörü üzerinde olması nedeniyle, riskleri farklı kurumlar arasında dağıtmak ve likit bir bankacılık sektörü oluşması amacıyla ABD'de 1970'lerde ikincil bir mortgage piyasası çalışmaları başlamıştır. Bu şekilde kredi veren kuruluşlar, ellerindeki mortgage kredilerini piyasada yer alan yatırımcılara direkt olarak satabileceklerdir. İkincil piyasaların kuruluşu özellikle Fannie Mae ve Freddie Mac isimli iki kuruluşun kurulması ile başlamıştır (Erol 2009).

İkincil mortgage piyasalarında yer alan yatırımcılar, bankalardan mortgage kredileri satın almaktadır. Banka, sattığı krediyi bilançosu dışına çıkarır ve komisyon alarak bu kredinin aylık taksitlerinin ödenmesi gibi bazı hizmetleri krediyi satın alan kuruluşlara sağlar. Kredi taksitleri ödendikçe banka bu taksitlerden alacağı komisyonu alarak kalan taksidi yatırımcılara iletir. Kurumsal yatırımcılar ise almış oldukları kredileri yapısını ve vadesine göre farklı şekillerde sınıflandırarak havuzlar oluşturur ve her kredi taksidi ödenmesi ile elde edilen nakit parayı teminat göstererek menkul kıymet ihraç eder. Bu sisteme İpoteğe Dayalı Menkul Kıymet Piyasası (İDMK) denmektedir (Erol 2009).

Fannie Mae, Freddie Mac gibi özel kuruluşların ihraç ettikleri menkul kıymetlerde temerrüt ve erken ödeme riski gibi riskler mevcuttur. Müşterilerin güvenilirliği, belgelerinin eksiksiz olması, kredinin miktarı ve kredi oranı gibi kriterleri sağlamadığı halde kullanılan uyumsuz krediler teminat gösterilerek ihraç edilen krediler nedeniyle bu kuruluşlarda birçok risk portföylerinde toplanmıştır. ABD Merkez Bankası'nın 2001-2006 yıllarında gevşek bir para politikası izlemesi ve faiz oranlarının oldukça düşmesi ile mortgage faizleri düşmüş, kredi almak kolaylaşmıştır. Bu da mortgage kredilerini ödemede sorun yaşayanların bile kredilerini koşullar itibariyle yeniden finanse edebileceklerini düşündürmüştür. Ancak takip eden yıllarda faiz oranlarının artması ile kişilerin kredilerini ödemekte zorlanması ile birçok konutun icra yolu ile satışa çıkmasına sebep olmuştur. Birincil mortgage piyasalarında yaşanan temerrüt artışı da ikincil piyasalardaki yatırımcıların fon alamamasına ve kriz oluşmasına sebep olmuştur (Erol 2009).

Özetlemek gerekirse, düşük faiz oranları ve kredi verme standartlarındaki esneklik ile yüksek faizli ipoteklerde keskin bir artış yaşanmıştır. Sürekli büyüme gösteren yüksek

faizli ipotek kredileri nedeniyle bankaların portföylerinde kalite kötüleşirken, yükselen konut fiyatları ile birlikte temerrüde düşen sayısı nispeten düşük kalmış, bu durum da emlak piyasasını cazip hale getirmiştir. Düşük temerrüt oranları ve artan konut fiyatları, bankaları kredi vermeye teşvik ederek bir emlak balonu oluşmasına neden olmuştur. Akabinde ise bu emlak balonunun çökmesi ile bankalar baskı altında kalmış, tahsili geciken kredi sayıları da artmaya başlamıştır. Ev fiyatlarının bu süreçte düşmesi neticesinde, düşük profile sahip olmasına rağmen kredi alan kesimi borcunu ödememeye itmiştir. Tüm bu süreç ile hacizler artmış, konut talebi düşmüş ve kredi piyasasının koşulları kötüleşmiştir. Bu da reel ekonomiyi ciddi şekilde etkilemiş ve tüm kredi kategorilerinde yüksek temerrüt oranlarını beraberinde getirmiştir (Tajik vd. 2015).

3.1.5 Türkiye'de konut kredisi ile ilgili uygulamalar

Türkiye'de ipotekli konut finansmanı sisteminde yapılan en köklü değişiklik 21.02.2007 tarihinde 5582 sayılı "*Konut Finansmanı Sistemine İlişkin Çeşitli Kanunlarda Değişiklik Yapılması Hakkında Kanun*"un yürürlüğe girmesi ile gerçekleşmiştir. Bu kanun ile hem ABD'de hem de Avrupa'da kullanılan ipotekli konut finansman sistemlerinin ülkemizde uygulanabilmesinin önü açılmıştır. Bu kanunun esas amaçları arasında; inşaat sektöründe kalitenin yükseltilmesi ve konut finansmanının geliştirilmesi sayılabilmektedir (Atasoy ve Tanrıvermiş 2021).

Türkiye'de bankacılık uygulamalarına ilişkin düzenlemeler 2001 yılından bu yana Bankacılık Düzenleme ve Denetleme Kurumu (BDDK) tarafından yürütülmektedir. 2010 yılında düşük faiz oranları ve nispeten olumlu kredi koşulları ile konut kredisi kullanımında %30 üzerinde bir büyüme izlenmiştir. Bu ortamda hem ihtiyati borç verme uygulamalarını desteklemek için hem de konut kredilerinde yüksek büyüme oranlarına engel olmak amacıyla Kredi Değer Oranı %75 olarak sınırlandırılmıştır. Kredi değer oranlarını dinamik bir politika aracı şeklinde kullanan diğer ülkelerde olduğu gibi, Türkiye'de de tüketici kredilerinde yaşanan yavaşlama döneminin ardından 27 Eylül 2016'dan geçerli olmak üzere KDO %80'e çıkarılmıştır (Baziki ve Çapacıoğlu 2021). Günümüzde ise Türkiye'de kredi kullanılacak gayrimenkulün değerine, gayrimenkulün ilk veya ikinci el satış olması durumuna ve bazı özelliklere dayalı olarak dönemsel bazda

kredi kullanım oranları değişiklik gösterebilmektedir. 2020 yılında, özellikle tüm dünyayı etkisi altına almış olan koronavirüs pandemisi devam etmekte iken, Türkiye'de kamu bankalarınca kredi paketleri açıklanmış, bu dönemde uygulanan faiz indirimleri ile konut, taşıt ve ihtiyaç kredisi gibi birçok alanda kredi kullanım oranları artmıştır. 2020 yılında açıklanan kampanya çerçevesinde ilk kez satışa konu olacak (sıfır) konutlar için 15 yıl vade imkanı sunulmuş, ilk yıl geri ödemesiz olmak üzere aşağıdaki örnek planda yer aldığı şekilde kredi şartları oluşturulmuştur.

Çizelge 0.2 Konutta ilk satış için örnek ödeme planı

Kredi Tutarı	Vade	Faiz Oranı	Aylık Ödeme	Toplam Ödeme
200.000 TL	14 Yıl	%0,64	1.946,47 TL	337.417,41 TL

İkinci el konutlar için ise yine 15 yıl vadeli, ilk yıl geri ödemesiz kredi imkanı sunulmuş olup, burada ise faiz oranı aylık %0,74 olarak belirlenmiş ve Çizelge 3.2'deki şekilde bir örnek ödeme planı oluşmuştur.

Çizelge 0.3 Konutta ikinci el satış için örnek ödeme planı

Kredi Tutarı	Vade	Faiz Oranı	Aylık Ödeme	Toplam Ödeme
200.000 TL	14 Yıl	%0,74	2.083,87 TL	350.090,21 TL

Özellikle pandemi döneminde azalan konut satışlarında, 2020 yılında gerçekleştirilen konut faizi indirimlerini ve vade değişikliklerini içeren kredi paketleri ile artış sağlandığı görülmekte olup, bir sonraki bölümde yer alan sayısal veriler incelendiğinde yaşanan artış gözlenebilmektedir.

Yine 2023 yılı başında açıklanan "Yeni Evim" kampanyası ile orta gelir düzeyindeki vatandaşların ev alımına yönelik çalışmalar gerçekleştirilmiş olup, bu kampanyanın şartları itibariyle 2020 yılındaki kampanyadan farklılaştığı görülmektedir. Bu kampanya yalnızca ilk kez satışa konu olacak konutları kapsamış ve konutu olmayan vatandaşlara yönelik olarak açıklanmıştır. Kredi faizinin %0,69'dan başladığı, ilk üç yılda devlet katkısı ile taksitlerin ödenebildiği bu kampanyada; Türkiye'nin üç bölgeye ayrıldığı

görülmektedir. İlk bölge İstanbul'dan oluşmakta olup, burada azami gelir 80.000 TL iken, ikinci bölgede ise Ankara, Bursa, İzmir, Antalya, Mersin ve Muğla yer almakta olup azami gelir 65.000 TL olarak belirlenmiştir. Diğer tüm iller ise üçüncü bölgede yer almakta olup, azami gelir ise 45.000 TL olmaktadır. İlk iki bölgede ev alacaklar için son bir yıldır o bölgede ikamet etmiş olma şartı aranırken, yine tüm bölgelerde son bir yıl içerisinde konut alacağı ilde konut satışı yapmamış olma şartı da aranmaktadır. Bu kampanya ile elde edilen konutlarda beş yıl satış yasağı şartı bulunmaktadır.

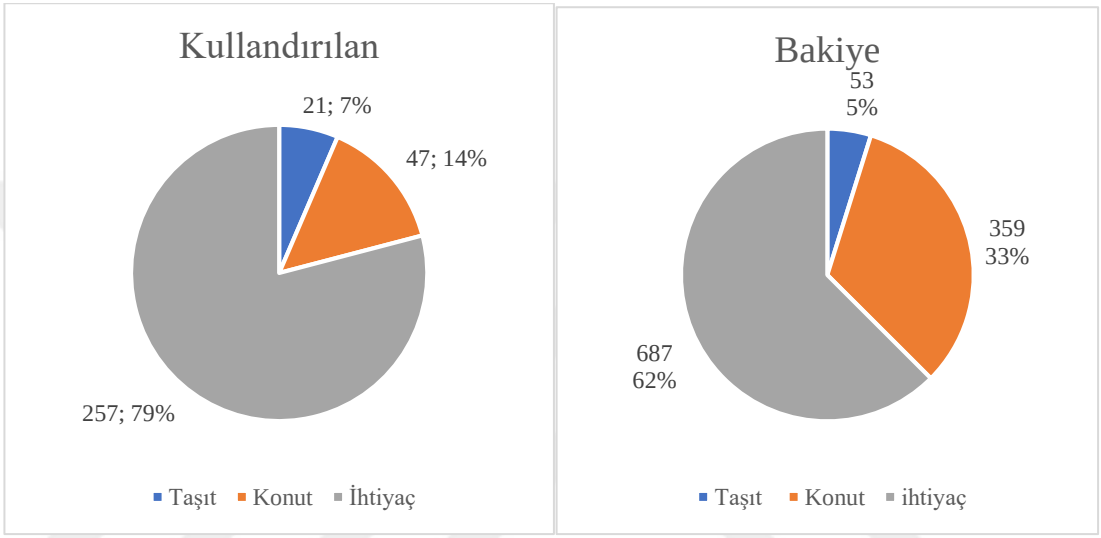
Son yıllarda konut edinimini kolaylaştırmak için devlet desteği ile gerçekleştirilen kampanyalar incelendiğinde, şartları itibariyle 2023 yılında gerçekleştirilen kampanyanın 2020 yılındaki kampanyaya göre daha sınırlı bir kitleye hitap ettiği ve etkisinin de bu çerçevede sınırlı kaldığı görülmektedir.

Ülkemizde konut kredilerini de içeren perakende krediler olarak adlandırılan konut ve tüketici kredilerinde kredi notlaması; müşterinin ödeme iştahının ve kapasitesini etkileyen faktörlerin matematiksel şekilde birleştirilmesi ve bir kredi notuna dönüştürülmesi şeklinde gerçekleştirilmektedir. Bu sistemde müşterinin kredi ödemesini etkileyeceği değerlendirilen özellikleri (yaş, gelir, cinsiyet, meslek, önceki borç ödeme alışkanlıkları gibi) ağırlıklandırılır ve bir nota ulaşılır. Bu sınır notunu oluşturan özellikler ve bu özelliklere verilen ağırlıklar, daha önceden müşterinin ödeme alışkanlıklarına ait verileri ve müşteri özelliklerine bakılarak diskriminant analizi ya da logit gibi bazı istatistiksel yöntemler ile hesaplanır. Sistem tarafından kabul edilen sınırdan altta kalan notlar için müşterinin kredisi reddedilirken, üzerinde ise talebi olumlu sonuçlandırılır. Bu sistem içerisinde kabul sınırını aşmış ancak düşük notlu müşterilere daha az kredi limiti tahsis edilirken daha yüksek notlu müşterilere daha yüksek kredi limiti tahsis edilebilmektedir (SPL Lisanslama Sınavları Çalışma Notları, 2023).

3.1.6 Türkiye'de kredi kullanımı ile ilgili bazı istatistikler

Bu kesimde, Türkiye Bankalar Birliği (TBB) tarafından Mart 2023'de açıklanmış ve "Tüketici Kredileri ve Konut Kredileri" başlıklı raporda üretilen istatistikler paylaşılmıştır. Bu çerçevede, Mart 2023 itibariyle, tüketici ve konut kredisi kullanan

toplam kişi sayısı 27 milyon 249 bin kişi olarak belirtilmiş, kullanılan toplam kredi miktarı ise 1 trilyon 99 milyar TL olarak tespit edilmiştir. Tüketici ve konut kredisi kullananların sayısı Ocak-Mart 2023 döneminde, bir önceki yılın aynı dönemine göre %58 artış gerçekleştiği görülmüştür. Yine aynı dönemde kredi türlerine göre dağılım ise Şekil 3.1'de sunulmuştur. Ocak-Mart 2023 döneminde %79 pay ile en büyük kredi payına ihtiyaç kredilerinin sahip olduğu görülmektedir (TBB, 2023).



Şekil 0.1 Mal ve hizmet gruplarına göre dağılım

Ocak-Mart 2023 itibariyle konut ve tüketici kredilerinden kanuni takibe alınan kredi miktarının yaklaşık 2 milyar 225 milyar olduğu ve bunun da aynı dönemde kullanılan konut ve tüketici kredilerinin %2,2'ine tekabül ettiği, takipteki kredilerin %2'sinin konut kredisi, %98'inin ise ihtiyaç kredisi olduğu raporda yer almaktadır. Bu çerçevede kullandırılan tüketici ve konut kredi miktarları ile kanuni takibe düşen tüketici ve konut kredi miktarları (Mart 2023-Mart 2023 dönemine ait) Çizelge 3.3 ve Çizelge 3.4'te gösterilmiştir (TBB, 2023).

2013-2020 yılı arasında gayrimenkul piyasalarına ait TÜİK verileri incelendiğinde 2017 yılına dek satışlarda bir artış olduğu, 2018 ve 2019 yıllarında ise geçmiş yıllara göre azalışlar olduğu görülmektedir. İpotekli satışların ise 2013-2017 arasında toplam satışlar içinde %33-40'lık bir paya sahip olduğu, 2019 yılında bu oranın %25'lere indiği ancak özellikle 2020 yılında konut kredisi faizlerinde yapılan indirim sonrası payın %41

seviyesine çıktığı görülmüştür. Yine veriler incelendiğinde toplam satış içerisinde ikinci el konutların payındaki artışın da ilk el konut fiyatlarının yüksek olması ve bundan dolayı alıcıların konuta erişimin sınırlı olması şeklinde değerlendirilebileceği düşünülmektedir. İlgili veriler Çizelge 3.5'te sunulmuştur (Atasoy ve Tanrıvermiş 2021).

Çizelge 0.4 Tüketici ve konut kredileri (Milyon TL)

		Kullandırılan			
Dönem		**Miktar**	**Kişi Sayısı (Adet)**	**İdari Takip**	**Kanuni Takip**
2022 Mart	**TP**	120.720	4.489.861	113	2.335
	YP	4	5	0	0
	Toplam	**120.724**	**4.489.866**	**113**	**2.335**
2022 Haziran	**TP**	215.264	6.115.668	36	1.875
	YP	8	12	0	0
	Toplam	**215.272**	**6.115.680**	**36**	**1.876**
2022 Eylül	**TP**	171.996	5.732.993	62	2.231
	YP	7	10	0	0
	Toplam	**172.003**	**5.733.003**	**62**	**2.231**
2022 Aralık	**TP**	258.294	7.084.859	65	2.338
	YP	4	12	0	0
	Toplam	**258.298**	**7.084.871**	**65**	**2.338**
2023 Mart	**TP**	323.929	7.115.630	165	2.225
	YP	11	12	0	0
	Toplam	**323.939**	**7.115.642**	**165**	**2.225**

Çizelge 0.5 Tüketici kredileri ve konut kredileri bakiye (Milyon TL)

Bakiye				
Dönem	**Miktar**	**Kişi Sayısı (Adet)**	**İdari Takip**	**Kanuni Takip**
2022 Mart	734.335	26.301.405	3.131	17.630
	258	353	0	22
	734.593	**26.301.758**	**3.131**	**17.651**
2022 Haziran	831.405	27.063.886	756	20.064
	265	332	0	23
	831.760	**27.064.218**	**756**	**20.087**
2022 Eylül	872.632	27.056.224	1.163	21.094
	256	313	0	25
	872.888	**27.056.537**	**1.163**	**21.119**
2022 Aralık	972.362	28.105.909	1099	20.882
	276	265	0	28
	972.638	**28.106.174**	**1099**	**20.910**
2023 Mart	1.098.621	27.248.622	1.183	23.483
	275	245	0	32
	1.098.897	**27.248.867**	**1.183**	**23.515**

Çizelge 0.6 2013-2020 döneminde Türkiye'de konut satışları

Yıllar	Toplam Satış	İpotekli Satış	Diğer Satış	İpotekli Satış (%)	Diğer Satış (%)	İlk Satış	İkinci El Satış	İlk Satış (%)	İkinci El Satış (%)
2013	1.157.190	460.112	697.078	39,76	60,24	529.129	628.061	45,73	54,27
2014	1.165.381	389.689	775.692	33,44	66,56	541.554	623.827	46,47	53,53
2015	1.289.320	434.388	854.932	33,69	66,31	598.667	690.653	46,43	53,57
2016	1.341.453	449.508	891.945	33,51	66,49	631.686	709.767	47,09	52,91
2017	1.409.314	473.099	936.215	33,57	66,43	659.698	749.616	46,81	53,19
2018	1.475.398	276.820	1.098.578	20,13	79,87	651.572	723.826	47,37	52,63
2019	1.348.729	332.508	1.106.221	24,65	75,35	511.682	837.047	37,94	62,06
2020	1.409.316	573.337	835.979	40,68	59,32	469.740	1.029.576	33,33	66,67

3.1.7 Türkiye Bankalar Birliği Risk Merkezi ve risk merkezi raporu

Finansal sektör için kredi risklerinin etkin yönetimi büyük bir öneme sahiptir. Etkin kredi risk yönetimi için yüksek sayıda gözleme dayanan; kapsamlı, doğru ve güncel bir bilgi setine ihtiyaç vardır. Banka ve diğer finansal kuruluşlardan risk yönetimine ilişkin toplanan bilgiler; kredi ve kredi kartı gibi hem finansal ürün ve hizmet taleplerinin değerlendirilmesinde hem de ekonominin gidişatına dair istatistiki raporların hazırlanmasında kullanılmaktadır. Türkiye'de risk yönetiminde kullanılan bu bilgilerin toplanması ve paylaşılmasını temin etmek maksadıyla 2013 yılında Türkiye Bankalar Birliği (TBB) Risk Merkezi hayata geçirilmiştir. Risk merkezi; banka, finansal kiralama, faktöring, varlık yönetimi şirketi ve çeşitli finansal kuruluşların bulunduğu 180 üyesinden müşterilerine ait kredi çek senet ve bunların ödenmelerine ilişkin bilgileri toplar ve bunları paylaşır. Toplanan bilgiler ile kişi ve şirketlere kredi borcu, çek ve senet bilgilerini gösteren bir risk merkezi raporu sunulur. Risk Merkezi Raporu anlaşmalı bankalarından veya ücretsiz olarak E-Devlet Kapısı üzerinden temin edilebilmektedir (TBB Risk Merkezi, 2023).

Rapor; gerçek kişilerin ve gerçek kişilerin ticari işletmelerinin finansal durumlarını takip edebilmeleri amacıyla oluşturulmuştur. Rapor üzerinden gerçek kişinin kredi limit ve borç, bireysel kredi başvuru, telefon ve internet gecikmeli fatura bilgileri, ihale yasaklılığı, senet ve çek bilgileriyle çek yasaklama ve kaldırı kararlarını içeren bilgilere

erişim sağlanabilmektedir. Raporda ilk sırada kredilere ilişkin özet bilgilerin bulunduğu "Kredi Limit ve Borç Tablosu" bilgileri mevcuttur. Varlık yönetim şirketlerine ait bilgiler ayrı gösterilmek üzere burada risk merkezi üyesi finansal kuruluşlardan kullanılmış kredilerin borç miktarı, kredi limiti ve kredi adedi gibi bilgiler bulunmaktadır. "Kredilerim Detay" kısmında ise raporda belirtilen tarihten geçerli olmak üzere açık bulunan kredilerin borç bilgileri, varsa yasal takip durumundaki bireysel ve ticari kredi borcuyla varlık şirketine ait borç bilgileri mevcuttur. "Bireysel Kredi Başvuru Bilgileri" kısmında ise rapor tarihi itibariyle başvurusu bulunan bir kredi varsa tarihi, hangi finansal kuruluşa ait olduğu, kredi türü ve miktarı gibi bilgiler bulunmaktadır. Rapor tarihine en yakın son 10 başvurunun görüntülendiği bu bölümü, "Telefon ve İnternet Gecikmeli Fatura Bilgileri" kısmı takip etmektedir. Bu kısımda ise risk merkezine bildirimi yapılan 1-30 gün, 21-60 gün, 61-90 gün, 91 gün üzeri geciken faturalara ait bilgiler yer almaktadır. "İhale Yasaklısı Bilgileri" bölümünde ise resmi gazetede yayımlanmış ve devam eden yasaklılık durumları yer almakta olup, "Senet Bilgileri" kısmında ise rapor tarihinden bir iş günü öncesini kapsayan ve bankaya teminata ya da tahsilata verilmiş, ileri vadeli ve vadesi gelmemiş, banka üzerinden ödenen, protesto edilen ya da protestosu kaldırılan senetlere ait bilgiler bulunmaktadır. Yine "Çek Bilgileri" bölümünde de karşılıksız çıkan ve ödenmemiş ya da karşılıksız çıkmış ve ödenmiş, çek bilgileri, çek hesabı bulunan banka bilgileri, son beş yıla ait benzer çek bilgileri mevcuttur. "Çek Yasaklama ve Kaldırı Karar Bilgileri" kısmında ise Ulusal Yargı Ağı Bilişim Sistemi UYAP tarafından Risk Merkezi'ne bilgisi verilen ve Risk Merkezi'nce bankalara bildirilen çek yasaklama ve kaldırı bilgileri mevcuttur (TBB Risk Merkezi, 2023).

3.2 Kredi Risk Puanı (Findeks Kredi Notu)

Ülkemizde kişilerin önündeki bir yıllık süreçte borçlarını ödeme olasılıklarını tahmin etmek amacıyla "Findeks Notu" adı verilen bir kredi risk puanı hesaplanmaktadır. Bu not finansal ilişkilerde kişiler hakkında fikir edinmek amacıyla kullanılmaktadır. Bu bakımdan yalnızca kredi kullanımında değil, ev kiralamada ya da mal ediniminde de gösterge amacıyla kullanımdadır (Findeks.com, 2023)

Findeks notunun hesaplanmasında şu bilgiler kullanılmaktadır.

- Kredili Ürün Ödeme Alışkanlıkları (Notun %45'i): Kredili ürün ödemelerin zamanlamasına göre nota etkisi bulunmaktadır. Son ödeme tarihinden sonra ödenen borçlar notu olumsuz etkilerken, zamanında yapılan ödemeler ise notu yükseltmektedir.
- Mevcut Hesap ve Borç Durumu (Notun %32'si): Bu kısımda kapanan kredilerin ne şekilde kapandığı (iyi ya da kötü), teminatlı ya da teminatsız mevcut borç bakiyeleri ve limitler not hesabına katılmaktadır.
- Yeni Kredili Ürün Açılışları (Notun %5'i): Burada ise henüz ödeme düzeni belli olmamış da olsa yeni açılan krediler bir risk artışı unsuru olarak notlamaya katılmaktadır.
- Kredi Kullanım Yoğunluğu (Notun %18'i): Hiç kredi kullanmamış ya da az kredi geçmişi olan bir kişiye kıyasla kredi kullanan ve düzenli ödemelerini gerçekleştiren kişilerin notlarının daha yüksek olması mümkündür (Findeks.com, 2023).

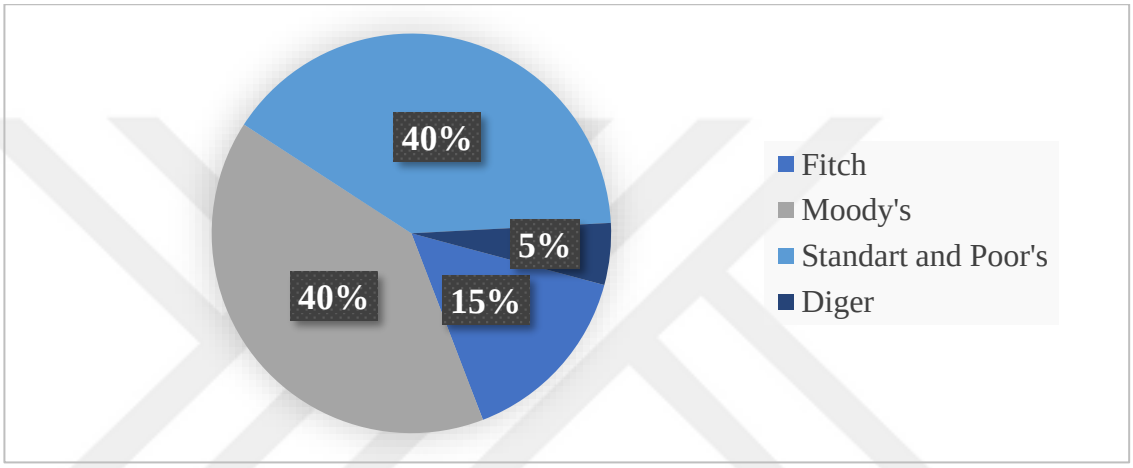
Ülkemizde özellikle konut kredileri çerçevesinde değerlendirildiğinde; konutun yatırım amaçlı görülmesi, geleceğin bir garantisi olarak değerlendirilmesi ve özellikle gelecek nesillere miras olarak bırakılmasının düşünülmesi gibi nitelikleri nedeniyle, konut kredilerin genellikle ödendiği ve kanuni takip oranının oldukça düşük olduğu görülmektedir. Öte yandan, konutun genellikle değerinin artış eğilimde olması ve özellikle enflasyonist ortamda uzun vadeye sahip konut kredilerinin kredi ödemelerinin sabit olması nedeniyle zaman içerisinde değerinin azalması da kredinin ödenmesini destekleyen önemli bir unsurdur.

3.3 Kredi Derecelendirme Sistemleri

Borçlu kişi ya da kuruluşun kredi riskinin kredi derecelendirme kuruluşları tarafından değerlendirilmesi işlemi kredi derecelendirme olarak ifade edilmektedir (Kılıçarslan ve Giten, 2016). Kredi derecelendirme bir sınıflandırma sistemi olup, firmaların veya bireylerin sayısal ya da sözel bilgilerine dayanmaktadır. Kredilerin vaktinde ve bütün olarak ödenmesi konusunda uluslararası sermaye piyasalarının kriterlerine uygun ve tarafsız bir ölçü sağlamaya çalışmaktadır (Kılıçarslan ve Giten, 2016).

Özellikle 1990'lar sonrasında kredi derecelendirme kuruluşlarının sayısı artmaya başlamıştır. Çok sayıda kredi kuruluşu faaliyette olmasına karşın global çapta kabul görmüş ve çalışmaları ağırlıklı olarak takip edilmiş olan kuruluşların sayısı daha azdır. Bunlardan başlıcaları ise Moody's, Standart and Poor's ve Fitch'tir (Kılıçarslan ve Giten, 2016).

Bu üç büyük kuruluşun piyasadaki hakimiyeti ise Şekil 3.2'de görülebilmektedir.



Şekil 0.2 Üç Büyük kredi derecelendirme kuruluşunun piyasa hakimiyeti

Kredi değerlendirme aşamasında; firmanın sektördeki durumu, sektörünün yapısı, makroekonomik olayların firmaya potansiyel etkisi, firmanın finans durumunun incelenmesi, firma kazanç ve varlıklarının borçlarını karşılayıp karşılayamayacağı gibi konular incelenir. İçsel risk değerlendirme sistemleri (IRRS) kapsamında, bankaların borç alan tarafın temerrüde düşme ihtimalini belirleyen bir borç temerrüt notu (Obligator Default Rating- ODR) ve temerrüt olur ise firmanın kayıp riskini derecelendiren "Temerrüt Durumu Kayıp Notu" (Loss Given Default Rating- LGDR) şeklinde iki şekilde derecelendirme yapılmaktadır. Bu süreç adımlar halinde şu şekilde özetlenebilmektedir (İldaş 2021).

i. Finansal İnceleme: Bu adımda kurumun finansal raporları incelenmektedir. Amaç; kazanç ve nakit akışların borcun geri ödenmesini karşılayıp karşılayamayacağını belirlemektir.

ii. Yönetim ve Diğer Nitel Faktörler: Bu kısımda ise günlük hesap işlemleri ve yönetim incelenmektedir. Şartlı yükümlülükler değerlendirilir ve kapsamlı bir çevresel değerlendirme yapılmaktadır.

iii. Sektör/Kuruluş Değerlendirmesi: Burada firmanın sektör içerisindeki konumu ve sektörel riskin analizi adımları izlenmektedir. Sektörel pozisyonu zayıf olması halinde firmanın notu düşürülebilmekte ancak güçlü olduğu tespit edildiğinde notu yükseltilmemektedir.

iv. Finansal Tablo Kalitesi: Burada analizi yapacak görevliye sağlanan finansal bilgilerin kalitesi ve doğruluğu test edilmektedir.

v. Ülke Riski: Ülke riskinin kredi notuna etkisi incelenmektedir. Bu adımlar sonucunda bir firma için Borç Temerrüt Notu (ODR) oluşturulmaktadır.

vi. Dışsal Kredi Notlarının Kararlaştırılması: Bir dış derecelendirilme kuruluşu tarafından firma değerlendirilmeye tabi ise, bu not ile oluşturulan ODR notu karşılaştırılmaktadır. Büyük ölçüde farklılık var ise bu noktada ODR'nin oluşturulması süreci yeniden değerlendirilmektedir. Bu adım oldukça önemlidir.

vii. Kredi Yapısı: Bu adımlar dışında eğer kredinin yapısının borçlunun temerrüde düşmesine bir etkisi olacağı düşünülüyorsa kredi notunda düşüş gerçekleştirilebilmektedir.

viii. Temerrüt Durumunda Kayıp Notu: Temerrüt olması ihtimali temerrüt halinde yaşanan kayıplar birbirinden bağımsız olarak incelenmektedir.

Bu adımlar takip edilerek her bir kredi borçlusu için tutarlı bir kredi notu belirlenmektedir (İldaş 2021).

4. SINIFLANDIRMA PROBLEMLERİNDE MAKİNE ÖĞRENMESİ ALGORİTMALARI

Makine öğrenmesi yöntemlerinin gelişmesi ve özellikle büyük hacimli müşteri verilerinin analize olan ihtiyacı neticesinde kredi notlandırılması sistemlerinin geliştirilmesi öncelikli hale gelmiştir. Bu modeller kabaca denetimli makine öğrenmesi ve denetimsiz makine öğrenmesi yöntemleri olarak ikiye ayrılabilmektedir. Bu iki yaklaşım arasındaki temel fark, makine öğrenmesi modellerine verilen girdilerin etiketlenmiş olup olmamasıdır. Denetimli makine öğrenmesi modelleri etiketli verilere uygulanmakta olup, geniş bir algoritma ağına sahiptir. Bunlardan bazıları Destek Vektör Makineleri (DVM), Karar Ağaçları (KA), Rasgele Orman (RO) ve Yapay Sinir Ağları (YSA) olarak sayılabilmektedir. Denetimsiz makine öğrenmesi yöntemlerinde ise girdiler etiketsiz olup, bu problemlerin çözümlenmesinde kullanılan yöntemlerden bazıları ise k-ortalama (k-means), Hiyerarşik Kümeleme (Hierarchical Clustering) olarak ele alınmaktadır (Bao vd. 2019).

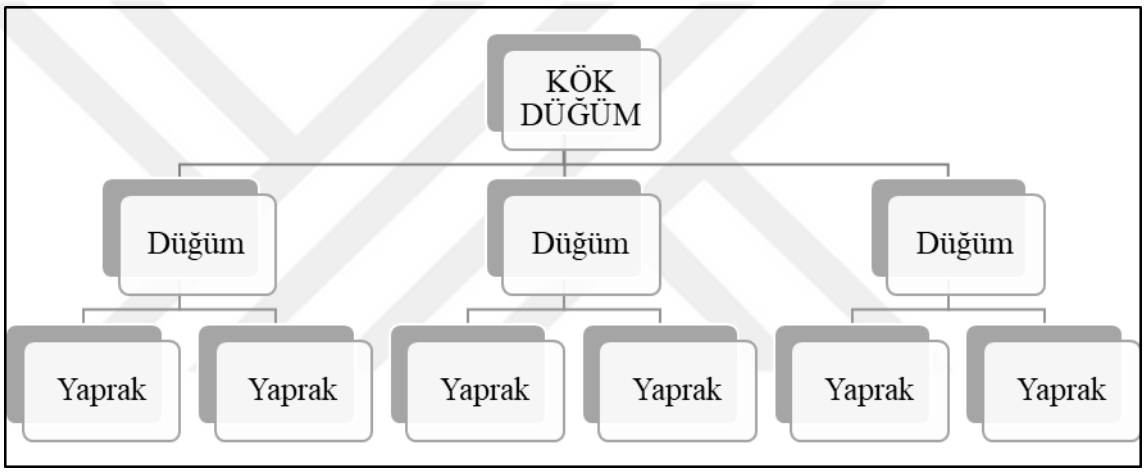
Denetimli ve denetimsiz makine öğrenmesi modellerinin her ikisi de kredi riski değerlendirmesinde yaygın şekilde kullanılmaktadır. Denetimli makine öğrenmesi algoritmaları, kredi notlandırma modellerinde müşterinin özellikleri ile kredi riski arasındaki ilişkiyi bulmak ve tahmin edilen sınıflandırmayı genellikle ikili formda sunmak için kullanılmaktadır (Bao vd. 2019). Literatürde genişçe yer aldığı üzere, denetimli öğrenme algoritmalarının uygulamaları, kredi notlandırmada oldukça iyi tahmin doğruluğuna sahiptir (David ve Frank, 2008). Denetimsiz makine öğrenmesi algoritmaları ise genel olarak kümeleme algoritmalarından oluşmakta olup, direkt bir tahmin sunmaktan çok benzer öğelerin bir kümeye toplanması amacıyla uygulanan bir veri madenciliği tekniği olarak kullanılmaktadır. Bu açıdan bakıldığında denetimsiz makine öğrenmesi algoritmaları sıklıkla denetimli makine öğrenmesi algoritmalarının tamamlayıcısı olarak kullanılmaktadır (Bao vd. 2019).

Çalışma kapsamında kredi riski modellemesinde en sık kullanıldığı tespit edilen ve bu çalışmada da kullanılacak olan bazı sınıflandırma algoritmalarına değinilmiştir.

4.1 Karar Ağaçları

Karar ağacı sınıflandırıcısı, çok aşamalı karar vermeye yönelik olası yaklaşımlardan biridir (Safavian ve Landgrebe 1991). Karar ağaçları bir denetimli öğrenme algoritmasıdır. İki aşamalı olarak gerçekleşmektedir. İlk aşamada eğitim verileri ile öğrenme sağlanarak bir model oluşturulur. İkinci aşamada ise test verileri ile modelin veriyi sınıflandırması gerçekleştirilir (Bilgin 2018).

Karar ağacı yapısı şekil 4.1'de gösterildiği biçimdedir.



Şekil 0.1 Karar Ağacı Yapısı

Bir karar ağacı; kök düğüm yapısı ile başlar ve ara düğümler yardımı ile dallanacağı yön belirlenir. Burada her bir sınıf bir yaprak tarafından gösterilmekte olup, her yaprağa giden bir yol olmalıdır. Ayrıca yapraklar arasında bağlantı da bulunmamaktadır. Kök düğümde yer alacak özellikler farklı birçok kriter tarafından belirlenebilmektedir. Bunlardan bazıları şöyledir (Bilgin 2018).

a. DKM kriteri: İki sınıfa haiz nitelikleri belirlemek için tasarlanmıştır. $M_y = d_{i,j}$ seçilen bir özelliğin $y = d_i$ durumunda olması halinde M veri kümelerinin alt kümelerini gösteriyor olsun. Buradaki i, seçilmiş olan özelliğin durum sayısını ve j sınıflara ayrılmak istenen özelliğin sınıf sayısını ifade etmektedir. Bu halde P_1 seçilen özellikte d_i durumu için ilk sınıfın gerçekleşme olasılığını ve P_2 ikinci sınıfın gerçekleşme

olasılığını göstermektedir. Bu şekilde DKM kriteri Eşitlik 4.1'deki gibi hesaplanır (Bilgin 2018).

$$DKM(M_{d_i}) = 2\sqrt{P_1 P_2} \tag{4.1}$$

b. Gini indeksi (GI): Veri kümesinde yanlış sınıflandırma olasılığı Gini indeksi olarak tanımlanmaktadır. Bu nedenle de Gini indeksi en küçük olan özellik seçilerek işlemlere devam edilir. $M_y = d_{i,j}$ seçilen bir özelliğin $y = d_i$ durumunda olması halinde M veri kümelerinin alt kümelerini gösteriyor olsun. Buradaki i, seçilmiş olan özelliğin durum sayısını ve j sınıflara ayrılmak istenen özelliğin sınıf sayısını ifade etmektedir. P_i, seçilen özellikteki d_i durumu için j sınıfının gerçekleşme ihtimalini göstermekte olsun. Bu şartlarda y özelliğinin d_i durumu için Gini indeksi hesaplanır ve en küçük Gini indeksine sahip özellik kök düğüme bağlanır (Bilgin 2018). Sınıflandırma ve regresyon problemleri karar noktaları oluşturmak için sınıflandırma görevlerinde Gini indeksini kullanmaktadır (Daniya vd. 2020). Gini indeksi Eşitlik 4.2'deki gibi hesaplanmaktadır.

$$GI(M_{d_i}) = 1 - \sum_j^k p_j^2 \tag{4.2}$$

4.1.1 Karar ağaçları algoritmaları

Bu kesimde, karar ağaçlarında en çok tercih edilen ve çalışma kapsamında kullanılan algoritmalara yer verilmiştir.

a. Chaid (chi-square automatic interaction) algoritması: Ağaç tabanlı segmentasyon tekniğinin bir k kategori (nominal veya sıralı) kriter değişkenini öngören anlamlı segmentler elde etmek için kullanılan etkili bir yaklaşımdır (Magidson ve Vermunt, 2004). Özellikleri ki-kare testi kullanılarak bölen algoritmada, kategorik ve sürekli veriler için çalışmak mümkündür. Bu algoritma çoklu ağaçlar üretmektedir ve bağımlı değişken için tüm bağımsız değişkenler ile çapraz tablolar oluşturulmaktadır. Bu tablodaki kare değerleri hesaplanır. Ki kare değeri en büyük olan özellik kök düğüme

yerleştirilir. B_{ij} beklenen değer, $T_i, i.$ satır değeri, $T_j; j.$ sütun değeri, $n;$ genel toplam, $G_{ij}; i.$ satır $j.$ sütundaki gözlem ve $B_{ij}; i.$ satır $j.$ sütundaki beklenen değer olmak üzere, kullanılan ki-kare değeri hesabı Eşitlik (4.3) ve (4.4)'te yer almaktadır (Bilgin 2018).

$$B_{ij} = \frac{T_i T_j}{n} \tag{4.3}$$

$$\chi^2 = \sum_i^r \sum_j^c \frac{(G_{ij} B_{ij})^2}{B_{ij}} \tag{4.4}$$

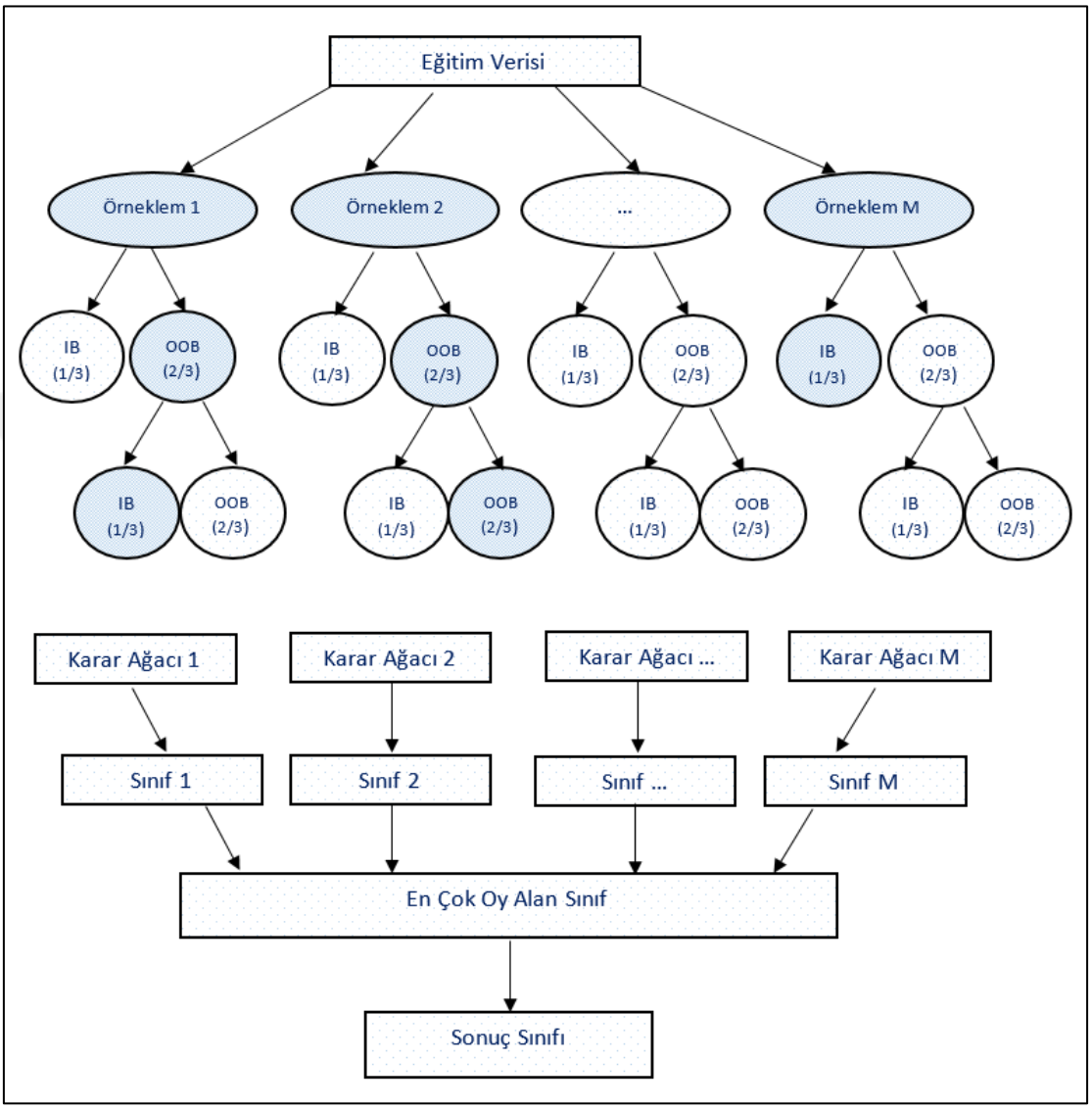
b. CART (classification and regression trees) algoritması: CART algoritmasında iki yaklaşım dikkate alınmaktadır. İlk olarak veriler uzamsal modellerine göre ağırlıklandırılarak örneklemin düzensizliği hesaba katılır. İkinci durumda ise algoritmanın her adımında diskriminant kuralının oluşturulmasında yer alan niceliklerin uzamsal tahminleri kullanılır (Bel vd 2009). Bu algoritma dallanma kriteri olarak Gini indeksinden faydalanmaktadır. Bölme işlemi sırasında eğer veriler kategorik ise Gini indeksi kullanılırken, sürekli ise en küçük kareler kullanılmaktadır.

4.2 Rasgele Orman

Rasgele orman sınıflandırıcısı, her sınıflandırıcının girdi vektöründen bağımsız olarak örneklenen rasgele bir vektör kullanılarak üretildiği ve her ağacın bir girdi vektörünü sınıflandırmak amacıyla en popüler sınıf için bir birim oy kullandığı ağaç sınıflandırıcılarının bir kombinasyonundan oluşmaktadır (Pal 2005). Bu algoritmada, bütün değişkenler içinden en iyi dal seçilerek her bir düğümü dallara ayırmak yerine, her bir düğümden rasgele olarak seçilmiş olan değişkenler içinden en iyi olan kullanılarak her bir düğümü dallara ayrılmaktadır (Bilgin 2018).

Veri setinden oluşturulacak eğitim veri setinden ağaçları oluşturmak için belirlenmiş olan ağaç sayısı, M ve her bir düğümde kullanılması planlanan özellik sayısı ise n ile gösterilsin. Eğitim veri setinin 2/3'ü InBag (IB) ismi verilen ön yükleme verisi olarak ayrılır. Eğitim kümesinden kalan 1/3'lük kısım ise OutofBag (OOB) ismi verilen ve test için kullanılacak test verisi olarak ayrılır. Her bir ön yükleme verisi ile bir ağaç modeli

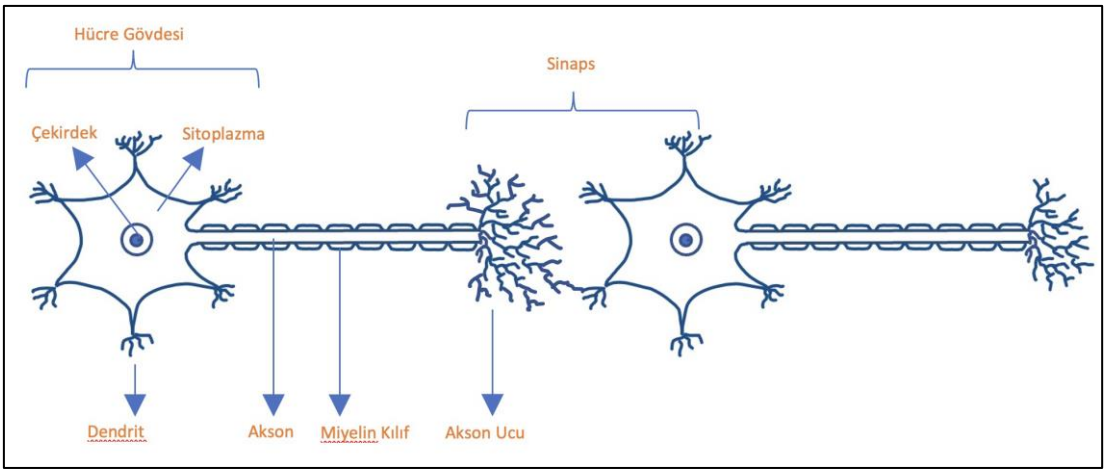
meydana getirilirken her bir düğüm noktasında n adet özellik tüm özellikler içinden rasgele olarak seçilir (Bilgin 2018). Şekil 4.2'de bir rasgele orman yapısı görülmektedir.



Şekil 4.2 Rasgele orman algoritması

4.3 Yapay Sinir Ağları

Yapay sinir ağları algoritmasında insanın öğrenme süreci taklit edilmektedir. İnsan beynindeki nöron yapısının modellenmesi ile oluşturulmuş bir algoritmadır.



Şekil 4.3 İnsan Sinir Hücresi Yapısı

Şekil 4.3'te insan sinir hücresi yapısı görülmektedir. Dendritler, sinaptik sinyalleri girdi şeklinde almaktadır. Hücre gövdesinde bu sinyaller işlenerek aksonlara iletilmekte ve aksonlardan hedefteki hücrelere iletilmektedir. Sinapslar ise akson uçlarıdır. Şekil 4.4 ise yapay sinir ağlarının yapısı gösterilmiştir (Bilgin 2018).

Şekil 4.4 Yapay Sinir Ağı Yapısı

Bir yapay sinir ağı temel olarak üç katmandan oluşmaktadır. Girdi katmanında, gelen girdiler sinir ağına alınmaktadır. Buradaki girdiler, model bağımsız değişkenlere karşılık gelmektedir. Çıktı ise bağımlı değişkeni temsil etmektedir. Girdi ve çıktı arasında yer alan katmanlar ise gizli katman olarak isimlendirilmektedir (Bishop 1994).

Ağırlıklar, hücreye giren verilerin etkisini göstermek üzere belirlenmekte olup, girdiler modelin eğitilmesi için kullanılacak örneklemdir (Sarıtaş ve Yaşar, 2019). Bu yöntemde başlangıç ağırlıklarının belirlenmesi, çıktının tahmin edilme durumu ve sürecini etkilemektedir. Model çalışırken, ağırlıklar çıktı ve tahmin arasındaki fark minimum olana dek yeniden düzenlenmektedir. Bunun için de farkların karelerinin toplamı şeklinde tanımlanan bir hata fonksiyonu tanımlanmaktadır. Bu hata fonksiyonunu minimuma indirmek için, "Gradient Descent Algorithm (Gradyan İnişi)" adı verilen bir algoritma kullanılmaktadır. Gradyan inişi, bir optimizasyon algoritması olup, türevi alınabilen bir fonksiyonun yerel minimumunu bulmak için birinci dereceden iterasyon içerir. Literatürde yapılan incelemelere göre bu ağırlıkların belirlenmesinde farklı yöntemler bulunmaktadır. İzlenen yollardan biri, ağırlıkların rasgele belirlenmesidir. Bu durum çıktının doğru şekilde tahmin edilememesine sebep olabilmektedir. Kolen ve Pollack (1990); çalışmalarında geri besleme yönetiminin başlangıç ağırlıklarını bulmada Monte Carlo Simülasyonuna bağlı olduğunu söylemektedir. Yam ve Chow (1995) ise çalışmalarında En Küçük Kareler Yöntemi temelli bir algoritma önermiş ve bu algoritma ile hem optimal başlangıç ağırlıklarının bulunabileceğini hem de modelin başlangıç ağırlıklarına bağımlılığının azaltılabileceğini göstermiştir.

Toplama fonksiyonu girdileri toplayan ve aktivasyonuna ileten bölüm olup, en sık kullanılan toplama fonksiyonu ise "ağırlıklı toplam fonksiyonu (ATF)"dur. Fonksiyon, (4.5) Eşitliğinde gösterilmiştir. Burada W: ağırlık, I: Girdi, n: hücreye gelen toplam girdi sayısını ifade etmektedir.

$$ATF = \sum_i^n W_i \, I_i \tag{4.5}$$

Aktivasyon fonksiyonu ise hücreye gelen girdinin işlenmesi ve çıktı değerinin hesaplanması ile ilgilidir. Doğrusal, adımsal, sinüs gibi bazı fonksiyonlar mevcut olup, en çok kullanılan aktivasyon fonksiyonu ise Sigmoid Fonksiyonudur. Sigmoid fonksiyonu Eşitlik (4.6)'da gösterilmiştir.

$$Sigmoid\ Fonksiyonu = \frac{1}{1+e^{-ATF}} \quad (4.6)$$

Yapay sinir ağları ikiye ayrılmaktadır (Bilgin, 2018):

(i) İleri Beslemeli Yapay Sinir Ağları: Aktarılan bilgilerin sadece ileri yönlü hareket ettiği yani girişten çıkış doğru ilerleyen bir süreç söz konusudur. Bu sinir ağlarında bir bilgi giriş katmanında iken herhangi bir değişikliğe uğramadan bir sonraki katmana aktarılır. Bu akış içerisinde bilgi, ara katmanlarda ve çıkış katmanında belirli bir ağırlığa sahip olur ve bu şekilde çıkış katmanına yönlendirilir.

(ii) Geri Beslemeli Yapay Sinir Ağları: En az bir hücrenin çıkışı başka bir hücrenin girişi olarak verilmektedir. Burada besleme işleminin katmandaki hücreler arasında olması şart olmadığı için doğrusal bir ilişki varlığından söz etmek mümkün olamayacaktır. Bu nedenle de sinir ağlarının yapısı veriye göre değişiklik göstermektedir.

4.4 Destek Vektör Makineleri

Destek vektör makineleri (DVM), bir makine öğrenmesi yöntemidir. Bu amaçla sonlu sayıda bir örüntü üzerinden iyi bir genelleme düzeyi bulmayı amaçlar. Bunun için "yapısal risk minimizasyonu (YRM) tümevarım prensibi" kullanmaktadır. Bu prensip ise VC (Vapnik–Chervonenkis) boyutunu ve deneysel riski en küçüklemek amacıyla aynı anda gerçekleşen girişimlerden oluşur. Bu teori AT&T laboratuvarlarında ortaya çıkmış olup, Vapnik ve arkadaşları tarafından ikili sınıflandırma problemi esaslı olarak geliştirilmiştir. Destek Vektör Makineleri, özellikle tanımlanması kolay olmayan örüntülerde ve karışık veri setlerinde kullanılması uygun olan bir algoritmadır. Bu algoritma ile örnekler içerisinden farklılıklarına göre örnekleri öğrenir ve daha önceden tanımadığı yeni verilerde sınıfların tahmininde bulunur.

DVM'nin regresyon ve sınıflandırma problemlerinin çözümü için Destek Vektör Regresyonu (DVR) ve Destek Vektör Sınıflandırması (DVS) olmak üzere iki türü mevcuttur. DVM, destek vektörlerin alt kümesi üzerine inşa edilen fonksiyon tahminleri sunar ve bunun için yüksek boyutlu nitelik uzayı kullanır.

4.4.1 Doğrusal sınıflandırıcılar

$\mathcal{X}$ girdi örüntülerinin uzayını (örneğin $\Re^d$) ve $\langle .,. \rangle$, $\mathcal{X}$ uzayında iç çarpımı göstermek üzere; n gözlem içeren eğitim verisi, $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$, $\mathbf{x} \in \mathcal{X}$, $y \in \{+1, -1\}$ olsun. Hiperdüzlem karar fonksiyonu, sınıflandırma problemlerinde;

$$D(\mathbf{x}) = \langle \mathbf{w}, \mathbf{x} \rangle + b \ , \quad \mathbf{w} \in \mathcal{X} \ , \quad b \in \Re \tag{4.7}$$

şeklindedir (Cristianini and Shawe-Taylor, 2000). Eğitim aşamasının başarı ile tamamlanması neticesinde, yeni gözlenen $\mathbf{x}$ örüntüleri için $\mathbf{w}$ ve b katsayı tahminleri aracılığı ile çıktılar $\operatorname{sign} D(\mathbf{x})$ işaret fonksiyonuna göre üretilmektedir.

Destek vektör sınıflandırmasının amacı, deneysel riski en aza indirmek ve hatasız sınıflandırılan verilerin payını ise en fazla yapmaktadır. Burada deneysel risk kavramı ile pay sınırının "yanlış" kısmında bulunan veriler için sınırdan sapmaların toplamları ifade edilmektedir. Pay sınırının "doğru" kısmında bulunan veriler ise hatasız sınıflandırmış verilerdir. Buradan hareketle girdi uzayını iki bölgeye ayıran ve adapte edilebilir bir kayıp fonksiyonu kullanımının gerekliliği söylenebilir. Bu kayıp fonksiyonu ise Δ pay büyüklüğüne bağlıdır. Girdi uzayının bu iki bölgesinin bir kısmı model tarafından (sıfır kayıp) açıklanabilirken, diğer bölüm ise bir miktar belirsizlik ile açıklanabilecektir. Bu işlem için tanımlanan kayıp fonksiyonu;

$$L_\Delta\big(y, D(\mathbf{x}, \mathbf{w})\big) = \text{maks}(\Delta - y\, D(\mathbf{x}, \mathbf{w}), 0) \tag{4.8}$$

şeklinde sunulur (Cherkassky and Mulier, 2007). Sınıflandırma problemleri için (4.8) eşitliğinde verilen DVM kayıp fonksiyonuyla deneysel risk

$$R_{emp}(\mathbf{w}) = \frac{1}{n} \sum_{i=1}^{n} L_\Delta\big(y_i, D(\mathbf{x}_i, \mathbf{w})\big) \tag{4.9}$$

ile belirtilir. $\xi_i = \text{maks}\ (1 - y_i D(\mathbf{x}_i, \mathbf{w}), 0)\ , i = 1, \ldots, n$ gevşek değişkenleri ile ifade edilen pay tabanlı kayıp ise pay sınırlarından sapmaların bir göstergesidir.

Optimum ayırma hiperdüzleminin bulunması, doğrusal kısıtlar ile tanımlı bir karesel optimizasyon problemidir. Buna göre, $(\mathbf{x}_i, y_i),\ i = 1, \dots, n;\ \mathbf{x} \in \Re^d$ eğitim verisi mevcut olmak üzere programlama problemi,

Amaç fonksiyonu:

$$\min_{\mathbf{w}\in\mathcal{X}, b\in\Re} \left[\frac{1}{2}\|\mathbf{w}\|^2 + \frac{C}{n}\sum_{i=1}^{n}\xi_i\right] \tag{4.10}$$

Kısıtlar:

$$y_i[\langle \mathbf{w}, \mathbf{x}_i\rangle + b] \geq 1 - \xi_i,\ i = 1, \dots, n$$

biçiminde oluşturulur (Cherkassky and Mulier, 2007). Deneysel risk ile hiperdüzlem karar fonksiyonunun karmaşıklığı arasındaki değişimi denetleyen C katsayısı ise kullanıcı eliyle tespit edilmektedir.

Yüksek girdi uzayları için çözümlenmesi durumunda bu optimizasyon problemi ikili (dual) biçimine dönüştürülmesine gerek vardır (Vapnik 1995). Optimizasyon teorisi uyarınca; kısıtların ve amaç fonksiyonun dışbükey olması halinde problemin dual biçimi mevcuttur. O halde, ilk problemin çöüzü ile dual problemin çözümleri de denk olacaktır. (4.10) ile verilen optimizasyon probleminin de bahse konu kriterleri yerine getirmesi sebebiyle dual biçimi mevcuttur. Kuhn-Tucker teoremi kullanılarak dual biçimine dönüştürülür (Strang 1986).

$(\mathbf{x}_i, y_i), i = 1, \dots, n$ eğitim verisi ve C de bir düzenleme sabiti ve $\alpha_i, i = 1, \dots, n$ Lagrange katsayıları için dual problem,

Amaç fonksiyonu:

$$\underset{\boldsymbol{\alpha}\in\Re^n}{\text{maks}} \left[\sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{n}\alpha_i\alpha_j y_i y_j \langle \mathbf{x}_i, \mathbf{x}_j\rangle\right] \tag{4.11}$$

Kısıtlar:

$$\sum_{i=1}^{n} y_i\alpha_i = 0$$

$$0 \leq \alpha_i \leq C/n,\ i = 1, \dots, n$$

biçiminde elde edilir (Cristianini and Shawe-Taylor, 2000). $\alpha_i, i = 1, \ldots, n$ dual problemin çözümü olmak üzere ve b ise

$$b = y_s - \sum_{i=1}^{n} \alpha_i \, y_i \langle \mathbf{x}_i, \mathbf{x}_s \rangle \tag{4.12}$$

eşitliği ile tespit edildiğinde, hiperdüzlem karar fonksiyonu,

$$D(\mathbf{x}) = \sum_{i=1}^{n} \alpha_i y_i \langle \mathbf{x}_i, \mathbf{x} \rangle + b \tag{4.13}$$

dir. α_i parametrelerinin sıfır olmadığı veri örneklerine destek vektör denmektedir (Kecman 2001).

4.4.2 Doğrusal olmayan sınıflandırıcılar

DVM; yüksek boyutlu bir girdi uzayında doğrusal olmayan sınıf sınırlarını uygulamak için nitelik uzayında doğrusal bir model kullanır. Bu amaçla, yüksek boyutlu bir nitelik uzayına bazı doğrusal olmayan dönüşümler uygulanır. Girdi uzayında doğrusal olmayan karar sınırlarını göstermek üzere nitelik uzayında oluşturulmuş bir doğrusal model mevcuttur. Bu şekilde DVM algoritması kullanılarak, karar sınırları arasında en yüksek pay gerçekleşecek şekilde, nitelik uzayında optimum ayırma hiperdüzlemi alınmış olur. Destek vektörler belirlenirken ise karar sınırlarına en yakın olan örnekler seçilir. İkili sınıflandırma karar sınırlarının belirlenmesi açısından diğer eğitim örneklerinin katkısı bulunmamaktadır (Gunn, 1998; Vapnik, 1998; Cristianini and Shawe-Taylor, 2000).

Makine öğrenmesinde kullanımı süregelen bir uygulama da eğitim verisinin bir başka uzayın alt kümesi olacak şekilde dönüştürülmesidir. Burada $\mathcal{X}$ girdi uzayı, $\mathcal{F} = \{\Phi(\mathbf{x}) : \mathbf{x} \in \mathcal{X}\}$ ise nitelik uzayıdır. $\Phi : \mathcal{X} \longrightarrow \mathcal{F}$ doğrusal olmayan bir dönüşümü göstermek üzere, doğrusal olmayan sınıflandırma problemlerinde karar fonksiyonu

$$D(\mathbf{x}) = \sum_{i=1}^{n} w_i \phi_i(\mathbf{x}) + b \tag{4.14}$$

biçiminde tanımlanır (Cristianini and Shawe-Taylor, 2000). Burada, m nitelik uzayının boyutunu göstermek üzere $\Phi(\mathbf{x}) = [\phi_1(\mathbf{x})\ \phi_2(\mathbf{x})\ \ldots\ \phi_m(\mathbf{x})]$'dir. Destek vektör algoritması yalnızca, $\mathbf{x}_i$ örneklerinin iç çarpımlarına bağlı olarak tanımlıdır. Buna göre, destek vektör optimizasyon problemini yeniden ortaya koymaya olanak veren, Φ fonksiyonundan ziyade, $k(\mathbf{x}, \mathbf{x}') = \langle \Phi(\mathbf{x}), \Phi(\mathbf{x}') \rangle$ iç çarpım çekirdeğinin bilinmesi yeterlidir.

Nitelik uzayında bir optimum ayırma hiperdüzleminin dizayn edilmesi ile bir doğrusal olmayan DVM sınıflandırıcısı oluşmuş olur. Pay hiperdüzleminin yaratılması ile bu süreç benzer özellikler taşır. $(\mathbf{x}_i, y_i),\ i = 1, \ldots, n$ eğitim verisi, k iç çarpım çekirdeği ve C de bir düzenleme sabiti olmak üzere doğrusal olmayan destek vektör sınıflandırması için dual optimizasyon problemi,

Amaç fonksiyonu:

$$\underset{\boldsymbol{\alpha} \in \Re^n}{\text{maks}} \left[\sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j\ k(\mathbf{x}_i, \mathbf{x}_j) \right] \tag{4.15}$$

Kısıtlar:

$$\sum_{i=1}^{n} y_i \alpha_i = 0$$

$$0 \leq \alpha_i \leq C/n\,,\ i = 1, \ldots, n$$

şeklindedir (Kecman, 2001). Burada $\alpha_i,\ i = 1, 2, \ldots, n$ Lagrange katsayılarını ifade etmektedir. Çekirdek fonksiyonunun tespit edilmesi ve bu çekirdek fonksiyonunun etkili şekilde kullanılabilecek bir fonksiyon olması esas amaçtır. Böylece, $D(\mathbf{x})$ karar hiperdüzlemi,

$$D(\mathbf{x}) = \sum_{i=1}^{n} \alpha_i y_i k(\mathbf{x}_i, \mathbf{x}) + b \tag{4.16}$$

ve $d > 3$ için aynı zamanda bir hiperdüzlem olan destek vektör sınıflandırıcısı,

$$\text{sign}\, D(\mathbf{x}) = \text{sign}\left(\sum_{i=1}^{n} \alpha_i y_i k(\mathbf{x}_i, \mathbf{x}) + b\right) \tag{4.17}$$

dir. Burada, $(\mathbf{x}_s, y_s)$ destek vektörlerden biri olmak üzere b parametresi,

$b = y_s - \sum_{i=1}^{n} \alpha_i \, y_i \, k(\mathbf{x}_i, \mathbf{x}_s)$ (4.18)
ile elde edilir.

İç çarpımın hesaplanması için çekirdek fonksiyonlarında yapılan değişik tercihlere göre kullanılan dönüşüm fonksiyonunun çeşidi de farklılık göstermektedir. Polinomiyal, Gauss radyal tabanlı fonksiyon ve Hiperbolik Tannjant çekirdek fonksiyonları DVM'de yaygın olarak kullanılmaktadır. Bu çekirdek fonksiyonlarının makine öğrenmesi ve sinir ağları alanlarında yaygın kullanıldığı görülmektedir (Kecman 2001).

4.5 K-En Yakın Komşu Algoritması

k en yakın komşu (kNN) algoritması; bir veri noktasının dahil olduğu sınıfın ve en yakın komşunun k değerine dayanarak belirlendiği bir yöntemdir (Taşçı ve Onan, 2016). kNN algoritmasında eğitim setinde mevcut örnekler n boyutlu sayısal niteliklerle belirtilmektedir. Her örnek n boyutlu uzayda bir noktayı gösterecek bicinde n boyutlu bir örnek uzayda saklanır. Bilinmeyen bir örnek geldiğinde, eğitim setinde gelen örneğe en yakın k adet en yakın örnek belirlenir ve yeni örneğin dahil olacağı sınıfın etiketi de bu en yakın k adet komşusunun sınıf etiketlerinin çoğunluk oylaması ile belirlenir (Taşçı ve Onan, 2016).

En yakın komşuluk tespitinde kullanılan farklı uzaklık hesaplama ölçütleri bulunmakta olup, en çok tercih edilenler Öklid, Manhattan ve Minkowski olarak sayılabilir (Bilgin 2018).

- Minkowski Uzaklığı: $X = (x_1, x_2, \dots, x_n)$ ve $Y = (y_1, y_2, \dots, y_n)$ biçiminde iki nokta arasındaki uzaklık Denklem (4.19)'de gösterildiği gibi hesaplanmaktadır.

$$(\sum_{i=1}^{n} |x_i - y_i|^p)^{1/p} \quad (4.19)$$

Denklemde yer alan uzaklık ölçüsü Minkowski Uzaklığı olarak adlandırılmaktadır.

- Öklid Uzaklığı: $X = (x_1, x_2, \dots, x_n)$ ve $Y = (y_1, y_2, \dots, y_n)$ biçiminde iki nokta arasındaki uzaklık hesaplansın. Minkowski uzaklığında $p = 2$ olmak üzere oluşan özel durum Öklid Uzaklığı olarak hesaplanmaktadır.

$$\left(\sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}\right) \tag{4.20}$$

Denklem (4.20)'de yer alan uzaklık ölçüsü Öklid Uzaklığı olarak adlandırılmaktadır.

- Manhattan Uzaklığı: $X = (x_1, x_2, \dots, x_n)$ ve $Y = (y_1, y_2, \dots, y_n)$ biçiminde iki nokta arasındaki uzaklık hesaplansın. Minkowski uzaklığında $p = 1$ olmak üzere oluşan özel durum Manhattan Uzaklığı olarak hesaplanmaktadır (Denklem (4.21)).

$$\left(\sum_{i=1}^{n}|x_i - y_i|\right) \tag{4.21}$$

k en yakın komşu algoritması regresyon problemleri için kullanıldığında, tahmin k-en benzer mesafelerin medyan ya da ortalamasına dayanmaktadır. Sınıflandırma problemlerinde kullanıldığında ise k-en benzer mesafelerin en yüksek frekanslı sınıfı olarak hesaplanabilmektedir (Pandey vd. 2017).

4.6 Lojistik Regresyon

n gözlem içeren $(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n)$, $\mathbf{x} \in \Re^d$, $y_i \in \{+1, -1\}$ eğitim verisi olsun. $\mathbf{x}_i$ gözlemini belirli iki sınıftan birine sınıflandırılan bir regresyon modeli dikkate alınmış olsun. Burada, $\mathbf{x}_i$ deneyi, $\mu(\mathbf{x}_i)$ ortalamayla bir Bernoulli denemesi şeklinde değerlendirilebilir. Bu şekilde, y_i, ortalaması $\mu(\mathbf{x}_i)$ ve varyansı $\mu(\mathbf{x}_i)\big(1 - \mu(\mathbf{x}_i)\big)$ olan bir Bernoulli rasgele değişkenidir. Lojistik Regresyon (LR) fonksiyonu burada her bir $\mathbf{x}_i$ deneyi ve bu deneye denk gelen çıktının beklenen değeri arasında mevcut ilişkinin modellenmesi için uygulanmaktadır. $\boldsymbol{\beta}$ parametre vektörünü göstermek üzere bu fonksiyon,

$$\mu(\mathbf{x},\boldsymbol{\beta}) = \frac{\exp(\boldsymbol{\beta}^T\mathbf{x})}{1+\exp(\boldsymbol{\beta}^T\mathbf{x})} \tag{4.22}$$

şeklindedir (Menard 1995, Komarek 2004). Böylece, ε hata terimi ile regresyon modeli,

$$y = \mu(\mathbf{x},\boldsymbol{\beta}) + \varepsilon \tag{4.23}$$

olarak yazılır. Burada hata terimi, $y = 1$ için $\varepsilon = 1 - \mu(\mathbf{x})$ ya da $y = 0$ için $\varepsilon = \mu(\mathbf{x})$ olacak şekilde yalnızca iki değer alabilecektir. y rasgele değişkeninin, $\mu(\mathbf{x})$ ortalama ve $\mu(\mathbf{x})\big(1-\mu(\mathbf{x})\big)$ varyans ile Bernoulli dağılımlı olmasından ötürü; ε hatası, 0 ortalama ve $\mu(\mathbf{x})\big(1-\mu(\mathbf{x})\big)$ varyansına sahiptir. Burada ortalamadan bağımsız σ^2 sabit varyansına sahip doğrusal regresyondan farklı bir durum mevcuttur (Komarek 2004).

$\boldsymbol{\beta}$ parametresine göre doğrusal olmayan lojistik regresyonda model tahminine hata kareler toplamını en az yaparak ulaşılması imkansızdır, zira bu doğrusal regresyonda kullanılan bir yöntemdir. Bunun sebebi ise hem olabilirlik fonksiyonunun en yükseğe eşdeğer olmamasından hem de hata kareler toplamının en düşük değerinin hesaplanmasının zor olmasından ileri gelmektedir. Bu amaçla LR modelini parametrelerine göre doğrusal bir fonksiyona dönüştürmek amacıyla bir lojit fonksiyonu kullanılır. Bunun için de (4.24)'te görülen fonksiyon kullanılır.

$$\mathrm{g}(\mu) = \ln\left[\frac{\mu}{1-\mu}\right] \tag{4.24}$$

lojit fonksiyonunun kullanımıyla mümkündür. (4.22) eşitliğinin ve bu eşitlikte verilen çıktı değişkenine ait lojit dönüşüm ve koşullu ortalama sonrası

$$\mathrm{g}(y) = \boldsymbol{\beta}^T\mathbf{x} + \tilde{\varepsilon}(\mathbf{x}) \tag{4.25}$$

ile verilen yeni bir model oluşmuş olur.

Oluşturulan bu yeni model parametrelerine göre doğrusal bir modeldir. Yine bağımsız değişkenlerin alacağı değerlerle ilişkili olarak, $-\infty$ ile $+\infty$ arasında değişen sürekli bir fonksiyondur. Fakat; bu modelde de parametreleri tahmin etmede en küçük kareler yönteminin kullanımı uygun değildir. Bunun sebebi ise varyansın ortalamadan bağımsız olmaması ve hata teriminin normal dağılıma uygunluk göstermemesidir. Bu nedenle, parametre tahmininde kullanımı önerilen yöntemler; en çok olabilirlik ya da yinelemeli yeniden ağırlıklandırılmış en küçük kareler yöntemidir (Komarek 2004).

LR yönteminin amacı; özellikle de veri bileşenleri ve parametrelerin doğrusal ilişkisi düşünüldüğünde ve doğrusal bir sınıflandırma olması da dikkate alındığında, negatif ve pozitif verileri birbirinden ayırmayı amaçlayan bir hiperdüzlem tespiti olarak düşünülebilmektedir.

4.7 Naive Bayes Algoritması

Naive Bayes sınıflandırıcısı temelde Bayes Teoremine dayanmaktadır. Bayes teoremi Eşitlik (4.26)'daki şekilde ifade edilmektedir.

$$P(A\backslash B) = \frac{P(B\backslash A)P(A)}{P(B)} \tag{4.26}$$

Burada;

$P(A\backslash B)$: B olayı gerçekleşmesi durumunda A olayının gerçekleşme olasılığı
$P(B\backslash A)$: A olayının gerçekleşmesi durumunda B olayının gerçekleşme olasılığı
$P(A)$ ve $P(B)$: A ve B olaylarının gerçekleşme olasılıkları.

biçiminde ifade edilir. Naive Bayes algoritmasının her bir özelliğin bağımsız ve çıktıya eşit derecede katkı yaptığı varsayımları bulunmaktadır (Chauhan ve Agrawal 2022).

y sınıf değişkeni, $X = (x_1, x_2, \dots, x_n)$ ise girdiyi göstermek üzere Bayes Teoremi kullanıldığında

$$P(y \mid x_1, x_2, \dots, x_n) = \frac{P(x_1 \mid y)P(x_2 \mid y)\dots P(x_n \mid y)}{P(x_1)P(x_2)\dots P(x_n)} \quad (4.27)$$

elde edilir. Verilen bir veri setinden girdiler bu denklemde yerine koyulmak suretiyle koşullu olasılık değerleri elde edilebilmektedir. Eşitliğin paydası veri setinden yapılan girdilere göre değişmeyeceğinden payda kaldırılabilmekte

$$P(y \mid x_1, x_2, \dots, x_n) \propto P(y) \prod_{i=1}^{n} P(x_i \mid y) \quad (4.28)$$

yazılabilmektedir. Sınıf değişkeni y, binary (ikili) bir değişken olabileceği gibi çoklu bir değişken de olabilir. Bu nedenle sınıf değişkenini maksimum olasılıkla bulmak gereklidir.

$$y = argmax_y P(y) \prod_{i=1}^{n} P(x_i \mid y) \quad (4.29)$$

ile verilen eşitlik kullanılarak özellikleri verildiğinde bir sınıflandırma elde edilebilmektedir.

$P(y|X)$ olasılığı, çıktı değişkenine karşı bir frekans tablosu oluşturularak elde edilebilmektedir. Bu frekans tabloları likelihood (olabilirlik) tablolarına dönüştürülür ve Naive Bayes eşitliği kullanılarak her bir sınıf için olasılıklar elde edilir. En yüksek olasılık elde edilen sınıf tahminin sonucu olarak elde edilmektedir (Chauhan ve Agrawal 2022).

5. BULANIK REGRESYON FONKSİYONLARI YAKLAŞIMI

Bulanık mantık ve bulanık küme kavramları Zadeh (1965, 1975) tarafından ortaya konulmuştur. Bulanık mantığın belirsizlik durumundan tahmin ve çıkarım yapmak amacıyla birçok alanda uygulamaları uygulamaları mevcuttur. Gürültü (noise) ve belirsizliklerin bir veri sisteminden tamamen çıkarılması mümkün olamadığından, belirsizlik halinde karmaşık sistemlerin modellenmesi için bulanık mantık ve teorisini kullanmak ise sıklıkla kullanılan bir yoldur. Bulanık kümeleme, üyelik dereceleri yardımıyla bir nesnenin hangi kümeye ait olduğunu belirlemek ve veri setindeki örtüşen küme yapısını tespit edebilmek amacıyla kullanılan bir yöntemdir.

Bulanık Regresyon Fonksiyonları (BRF) yaklaşımına göre gözlemlerin benzerliklerine göre kümelenmesini sağlamak amacıyla Bulanık k-Ortalama (BKO) (Bezdek 1981) kümeleme algoritmasından faydalanılır. Çalışmanın bu bölümünde, bulanık kümeleme analizine ilişkin genel bir değerlendirme yapılacak olup, BRF modelleme süreci detaylı bir biçimde ele alınacaktır.

5.1 Bulanık K-Ortalama Kümeleme Algoritması

Esnek hesaplama ile bir sistem, yerel ve genel sistem modelleme gibi iki farklı yaklaşıma bağlı olarak modellenebilmekedir (Babuška ve Verbruggen 1997). Genel modelleme bölümünde, varolan ilişkilerin tespit edilmesi için sistemin tamamının analiz edilmesi ve çözümşlenmesi söz konusudur. Yerel modellemede ise sistem başlangıçta anlam ifade eden parçalara bölünür. Akabinde doğrusal ya da doğrusal olmayan yöntemler aracılığı ile alt modeller oluşturulur. Oluşan yerel modellerin özelliklerinin saptanmasında bulanık kümeleme algoritmaları kullanılır.

Kümeleme yapısı itibariyle bulanık kümeleme algoritmaları; bir amaç fonksiyonu ile kovaryans matrisine veya bir bulanık ilişkiye dayanan, sinir-bulanık kümeleme veya parametrik olmayan sınıflayıcı yapısına göre birbirinden ayrışmaktadır. Bu çalışma kapsamında “amaç” tabanlı bulanık kümeleme algoritmaları temel alınacaktır. Amaç tabanlı bulanık kümeleme algoritmalarında; olası küme parçalanmaları için bir amaç

fonksiyonu ile hem toplam hata hem de sayısal bir uzaklık ölçüsü en aza indirilmeye çalışılır. Amaç fonksiyonunun optimal değeri, küme parçalanmalarının da olabilecek en iyi durumuna eriştiğini ifade eder. Bu da amaç tabanlı kümeleme algoritmalarını bir optimizasyon probleminin çözülmesine bağlı hale getirir (Çelikyılmaz ve Türkşen 2009).

BKO algoritmasında amaç; küme yapısına bağlı olarak tanımlanan kısıtlara göre amaç fonksiyonunun global minimum veya maksimumunu tespit etmektir (Çelikyılmaz ve Türkşen 2009). $X = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n\}$ nesneler kümesini göstersin. Her bir i nesnesi ($i = 1, 2, \dots, n$), d boyutlu $\mathbf{x}_i = \left[x_{1,i}\ x_{2,i} \dots\ x_{d,i}\right]^T \in \Re^d$ vektörü ile temsil edilsin. Buna göre, $n \times d$ boyutlu veri matrisi,

$$\mathrm{X} = \begin{bmatrix} x_{1,1} & x_{1,2} & \dots & x_{1,d} \\ x_{2,1} & x_{2,2} & \dots & x_{2,d} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n,1} & x_{n,2} & \dots & x_{n,d} \end{bmatrix} \tag{5.1}$$

biçiminde tanımlanır. Her bir verinin kümelere olan üyelik derecelerini gösteren U parçalanma matrisi tasarımıyla bir bulanık kümeleme algoritması, veri kümesini c sayıda örtüşen kümeye parçalar (Çelikyılmaz ve Türkşen 2009).

Bulanık parçalanma matrisi, U, her k $(k = 1, 2, \dots, c)$ kümesinde yer alan $\mathbf{x}_i$ $(i = 1, 2, \dots, n)$ nesnelerinin üyelik derecelerinden oluşan bir matristir. k kümesindeki $i.$ vektörün üyelik derecesi $\mu_{k,i} \in$ U ile gösterilir. Buna göre de parçalanma matrisi,

$$\mathrm{U} = \begin{bmatrix} \mu_{1,1} & \mu_{2,1} & \dots & \mu_{c,1} \\ \mu_{1,2} & \mu_{2,2} & \dots & \mu_{c,2} \\ \vdots & \vdots & \ddots & \vdots \\ \mu_{1,n} & \mu_{2,n} & \dots & \mu_{c,n} \end{bmatrix} \tag{5.2}$$

ile verilir. Bulanık kümeleme algoritmasında; c sayıda örtüşen küme, merkez (prototip) vektörleri ile temsil edilir (Çelikyılmaz ve Türkşen 2009). $V = \{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_c\} \in \mathcal{R}^{c \times d}$ olmak üzere her bir küme merkezi genellikle, d sayıda nesnenin ağırlık merkezi olarak ifade edilir (Çelikyılmaz ve Türkşen 2009).

Bu tez kapsamında "amaç fonksiyonu tabanlı noktasal (uzaklık ölçütlü) kümeleme algoritmalar" ele alınacaktır. BKO kümeleme algoritmasını kullanan Bulanık Regresyon Fonksiyonları da bu özelliği nedeniyle bir sonraki bölümde detaylı şekilde sunulacaktır.

5.1.1 Bulanık k-ortalama kümeleme algoritması

BKO kümeleme (Bezdek 1981) yönteminde $X = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n\}$ veri setinin kaç adet kümeye ayrılacağını belirten c sayısının bilindiği ya da tespit edebileceği varsayımı mevcuttur. BKO kümeleme algoritmasındaki bu c sayısının tespit edilmesi için bazı yöntemler bulunmuştur. Bunlardan biri de "Küme Geçerlilik İndeksi Analizi"dir.

c küme sayısını ve m bulanıklık parametresini göstermek üzere BKO kümeleme algoritmasında amaç

$$\min J\,(\mathrm{X};\mathrm{U},\mathrm{V}) = \sum_{k=1}^{c}\sum_{i=1}^{n}\left(\mu_{k,i}\right)^{m} d^2(\mathbf{x}_i, \mathbf{v}_k) \tag{5.3}$$

fonksiyonunun en küçük değerini veren küme yapısına ulaşmaktır. Küme merkez vektörü $\mathbf{v}_i$ ile her bir küme temsil edilmektedir. Burada $d^2(\mathbf{x}_i, \mathbf{v}_k)$ ise $i.$ nesnenin $k.$ küme merkezine olan uzaklığıdır. Bulanıklaştırıcı (fuzzifier) veya bulanıklık; (4.3) denkleminde yer alan $m \in (1, \infty)$ değeridir. m değeri kümelerin örtüşme derecesini belirler. Kümelerin örtüşmemesi durumu ise "$m = 1$" durumu ile ifade edilmekte olup, bir kesin (crisp) kümeleme yapısını temsil etmektedir. Küme merkezi $\mathbf{v}_i$'lerden veri nesnelerinin uzaklaşması durumunda amaç fonksiyonunun değeri büyüyecektir. O halde amaç fonksiyonun değerini etkileyenler; küme merkezlerinin sayısı ve yeridir. Optimum çözüm bulunduğunda ise amaç fonksiyonun değeri minimum olmalı ve global minimum için çözüm aranmalıdır. U parçalanma matrisine aşağıda yer alan kısıtların eklenmesi ile problemde ulaşılabilecek gereksiz çözümlerden kaçınılması amaçlanır.

$$\sum_{k=1}^{c}\mu_{k,i} = 1 \;,\quad \forall i > 0 \tag{5.4}$$

$$0 < \sum_{i=1}^{n}\mu_{k,i} < n \;,\quad \forall k > 0 \tag{5.5}$$

Parçalanma matrisinde her bir satır toplamının 1'e eşit olduğu eşitlik (5.4) ile verilen kısıtta görülebilmektedir. Krishnapuram ve Keller (1993) yaptıkları çalışmada (5.4)'te yer alan denklemi kümelemeye olabilirlik yaklaşımı şeklinde açıklamışlardır. (5.5) ise parçalanma matrisindeki her bir sütundaki üyelik değerlerinin sıfırdan büyük olacağını ve n (gözlem sayısı) değerinin üzerine çıkamayacağını göstermektedir. Bu kısıt ile birlikte her kümede en azından bir eleman olması garantilenmiş olur. Ayrıca üyelik değerlerine ilişkin belirli bir dağılım bulunması gerektiğine dair bir kısıt olmadığı da görülmektedir.

Uzaklık ölçüsü için genel formül ise

$$d^2(\mathbf{x}_i, \mathbf{v}_k) = (\mathbf{x}_i - \mathbf{v}_k)^T \, \mathrm{A}_k \, (\mathbf{x}_i - \mathbf{v}_k) \geq 0 \tag{5.6}$$

biçimindedir. (5.6) eşitliğinde $\mathrm{A}_k, k = 1, 2, \ldots, c$ norm matrisi, pozitif tanımlı simetrik bir matristir. Bulanık kümeleme algoritmalarında Öklid, Minkowski, Mahalanobis Uzaklığı gibi uzaklık ölçütlerinin kullanımı da mümkündür (Çelikyılmaz ve Türkşen 2009). BKO kümeleme algoritmasında Öklid uzaklığı yaygın olarak kullanılmaktadır.

Buna göre, BKO kümeleme algoritması bir optimizasyon problemi olarak;

$$\begin{aligned} &\min J(\mathrm{X}; \mathrm{U}, \mathrm{V}) = \textstyle\sum_{k=1}^{c} \sum_{i=1}^{n} \left(\mu_{k,i}\right)^m d^2(\mathbf{x}_i - \mathbf{v}_k)_{\mathrm{A}} \\ &0 \leq \mu_{k,i} \leq 1 \ , \quad \forall i, k \\ &\textstyle\sum_{k=1}^{c} \mu_{k,i} = 1 \ , \quad \forall i > 0 \\ &0 < \textstyle\sum_{i=1}^{n} \mu_{k,i} < n \ , \quad \forall k > 0 \end{aligned} \tag{5.7}$$

matematiksel modeli ile tanımlanır (Bezdek 1981). Bu problem Lagrange Çarpanları yöntemi ile çözüldüğünde (Khuri 2003) kısıtsız bir optimizasyon problemi ve amaç fonksiyonundan oluşan bir model elde edilmiş olur. (4.7) ile verilen primal problem Lagrange çarpanları (λ) olarak bilinen parametreler kullanılarak kısıtsız bir problem şekline getirilir. Böylece,

$$\text{maks } W(\mathrm{U},\mathrm{V}) = \sum_{k=1}^{c}\sum_{i=1}^{n}\left(\mu_{k,i}\right)^{m} d^2(\mathbf{x}_k,\mathbf{v}_i)_{\mathrm{A}} - \lambda\left(\sum_{k=1}^{c}\mu_{k,i} - 1\right) \tag{5.8}$$

ile verilen amaç fonksiyonuna göre optimum üyelik değerleri ve küme merkezleri,

$$\mu_{k,i}^{(t)} = \left[\sum_{l=1}^{c}\left(\frac{d\left(\mathbf{x}_i,\mathbf{v}_k^{(t-1)}\right)}{d\left(\mathbf{x}_i,\mathbf{v}_l^{(t-1)}\right)}\right)^{\frac{2}{m-1}}\right]^{-1} \tag{5.9}$$

$$v_k^{(t)} = \frac{\sum_{i=1}^{n}\left(\mu_{k,i}^{(t)}\right)^{m}\mathbf{x}_i}{\sum_{i=1}^{n}\left(\mu_{k,i}^{(t)}\right)^{m}}, \quad \forall k = 1,2,\ldots,c \tag{5.10}$$

biçiminde belirlenir (Çelikyılmaz ve Türkşen 2009). (5.9) denkleminde $\mathbf{v}_k^{(t-1)}$, $(t-1)$. yinelemede k. küme için elde edilen küme merkez vektörünü göstermektedir. (5.9) ve (5.10) denklemlerinde birlikte görülen $\mu_{k,i}^{(t)}$ ise t. yinelemede hesaplanılan optimum üyelik değerlerini göstermektedir. Buradan elde edilen sonuçta küme merkezlerinin ve üyelik değerlerinin birbirine bağımlı oldukları tespit edilir. Bu nedenle Bezdek (1981) tarafından bir yinelemeli formül öne sürülmüş olup, burada amaç üyelik değerleri ve küme merkezlerinin tespit edilmesidir. Buna göre her bir t iterasyonunda, $J^{(t)}$ amaç fonksiyonu,

$$J^{(t)} = \sum_{k=1}^{c}\sum_{i=1}^{n}\left(\mu_{k,i}^{(t)}\right)^{m} d^2\left(\mathbf{x}_i,\mathbf{v}_k^{(t)}\right) > 0 \tag{5.11}$$

ile belirlenir.

BKO algoritmasının durdurma kuralı iki şekildedir. Bir durumda belli bir yineleme sayısına erişilir ve algoritma durur. Diğer durumda ise ε gibi bir değer tanımlanır ve algoritmanın iki en yakın kümenin ayrılma değerinin bu ε değerinden küçük olması halinde algoritmanın sonlanması şeklinde tanımlama yapılır. BKO kümeleme algoritması adımsal olarak aşağıda verilmiştir:

Adım 1. (5.10) denklemi ile başlangıç küme merkezleri tespit edilir. Bu tespit için başlangıç parçalanma matrisinin üyelik değerleri kullanılır.

Adım 2. Başlangıçta belirlenen yenileme sayısına dek sürmek üzere t=1 den bir yineleme başlatılır.

Adım 2.1. $\mathbf{x}_i$ girdi veri vektörünü ve $(t-1)$. yinelemede elde edilen $\mathbf{v}_k^{(t-1)}$ küme merkezlerini kullanarak, (5.9) ile verilen denklem ile, k. kümede her bir i girdi veri nesnesinin üyelik değeri, $\mu_{k,i}^{(t)}$ hesaplanır.

Adım 2.2. $\mu_{k,i}^{(t)}$ üyelik değeri ve $\mathbf{x}_i$ girdisi kullanılarak, (5.10)'da gösterilen küme merkez fonksiyonu kullanılarak, t yinelemesinde her bir k kümesinin küme merkezleri hesaplanır.

Adım 2.3. $\left|\mathbf{v}_k^{(t)} - \mathbf{v}_k^{(t-1)}\right| \leq \varepsilon$ şeklinde bir durdurma kuralı verilir ve kural yerine getirildiğinde algoritma durur. Eğer işlem sağlanamazsa *Adım 2*'ye dönülür.

Bulanıklık derecesi olan (m)'nin algoritmaya olan etkisini incelemek için (5.9)'da verilen denklemde aşağıdaki gibi limitler alınır.

$$\lim_{m\to\infty} \mu_{k,i}(\mathbf{x}) = \lim_{m\to\infty} \left[\sum_{l=1}^{c} \left(\frac{d^2(\mathbf{x}_i, \mathbf{v}_k)}{d^2(\mathbf{x}_i, \mathbf{v}_l)}\right)^{\frac{1}{m-1}}\right]^{-1} = \frac{1}{c} \quad , \forall k, l = 1, 2, \dots, c. \tag{5.12}$$

Küme merkezlerinin birbirinden farklı olduğu varsayımına göre

$$\lim_{m\to 1} \mu_{k,i}(\mathbf{x}) = \begin{cases} 1, & d^2(\mathbf{x}_i, \mathbf{v}_k) < d^2(\mathbf{x}_i, \mathbf{v}_l)\ ;\ \forall k, l = 1, 2, \dots, c\ ;\ k \neq l \\ 0, & diğer\ yerlerde \end{cases} \tag{5.13}$$

olarak elde edilir. Buna göre m değeri arttıkça; $\mu_{k,i}$ üyelik değeri sıfıra yakınsayacaktır.m değerinin büyümesiyle elde edilen sonuçların daha bulanık ve kümelerdeki örtüşmenin daha fazla olduğu görülür. Bunun sebebi ise m parametresinin örtüşme derecesini göstermesidir. Bağlantılı şekilde m değerinin küçülmesi durumunda ise bulanık

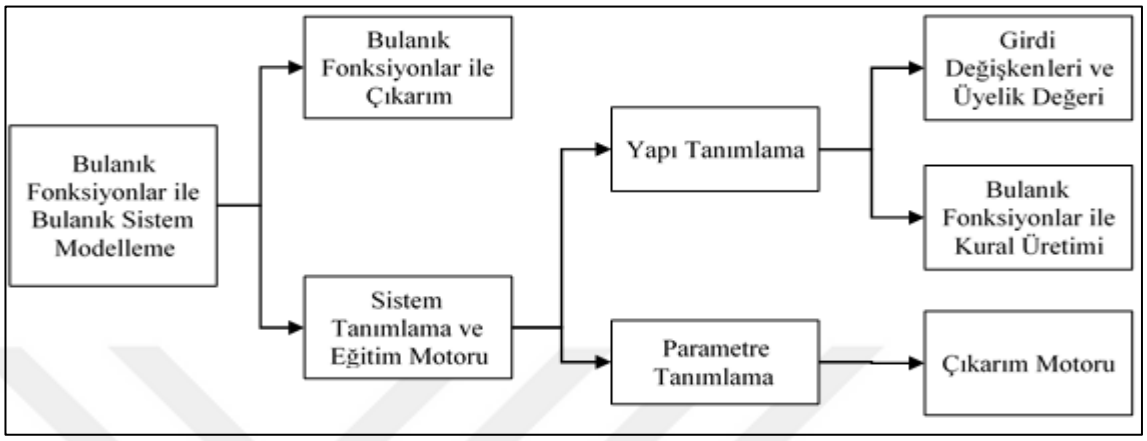
kümeleme sonuçları bir kesin kümeleme modeline daha yakın olarak görülecektir. $m = 1$ olduğunda ise kümeler arasında hiçbir örtüşme durumunun olmadığı ve yöntemin bir kesin kümeleme yöntemi ile aynı olduğu görülür. Bu durumda da bütün üyelik değerleri $\mu_{k,i} \in \{0, 1\}$'dir. Türkşen (1999) çalışmasında m değerinin bu sistem modellemesi içerisinde 2 olarak alınmasının uygun olduğu görüşünü ileri sürmüştür. En ideal küme sayısının tespiti için ise daha önce de bahsedildiği üzere bir Küme Geçerlilik İndeksi kullanılması önemlidir.

5.2 Sınıflandırma Problemlerinde Bulanık Regresyon Fonksiyonları

BRF, değişkenler arasındaki bulanık ilişkileri modellemek amacıyla bulanık kümeleme sonucunda elde edilen üyelik değerlerini sistem davranışını açıklamak üzere kullanan bir yaklaşımdır. BRF, üyelik değerlerinin bu kullanımı itibariyle diğer bulanık sistem modelleme yaklaşımlarından daha farklıdır. Bu çerçevede yerel bulanık regresyon fonksiyonlarının tahmin performansını arttırmak için kümeleme algoritmalarından faydalanılmaktadır. Girdi/çıktı davranışının yerel bulanık regresyon fonksiyonları ile açıklanabilir olması da BRF sistem modellemesinin performansını doğrudan belirlemektedir.

Şekil 5.1'de gösterildiği gibi, BRF sistem modelleri ile bulanık kural tabanlı yapılar ya da bunların genişlemelerine dair klasik bulanık sistem modellerinin sistem tasarım adımları benzerdir. Bu modelleri farklılaştıran kısım ise yapı tanımlama kısmıdır. Yapı tanımlama teknolojileri ise her bir örüntü ve çıkarım yöntemi için bulanık kuralların geliştirilmesi şeklinde tanımlanabilmektedir. BRF yaklaşımında veriler birçok örtüşen bulanık kümelere bölünür. Bu kümeler ise birbirinden farklı karar kurallarını belirtmek için kullanılır. Bu yöntem ilk etapta bu bulanık parçalanmalar için BKO kümeleme algoritmasından faydalanmakta idi. BRF yaklaşımı ile yeni olarak; üyelik değerleri ve bunların dönüşümlerinin ek bulanık tanımlayıcılar şeklinde kullanıma katılması ve nesneler arasındaki belirsizliğin bu şekilde daha net olarak ortaya konulması sağlanmıştır. Bu çerçevede her bir küme için farklılaşan veri setlerinin yapısını anlayabilmek için üyelik değerleri ve kullanıcı tarafından tanımlanmış olası dönüşümleri esas veri setine girdi olarak ilave edilir. Lokal girdi çıktı ilişkilerini tespit etmek amacıyla her bir kümeye

ait veri setleri kullanılarak lokal fonksiyonlar belirlenir. “Bulanık Regresyon Fonksiyonları” olarak isimlendirilen bu yaklaşım ilk olarak Türkşen (2008) tarafından önerilmiştir.



Şekil 0.1 BRF ile bulanık sistem modelleme

Üyelik değerleri; özellikle benzerliklerin vektör neseneleri arasındaki uzaklıklar ile izah edildiği sistem modelleme yaklaşımlarında kritik bir rol üstlenmektedir. Klasik bulanık kural tabanlı yaklaşımlara göre sistem çıktısı ve model çıktısı arasındaki hatayı minimize edebilmesi açısından BRF’nin daha iyi sonuçlar verdiği tespit edilmiştir.

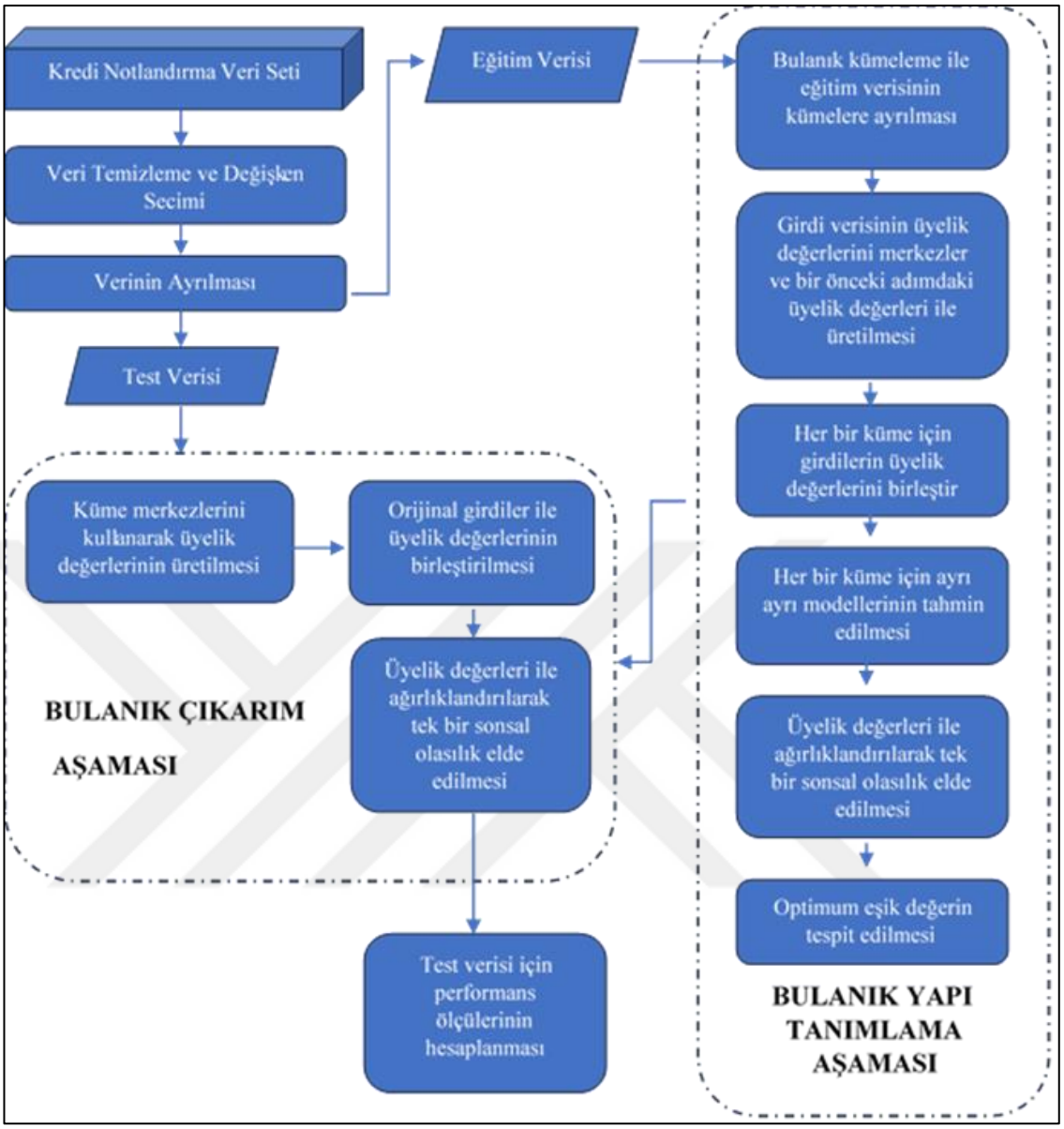
5.1.2 Yapı tanımlama ve çıkarım

Çeşitli veri madenciliği problemlerinde çok sayıda uygulamaları bulunan sınıflandırma problemleri, kategorik bir çıktı değişkeni ile girdi değişkenleri arasındaki fonksiyonel ilişkiyi öğrenmek ile ilgilenmektedir. Sınıflandırma problemlerine ilişkin daha fazla detay Mitchell (1997); Stork vd. (2001); Hastie vd. (2009) çalışmalarında bulunabilmektedir. Ayrıca, sınıflandırma modelleri; test örneklerini sınıf etiketleriyle tanımlanan gruplara ayırmakta kullanılmaktadır. Gözlemlerin gruplara ayrılması hem kümeleme hem de sınıflandırma ile yapılırken, bu ikisi arasında temel bir fark vardır. Kümeleme problemlerinde ayırma, gözlemler arasındaki benzerlikleri temel alarak gerçekleştirilir. Buna göre; istatistiksel öğrenme algoritmaları arasında bulunan kümeleme ve sınıflandırma, denetimsiz ve denetimli algoritmalar olarak tanımlanmaktadır. Denetimli

öğrenmede, sınıf etiketleri önemli özellikleri temsil ettiği için modelleme süreci genelde uygulamaya özel denetim gerektirir.

Bulanık üyelik değerleri, nesnelerin benzerliklerinin vektörler arasında mesafeye bağlı olarak tanımlandığı sistem modelleme yaklaşımları için önemli bir rol oynamaktadır (Türkşen and Çelikyılmaz 2006). Kümeleme tabanlı bulanık model olarak tanımlanan BRF, sistem ve model çıktıları arasındaki hatayı en aza indirmede geleneksel bulanık kural tabanlı yaklaşımlara göre daha iyi sonuçlar verdiği bulunmuştur (Başer ve Demirhan 2017, Chakravarty vd. 2020). BRF, sistemin yapısını tanımlayabilmek için Bezdek (1981) tarafından önerilen BKO kümeleme algoritmasını kullanmaktadır. BKO algoritmasının bir sonucu olarak elde edilen üyelik değerleri, girdi ve çıktı arasındaki fonksiyonel ilişkiyi açıklama için orijinal veri setine girdi olarak eklenirler.

Türkşen ve Çelikyılmaz 2006 tarafından önerilen ve Başer vd (2023) tarafından geliştirilen sınıflandırma problemleri için BRF yaklaşımı, eğitim ve çıkarım aşamalarından oluşmaktadır. Eğitim aşamasında model tahmini, tüm veri setinden rasgele olarak seçilmiş eğitim verisi ile gerçekleştirilir. Daha sonra, çıkarım aşamasında test ya da doğrulama verisi kullanılarak model tahmin doğruluğu değerlendirilir. Bu akış şemasındaki adımların detaylı açıklaması algoritmada sunulmuştur. BRF uygulamasının önemli adımları, girdi/çıktı (kümeleme adımı) ve küme parametrelerinin tahmini (tahmin adımı) şeklindedir.



Şekil 5.2 BRF Algoritması Akış Şeması

FKM kümeleme yöntemini kullanan BRF yaklaşımı, bulanık yapı tanımlama (eğitim) ve bulanık çıkarım (test) olarak iki aşamada açıklanmaktadır. İlk aşamada tamamlanması gereken yedi adım bulunmaktadır ve bunların sonuçları beş adımdan oluşan ikinci aşamanın girdileridir. Şekil 5.2'de yer alan akış şemasında açıklanmış olan adımların; her bir girdinin temerrüde düşme olasılığını tahmin etmek için kredi veri setine uygulanması hedeflenmektedir. BRF algoritmasının yapı tanımlama aşamasının yedi adımı ve çıkarım aşamasının beş adımı aşağıda açıklanmıştır.

Adım 1. $\{(\mathbf{x}_i, y_i);\ i = 1, \dots, n\}$ eğitim veri çiftlerinde, $m \geq 1.1$ ve $c > 1$ için FKM kümeleme algoritması uygulanır. Burada c ve m sırasıyla küme sayısını ve bulanıklığın derecesini göstermektedir.

$\mathrm{X} = \left[x_{j,i} \middle| i = 1, 2, \dots, n; j = 1, 2, \dots, d\right]$ olacak şekilde $\mathrm{Z} = (\mathrm{X}, \mathbf{y})$, $\left(n \times (d+1)\right)$ girdi-çıktı veri seti (matris) olsun. Burada d seçilen girdi sayısını ve n veri vektörlerinin toplam sayısını ifade etmektedir. Böylece, $\mathbf{x}_i = \left(x_{1,i}, x_{2,i}, \dots, x_{d,i}\right) \in \Re^d$ ve çıktı $y_i \in \Re$ olmak üzere $\mathbf{z}_i = (\mathbf{x}_i, y_i) \subset \Re^{d+1}$, $i = 1, 2, \dots n$ eğitim veri setinden herhangi bir veri noktasını (vektör) belirtmektedir ve her veri noktası $(d+1)$ boyutundaki girdi vektörlerinden oluşmaktadır. Ayrıca, $\mu_{k,i}(\mathbf{x}, y) \in [0, 1]$, k kümesi içindeki i. verinin üyelik değerini göstersin.

Buna göre; FKM'yi (Bezdek 1981) aşağıda verilen şekilde uygulayarak; $\mathbf{v}_k(\mathbf{x}, y) = \left(x_{1,k}^c, x_{2,k}^c, \dots, x_{d,k}^c, y_k^c\right) \in \Re^{d+1}$ optimum küme merkezleri ve $\mu_{k,i}(\mathbf{x}, y)$ etkileşimli (girdi-çıktı) üyelik değerleri yinelemeli olarak hesaplanır.

$$\mathbf{v}_k(\mathbf{x}, y) = \frac{\sum_{i=1}^{n}\left(\mu_{k,i}(\mathbf{x},y)\right)^m \mathbf{z}_i}{\sum_{i=1}^{n}\left(\mu_{k,i}(\mathbf{x},y)\right)^m} \tag{5.14}$$

$$\mu_{k,i}(\mathbf{x}, y) = \left(\sum_{l=1}^{c}\left(d_{k,i}(\mathbf{x}, y) \Big/ d_{l,i}(\mathbf{x}, y)\right)^{2/m-1}\right)^{-1} \tag{5.15}$$

$$d_{k,i}(\mathbf{x}, y) = \|(\mathbf{x}_i, y_i) - \mathbf{v}_k(\mathbf{x}, y)\|;\ i = 1, 2, \dots, n;\ k = 1, 2, \dots, c \tag{5.16}$$

m ve c'nin optimum değerlerini tespit etmek için, küme geçerlilik indeksi (CVI) analizi kullanılabilir (Pal ve Bezdek 1995, Kim ve Ramakrishna 2005). Türkşen'e (1999) dayalı sistem modelleme analizinde makul kümeleme sonuçları elde edebilmek için $m = 2$ olarak alınır.

Adım 2. Bir önceki adımda elde edilen küme merkez vektörleri kullanılarak girdi uzayı için $\mathbf{v}_k(\mathbf{x}) = \left(x_{1,k}^c, x_{2,k}^c, \dots, x_{d,k}^c\right) \in \Re^d$ olarak elde edilir ve

$$\mu_{k,i}(\mathbf{x}) = \left(\sum_{l=1}^{c} \left({d_{k,i}(\mathbf{x})} \Big/ {d_{l,i}(\mathbf{x})} \right)^{2/m-1} \right)^{-1} \tag{5.17}$$

$$d_{k,i}(\mathbf{x}) = \|\mathbf{x}_i - \mathbf{v}_k(\mathbf{x})\|;\ i = 1, 2, \ldots, n;\ k = 1, 2, \ldots, c \tag{5.18}$$

eşitlikleri kullanılarak girdi uzayı için üyelik değerleri hesaplanır.

Adım 3. $\mu_{k,i}$ üyelik değerleri ve $\mu_{k,i}^2$, $\exp \mu_{k,i}$, $\ln \mu_{k,i}$ gibi bunların kullanıcı tanımlı matematiksel dönüşümleri, her bir küme için orijinal girdiler ile birleştir. Burada, her bir k için,

$$\mu_{k,i} > \alpha\text{–kesim},\ k = 1, 2, \ldots, c;\ i = 1, 2, \ldots, l;\ l < n \tag{5.19}$$

kısıtı ile farklı $\alpha -$ kesim seçimi için farklı bir alt küme elde edilebilir.

α–kesim> 0 kullanımı ile belirli bir kümeye olan üyelik derecesi küçük olan gözlemler o kümede ihmal edilmiş olur böylece kümelerde model tahmininde bazı gözlemler kullanılmaz. Bu durum özellikle büyük veri durumunda kümelere düşen gözlem sayılarını azaltmak üzere etkili bir yöntem sağlar. Deneysel sonuçlara göre, eğer kümedeki gözlemlerin sayısı n/c'nin altında kalırsa, genellikle α-kesim = 0 olarak alınır (Çelikyılmaz ve Türkşen 2009).

Adım 4. Her bir küme için bulanık sınıflandırma modelleri LR, DVM, kNN, KA, RO, GNB ve ANN makine öğrenmesi yöntemleri ile tahmin edilir.

Adım 5. Her bir küme için bulanık sınıflandırıcıdan elde edilen çıktı değerleri, bir sigmoid fonksiyonu kullanılarak sonsal olasılıklara (posterior probability) dönüştürülür (Kuhn 2008).

Adım 6. Her bir küme için elde edilen bulanık çıktılar, bu çıktılara karşışık gelen üyelik değerleri ile ağırlıklandırılarak tek bir çıktı değeri,

$$\hat{P}_i = \frac{\sum_{k=1}^{c} \hat{P}_{k,i}\, \mu_{k,i}}{\sum_{k=1}^{c} \mu_{k,i}},\ k = 1, 2, \ldots, c;\ i = 1, 2, \ldots, n \tag{5.20}$$

biçiminde hesaplanır.

Adım 7. Sonsal olasılıklara dayalı olarak belirli bir BRF modeli için optimum eşik değeri hesaplanır. Optimum eşik değeri, her model için eğitim veri setinde doğru sınıflandırma oranı en yüksek olacak şekilde iteratif olarak belirlenir.

BRF yaklaşımı için önerilen test aşaması, yapı tanımlaması sırasında belirlenen modellerin test veri seti üzerindeki genelleme yeteneklerini değerlendirmek için aşağıdaki beş adım ile sunulmuştur:

Adım 1. Küme merkezleri kullanılarak her girdi için üyelik değerleri hesaplanır.

Adım 2. Bir önceki adımda elde edilen üyelik değerlerinin dönüşümleri ile girdiler birleştirilir.

Adım 3. Bulanık sınıflandırıcı tahminleri kullanılarak yeni girdiler için sınıf olasılıkları hesaplanır.

Adım 4. Bir girdi için kümelerden elde edilen sonsal olasılıklar, karşılık gelen üyelik değerleri ile ağırlıklandırılarak tek bir çıktı hesaplanır.

Adım 5. İkili çıktıdan birini temsil eden her veri nesnesi için bir sınıf etiketi, $\hat{y}_i = 0$ ya da $\hat{y}_i = 1$, tahmin etmek için yapı tanımlaması sırasında elde edilen optimum eşik değerleri kulanılır. Böylece modelin doğruluğu açısından performansları, tahmin edilen sınıf etiketlerine göre değerlendirilir.

6. İKİLİ SINIFLANDIRMA PROBLEMLERİNDE DENGESİZ VERİ SETLERİ İÇİN YENİDEN ÖRNEKLEME STRATEJİLERİ

Bilim ve teknoloji dünyasında yaşanan gelişmeler ham verilerin büyümesinde ve kullanılabilirliğinin artmasında etkin bir rol oynamıştır. Bu da dengesiz öğrenme sorununun akademi, endüstri ve hükümet finansman kurumlarında önemli miktarda ilgili görmesini sağlamıştır (He ve Garcia 2009). Gerçek dünya problemlerinde ve literatürde, genellikle ikili sınıflandırmanın tespit edilecek iki sınıftan birine ait örneklerinin sayısı diğerinden çok daha az olduğu ve dengesiz olarak adlandırılan bir veri seti üzerinde yapılması gerekir (Cateni vd. 2014). Nadir olaylar, özellikle de toplumu olumsuz etkileme potansiyeline sahip olanlar, genellikle insanların karar verme tepkilerini gerektirmektedir. Bu olayların tespit edilmesi, veri madenciliği ve makine öğrenmesi topluluklarında bir tahmin görevi şeklinde görülebilir. Bu olaylar günlük hayatta nadir olarak görüldüğünden, bu tahmin görevlerinde veri eksikliği sorunu oluşmaktadır (Haixiang vd. 2016).

Dengesiz veri, sınıflardan birinin veya bazılarının diğerlerinden çok daha fazla örneğe sahip olduğu veri kümesini ifade eder (Haixiang vd. 2017). Sınıf dengesizliği, destek vektör makineleri, karar ağaçları ve sinir ağları gibi tüm yaygın sınıflandırıcıların performansını olumsuz etkilemektedir. Aslında bu sınıflandırıcıların standart versiyonları, tüm veri setindeki genel performansı optimize etmek için tasarlanmıştır (Cateni vd. 2014). Çoğu standart algoritma, dengeli sınıf dağılımları veya eşit yanlış sınıflandırma maliyetleri olduğunu varsayar. Bu nedenle karmaşık dengesiz veri kümeleri ile çalıştığında, bu algoritmalar verilerin dağılım özelliklerini doğru şekilde temsil etmekte başarısız olur ve sonuç olarak veri sınıfları arasında dengesiz sınıflandırma doğrulukları ortaya çıkarır (He ve Garcia 2009). Veri madenciliği yaklaşımları, ticari ve yönetimsel karar vermeye rehberlik edecek sınıflandırma modelleri oluşturmak için yaygın şekilde kullanılmasına karşın, dengesiz verilerin sınıflandırılması geleneksel sınıflandırma modellerini önemli ölçüde etkilemektedir (Haixiang vd. 2017).

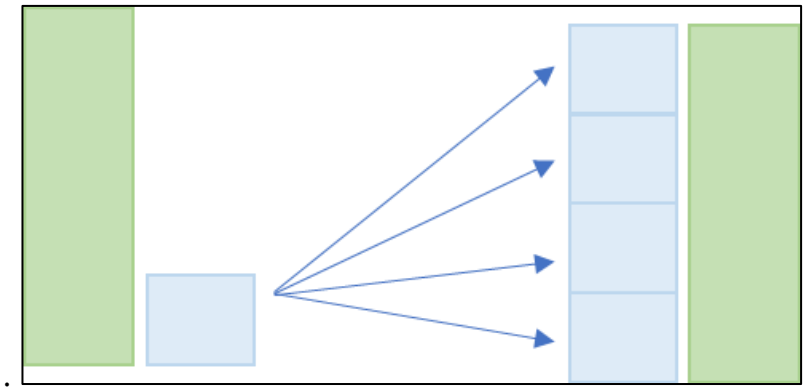
6.1 Dengesiz Öğrenme için Örnekleme Yöntemleri

Tipik olarak, dengesiz öğrenme uygulamalarında örnekleme yöntemlerinin kullanımı, dengeli bir dağılım sağlamak için dengesiz veri setinin bazı mekanizmalar tarafından değiştirilmesinden oluşur. Çalışmalar birkaç temel sınıflandırıcı için dengeli bir veri setinin, dengesiz bir veri setine kıyasla gelişmiş genel sınıflandırma performansı sağladığını göstermiştir. Bu da dengesiz öğrenmede örnekleme yöntemlerinin kullanılmasının gerekliliğini ortaya koymaktadır (He ve Garcia 2009).

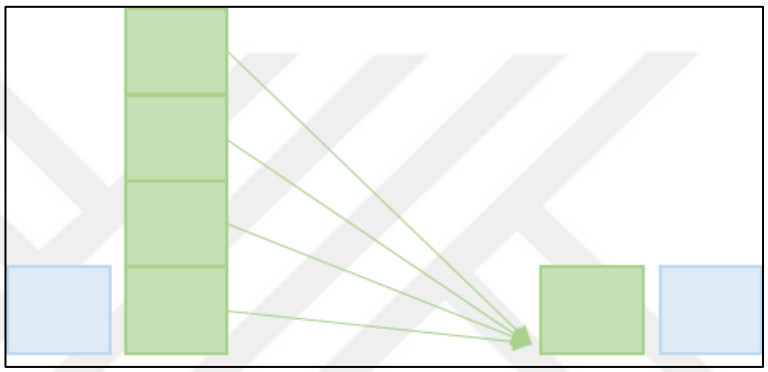
6.1.1 Rasgele aşırı örnekleme ve eksik örnekleme yöntemi

Dengesiz veri setlerinde örnekleme yöntemi, rasgele eksik örmekleme ve rasgele aşırı örnekleme olmak üzere iki farklı yaklaşım ile uygulanabilmektedir. Eksik örnekleme, her sınıfın gözlem sayılarını eşitlemek amacıyla çoğunluk sınıfına ait gözlemleri azaltarak veri sayısını düşürmeyi hedeflemektedir. Aşırı örnekleme ise; azınlık sınıfının ağırlığını arttırabilmek için yeni rasgele gözlemler üretmeyi veya çoğaltmayı amaçlayan bir yöntemdir (Fernandez vd. 2018).

Sınıflandırma problemlerinde klasik yöntemlerden olan rasgele eksik örnekleme ve rasgele aşırı örnekleme sezgisel olmayan yöntemlerdir. Rasgele eksik örnekleme, dengeli bir veri seti oluşması için çoğunluk sınıfından rasgele elemeler yaparak sınıf dağılımını dengelemeyi hedefleyen bir yöntemdir. Burada en son dengeleme oranı ayarlanabilmektedir (Fernandez vd. 2018, Masood ve Begum 2022). Rasgele aşırı örneklemede ise; azınlık sınıfı örneklerinin rasgele kopyalanması yoluyla sınıf dağılımının dengelenmesi amaçlanmaktadır (Fernandez vd. 2018). Aşırı örnekleme yöntemi Şekil 6.1'deki gibi, eksik örnekleme yöntemi ise Şekil 6.2'deki gibi görselleştirilebilir.



Şekil 6.1 Aşırı Örnekleme Yöntemi Gösterimi



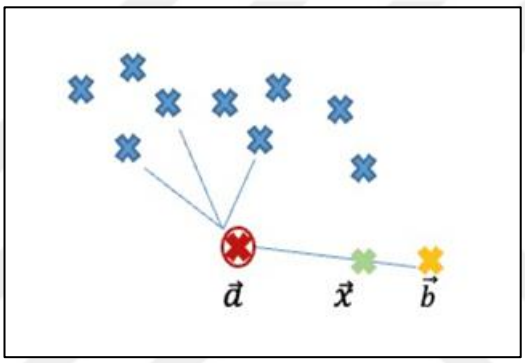
Şekil 6.2 Eksik Örnekleme Yöntemi

Rasgele eksik örnekleme yaklaşımının devamı olarak; çoğunluk sınıfından gözlemler çıkarma konusunda başka yaklaşımlar mevcuttur. Bunlarda çıkarmak için gereksiz gözlemleri ya da çıkarmamak için gerekli gözlemleri tanımlamaya çalışan yöntemler veya modeller de bulunmaktadır.

6.1.2 SMOTE (synthetic minority oversampling technique-sentetik azınlık aşırı örnekleme tekniği)

SMOTE, tahmin modelinin sınıflandırma doğruluğunu arttırmak için yaygın kullanılan örnekleme tekniklerinden biridir (Dang vd. 2021). Bu teknik, azınlık sınıfının mevcut örneklerini çoğaltmak için kullanılır ve azınlık sınıfındaki her bir örneği çevreleyen komşular hakkında bir yapay örnekler oluşturmasını sağlar. SMOTE, azınlık sınıfı örneklerinin rasgeleleştirilmesi, her bir azınlık sınıfı örneği için k-en yakın komşunun

hesaplanması ve sentetik örnekler üretilmesi olarak bilinen üç bölüme ayrılmıştır (Dang vd. 2021). Yöntemin ana fikri temel olarak Şekil 6.2'de gösterilmiştir. Azınlık sınıfından rasgele bir ($\vec{a}$) girdi vektörü (örnek) alınır ve yine azınlık sınıfından bu girdinin en yakın ($\vec{b}$) komşusu seçilir. Bu iki özellik arasındaki fark alınır ve 0-1 arasında rasgele bir ağırlık (w) ile çarpılır ve dikkate alınan ($\vec{a}$) özellik vektörüne eklenir. Bu şekilde seçilen sentetik örnek ($\vec{x}$), azınlık sınıfından seçilen bir rasgele örneklem olan ($\vec{a}$) ve ($\vec{b}$) nin doğrusal bir kombinasyonu şeklinde hesaplanmış olur ve belirli iki özellik arasındaki çizgi boyunca rasgele bir noktanın seçilmesini sağlar (Chawla vd. 2002, Aydın 2022). Şekil 6.2'de SMOTE'nin k=4 ve w =0,7 rasgele seçilmiş ağırlık için gösterimi yer almaktadır.



Şekil 6.3 SMOTE Gösterimi (k=4, w=0,7)

6.1.3 ADASYN (adaptive synthetic sampling- uyarlanabilir sentetik örnekleme) tekniği

He vd. (2008) tarafından SMOTE ve SMOTE tabanlı metotların başarısından yola çıkılarak, dengesiz veri setlerinde öğrenmeyi kolaylaştırmak için uyarlanabilir yeni bir yöntem önerilmiştir. Bu yöntemin iki temel amacı vardır: yanlılığı azaltmak ve uyarlanabilir bir öğrenme sağlamak. İki sınıflı bir sınıflandırma problemi için önerilen ADASYN tekniğinin algoritma adımları aşağıda belirtildiği şekildedir (He vd. 2008).

x_i, n boyutlu özellik uzayı X içinde bir örnek ve $y_i \in Y = \{1, -1\}$ x_i ile ilişkili sınıf kimlik etiketi olmak üzere D_{tr}; m örnekli $\{x_i, y_i\}$ i=1,…,m bir eğitim veri seti. m_l çoğunluk sınıfının, m_s ise azınlık sınfının örnek sayısını göstersin. Bunun için; $m_s \leq m_l$ ve $m_s + m_l = m$ dir.

Adım 1. Sınıf dengesizliği derecesi

$$d = \frac{m_s}{m_l} \tag{6.1}$$

biçiminde hesaplanır. Burada $d \in (0,1]$ dir.

Adım 2. Eğer $d < d_{th}$ (d_{th} önceden belirlenmiş kabul edilebilecek en yüksek sınıf dengesizlik oranı) ise;

a. Azınlık sınıfı için üretilmesi gereken sentetik veri örneklerinin sayısı

$$G = (m_l - m_s)x\ \beta \tag{6.2}$$

eşitliğindeki gibi hesaplanır. Burada $\beta \in [0,1]$ sentetik veri üretiminden sonra arzu edilen denge seviyesini tanımlamak için kullanılmaktadır. Eğer $\beta = 1$ ise, veri üretme sürecinden sonra tamamen dengeli bir veri setinin oluşturulmuş demektir.

b. Her bir $x_i \in azınlıksınıfı$ örneği için, n boyutlu bir uzayda Öklid uzaklığı ile k en yakın komşular bulunur ve

$$r_i = \frac{\Delta_i}{K},\ i = 1, \dots, m_s \tag{6.3}$$

eşitliği ile r_i oranı hesaplanır. Burada, Δ_i, çoğunluk sınıfına ait olan x_i'nin k en yakın komşuluklarındaki örnek sayısını göstermekte olduğundan $R_i \in [0,1]$.

c. r_i'nin normalleştirilmesi

$$\hat{r}_i = \frac{r_i}{\sum_{i=1}^{m_s} r_i} \tag{6.4}$$

eşitliğindeki gibidir. Böylece $\hat{r}_i$ bir yoğunluk fonksiyonudur ($\sum_i \hat{r}_i = 1$).

d. Her bir azınlık örneği için x_i için üretilmesi gereken sentetik veri örneklerinin sayısı

$$g_i = \hat{r}_i \; x \; G \tag{6.5}$$

eşitliğindeki gibi hesaplanır. Burada G; (6.2) eşitliğinde tanımlandığı gibi azınlık sınıfı için üretilmesi gereken toplam sentetik veri örneklerinin sayısıdır.

e. Her bir azınlık sınıfı veri örneği x_i için, g_i sentetik veri örnekleri aşağıdaki adımlara göre hesaplanır. 1'den g_i'ye kadar döngü yapılmak üzere:

i. Veri x_i için K en yakın komşuluklardan rasgele bir azınlık veri örneği, x_{zi}, seçilir.
ii. Sentetik veri örneği

$$s_i = x_i + (x_{zi} - x_i) \; x \; \lambda \tag{6.6}$$

eşitliğindeki gibi üretilir. Burada $\lambda \in [0,1]$ olmak üzere rasgele bir sayı ve $(x_{zi} - x_i)$ n boyutlu uzayda bir fark vektörü olarak tanımlanmaktadır.

ADASYN algoritmasındaki ana fikir $\hat{r}_i$ yoğunluk dağılımını, her bir azınlık veri örneği için üretilmesi gereken sentetik örneklerin sayısına otomatik olarak karar vermede bir kriter olarak kullanmaktadır. Fiziksel olarak; $\hat{r}_i$ farklı azınlık sınıf örnekleri için ağırlık dağılımının öğrenmedeki zorluk seviyelerine göre bir ölçümüdür. ADASYN sonrası ortaya çıkan veri seti, yalnızca veri dağılımının dengeli bir temsilini sağlamakla kalmayacak (β katsayısı tarafından tanımlanan arzu edilen denge seviyesine göre) aynı zamanda öğrenme algoritmasını, öğrenmesi zor örneklere odaklanmaya zorlayacaktır. Bu da; her bir azınlık veri örneği için eşit sayıda sentetik örnekler üreten SMOTE algoritması ile ADAYSN algoritması karşılaştırıldığında en temel farktır.

6.1.4 Bilgilendirilmiş eksik örnekleme

Liu vd. (2009) tarafından önerilmiş olan bilgilendirilmiş eksik örneklemenin iyi sonuçlar veren iki örneği olarak Kolay Topluluk (KT - Easy Ensemble) ve Kademeli Denge (KD

- Balance Cascade) algoritmaları mevcuttur. Bu iki yöntemin amacı, geleneksel rasgele düşük örnekleme yönteminde ortaya çıkan bilgi kaybı eksikliğinin üstesinden gelmektedir (Liu vd. 2009)

KT algoritmasının uygulamasında; çoğunluk sınıfından birkaç alt kümeyi bağımsız olarak örnekleyerek ve her bir alt kümenin azınlık sınıf verisi ile kombinasyonuna dayalı çoklu sınıflandırıcılar geliştirerek bir topluluk öğrenme sistemi elde edilir. Bu şekilde KT, yer değiştirmeli bağımsız rasgele örnekleme kullanılarak çoğunluk sınıfı verilerini araştıran denetimsiz bir öğrenme algoritması olarak düşünülebilir. Öte yandan KD algoritması, hangi çoğunluk sınıfı örneklerinin yetersiz örnekleneceğini sistematik olarak seçmek için bir sınıflandırıcı topluluğu geliştiren denetimli bir öğrenme yaklaşımı benimsemektedir (He vd. 2008). Bu yöntem sırayla temel öğrenicileri yaratır ve önceki öğreniciler tarafından filtrelenen örnekler üzerinde yeni öğrenicileri oluşturur. İlk etapta, çoğunluk sınıfının bir kısmını ve azınlık sınıfının tamamını içeren bir alt küme örneği üzerinde bir öğrenici eğitilir. Akabinde çoğunluk sınıfından örneklenen alt küme elenir. Böylece arttırılan çoğunluk sınıfından elde edilen bir alt küme ile azınlık sınıfı kullanılarak yeni bir topluluk öğrenicisi yaratılır. Süzülen bu yeni veri seti örneklerinde yeni öğreticiler yinelemeli olarak yaratılır ve son kısımda tüm öğreniciler birleştirilir. Bu yöntem, doğru şekilde modellenen örneklerin bir sonraki sınıflandırıcıyı oluşturmada bir faydasının olmadığını varsayar (Sarmanova 2013).

6.1.5 Küme tabanlı örnekleme metodu

Küme tabanlı örnekleme algoritmaları, çoğu örnekleme algoritmalarında bulunmayan ek bir esneklik unsuru sağladıkları için özellikle ilgi çekmektedir. Jo ve Japowicz (2004); küme tabanlı aşırı örnekleme yöntemini; hem sınıf içi hem sınıflar arası dengesizlik problemi ile etkili şekilde başa çıkabilmek için önermiştir. Küme tabanlı aşırı örnekleme algoritması, k-ortalama kümeleme tekniğini kullanmaktadır (Jo ve Japowicz 2004). Bu yöntem, her kümeden (her iki sınıf için) rasgele bir k örnek kümesi alır ve bu örneklerin küme merkezi olarak belirlenen ortalama özellik vektörünü hesaplar. Sonra, kalan eğitim örnekleri birer birer sunulur ve her örnek için kendisi ile her bir küme merkezi arasındaki Öklid uzaklık vektörü hesaplanır. Her eğitim örneği daha sonra en küçük mesafe

vektörünü içeren kümeye atanır. Son olarak tüm küme araçları güncellenir ve süreç tüm örnekler bitene dek tekrar eder (her örnek için yalnızca bir küme ortalaması güncellenir (He ve Garcia 2009).

6.1.6 Dengesiz öğrenme için maliyete duyarlı metotlar

Örnekleme yöntemleri, dağılımdaki sınıf örneklerinin temsili oranlarını dikkate alarak dağılımları dengelemeye çalışırken, maliyete duyarlı öğrenme yöntemleri örneklerin yanlış sınıflandırılmasıyla ilişkili maliyetleri dikkate alır (Elkan 2001). Maliyete duyarlı öğrenme, farklı örnekleme stratejileri yoluyla dengeli veri dağılımları oluşturmak yerine, herhangi bir veri örneğini yanlış sınıflandırmanın maliyetlerini tanımlayan farklı maliyet matrisleri kullanarak dengesiz öğrenme problemini çözümler (Elkan 2001, Ting 2002). Son araştırmalar, maliyete duyarlı öğrenme ile dengesiz verilerden öğrenme arasında güçlü bir bağlantı olduğunu göstermektedir; bu nedenle, maliyete duyarlı yöntemlerin teorik temelleri ve algoritmaları doğal olarak dengesiz öğrenme problemlerine uygulanabilir (He ve Garcia 2009). Ayrıca çeşitli çalışmalarda, belirli dengesiz öğrenme alanları dahil olmak üzere bazı uygulama alanlarında, maliyete duyarlı öğrenmenin örnekleme yöntemlerinden daha üstün olduğunu göstermiştir (He ve Garcia 2009).

6.1.7 Çekirdek tabanlı öğrenme

Çekirdek tabanlı öğrenmenin ilkeleri, istatistiksel öğrenme teorileri ve Vapnik-Chervonenkis (VC) boyutlarına odaklanır (He ve Garcia 2009). Temsili çekirdek tabanlı öğrenme paradigması, destek vektör makineleri (DVM), dengesiz veri kümelerine uygulandığında nispeten sağlam (robust) sınıflandırma sonuçları sağlayabilir. DVM'ler, destek vektörleri ile varsayılan kavram sınırı (hiperdüzlem) arasındaki ayrım marjını (yumuşak marj maksimizasyonu) en üst düzeye çıkarmak ve bu arada toplam sınıflandırma hatasını en aza indirmek için kavram sınırlarının (destek vektörleri) yakınında belirli örnekler kullanarak öğrenmeyi kolaylaştırır (He ve Garcia 2009). Dengesiz verilerin DVM'ler üzerindeki etkileri, yumuşak marj maksimizasyon paradigmasının yetersizliklerinden yararlanır (Akbani vd. 2004, Raskuuti ve Kowalczyk 2004). DVM'ler toplam hatayı en aza indirmeye çalıştıklarından, doğal olarak çoğunluk

kavramına doğru önyargılıdırlar. En basit durumda, iki sınıflı bir mekan, çoğunluk kavramının komşuluğunda "ideal" bir ayırma çizgisiyle doğrusal olarak ayrılır. Bu durumda azınlık kavramını temsil eden destek vektörlerinin bu ideal çizgiden "uzak" olması ve sonuç olarak nihai hipoteze daha az katkı sağlaması söz konusu olabilir (He ve Garcia 2009). Ayrıca, azınlık kavramını temsil eden veri eksikliği varsa, temsili destek vektörlerinde de performansı düşürebilecek bir dengesizlik olabilir. Bu aynı özellikler, doğrusal olmayan sınıflandırma için oldukça belirgindir. Bu amaçla, yüksek boyutlu bir nitelik uzayına bazı doğrusal olmayan dönüşümler uygulanır. Girdi uzayında doğrusal olmayan karar sınırlarını göstermek üzere nitelik uzayında oluşturulmuş bir doğrusal model mevcuttur. Ancak bu durumda, sınıfları ayıran optimal hiperdüzlem, daha yaygın çoğunluk sınıfının yanlış sınıflandırılmasından kaynaklanan yüksek hata oranlarını en aza indirmek için çoğunluk sınıfına doğru eğimli olacaktır. En kötü durumda, DVM'ler tüm örnekleri çoğunluk sınıfına ait olarak sınıflandırmayı öğrenecektir; bu teknik, dengesizlik şiddetliyse, veri alanı genelinde minimum hata oranını sağlayabilmektedir (He ve Garcia 2009).

7. UYGULAMA

Klasik BKO kümeleme yöntemini kullanan BRF yaklaşımı; bulanık yapı tanımlama (eğitim) ve bulanık çıkarım (test) olarak iki aşamada ele alınmaktadır. BRF yaklaşımının algoritmik olarak sunumu Şekil 5.2'de yer almaktadır. Bu bölümde, konut kredilerinde risk notlandırılmasını gerçekleştirmek amacıyla BRF modellerinden ve klasik makine öğrenmesi algoritmalarından elde edilen bulgular sunulmuştur. Ayrıca, %10 oranında riskli (default) gözlem içeren bir test veri setinde model performansının iyileştirilmesi amacıyla yeniden örnekleme stratejileri arasında yer alan eksik örnekleme sürecinden de faydalanılmıştır. Konut kredisi veri setinde yer alan değişkenler ve özelliklerinin incelenmesinin ardından değişken seçimi, model parametrelerinin seçimi ve model tahminine ilişkin elde edilen bulgular özetlenmiştir.

7.1 Konut Kredisi Verisi

Çalışmada açık erişimde bulunan ve Kaggle tarafından sunulan Home Credit tarafından sağlanan veriler kullanılmıştır. (Veri Kaynak: https://www.kaggle.com/competitions/home-credit-default-risk/data). Home Credit grubu 1997 yılında kurulan ve hâlihazırda 9 ülkede operasyonlarını yürütmekte olan bir kuruluştur. Kuruluşun temel amacı, kredi geçmişi olmayan veya çok az olan kişilere kredi sağlamak olup, bu çerçeveden değerlendirildiğinde müşteri profillerinin daha riskli bir grup olduğu söylenebilecektir. Kuruluş, faaliyetleri kapsamında ellerinde bulunan müşterilerine ait gerçek verileri erişime açık bir platform olan Kaggle üzerinden bir yarışma ile sunmuştur. Veri seti 307511 gözlemden, 50 adet kategorik ve 70 adet sürekli olmak üzere toplam 120 adet özellikten oluşmaktadır. Home Credit tarafından sunulan veri setinde müşterilerine ait birçok bilgi yer almakta olup, bunlardan bazıları kişilerin arabasının/evinin olup olmadığı, cinsiyeti, gelir düzeyi vb. şeklindedir.

7.2 Veri Keşfi ve Değişken Seçimi

Bankalar kredi kullandırdıkları müşterileri hakkında oldukça titiz bir çalışma yürütmekte ve başvuru sahibi hakkında birçok bilgiyi kayıt altına almaktadır. Bu verilerin dijital

ortama aktarımı sırasında birçok kişisel veri ise gizlenmektedir. Böylece kayıp gözlemlerin sayısı artmakta ve bu nedenle temin edilen verilerde kapsamlı bir veri temizleme aşaması gerçekleştirilmesi gerekebilmektedir.

Bu çalışmada konut kredisi veri setinde birçok eksik ve tutarsız veri olması nedeniyle bir veri temizliği çalışması gerçekleştirilmiştir. Bir değişken için eksik veri oranı %95'in üzerinde ise o değişken veri setinden çıkarılmıştır. %95'in altında ise boş hücreler 1, dolu hücreler 0 olacak şekilde kod içeren bir değişken yeniden tanımlanmıştır. Değişkenler ve gözlemler arasında kontroller vasıtasıyla mümkün olduğunca daha fazla örneklem hacmine ulaşmak üzere veri temizliği süreci gerçekleştirilmiştir. Böylece 167732 sayıda gözlem içeren bir veri setine ulaşılmıştır. Veri setinde yer alan kategorik değişkenler için model tahmininde kullanılmak üzere kukla değişkenler tanımlanmıştır.

Sınıflandırıcıların tahmin performansını arttırmak, daha uygun maliyetli ve hızlı sınıflandırıcılar sağlamak amacıyla değişken seçiminde Filtre Yöntemi ve Geriye Doğru Eleme Yöntemi kullanılmıştır. Bu adım özellikle gerçek süreçte uygulanabilecek bir risk değerlendirme modeli oluşturmak için kullanılan önemli değişkenlerin neler olabileceği hakkında fikir vermesi açısından da önem arz etmektedir.

Veri özelliklerine göre değişken seçimi gerçekleştiren filtre yönteminde, önsel olarak herhangi bir sınıflandırma algoritması kullanımına ihtiyaç yoktur (Liu ve Motoda, 2007). Fisher Skoru, denetimli bir değişken (öznitelik) seçim algoritmasıdır ve birçok modelleme probleminde geniş çapta uygulanmıştır (Gu vd. 2011). Aynı sınıftaki gözlemler için değerleri daha düzgün dağılan; farklı sınıflardaki gözlemler için benzer olmayan değişkenleri seçmek için tasarlanmıştır. Bu çalışmada, Fisher skoruna göre önemli değişkenleri belirlemek için bir eşik değeri kullanılmıştır. Ardından, geriye doğru eleme yöntemine dayalı olarak kredi riski üzerinde etkisi istatistiksel olarak anlamlı değişkenler seçilmiştir.

Bu çalışmada kullanılan konut kredisi verisi yapılan veri temizliği sonrası 167732 adet gözlemden oluşmakta olup, özelliklerin 16'sı kategorik, 14'ü ise süreklidir. Bu verilerin 153535 adedi 0 (temerrüde düşmeyen) ve 14207 adedi ise 1 (temerrüde düşen)

gözlemlerden oluşmaktadır. Temerrüde düşen (1) veri sayısının, temerrüde düşmeyenlere (0) oranının %9,25 olduğu görülmüştür. Veri seti incelendiğinde asıl tespiti istenen ve temerrüde düşme riski taşıyan verilerin azınlıkta olduğu ve veri setinin dengesiz olduğu görülmektedir. Bu çalışmada Çelikyılmaz ve Türkşen (2009) tarafından önerilen ve Başer vd. (2023) tarafından geliştirilen bulanık sınıflandırma yöntemi, yeniden örnekleme ile entegre edilmiş dengesiz dağılımın model performansı üzerindeki etkisi değerlendirilmiştir. Değişken seçimi sonucunda belirlenen değişkenler ve karakteristikleri Çizelge 7.1'de sunulmuştur.

Çizelge 0.1 Değişken seçimi sonucunda belirlenen değişkenler ve özellikleri

Değişken	**Tür**	**Değişken**	**Tür**
Kontrat Tipi	Kategorik	Kimlik düzenlenme zamanından geçen süre (gün)	Sürekli
Cinsiyet	Kategorik	İş Telefonu	Kategorik
Araba Sahipliği	Kategorik	Telefon	Kategorik
Mülkiyet Durumu	Kategorik	Eposta	Kategorik
Toplam Geliri (log)	Sürekli	Müşterinin yaşadığı bölgenin derecesi	Sürekli
Kredi Tutarı	Sürekli	Adresin statü ile eşleşmesi (bölge)	Kategorik
Yıllık Ödeme	Sürekli	Adresin statü ile eşleşmesi (şehir)	Kategorik
Mal Fiyatı	Sürekli	Dış veri kaynağından alınan skor 2	Sürekli
Eğitim Düzeyi	Kategorik	Dış veri kaynağından alınan skor 3	Sürekli
Ev Tipi	Kategorik	Müşterinin Sosyal Çevresinde Temerrüde Düşenlerin Sayısı	Sürekli
Yaş	Sürekli	En son telefon numarasu değiştirme zamanı (gün)	Sürekli
Çalıştığı Gün Sayısı	Sürekli	Sağlanan Dökümanlar (3,5,8,16,18)	Kategorik
Kayıtlı Gün Sayısı	Sürekli	Kredi Bürosunda Yapılan Sorgulamaların Sayısı	Sürekli

7.3 Model Seçim Kriterleri

Bir sınıflandırıcının gözlemlerin sınıf etiketlerinin tahmin etmede ne kadar iyi ya da ne kadar doğru olduğunu değerlendirmek için doğruluk oranı (accuracy ratio-AR), duyarlılık (sensitivity) ve özgüllük (specificity) oranları, Eğri Altında Alan (Area Under Curve – AUC) yaygın olarak kullanılan bazı kriterlerdir. Bu kriterleri açıklamak için öncelikle sınıflandırma doğruluğu matrisinin ne olduğu incelenmelidir (Çizelge 7.2).

Çizelge 0.2 Sınıflandırma hatası matrisi

		Gerçek Sınıf (True Class)	
		Pozitif	**Negatif**
Sınıf Kestirimi (Predicted Class)	**Pozitif**	**TP (Gerçek Pozitif)**	**FP (Yanlış Pozitif)**
	Negatif	**FN (Yanlış Negatif)**	**TN (Gerçek Negatif)**

Gerçek pozitif (TP): Veri setinde, sınıflandırma modeli tarafından pozitif olarak sınıflandırılan ve gerçekte de pozitif sınıfta yer alan gözlem sayısıdır.

Gerçek negatif (TN): Veri setinde, sınıflandırma modeli tarafından negatif olarak sınıflandırılan ve gerçekte de negatif sınıfta yer alan gözlemlerin sayısıdır.

Yanlış pozitif (FP): Veri setinde, sınıflandırma modeli tarafından pozitif olarak sınıflandırılan ancak gerçekte negatif sınıfa ait olan gözlemlerin sayısıdır.

Yanlış negatif (FN): Veri setinde, sınıflandırma modeli tarafından negatif olarak sınıflandırılan ancak gerçekte pozitif sınıfa ait olan gözlemlerin sayısıdır.

Doğruluk (Accuracy): Eğitim verileri kullanılarak kurulan modelin, test verilerini doğru sınıflandırma oranıdır. Doğruluk oranı,

$$\text{Doğruluk} = \frac{TP+TN}{TP+TN+FP+FN} \tag{7.8}$$

biçiminde tanımlanır.

Duyarlılık (Sensitivity): Gerçekte pozitif olan verilerin pozitif sınıfa atanması yani doğru sınıflandırılması oranıdır. Duyarlılık oranı,

$$\text{Duyarlılık} = \frac{TP}{TP+FN} \tag{7.9}$$

biçiminde tanımlanır.

Özgüllük (Specificity): Gerçekte negatif olan verilerin negatif sınıfa atanması yani doğru sınıflandırılması oranıdır. Özgüllük oranı,

$$\text{Özgüllük} = \frac{TN}{TN+FP} \tag{7.10}$$

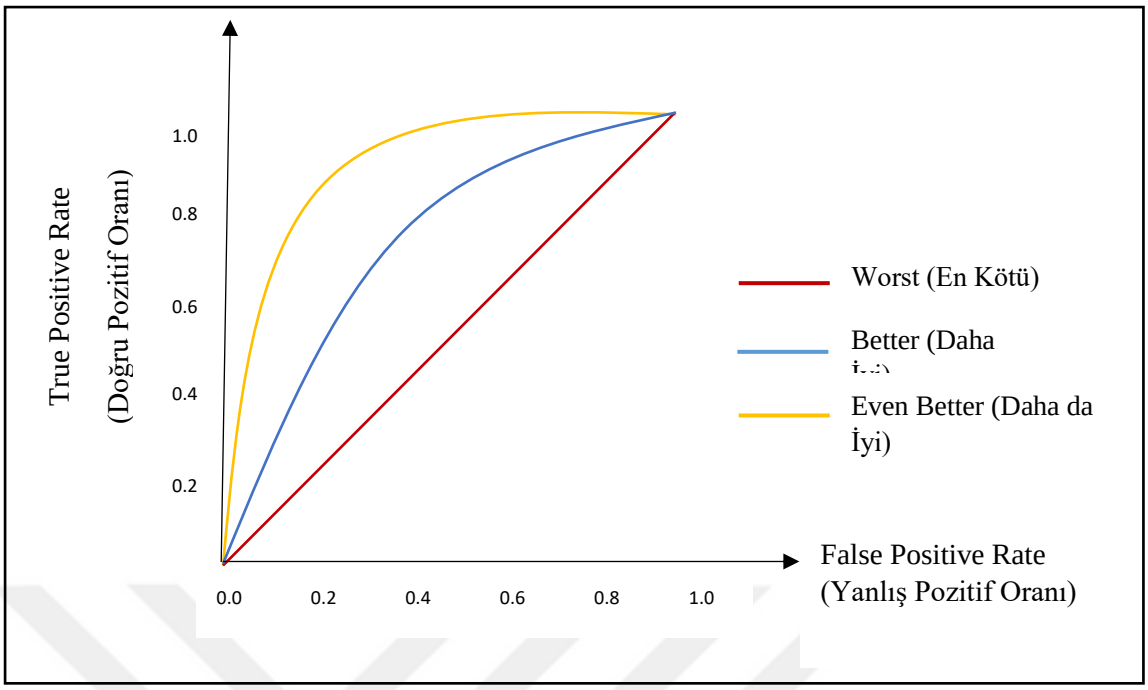
biçiminde tanımlanır.

G-ortalama (G-mean): G-ortalama, duyarlılık ile özgüllük oranının geometrik ortalamasıdır. Bu ölçüt, duyarlılık ve özgüllük ölçütlerini dengede tutmakta; bir yandan da sınıftaki doğruluğu en üst noktaya taşımaya çalışmaktadır. Özellikle, dengesiz sınıflandırma problemlerinde duyarlılık ve özgüllük değerlerinde yaşanan olumsuzlukları dengeleyerek tek bir kriter ile değerlendirilmesini sağlamaktadır. G-Ortalama,

$$\text{G} - \text{Ortalama} = \sqrt{\text{Duyarlılık x Özgüllük}} \tag{7.11}$$

biçiminde hesaplanmaktadır.

ROC eğrisi ve AUC (receiver operating characteristic curve – area under curve): İkili sınıflandırma sistemi içerisinde ROC Eğrisi doğru pozitiflerin yanlış poazitiflere oranı olarak ifade edilmektedir. ROC eğrisi altında kalan alan (AUC, Area Under Curve) bir parametrenin iki grubu ne kadar iyi ayırabildiğinin bir ölçüsüdür. ROC değeri 1'e yaklaştıkça pozitiflerin negatiflerden daha iyi şekilde ayrıldığı anlaşılır. Şekil 7.1'te ROC eğrisi gösterilmiştir.



Şekil 0.1 ROC Eğrisi

7.4 Model Parametrelerinin Optimum Değerlerinin Belirlenmesi

Bu çalışmada analizler R programlama dili kullanılarak gerçekleştirilmiştir. BKO kümeleme algoritmasının uygulaması "fclust" paketi ile gerçekleştirilmiştir (Ferraro vd. 2019). BKO yöntemi için küme sayısı 5; bulanıklık derecesi varsayılan değer olan 2 olarak seçilmiştir.

Denetimli sınıflandırma modelleri Rasgele Orman, Karar Ağaçları ve Destek Vektör Makineleri için sırasıyla "randomForest", "rpart" (Breiman 2001), "e1071" (Meyer vd. 2023) paketleri kullanılmıştır. Ayrıca yapay sinir ağları, k en yakın komşuluk ve lojistik regresyon model tahmini için de "caret" (Kuhn 2008) paketinden faydalanılmıştır. Optimum parametre değerleri, ızgara arama kullanılarak 10 tekrarlı ve 5 kat çapraz doğrulama ile belirlenmiş ve doğrulama metriği olarak ROC değeri seçilmiştir. Çizelge 7.3'te hiper parametrelerin açıklamaları ve arama uzayları yer almaktadır.

Çizelge 0.3 Hiper-parametrelerin tanımı ve seçimi

Algoritma	Hiper Parametre	Açıklama	Arama Uzayı
Rasgele Orman (RO)	İterasyon Sayısı	Ormandaki ağaç sayısı	[5, 40]
	Maksimum Derinlik	Tek bir ağacın maksimum derinliği	6
Karar Ağaçları (KA)	Karmaşıklık Parametresi	Genel uyum eksikliğini bir karmaşıklık parametresi kadar azaltmayan herhangi bir ayırma denenmemiştir.	[3.21e-4, 6.76e-3]
	Maksimum Derinlik	Ağacın maksimum derinliği	15
Yapay Sinir Ağları (YSA)	Gizli birim sayısı	Gizli katmandaki nöronların sayısı	10
	Bozulma	Düzenlileştirme parametresi	0.5
Destek Vektör Makineleri (DVM)	Çekirdek Fonksiyonu	Verileri daha fazla sayıda boyut uzaylarına dönüştürme fonksiyonu	{Radyal, Lineer}
	Gamma	Tek bir eğitim örneğinin hiper düzlem üzerindeki etkisini kontrol eder	1
K En Yakın Komşuluk (kNN)	k	Komşuların sayısı	[2, 43]
Naive Bayes (GNB)	Çekirdek	Sınıf tahmini için koşullu yoğunlukların kullanımı	{DOĞRU, YANLIŞ}

7.5 Modellerden Elde Edilen Bulgular

Çalışmanın bu kesiminde konut kredisi verileri üzerinde FKM bulanık kümeleme algoritmasına dayanan BRF yöntemi ile klasik makine öğrenmesi yöntemlerinden elde edilen bulgular değerlendirilecektir. Ayrıca çalışmada ele elınan problemlerden bir diğeri olan yeniden örnekleme yaklaşımınlarından eksik örnekleme durumunda iki yöntemden elde edilen model performans göstergelerinin değişimi de incelenecektir. Kredi riski modelleme yaklaşımları değerlendirildiğinde esasında kredi riskli bireyleri içeren azınlık sınıfına ilişkin doğru kestirimler model performansında belirleyici unsuru oluşturmaktadır.

Belirtilen amaçlar doğrultusunda gerçekleştirilen analizin ilk aşamasındaki veri keşfi ve değişkenlerin seçimi süreci sonucunda elde edilen veri seti özellikleri Çizelge 7.4'te sunulmuştur. Buna göre, kredi riskliliği üzerinde etkili olan değişkenlerin seçimi

sonucunda model tahmininde kullanılacak 30 bağımsız değişkenden 16'sı kategorik, 14'ü ise sürekli değişken tipindedir (Çizelge 7.1). Modelde yer alan kategorik değişkenler için kukla (dummy) değişken tanımlaması yapılmıştır. Veri setinde kredi riskli olan birimlerin ($y = 1$) toplam birim sayısına oranı %8.47 olarak elde edilmiş; bu durum verilerin kredi riskliliği durumuna göre dengesiz bir dağılım sergilediğini göstermektedir.

Çizelge 0.4 Veri özellikleri

Gözlem Sayısı	**Değişken Sayısı**	**Sınıf Sayısı ve Dağılımı**	**Kredi Riskli Birim Oranı**
167732	30	2 [153525/14207]	8.47%

Makine öğrenmesi algoritmaları test ve eğitim verisi olarak rasgele seçilmiş iki veri kümesi için tasarlanmış ve uygulanmıştır. Literatürde yer alan benzer çalışmalarda eğitim ve test verilerinin bölümlenmesi yaygın şekilde sırasıyla %80-%20 şeklinde olduğundan çalışmada da bu şekilde bölümlendirilmiştir. Eğitim verileri kullanılarak elde edilen model tahminleri, test verisi üzerinden ulaşılan model uyum iyiliği ölçütleri açısından karşılaştırılmıştır.

Yeniden örnekleme stratejilerinden olan eksik örnekleme yaklaşımına göre model eğitimi, rasgele seçilen beş farklı eğitim veri seti kullanılarak gerçekleştirilmiş ve model tahminleri tek bir test veri seti üzerinde değerlendirilmiştir. Buna göre ortaya çıkan beş farklı uygulamada, eğitim veri setlerindeki riskli ve riskli olmayan birim dağılımı Çizelge 7.5'te sunulmuştur. Bu dağılımlarda birinci uygulama riskli birim durumuna göre en dengesiz dağılımı gösterirken; beşinci uygulama ise dengeli dağılımı göstermektedir. Beş farklı eğitim veri seti kullanılarak gerçekleştirilen model tahminleri riskli birim oranı %10 olan bir test veri setinde değerlendirilmiştir. Burada elde edilen model bulgularına göre her bir uygulamada riskli birimlerin doğru kestirimi, model performansı açısından önemli bir gösterge olacaktır.

Birinci uygulamada, eğitim veri setinde riskli birim oranı %10 olacak biçimde örnekleme yapılmış ve BRF ve klasik makine öğrenmesi yöntemlerinin test verisi üzerindeki model doğruluk ölçütleri Çizelge 7.6'da sunulmuştur. Elde edilen model tahminlerinde AUC değerlerine göre LR ve BRF-LR yöntemlerinin; Doğruluk ölçütüne göre ise LR ve BRF-

kNN yöntemlerinin daha iyi sonuç verdiği gözlemlenmiştir. Ancak Özgüllük ve Duyarlılık değerleri incelendiğinde özellikle klasik yöntemlerin test verisinde sadece riskli olmayan bireyleri doğru sınıflandırabildiği; riskli bireyleri ise doğru sınıflandıramadığı görülmüştür. Özgüllük ölçütüne göre BRF-DVM_Radyal ve BRF_RO yöntemlerinin; G-Ort. değerlerine göre ise BRF-DVM_Radyal, BRF_RO ve BRF-LR yöntemlerinin iyi sonuç verdiği belirlenmiştir. Buna göre dengesiz veri setlerinde BRF yönteminin riskli bireylerin belirlenmesinde daha başarılı olduğu sonucuna ulaşılmıştır.

Çizelge 7.5 Yeniden örnekleme yapısı

	Eğitim Verisi (n = 20000)		**Test Verisi (n = 5000)**	
	Riskli birim sayısı	**Riskli olmayan birim sayısı**	**Riskli birim sayısı**	**Riskli olmayan birim sayısı**
Uygulama-1	2000 **(%10)**	18000 **(%90)**	500 **(%10)**	4500 **(%90)**
Uygulama-2	4000 **(%20)**	16000 **(%80)**		
Uygulama-3	6000 **(%30)**	14000 **(%70)**		
Uygulama-4	8000 **(%40)**	12000 **(%60)**		
Uygulama-5	10000 **(%50)**	10000 **(%50)**		

İkinci uygulamada, eğitim veri setinde riskli birim oranı %20 olacak biçimde örnekleme yapılmış ve BRF ve klasik makine öğrenmesi yöntemlerinin test verisi üzerindeki model doğruluk ölçütleri Çizelge 7.7'de sunulmuştur. AUC değerlerine göre RO ve BRF-RO yöntemlerinin iyi sonuç verdiği belirlenmiştir. Birinci uygulamada olduğu gibi klasik yöntemlerden YSA, kNN, GNB ve DVM_Lineer'in riskli olmayan bireyleri doğru sınıflandırırken; riskli bireylerin ise tamamını yanlış sınıflandırdığı belirlenmiştir. Özgüllük ve G-Ort. ölçütlerine göre BRF-RO ve BRF-DVM_Radyal yöntemlerinin yüksek model performansına ulaştığı görülmektedir.

Üçüncü uygulamada, eğitim veri setinde riskli birim oranı %30 olacak biçimde örnekleme yapılmış ve BRF ve klasik makine öğrenmesi yöntemlerinin test verisi üzerindeki model doğruluk ölçütleri Çizelge 7.8'de sunulmuştur. AUC değerlerine göre ikinci uygulamaya benzer şekilde BRF-RO ve RO yöntemlerinin model performansı açısından güçlü olduğu belirlenmiştir. Özgüllük ve Duyarlılık ölçütleri birlikte değerlendirildiğinde YSA, kNN ve GNB yöntemlerinin yine riskli bireylerin sınıflandırılmasında başarısız oldukları belirlenmiştir. Bu durum önceki uygulamalardan elde edilen bulgular ile birleştirildiğinde YSA, kNN ve GNB modellerinin, riskli olmayan bireylerin doğru sınıflandırılması

açısından aşırı uyum sorunu barındırdığı biçiminde yorumlanabilir. G-Ort ölçütüne göre yine önceki uygulamalarda olduğu gibi BRF-RO ve BRF-DVM_Radyal yöntemlerinin yüksek sınıflandırma başarısına ulaştığı belirlenmiştir.

Çizelge 0.6 Riskli birim oranı %10 olan eğitim veri setinden (uygulama-1) elde edilen model tahminlerinin test veri seti üzerindeki performansı

Model	**Doğruluk**	**AUC**	**Duyarlılık**	**Özgüllük**	**G-Ort.**
BRF-DVM_Radyal	0.7635	0.6464	**0.4050**	0.8033	**0.5704**
BRF-RO	0.8565	0.7128	**0.2700**	0.9217	**0.4988**
BRF-LR	0.8995	**0.7497**	**0.0450**	0.9944	**0.2115**
LR	**0.9010**	**0.7535**	0.0300	0.9978	0.1730
BRF-KA	0.8960	0.6397	0.0150	0.9939	0.1221
DVM_Radyal	0.8995	0.6359	0.0100	0.9983	0.0999
BRF-kNN	**0.9005**	0.5865	0.0050	1.0000	0.0707
BRF-GNB	0.8995	0.7011	0.0006	0.9994	0.0245
BRF-DVM_Lineer	0.8995	0.5081	0.0006	0.9994	0.0245
RO	0.9000	**0.7257**	0.0000	1.0000	-
GNB	0.9000	0.7256	0.0000	1.0000	-
kNN	0.9000	0.5897	0.0000	1.0000	-
BRF-YSA	0.9000	0.5819	0.0000	1.0000	-
YSA	0.9000	0.5675	0.0000	1.0000	-
DVM_Lineer	0.9000	0.5514	0.0000	1.0000	-
DT	0.9000	0.5000	0.0000	1.0000	-

Çizelge 0.7 Riskli birim oranı %20 olan eğitim veri setinden (uygulama-2) elde edilen model tahminlerinin test veri seti üzerindeki performansı

Model	**Doğruluk**	**AUC**	**Duyarlılık**	**Özgüllük**	**G-Ort.**
BRF-RO	0.8035	**0.7647**	**0.5550**	0.8311	**0.6792**
BRF-DVM_Radyal	0.8375	0.6819	**0.3300**	0.8939	**0.5431**
RO	**0.9030**	**0.7679**	**0.2000**	0.9811	**0.4430**
BRF-LR	0.8875	0.7530	0.1900	0.9650	0.4282
LR	0.8875	**0.7571**	0.1700	0.9672	0.4055
KA	0.8895	0.6078	0.1200	0.9750	0.3421
BRF-KA	0.8895	0.6012	0.1200	0.9750	0.3421
DVM_Radyal	0.8930	0.6730	0.1000	0.9811	0.3132
BRF-kNN	0.8940	0.5954	0.0300	0.9900	0.1723
BRF-YSA	0.8955	0.7135	0.0250	0.9922	0.1575
kNN	**0.9010**	0.5990	0.0100	1.0000	0.1000
BRF-GNB	0.8995	0.7043	0.0100	0.9983	0.0999
GNB	0.9000	0.7306	0.0000	1.0000	-
YSA	0.9000	0.6004	0.0000	1.0000	-
BRF-DVM_Lineer	0.8995	0.5596	0.0000	0.9994	-
DVM_Lineer	0.9000	0.5229	0.0000	1.0000	-

Çizelge 0.8 Riskli birim oranı %30 olan eğitim veri setinden (uygulama-3) elde edilen model tahminlerinin test veri seti üzerindeki performansı

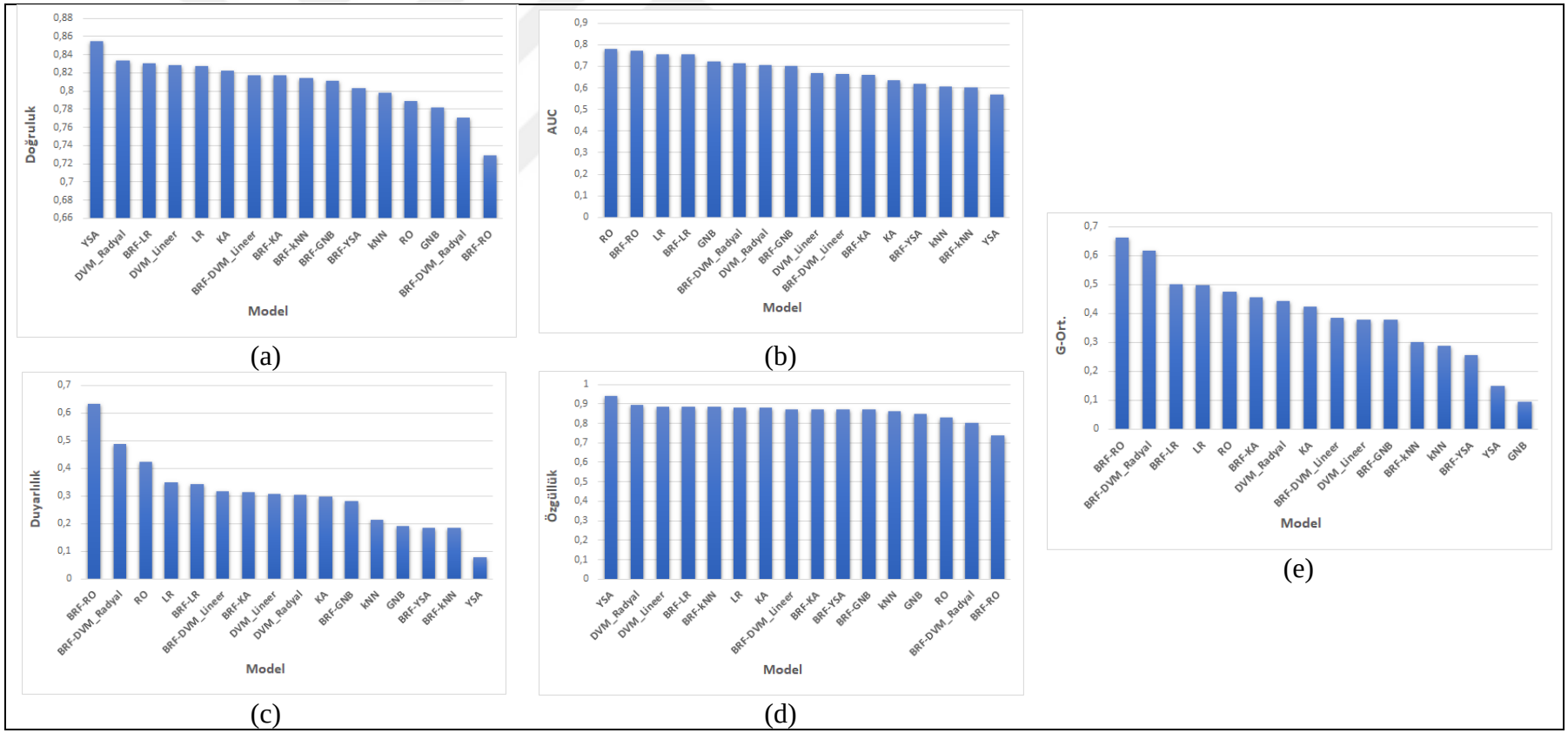
Model	Doğruluk	AUC	Duyarlılık	Özgüllük	G-Ort.
BRF-RO	0.7055	**0.7912**	**0.7100**	0.7050	**0.7075**
RO	0.8620	**0.7910**	**0.4050**	0.9128	**0.6080**
BRF-DVM_Radyal	0.8110	0.7292	**0.4250**	0.8539	**0.6024**
BRF-DVM_Lineer	0.8415	0.7551	0.3650	0.8944	0.5714
LR	0.8540	**0.7573**	0.3250	0.9128	0.5447
DVM_Lineer	0.8590	0.7564	0.3150	0.9194	0.5382
BRF-KA	0.8235	0.6525	0.3100	0.8806	0.5225
BRF-LR	0.8600	0.7549	0.2850	0.9239	0.5131
BRF-GNB	0.8510	0.7080	0.2550	0.9172	0.4836
KA	0.8425	0.6672	0.2550	0.9078	0.4811
DVM_Radyal	0.8660	0.7170	0.2400	0.9356	0.4738
kNN	0.8855	0.6050	0.0500	0.9783	0.2212
BRF-kNN	0.8855	0.6051	0.0450	0.9789	0.2099
BRF-YSA	**0.8980**	0.5996	0.0050	**0.9972**	0.0706
GNB	**0.9000**	0.7324	0.0000	**1.0000**	-
YSA	**0.9000**	0.5371	0.0000	**1.0000**	-

Çizelge 0.9 Riskli birim oranı %40 olan eğitim veri setinden (uygulama-4) elde edilen model tahminlerinin test veri seti üzerindeki performansı

Model	Doğruluk	AUC	Duyarlılık	Özgüllük	G-Ort.
RO	**0.8125**	**0.7971**	**0.5800**	0.8383	**0.6973**
BRF-RO	0.5825	**0.8028**	**0.8400**	0.5539	**0.6821**
LR	0.7895	**0.7560**	0.5250	0.8189	**0.6557**
BRF-DVM_Radyal	0.7575	0.7488	**0.5500**	0.7806	0.6552
DVM_Lineer	0.7915	0.7548	0.5000	0.8239	0.6418
BRF-DVM_Lineer	0.8085	0.7528	0.4800	0.8450	0.6369
BRF-LR	0.8075	0.7549	0.4750	0.8444	0.6333
BRF-GNB	0.7230	0.7043	0.5250	0.7450	0.6254
DVM_Radyal	0.8105	0.7512	0.4600	0.8494	0.6251
BRF-KA	0.7925	0.6979	0.4500	0.8306	0.6114
KA	0.7915	0.6899	0.4500	0.8294	0.6109
kNN	0.7615	0.6121	0.3200	0.8106	0.5093
BRF-kNN	0.8000	0.6114	0.2450	**0.8617**	0.4595
BRF-YSA	0.7745	0.5952	0.2500	0.8328	0.4563
YSA	**0.8675**	0.5844	0.0700	**0.9561**	0.2587
GNB	**0.9000**	0.7312	0.0000	**1.0000**	-

Çizelge 0.10 Riskli birim oranı %50 olan eğitim veri setinden (uygulama-5) elde edilen model tahminlerinin test veri seti üzerindeki performansı

Model	**Doğruluk**	**AUC**	**Duyarlılık**	**Özgüllük**	**G-Ort.**
BRF-RO	0.6975	**0.7916**	**0.7800**	0.6883	**0.7327**
BRF-LR	0.6970	**0.7556**	0.7250	0.6939	**0.7093**
BRF-DVM_Radyal	0.6845	0.7554	0.7350	0.6789	**0.7064**
DVM_Lineer	0.6935	0.7540	0.7200	0.6906	0.7051
LR	**0.7060**	0.7541	0.7000	**0.7067**	0.7033
DVM_Radyal	**0.7000**	0.7552	0.7050	**0.6994**	0.7022
BRF-DVM_Lineer	0.6385	0.7526	0.7450	0.6267	0.6833
BRF-KA	0.6835	0.7162	0.6800	0.6839	0.6819
KA	0.6885	0.7193	0.6700	0.6906	0.6802
BRF-GNB	0.6830	0.6859	0.6100	0.6911	0.6493
RO	0.4695	**0.8218**	**0.9300**	0.4183	0.6237
kNN	0.5405	0.6186	0.6900	0.5239	0.6012
BRF-kNN	0.5890	0.6175	0.5950	0.5883	0.5917
BRF-YSA	0.5475	0.6079	0.6400	0.5372	0.5864
YSA	**0.7055**	0.5469	0.3250	**0.7478**	0.4930
GNB	0.3090	0.6976	**0.9550**	0.2372	0.4760



Şekil 7.2 Riskli birim dağılımına göre gerçekleştirilen beş farklı uygulamadan elde edilen bulguların değerlendirilmesi

Dördüncü uygulamada, eğitim veri setinde riskli birim oranı %40 olacak biçimde örnekleme yapılmış ve BRF ve klasik makine öğrenmesi yöntemlerinin test verisi üzerindeki model doğruluk ölçütleri Çizelge 7.9'da sunulmuştur. Model Doğruluk ölçütü ve Özgüllük değerlerine göre YSA, kNN ve GNB modelleri ön sıralarda yer alırken; AUC, Duyarlılık ve G-Ort. ölçütlerine göre RO, BRF-RO, LR, BRF-DVM_Radyal modelleri iyi bir sınıflandırma başarısına ulaşmıştır.

Beşinci uygulamada, eğitim veri setinde riskli birim oranı %50 olacak biçimde örnekleme yapılmış ve BRF ve klasik makine öğrenmesi yöntemlerinin test verisi üzerindeki model doğruluk ölçütleri Çizelge 7.10'da sunulmuştur. Bu uygulama riskli ve riskli olmayan birim sayısına göre eğitim veri setinde dengeli dağılımı göstermektedir. Bu duruma göre modellerden elde edilen performans göstergeleri değerlendirildiğinde BRF ve klasik makine öğrenmesi yöntemlerinin birbirine benzer sonuçlar ürettiği görülmüştür. Genel olarak ölçütler değerlendirildiğinde BRF-RO, BRF-LR ve BRF-DVM_Radyal modellerinin diğerlerine göre daha yüksek performansa sahip olduğu belirlenmiştir.

Riskli sınıf dağılımlarına göre oluşturulan beş farklı uygulamadan elden edilen her bir model performans ölçüt değerinin ortalaması alınarak Şekil 7.2 ile verilen değerlendirmeye ulaşılmıştır. Şekil 7.2 (a) ve (d)'de verilen grafikler modellerin Doğruluk ve Özgüllük ölçütlerine göre karşılaştırmasını göstermektedir. Her iki ölçüte göre ilk sırada klasik YSA yöntemi yer almakta ve bu durum esasında riskli bireylerin sınıflandırılmasındaki aşırı uyum sorunundan kaynaklandığı düşünülmektedir. Şekil 7.2 (b)'de AUC ölçütüne göre karşılaştırma yapılmakta; RO, BRF-RO, LR ve BRF-LR ilk sıralarda yer almaktadır. Şekil 7.2 (c) ve (e)'de sunulan grafikler incelendiğinde ise BRF-RO, BRF-DVM_Radyal ve BRF-LR yöntemlerinin ilk sıralarda yer aldığı görülmektedir. Bu durumun BRF yaklaşımının dengesiz veri setlerinde riskli bireyleri sınıflandırmasındaki başarısından kaynaklandığı düşünülmektedir.

7. SONUÇ

Kredi riskinin tespit edilmesi tüm finansal kuruluşlar açısından hayati bir öneme sahiptir. Bu amaçla yapılan çalışmalarda hedeflenen, başvuranların kredibilitesine göre iyi veya kötü şeklinde sınıflandırılmasını sağlamaktadır. Bu şekilde finansal kuruluşlar kredi verecekleri kişi ya da kurumlara daha doğru şekilde karar verebilecek ve risklerini azaltabileceklerdir. Uzun yıllardır kredi riskinin tespitinde birçok farklı yöntem kullanılmış olup, özellikle makine öğrenmesi yöntemlerinin gelişmesi ile birlikte kredi riskinin belirlenmesinde kullanılmaya başlandığı görülmektedir. Bu çalışmada, BRF yaklaşımı ve klasik makine öğrenmesi yöntemleri kullanılarak kredi riskinin kestiriminin elde edilmesi amaçlanmıştır. Bu amaçla özellikle BRF yöntemi ile yeniden örnekleme stratejisi birleştirilmiş ve bu şekilde model çıktılarının performansının arttırılması hedeflenmiştir.

Çalışmada Home Credit tarafından sağlanan ve erişime açık bir veri seti kullanılmıştır. Veri setinde tespiti istenen riskli birimlerin veri seti içindeki toplam birim sayısına oranı %8,47 olarak hesaplanmıştır. Bu durum ise veri setinin dengesiz olduğunu göstermektedir. Çalışmada bir yeniden örnekleme stratejisi olan eksik örnekleme yaklaşımı kullanılmış ve riskli birimlerin farklı dağılımlarına göre model performansları değerlendirilmiştir. Beş farklı uygulamaya göre oluşturulan eğitim verileri kullanılarak elde edilen model tahminlerinin sınıflandırma başarısı bir test verisi üzerinde karşılaştırılmıştır. Model performanslarının karşılaştırılmasında doğruluk, özgüllük, AUC, duyarlılık ve G-ortalama ölçütlerinden faydalanılmıştır.

Doğruluk ve özgüllük kriterleri açısından bakıldığında ilk sırada yer alan klasik YSA yönteminin aşırı uyum sorunu olduğu değerlendirilmiştir. AUC ölçütüne göre ise ilk sıralarda RO, BRF-RO, LR ve BRF-LR yöntemlerinin bulunduğu görülmektedir. Duyarlılık ve G-ortalama kriterleri açısından ise BRF-RO, BRF-DVM_Radyal ve BRF-LR yöntemlerinin ilk sırada olduğu anlaşılmaktadır. Bu durumun ise dengesiz veri setlerinde riskli bireylerin sınıflandırılmasındaki başarısının bir sonucu olduğu değerlendirilmiştir.

Çalışmanın bundan sonraki aşamalarında, kredi riski veri setlerinde mevcut dengesizlik sorununun giderilmesi amacıyla farklı yeniden örnekleme stratejileri kullanılabilir. Çalışmada literatürde de sıklıkla tercih edilen bazı makine öğrenmesi yöntemlerine yer verilmiş olup, farklı makine öğrenmesi algoritmaları kullanılarak BRF yaklaşımının performansının geliştirilmesi mümkün olabilecektir.

KAYNAKLAR

Akbani,R., Kwek, S., Japkowicz, N. 2004. Applying Support Vector Machines to Imbalanced Data Sets," Lecture Notes in Computer Science, vol. 3201, pp. 39-50.

Alpaydin, E. 2016. Machine learning: The new AI. MIT Press.

Anderson, R. 2007. The credit scoring toolkit: Theory and Practice for Retail Credit Risk Management and Decision Automation. Oxford University Press

Atasoy, T., Tanrıvermiş, 2021. H. Türkiye'de Konut Kredisi Hacmi ile Seçilmiş Makroekonomik Faktörler Arasındaki İlişkinin Değerlendirilmesi, Erciyes Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi, Sayı: 59, Mayıs-Ağustos 2021, ss: 461-484

Aydın, A.M. 2022. Müşteri Kaybı Tahmininde Sınıf Dengesizliği Problemi, Politeknik Dergisi, 25(1):351,360

Babuška, R.,Verbruggen H.B. 1997. Constructing Fuzzy Models by Product Space Clustering, Model Identification, Springer-Verlag Berlin Heidelberg

Bai, C. G., Shi, B. F., Liu, F., Sarkis, J. 2019. Banking credit worthiness: Evaluating the complex relationships. Omega, 83, 26–38.

Bao, W., Lianju, N., Yue, K. 2019. Integration of unsupervised and supervised machine learning algorithms for credit risk assessment. Expert Systems with Applications, 128, 301–315.

Başer, F., Demirhan, H. 2017. A Fuzzy Regression with Support Vector Machine Approach to the Estimation of Horizontal Global Solar Radiation. Energy, 123, 229–240.

Başer, F., Koç, O., Kestel, A.S.S. 2023. Credit Risk Evaluation Using Clustering Based Fuzzy Classification Method, Expert Systems With Applications 223 (2023) 119882, 1-12.

Baziki, B.S., Çapacıoğlu, T. 2021. Loan-to-Value Caps, Bank Lending, and Spillover to General-Purpose Loans, Working Paper No: 21/23, Central Bank of the Republic of Turkey

Bezdek, J. C. 1981. Objective Function Clustering in Pattern Recognition with Fuzzy Objective Function Algorithms (ss. 43-93). Springer, Boston, MA.

Bilgin, M., Veri Biliminde Makine Öğrenmesi, Papatya Yayıncılık Eğitim,m Sertifika No: 11218, 160 s., İstanbul

Bishop, M.C. 1994. Neural Networks and Their Applications, Rev Sci Instrum 65, 1803-1832

Boughaci, D., Alkhawaldeh, A. A., Jaber, J. J., Hamadneh, N. 2021. Classification with Segmentation for Credit Scoring and Bankruptcy Prediction. Empirical Economics, 61(3), 1281–1309.

Breiman, L. 2001. Random Forests. Machine Learning, 45(1), 5–32.

Budak, H., Erpolat, S. 2012. Kredi Riski Tahmininde Yapay Sinir Ağları ve Lojistik Regresyon Analizi Karşılaştırılması, AJIT-e: Online Academic Journal of Information Technology Fall/Güz 2012 – Cilt/Vol: 3 - Sayı/Num: 9, DOI: 10.5824/1309-1581.2012.4.002.x, Türkiye

Cateni, S., Colla,V., Vannuci, M. A. 2014. Method for Resampling Imbalanced Datasets in Binary Classification Tasks for Real-World Problems, Neurocomputing 135 (2014) 32-41

Chakravarty, S., Demirhan, H., Başer, F. 2020. Fuzzy regression functions with a noise cluster and the impact of outliers on mainstream machine learning methods in the regression setting, Applied Soft Computing Journal 96 (2020) 106535

Chang, Y.C., Chang, K.H., Huang Y.H. 2020. A Novel Fuzzy Credit Risk Assessment Decision Support System Based on the Python Web Framework, Journal of Industrial and Production Engineering, 75:5, 229-244.,

Chauhan, N.,Agrawal R. 2022. Probabilistic Optimized Kernel Naive Bayesian Cloud Resource Allocation System, Wireless Personal Communications, 124:2853–2872

Chawla, N.V.,Bowyer W.K.,Hall,O.L.,Kegelmeyer, W.P. 2002. SMOTE: Synthetic Minority Over-Sampling Technique, Journal of Artifical Intelligence Research 16, 321-357

Cherkassky, V., Mulier, F. 2007. Learning From Data-Concepts, Theory and Methods, IEEE Press

Cristianini, N., Shawe-Taylor, J. 2000. An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods, Cambridge University Press.

Correa, A., Gonzalez, A., Nieto, C., Amezquita, D. 2012. Constructing a credit risk scorecard using predictive clusters. SAS Global Forum, 128.

Çelikyılmaz, A., Türksen, I. B. 2007. Fuzzy Functions with Support Vector Machines. Information Sciences, 177(23), 5163–5177.

Celikyilmaz, A., Türkşen, I. B. 2008a. Enhanced Fuzzy System Models with Improved Fuzzy Clustering Algorithm. IEEE Transactions on Fuzzy Systems, 16(3), 779–794.

Çelikyılmaz, A., Türkşen, I. B. 2008b. Uncertainty modeling with evolutionary improved fuzzy functions approach. IEEE Systems, Man, and Cybernetics-Part B, 38(4), 1098–1110.

Çelikyılmaz, A., Türkşen, I. B. 2009. Modeling Uncertainty with Fuzzy Logic with Recent Theory and Applications Introduction Modeling Uncertainty with Fuzzy Logic: With Recent Theory And Applications. Springer-Verlag Berlin.

Dang, K.T., Tran, T.C., Tuan, L.M., Tiep, M.V. 2021. Machine Learning Based on Resampling Approaches and Deep Reinforcement Learning for Credit Card Fraud Detection Systems, Applied Sciences, 11, 10004

Daniya, T., Geetha, M., Kumar, S.K. 2020. Classification and Regression Trees with Gini Index, Advance in Mathematics: Scientific Journal 9 (2020), No: 10, 8237-8247, ISSM: 1858-8365 (printed); 1858-8438 (electronics)

David, A.B., Frank, E. 2009. Accuracy of Machine Learning Models Versus "Hand Crafted" Expert Systems – A Credit Scoring Case Study, Expert Systems with Aplications 36 (2009) 5264-5271

Elkan, C. 2001. The Foundations of Cost-Sensitive Learning, Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI'01), 2001 Joint Conf. Artificial Intelligence, pp. 973-978.

Erol, I. 2009. ABD Mortgage Krizi ve Türkiye'de İkincil Mortgage Piyasası, Eğitim Bilim Toplum Dergisi, Cilt: 7, Sayı: 27, Güz: 2009 Sayfa: 16-34

Fernandez, A., Garcia, S., Galar, M., Prati, C.R., Krawczyk, F.H. 2018. Learning From Imbalanced Data Sets. Springer Nature Switzerland AG eBook.

Ferraro, M. B., Giordani, P., Serafini, A. 2019. Fclust: An R Package for Fuzzy Clustering. The R Journal, 11(1).

Findeks, Kredi Kayıt Bürosu. 2023. Web Sitesi: https://www.findeks.com/ Erişim Tarihi: 04.10.2023.

Galar, M., Fernandez, A., Barrenechea, E., Bustince, H., Herrera, F. (2011). 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.

Gatzert, N., Wesker, H. 2012. A Comparative Assessment of Basel II/III and Solvency II. The Geneva Papers on Risk and Insurance-Issues and Practice, 37(3), 539–570.

Ghanbari, S., Pashazadeh, S., Bevrani, H. 2014. Credit risk prediction using clustered classification. International Journal of Artificial Intelligence and Mechatronics, 3(5), 247–253.

Golbayani, P., Florescu, I., Chatterjee, R. 2020. A Comparative Study Of Forecasting Corporate Credit Ratings Using Neural Networks, Support Vector Machines, And Decision Trees. The North American Journal Of Economics And Finance, 54, Article 101251.

Gu, Q., Li, Z., Han, J. (2011). Generalized fisher score for feature selection. Proceedings of the International Conference on Uncertainty in Artificial Intelligence.

Gunn, S.R. 1998. Support Vectır Machines for Classification and Regression, Technical Report, University of Sputhampton, Faculty of Engineering and Applied Science, Departmant of Electronics and Computer Science.

Haixiang, G., Yijimg, L., Shang, J., Mingyun G., Yuanyue, H., Bing, G. 2017. Learning From Class-Imbalanced Data: Review of Methods and Applications, Expert Systems with Apllications 73 (2017) 220-239

Hand, D. J., Henley, W. E. 1997. Statistical Classification Methods İn Consumer Credit Scoring: A Review. Journal Of The Royal Statistical Society: Series A (Statistics in Society), 160(3), 523–541.

Harris, T. 2015. Credit Scoring Using The Clustered Support Vector Machine. Expert Systems with Applications. 2015. 42(2), 741–750.

Hastie, T., Tibshirani, R., Friedman, J. H., Friedman, J. H. 2009. The Elements of Statistical Learning: Data Mining, Inference and Prediction (Vol. 2,, 1–758.

He, H.,Garcia, E.A. 2009. Learning From Imbalanced Data, IEEE Transactions on Knowledge and Data Engineering, Vol.21, No:9, September 2009, 1263-1284.

He, H., Bai, Y., A. Edwardo, Garcia, A., Li, S. 2008. ADAYSN: Adaptive Synthetic Approach for Imbalanced Learning, 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)

Huang, Z., Chen, H., Hsu C., Chen, W., Wu, S. 2004. Credit Rating Analysis with Support Vector Machines and Neural Networks: A Market Comparative Study, Decision Support Systems 38 (2004) 543-558.

İldaş, T. 2021. Kredi Riski Ölçüm Modellerinin Değerlendirilmesi, Finansal Araştırmalar ve Çalışmalar Dergisi, Cilt: 13, Sayı: 25, Temmuz 2021 ISSN: 2529-0029, ss. 516-547, Türkiye

Jo, T., Japowicz, N. 2004. Class Imbalances Versus Small Disjuncts, ACM SIGKDD Explorations Newsletter, vol. 6, no. 1, ss. 40-49.

Kecman, V. 2001. Learning and Soft Computing Support Vector Machines, Neural Networks and Fuzzy Logic Models, A Bradford Book, The MIT Press, Cambridge, Massachussetts.

Khandani E.A, Kim, J., Lo, W.A. 2010. Consumer Credit-Risk Models Via Machine Learning Algorithms, Journal of Banking Finance 34(2010) 2767-2787, 2010, USA

Khuri, A. I. 2003. Advanced Calculus with Applications in Statistics. Wiley Interscience, Hoboken.

Kılıç, S. 2007. Konut Finansman Modeli Olarak Yapı Tasarruf Sandıkları; Almanya ve Türkiye'deki Uygulamaları, Celal Bayar Üniversitesi İ.İ.B.F. Yönetim ve Ekonomi Dergisi, Yıl: 2007, Cilt: 14, Sayı: 1, ss: 231-246

Kılıçarslan, H., Giter, M. Kredi Derecelendirme ve Ortaya Çıkan Sorunlar, Maliye Araştırmaları Dergisi 2016, Yıl: 2, Cilt: 2, Sayı: 1, ss: 61-81

Kim, M., Ramakrishna, R. S. 2005. New Indices for Cluster Validity Assessment. Pattern Recognition Letters, 26(15), 2353–2363.

Koc, O. 2019. Comparison of Machine Learning Algorithms on Consumer Credit Classification. Middle East Technical University). Master's Thesis.

Kolen, J.F.; Pollack, J.B. 1990. Back Propagation Is Sensitive to Initial Conditions. In Proceedings of the Advances in Neural Information Processing Systems 3, Denver, CO, USA, 26–29 November 1990;

Komarek, P. 2004. Logistic Regression for Data Mining and High-Dimensional Classification, Carnegie Mellom University, The Phd Thesis

Krishnapuram R., and Keller M.J. 1993. A Possibilistic Approach to Clustering. IEEE Transactions on Fuzzy Systems. 1(2) 98-110

Kuhn, M. 2008. Building Predictive Models in R Using the Caret Package. Journal of Statistical Software, 28, 1–26.

Lessmann, S., Baesens, B., Seow, H. V., Thomas, L. C. 2015. Benchmarking State-of-the-Art Classification Algorithms for Credit Scoring: An update of research. European Journal of Operation

Liang, L., Cai, X. 2020. Forecasting Peer-to-Peer Platform Default Rate with LSTM Neural Network, Electronic Commerce Research and Applications 43 (2020) 100997

Lim, M. K., Sohn, S. Y. 2007. Cluster-Based Dynamic Scoring Model. Expert Systems with Applications, 32(2), 427–431.

Liu, H., Motoda, H. 2007. Computational Methods of Feature Selection, Chapman&Hall/CRC Data Mining and Knowledge Discovery Series

Liu, X.Y., Wu, J. ve Zhou, Z.H., 2009. Exploratory Undersampling for Class Imbalance Learning, IEEE Transactions on Systems, Man and Cybernetics, 39(2):539-550.

Magidson, J., Vermunt., J.K. 2004. An Extension of the CHAID Tree-Based Segmentation Algorithm to Multiple Dependent Variables, Conference: Classification - the Ubiquitous Challenge, Proceedings of the 28th Annual Conference of the Gesellschaft für Klassifikation e.V., University of Dortmund, March 9-11

Malhotra, R., Malhotra, D. K. 2002. Differentiating Between Good Credits and Bad Credits Using Neuro-Fuzzy Systems, European Journal of Operational Research, 136(1), 190–211

Marques, A. I., García, V., S´anchez, J. S. 2012. Exploring the Behaviour of Base Classifiers in Credit Scoring Ensembles, Expert Systems with Applications, 39(11), 10244–10250

Masood, W.S., Begum, S.A. 2022. Comparison of Resampling Techinques for Imbalanced Datasets in Student Dropout Prediction, 2022 IEEE Silchar Subsection Conference

Menard, S. 1995. Applied Logistic Regression Analysis, a Sage University Paper

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., Leisch, F., Chang, C. C., Lin, C.C. (2023). Package 'e1071'. https://cran.r-project.org/web/packages/e1071/e1071.pdf . Erişim Tarihi: 14.08.2023

Mitchell, T. M. 1997. Machine Learning (Vol. 1, No. 9). New York: McGraw-Hill.

Öztürk, N.,Fitöz, E. Türkiye'de Konut Piyasasının Belirleyicileri: Ampirik Bir Uygulama, ZKÜ Sosyal Bilimler Dergisi, Cilt 5 Sayı 10, 2009 ss 21-46.

Pal, M. 2005. Random Forest Classifier for Remote Sensing Classification, International Journal of Remote Sensing, Vol. 26, No: 1, 10 Jenaury 2005, 217-222

Pal, N. R., Bezdek, J. C. 1995. On cluster validity for the fuzzy c-means model. IEEE Transactions on Fuzzy Systems, 3(3), 370–379.

Pandey, N.T., Mohapatra, K.S. 2017. Credit Risk Analysis Using Machine Learning Classifiers, International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS-2017), India

Ramkumar, M. 2016. A modified ANP and Fuzzy Inference System Based Approach for Risk Assessment of in-House and Third Party E-Procurement Systems, Strategic Outsourcing: An International Journal, 9(2), 159–188.

Raskutti, B., Kowalczyk, A. 2004. Extreme Re-Balancing for SVMs: A Case Study, ACM SIGKDD Explorations Newsletter, vol. 6, no. 1, ss. 60-69,

Safavian, R.S.,Landgrebe, D. 1991. A Survey of Decision Tree Classifier Methodology, IEEE Transaction on System, Man, and Cybernetics, Vol.21, No: 3, May/June

Sarıtaş, M.M., Yaşar, A. 2019. Performance Anaylsis of ANN and Naive Bayes Classification Algorithm for Data Classification, International Journal of Intelligent Systems and Applications in Engineering, IJISAE, 2019, 7(2), 88-91

Sarmanova, A. 2013. Veri Madenciliğindeki Sınıf Dengesizliği Sorununun Giderilmesi, Yıldız Teknik Üniversitesi Bilgisayar Mühendisliği Anabilim Dalı Yüksek Lisans Tezi,

Scitovski, S., Šarlija, N. 2014. Cluster Analysis in retail segmentation for credit Scoring, Croatian Operational Research Review, 5(2), 235–245.

Selcuk, M., Koc, O., Kestel, A. S. 2022. The Prediction Power of Machine Learning on Estimating the Sepsis Mortality in the Intensive Care Unit. Informatics in Medicine Unlocked, 28, Article 100861.

Sermaya Piyasası Lisanslama, 2023. Web Sitesi: https://www.spl.com.tr/docs/other/10ec2a96-649b-4a.pdf Erişim Tarihi: 14.08.2023

Sezer, A. 2011. Konut Finansman Sistemleri ve Türkiye Uygulamaları, Ankara Üniversitesi Fen Bilimleri Enstitüsü Taşınmaz Geliştirme Anabilim Dalı Yüksek Lisans Tezi.

Shi, J., Xu, B. 2016. Credit Scoring by Fuzzy Support Vector Machines with a Novel Membership Function, Journal of Risk and Financial Management, 9(4), 13

Shieh, M. D., Yang, C. C. 2008. Classification Model for Product Form Design Using Fuzzy Support Vector Machines, Computers Industrial Engineering, 55(1), 150–164.

Sohn, S. Y., Kim, D. H., Yoon, J. H. 2016. Technology Credit Scoring Model With Fuzzy Logistic Regression, Applied Soft Computing, 43, 150–158.

Strang, G. 1986. Introduction to Applied Mathematics, Wellesley, Wellesley-Cambridge Press.

Stork, D. G., Duda, R. O., Hart, P. E., Stork, D. 2001. Pattern Classification, A Wiley-Interscience Publication.

Sun, Y., Chai, N., Dong, Y., Shi, B. 2022. Assessing and Predicting Small Industrial Enterprises' Credit Ratings: A Fuzzy Decision-Making Approach, International Journal of Forecasting, 38(3), 1158–1172.

Syau, Y. R., Hsieh, H. T., Lee, E. S. 2001. Fuzzy Numbers in the Credit Rating of Enterprise Financial Condition, Review of Quantitative Finance and Accounting, 17(4), 351–360.

Tajik, M., Aliakbari, S., Ghalia, T., Kaffash, S. 2015. House Prices and Credit Risk: Evidence from the United States, Economic Modelling 51 (2015) 123-135

Taşçı, E., Onan, A. 2016. K-En Yakın Komşu Algoritması Parametrelerinin Sınıflandırma Performansı Üzerine Etkisinin İncelenmesi, 18. Akademik Bilişim Konferansı Bildirisi

Teles, G., Rodrigues, J. J., Saleem, K., Kozlov, S., Rabelo, R. A. 2020. Machine Learning and Decision Support System on Credit Scoring, Neural Computing and Applications, 32(14), 9809–9826.

Thomas, C.L., Edelman B.D., Crook, N.J. 1997. Credit Scoring and It's Applications, Siam Yayınları, USA

Tian, Z., Xiao, J., Faxeng, H., Wei, Y. 2019. Credit Risk Assessment Based on Gradient Boosting Decision Tree, 2019 International Conference on Identification, Information and Knowledge in the Internet of Things (IIKI2019), USA and China

Ting, K.M. 2002. An Instance-Weighting Method to Induce Cost-Sensitive Trees, IEEE Transcations on Knowledge and Data Engineering, Vol. 14, No: 3, May/June 2002.

Türkiye Bankalar Birliği (TBB) Risk Merkezi. 2023. Web Sitesi: https://www.riskmerkezi.org/tr/risk-merkezi/hakkinda/8 Erişim Tarihi: 14.08.2023

Türkiye Bankalar Birliği (TBB), 2023. Web Sitesi: https://www.tbb.org.tr/Content/Upload/istatistikiraporlar/ekler/4101/Tuketici_Kredileri_Raporu-Mart_2023.pdf Erişim Tarihi: 14.08.2023

Türkşen, I.B, 1999. Type 1 and Type 2 Fuzzy System Modeling, Fuzzy Sets and Systems 106 (1999) 11-34.

Türkşen, I. B., Celikyilmaz, A. 2006. Comparison of Fuzzy Functions with Fuzzy Rule Base Approaches, International Journal of Fuzzy Systems, 8(3), 137–149.

Türkşen, I. B. 2008. Fuzzy functions with LSE. Applied Soft Computing, 8, 1178–1188.

Türkşen, I. B. 2009. Review of fuzzy system models with an emphasis on fuzzy functions. Transactions of the Institute of Measurement and Control, 31(1), 7–31.

Wang, G., Hao, J., Ma, J., Jiang, H. 2011. A Comparative Assesment of Ensemble Learning For Credit Scoring, Expert System With Applications 38 (2011) 223-230

Vapnik, V. (1995). The Nature of Statistical Learning Theory. Newyork: Springer.

Vapnik, V. (1998). Statistical Learning Theory. Newyork: John Wiley & Sons.

Yam, Y.F.,Chow,T.W.S. 1995. Determining Initial Weights of Feedforward Neural Networks Based on Least Square Method, Neural Processing Letters, Vol. 2, No. 2, 13-17.

Yu, L., Yao, X. Wang, S., Lai, K.K. 2011. Credit Risk Evalution Using a Weighted Least Squares SVM Classifier With Design of Experiment for Parameter Selection, Expert Systems with Applications 38 (2011) 15392-15399

Zadeh, L.A. 1965. Fuzzy Sets," Inf. Control, vol. 8, no. 3, pp. 338–353,Jun. 1965.

Zadeh, L.A. 1975. The Concept of Linguistic Variable and Its Application to Approximate Reasoning—I," Inf. Sci., vol. 8, no. 3, pp. 199–249.

Zhang, H., He, H., Zhang, W. 2018. Classifier Selection and Clustering with Fuzzy Assignment in Ensemble Model for Credit Scoring, Neurocomputing, 316, 210–221.

Zhao, Z., Xu, S., Kang, B. H., Kabir, M. M. J., Liu, Y., Wasinger, R. 2015. Investigation and Improvement of Multi-Layer Perceptron Neural Networks for Credit Scoring, Expert Systems with Applications, 42(7), 3508–3516.