TÜRKİYE CUMHURİYETİ
ANKARA ÜNİVERSİTESİ
SAĞLIK BİLİMLERİ ENSTİTÜSÜ

SAĞLIK ALANINDA BÜYÜK VERİ VE VERİ MADENCİLİĞİ YÖNTEMLERİNİN KULLANIMI

Batuhan BAKIRARAR

BİYOİSTATİSTİK ANABİLİM DALI
YÜKSEK LİSANS TEZİ

DANIŞMAN
Prof. Dr. Yasemin YAVUZ

ANKARA
2016

Ankara Üniversitesi

Sağlık Bilimleri Enstitüsü Müdürlüğü'ne,

Yüksek Lisans tezi olarak hazırlayıp sunduğum "Sağlık Alanında Büyük Veri ve Veri Madenciliği Yöntemlerinin Kullanımı" başlıklı tez; bilimsel ahlak ve değerlere uygun olarak tarafımdan yazılmıştır. Tezimin fikir/hipotezi tümüyle tez danışmanım ve bana aittir. Tezde yer alan araştırma tarafımdan yapılmış olup, tüm cümleler, yorumlar bana aittir.

Yukarıda belirtilen hususların doğruluğunu beyan ederim.

Öğrencinin Adı Soyadı: Batuhan BAKIRARAR

Tarih:

İmza:

Ankara Üniversitesi Sağlık Bilimleri Enstitüsü Biyoistatistik Anabilim Dalında Batuhan BAKIRARAR tarafından hazırlanan "Sağlık Alanında Büyük Veri ve Veri Madenciliği Yöntemlerinin Kullanımı" adlı tez çalışması aşağıdaki jüri tarafından YÜKSEK LİSANS TEZİ olarak OY BİRLİĞİ ile kabul edilmiştir.

Tez Savunma Tarihi:

Prof. Dr. Atilla Halil ELHAN

Ankara Üniversitesi

Tıp Fakültesi

Biyoistatistik Anabilim Dalı

Jüri Başkanı

Prof. Dr. Yasemin YAVUZ

Ankara Üniversitesi

Tıp Fakültesi

Biyoistatistik Anabilim Dalı

Tez Danışmanı

Doç. Dr. Erdem KARABULUT

Hacettepe Üniversitesi

Tıp Fakültesi

Biyoistatistik Anabilim Dalı

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

Prof. Dr. K. Zafer KARAER

Sağlık Bilimleri Enstitüsü Müdürü

İÇİNDEKİLER

ÖNSÖZ

Büyük veri, aslında artan veri saklama ve veri işleme maliyetleri sonucu ortaya çıkmıştır. Sektörlerdeki anlık değişimler ve artan talepler sonucu daha farklı türlerde veri saklama gereksinimi doğmuştur. Özellikle sağlık alanında klasik veritabanı yöntemleriyle saklanamayacak çeşitte ve yönetilemeyecek büyüklükte veri üretilmektedir ve üretilen verinin hızı çok fazladır. Bu çalışmada diyabet şüphesi ile hastaneye başvuran 101766 kişiye ait veriler kullanılarak diyabet teşhis tahmini yapılmaya çalışılmıştır.

Tez çalışmam ve yüksek lisans öğrenimim süresince bilgi ve tecrübeleri ile yaptığı yardımlardan ötürü danışmanım Sayın Prof. Dr. Yasemin YAVUZ'a ve akademik anlamda desteğini esirgemeyen Anabilim Dalımız öğretim üyelerinden Sayın Prof. Dr. Atilla Halil ELHAN'a teşekkürlerimi ve saygılarımı sunarım.

Yüksek lisans öğrenimime katkı sağlayan Anabilim Dalımız öğretim üyesi hocalarıma teşekkür ederim.

Eğitim hayatım boyunca maddi ve manevi desteklerini bir an olsun esirgemeyen aileme sonsuz teşekkürü borç bilirim.

SİMGELER VE KISALTMALAR

CAT	Computer Aided Translation
EB	Ekzabayt
GB	Gigabayt
HDFS	Hadoop Distributed File System
HIV	Human Immunodeficiency Virus
MRI	Magnetic Resonance Imaging
PB	Petabayt
RAID	Redundant Array of Independent Disks
SAP	Systems Analysis and Program Development
SPSS	Statistical Package for the Social Sciences
SQL	Structured Query Language
TB	Terabayt

ŞEKİLLER

ÇİZELGELER

1. GİRİŞ

1.1. Büyük Veri (Big Data) Nedir?

Büyük veri ile ilgili çalışan bilim insanları bu konuda tek bir ortak tanım olamayacağına, kullanılan alana göre farklı tanımlamalar yapılabileceğine vurgu yapmışlardır. Bu alanda yapılan tanımlardan birkaçı üzerinde durmak yararlı olacaktır. Vinod (2013)'a göre büyük veri, tipik olarak verinin büyüklük olarak terabit veya petabitin yüzlerce katı olmasını tanımlayan bir kavramdır. Kavramı operasyonel ve uygulama bakımından ele alan Rubinstein (2013)'a göre ise büyük veri "işletme, devlet veya organizasyonların farklı dijital veri setlerini bütünleştirerek istatistik ve veri madenciliği teknikleriyle gizli kalmış bilgileri ve sürpriz korelasyonları kullanmalarını tanımlar" (Demirtaş ve Argan, 2015).

LCIA (2011), büyük veriyi "çoğunluğu yapılandırılmamış olan ve sonu gelmez bir şekilde birikmeye devam eden, geleneksel ilişki bazlı veri tabanı teknikleri yardımıyla çözülemeyecek kadar yapısallıktan uzak, çok çok büyük, çok ham ve üstel bir şekilde büyümekte olan veri setleri şeklinde tanımlamaktadır. Bir başka yaklaşıma göre ise büyük veri "standart veritabanı yönetim veya analitik araçlar yardımıyla çözülmesi zor veya imkânsız olan çok büyük, çok karmaşık veya çok hızlı analiz edilmesi gereken veri setlerini tanımlayan bir terimdir" (Partners, 2012).

Dumbill (2013) ise daha kapsamlı ve terminolojik bir tanım ortaya koymuştur: "Büyük veri, geleneksel veri tabanı sistemlerinin işlenme kapasitesini aşan veridir. Bu veri çok büyüktür, çok hızlı hareket eder veya veri tabanınızdaki altyapı alanına uygun değildir. Bu veriden değer elde etmek ve onu işlemek için alternatif bir yol bulmanız gerekir".

Çizelge 1.1.'de bu kavramın daha iyi anlaşılabilmesi için tüm hacimsel büyüklükler bayt cinsinden verilmiş ve kapasiteleri günlük karşılaşabileceğimiz

örneklerle anlatılmaya çalışılmıştır. Günümüzde veri tabanları terabayt (TB), petabayt (PB) ve ekzabayt (EB) gibi terimler kullanılarak tanımlanır (Altunışık, 2015).

Çizelge 1.1. Büyük Veri Depolama Kapasitelerine İlişkin Terimler

Terim	Boyut	Örnek Kapasite
GB (Gigabayt)	1 milyar bayt	1GB=2 saatlik CD kalitesinde ses veya 7 dakikalık HDTV
TB (Terabayt)	1 trilyon bayt	1TB=2000 saatlik CD kalitesinde ses veya 5 günlük HDTV
PB (Petabayt)	1 quadrilyon bayt	1PB=7 haftalık HDTV veya 1.5 milyon 64GB'lık iPod
EB (Ekzabayt)	1 quintilyon bayt	1EB=16 aylık HDTV veya 15 milyon 64GB'lık iPod

Günümüzde terabayt büyüklüğündeki veriler harddiskler yardımıyla saklanabilirken, petabayt, ekzabayt ve zetabayt büyüklüğündeki veriler ancak sanal sunucular aracılığıyla saklanabilmektedir. 2000 yılında tüm dünyada saklanan verinin büyüklüğü 800,000 petabayt iken 2020 yılında bu verinin büyüklüğünün 35 zetabayt olacağı tahmin edilmektedir. Bu verinin de %80′inin gri veri yani veritabanlarında saklanamayan resim, video, müzik ve ofis belgeleri olacağı düşünülmektedir (Demirtaş ve Argan, 2015).

1.2. Büyük Veriye Gerek Duyulma Nedenleri

Sektörlerdeki ve kurumlardaki anlık değişimler ve artan talepler sonucunda daha çok ve çeşitli veri saklanması ihtiyacı doğmuştur. Buna bağlı olarak artan veri saklama ve veri işleme maliyetleri sonucunda "Büyük Veri" çözümlerine ihtiyaç duyulmuştur. Bu konuyla ilgili şöyle bir örnek verilebilir. 2 milyarlık müşteri

hareketinden oluşan bir veri havuzunda sadece ödemesini zamanında yapmamış müşterileri bulmak için bile yazılan bir sorgudan ilişkisel veritabanları ile 70 dakikada sonuç alınabilirken büyük veri çözümleri ile bu süre 13 saniyeye kadar düşürülebilmektedir. Büyük veri çözümlerinin kullanımı basit bir sorguda bile bu düzeyde zaman ve maliyet tasarrufu sağlanmasına olanak tanımaktadır. Bunun yanında gerçek zamanlı analiz yapmaya olanak sağlaması da büyük avantajlarından biridir (Altunışık, 2015).

Multimedya verilerindeki artış büyük verinin doğmasında bir diğer önemli faktördür. İnternette bulunan veri, tüm verilerin yaklaşık %70'ini oluşturmaktır. Sağlık sektöründe kaydedilen klinik verilerin en az %95'inin video formatında olması da multimedya verilerinin büyük veri içindeki önemini göstermektedir. (Manyika ve ark., 2011).

Makineler (telefonlar, sağlık alanında kullanılan mamografi, ultrason, EKG cihazları gibi cihazlar) üzerindeki sensörler yardımıyla toplanan verilerin gönderilip alınması büyük bir veri trafiğine neden olmaktadır. Cihazlar arası etkileşim sonucu oluşan bu verileri kaydetme yoluna gidilmesi bu trafiğin daha da artmasına neden olmaktadır. "Nesnelerin İnterneti" olarak tanımlanan bu kavram sayesinde üretilecek verinin her yıl üssel şekilde büyüyeceği tartışılmaz bir gerçektir (Court, 2015).

En büyük veri yükünü ise sosyal medya içerikleri oluşturmaktadır. Dünya üzerinde 2015 yılı itibariyle yaklaşık 3.65 milyar sosyal medya hesabı bulunmaktadır. Sadece Facebook'ta 1.4 milyar aktif kullanıcının bulunduğu ve yaklaşık 4.5 milyar beğeni yapıldığı düşünüldüğünde sosyal medyanın veri büyüklüğüne katkısı daha iyi anlaşılacaktır. (McKinsey, 2015).

Tüm bu verilerin saklanmasının standart veritabanı çözümleriyle mümkün olmadığının anlaşılmasıyla yeni çözümler aranmaya başlanmıştır. İki konuda yaşanan gelişme büyük veri setlerinin oluşmasında önemli rol oynamıştır. Birincisi, veri depolama maliyetlerini çok düşüren ve veritabanı yönetimlerini daha profesyonel ve ticari hale getiren bulut tabanlı çözümlerin ortaya çıkması iken (Lohr,

2012), ikinci gelişme büyük hacimli verileri birden çok sanal sunucuya dağıtmaya imkan tanıyan NoSQL ve benzeri veritabanı tasarımlarının ve Hadoop teknolojisinin ortaya çıkmasıdır. Bu gelişmeler paralelinde büyük veri setleri oluşmaya başlamıştır (Altunışık, 2015).

1.3. Büyük Verinin Riskleri

Büyük veri birçok avantaja sahip olmasına rağmen dezavantajları da mevcuttur:

-Kişilere ait tüm veriler depolandığı için kişisel bilgilere ulaşmak daha kolay hale gelmiş, kişisel gizlilik daha fazla ihlal edilmeye başlanmıştır (Snoad, 2011).

-Kişinin istediği ve istemediği tüm bilgiler akıllı telefonlardaki uygulamalar aracılığıyla veri olarak saklanmakta ve kullanılmaktadır. Bu da büyük verinin etik konusu ile örtüşmemektedir (Snoad, 2011).

-Saklanan tüm bu kişisel veya devlete ait bilgilerin kötü amaçlı şahıs ve örgütlerin eline geçmesi belki de en büyük tehdit unsurudur. Bu veriler ileride terörizm için kullanılabilir veya başka ülkelerle paylaşılabilir. Basit bir örnek ile açıklamak gerekirse, eski NSA veri analisti Edward Snowden Yahoo, Google ve Facebook gibi şirketlerin sakladığı tüm verilere erişmiştir ve insanların en gizli bilgilerine kadar birçok veriyi sızdırmıştır (Kelion, 2013).

-İnsanlar sosyal medyayı daha aktif kullanmaya başladıkça, internet üzerinden daha fazla alışveriş yaptıkça yani daha fazla çevrim içi yaşadıkça tüm bu bilgiler depolandığı için kişilere ait birçok bilginin tahmin edilmesi de gün geçtikçe o kadar kolay olacaktır (Snoad, 2011).

1.4. Büyük Veride 5V Kavramı

Veride büyüklük sadece hacimsel (Volume) olarak ifade edilemez. Verinin aynı zamanda çok hızlı hareket ediyor olması (Velocity) veya birçok çeşitlilikte

(Variety)(yapılandırılmış/yapılandırılmamış) olması da büyük olarak adlandırılmasına neden olur. (Gobble, 2003). Bu üç kavrama ek olarak verinin doğrulanabilir olması (Verification) ve anlamlı bir değere (Value) sahip olması da büyük veriyi tanımlayan özelliklerdir.

1.4.1. Hacim (Volume)

Hacim, verinin büyüklüğünü ifade etmekte kullanılır (Hoy, 2014). Aşağıda verilen bazı bulgular günümüzde verinin hacimsel büyüklüğünü çarpıcı bir biçimde göstermektedir:

-Dünyanın varoluşundan 2003 yılına kadar üretilen veri günümüzde sadece iki günde üretilmektedir.

-Dünya üzerindeki veri akışının %90'ı son iki yılda gerçekleşmiştir.

-Firmalar tarafından saklanan veri miktarı her yıl yaklaşık 1.6 katına çıkmaktadır.

-Dünyada bir dakika içinde yaklaşık 204 milyon e-posta gönderilmekte, 270 bin tweet atılmakta ve Facebook'a 200 bin resim yüklenmektedir.

-Google'da günde yaklaşık 3.5 milyar arama yapılmaktadır.

-Youtube'a dakikada yaklaşık 100 saatlik video yüklenmektedir.

-Dünyada bir günde depolanan veriler DVD'ye kaydedilerek DVD'ler üst üste koyulursa Ay'a iki defa erişilebilinir.

-Büyük veri çözümlerinin akışkan veriyi işleyebilme özelliği sayesinde işlenebilecek suçların çoğu önceden öngörülüp engellenebilmektedir.

-Sağlık sektöründe büyük verilerin analiz edilmesi sonucunda yaklaşık 300 milyar dolarlık tasarruf sağlanabileceği tahmin edilmektedir (Altunışık, 2015).

1.4.2. Hız (Velocity)

Büyük veride en önemli kavramlardan biri de verilerin işlenme hızıdır. Veriler ne kadar hızlı işlenirse o kadar çabuk bilgi üretilir, anlamlandırılır ve kullanılmaya başlanır. Özellikle de akışkan verilerle çalışmak güncellik, hızlı karar verme ve rakiplere göre öne geçme avantajı sağlayabilir (Altunışık, 2015). Örneğin marketler, akıllı telefonlardaki konum belirleme özelliğinden faydalanarak otoparktaki kişileri algılayıp hızlı bir şekilde alışveriş bilgilerine ulaşarak kişilere özel satış tahminleri yapabilmektedir (Demirtaş ve Argan, 2015).

1.4.3. Çeşitlilik (Variety)

Çeşitlilik, verinin heterojen yapısını anlatmaktadır. Büyük veri içeren sistemlerde çok farklı kaynaklar üzerinden veri akışı sağlandığı için saklanan veriler farklı tiplerde olabilmektedir. Örneğin telefonlardan, tabletlerden, sosyal ağlardan ve entegre devrelerden gelen veriler çok farklı formatlarda, dillerde ve kodlarda kaydedilmektedir (Dumbill, 2013). Büyük veri olarak adlandırılan bu verilerin yaklaşık %80'i yapılandırılmamış olup kullanıma hazır değildir. Bunun en büyük nedeni ise farklı teknolojilerin ürettiği verilerin farklı formatlarda olmasıdır. Tüm bu faktörler düşünüldüğünde bu verilerin büyük veri çözümleri olmadan birleştirilmesi imkansızdır (Göksu, 2014).

1.4.4. Doğrulama (Verification)

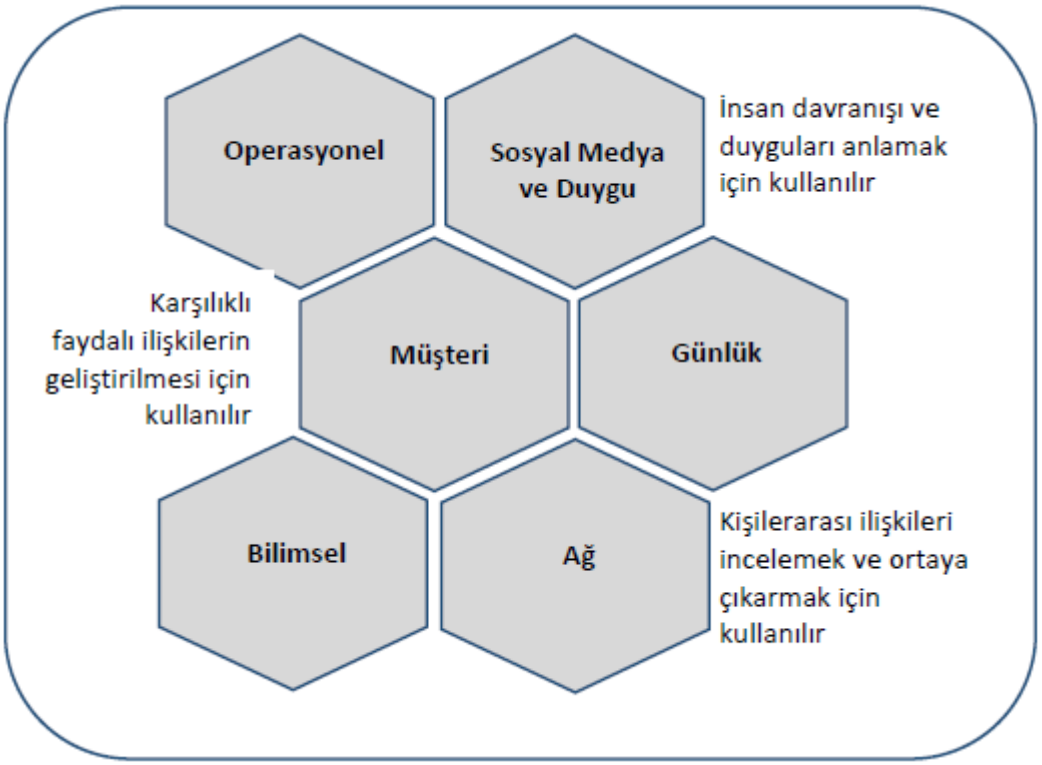
Büyük verilerin güvenli şekilde depolanması da oldukça önemlidir. Verilerin sadece yetkili kişiler tarafından görülebilmesi, işlenebilmesi ve gizli kalması gerekmektedir (Göksu, 2014).

1.4.5. Değer (Value)

Büyük veri işlenerek değer yaratılmalıdır. Örneğin kurumların karar verme süreçlerinde hızlı ve doğru karar verebilmeleri için bilgi üretilmesi değer yaratılması olarak algılanabilir. Örneğin Sağlık Bakanlığı'nın belli kararlar alırken hastalık çeşitleri, ilaç tüketimi, hasta başına düşen doktor sayısı gibi tüm kayıtlı verilere anlık olarak ulaşabilmesi gerekmektedir (Göksu, 2014).

1.5. Büyük Veride Kullanılan Veri Türleri

Büyük veride veriler çok farklı şekillerde sınıflandırılabiliyor olsa da veri türlerini 6 ana başlıkta toplamak mümkündür (Şekil 1.1.). Operasyonel veriler sensörler, makineler, bazı ölçüm aygıtları ve otomasyon süreçlerinden elde edilen verilerdir. Müşteri hizmet anlaşmalarının kapsamı, tesis kurulumu ve yönetimi gibi süreçleri yönetme ile ilgili kararlar almak için kullanılabilirler. Çeşitli internet sitelerindeki müşteri profilini ve hareketlerini inceleyerek daha iyi hizmet sunmaya ve pazarlama stratejilerini müşteri bazlı uygulamaya imkan sağlar. Bilimsel veri; yeni bilgiler elde etmek ve mevcut bilgiyi doğrulamak için kullanılabilir. Yeni hastalıklara ait gen tespiti, hastalık salgınlarının tahmini bilimsel veriye örnek olabilir. Ağ verileri ise ağ üzerindeki veri alışverişi sayesinde kişiler, firmalar hakkında genel bilgiye sahip olmaya ve davranış tespitine olanak sağlar (Narayanan, 2014).



Şekil 1.1. Büyük Veride Kullanılan Veri Türleri

Sağlık alanında ise büyük veri kaynakları dahili (bilgisayarlı talimat girişi, klinik karar destek sistemleri, elektronik tıbbi kayıtlar) ve harici (sigorta kurumları/şirketleri, laboratuvarlar, eczaneler, devlet kaynakları vs.) olmak üzere iki tiptedir ve bu veriler farklı format ve konumlarda bulunurlar. Sağlık kurumlarına ait web siteleri, akıllı telefonlarda kullanılan sağlık hizmetleri uygulamalarından alınan veriler web ve sosyal medya kaynaklarına, sensörler, yaşamsal bulgu cihazlarından alınan veriler makine kaynaklarına, parmak izi, puls-oksimetre okumaları, el yazısı, röntgen ve diğer tıbbi görüntüler, kan basıncı, nabız, genetik, retinal tarama vs tip veriler biyometrik veri kaynaklarına, hekim notları, kağıt belgeler ve elektronik sağlık kayıtları gibi yapılandırılmış ya da yapılandırılmamış veriler ise insan üretimi veri kaynaklarına örnek olarak gösterilebilir (Beyan, 2016).

1.6. Büyük Veri Teknolojisinin Getirdiği Yenilikler

Büyük veri teknolojisinin en önemli yanı maliyetleri düşürmek, veri işleme hızını artırmak, yeni ürün ve hizmetler geliştirmek veya daha iyi karar verme için yeni veri ve modellerin kullanıma imkan tanımaktır (Davenport, 2014).

Büyük veri teknolojilerindeki yenilik esas olarak verinin geleneksel veri tabanı yazılımlarıyla veya tekli sunucularla iyi idare edilememesidir. Geleneksel ilişkisel veritabanları veriyi basit sütun satır formatında rakamlar olarak ele alırken büyük veri farklı formatlarda verileri saklayabilir (Davenport, 2014).

Büyük verileri tek bir sunucu ile işlemek imkansız olduğu için Hadoop kullanılarak veriler birden çok sunucuya dağıtılıp eş zamanlı işlenir ve sonuçları gösterilir. Hadoop bir hesaplama görevini birden fazla sunucu arasında bölerek veri işleme süresini yüz kat veya daha fazla düşürebilir. Büyük verinin yükselişi birçok (binlerce) işlemcisi olan ucuz fiyatlı ticari sunucuların yükselişiyle aynı zamanlı olmuştur ve bu sayede büyük veri daha çok kullanılmaya başlanmıştır (Davenport, 2014).

Kuruluşların incelemesi gereken yeni teknolojiler sadece bunlarla sınırlı değildir. Aslında geçtiğimiz birkaç sene içinde büyük veri teknoloji ortamı çarpıcı bir şekilde değişmiştir ve değişmeye devam edecektir. Yeni teknoloji ile birlikte sütunlu (dikey) veritabanları gibi yeni veritabanı şekilleri, yeni programlama dilleri (Pig, Hive) ve büyük veri gereçleri (özel sunucular) ile bellek içi analitikler (kesintili olarak sürücü depoları üzerinde değil, tamamen bir bilgisayar belleğinde çalışan bilgisayar analitikleri) gibi veri işleme için yeni donanım mimarileri ortaya çıkmıştır (Davenport, 2014).

Büyük veri teknoloji ortamının geleneksel bilgi yönetiminden farklı kritik bir yanı daha vardır. Önceden veri analizinin amacı, veriyi analiz için veri ambarı gibi farklı havuzlara ayırmaktı. Ancak büyük verinin hızı ve hacmi verinin herhangi bir ayırma yaklaşımını süratle aşabileceği anlamına geliyor. Bir örnek verecek olursak,

müşterilerinden muazzam bir sayıda tıklama verisi toplayan EBay, veri ambarında 40 petabayt'tan fazla veri bulunduruyor. Bu nedenle büyük veri teknolojileri ortamında birçok kuruluş, yüksek veri miktarlarını bir süreliğine depolayıp yeni yığınlar gelince çıkarmak için Hadoop ve benzeri teknolojileri kullanıyor. Verinin süreğenliği, bu veriler üzerinde bazı analiz ve keşifler yapmayı çok zor hale getirmekte, bu analizleri yalnızca büyük veri ortamında mümkün kılmaktadır. Bu veri yönetim yaklaşımı kurumsal veri ambarlarını tamamlayıcı olacak gibi görünmektedir (Davenport, 2014).

Büyük verinin o kadar da yeni olmayan yanı ise verinin analiz aşamasıdır. Şimdiye kadar konuştuğumuz teknolojiler büyük veriyi depolamak ve onu yapılandırılmamış formattan tipik rakam satır ve sütunlarına çevirmek için kullanılmaktadır. Veri istenen formata dönüştürüldükten sonra büyük de olsa klasik analiz yöntemleri yardımıyla analiz edilmektedir. Analiz için çoklu ticari sunucu kullanımı gerekse de kullanılan temek istatistik ve matematiksel algoritmalarda çok büyük değişiklikler olmamıştır (Davenport, 2014).

Kuruluşların büyük veri analizinde kullandıkları araçlar geçmişte veri analizinde kullanılanlardan çok farklı değildir. Bunlar arasında açık kaynak kodlu (örneğin R) veya tescilli (SAS, SPSS gibi) istatistik programlarıyla gerçekleştirilen temel istatistik uygulamaları mevcuttur. Ancak analist veya karar vericilerin bir hipotezden yola çıkarak bunun veriyle uyumunu test ettikleri geleneksel hipoteze dayalı analiz yaklaşımı yerine büyük veri analistlerinin veri madenciliği tekniklerini (makine öğrenmesi) daha çok tercih ettikleri görülmektedir (Davenport, 2014).

1.7. Büyük Veri Teknolojileri

Yeni geliştirilen teknolojiler sayesinde büyük verilerin işlenebilmesi bu verilerden değer elde edilebilmesine imkan tanımıştır. Bilindiği gibi büyük veri sadece verilerin saklanmalarına olanak sağlamakla kalmaz aynı zamanda bu verilerin işlenmesine ve analiz edilmesine de olanak sağlar. Kullandığı yeni teknolojiler

sayesinde metin, ses ve video analizi gibi standart yöntemlerle analizi mümkün olmayan veri tiplerini de analiz etme imkanı sunar. En yaygın kullanılan büyük veri teknolojileri Çizelge 1.2.'de verilmiştir (Beyan, 2016).

Çizelge 1.2. En Sık Kullanılan Büyük Veri Teknolojileri

Teknoloji	Tanım
Hadoop	Çoklu paralel sunucular üzerinden büyük veri işlemesi yapan açık kaynak kodlu yazılım
MapReduce	Hadoop'un üzerinde çalıştığı mimari altyapı
Betik diller	Büyük veriyle uyumlu çalışan programlama dilleri (ör. Python, Pig, Hive)
Makine öğrenmesi	Bir veri kümesine en iyi uyan modeli hızla bulmak için kullanılan yazılım
Doğal dil işleme	Metin analizi yazılımı-frekans, anlam vb.
Görsel analitik	Analitik sonuçların görsel veya grafik formatta gösterimi
Bellek içi analitik	Daha hızlı sonuç almak için büyük verinin bilgisayar belleğinde işlenmesi

1.7.1. Hadoop

Hadoop, birden fazla sunucunun oluşturduğu küme üzerinde büyük verileri işlemek ve analiz etmek için gerekli uygulamaları çalıştıran Java ortamında geliştirilmiş açık kaynaklı bir kütüphanedir. Hadoop, Hadoop Distributed File System (HDFS) ve MapReduce bileşenlerinden oluşur (Enacore, 2016).

HDFS birden fazla sunucunun disklerini birleştirerek tek bir sanal disk gibi kullanır, böylece tek bir sunucu üzerinde depolanması imkansız büyüklükteki verileri saklayabilir. Bu veriler birden fazla sunucu üzerine (varsayılan format olarak 3 kopya) dağıtılır ve RAID benzeri bir yapıyla yedeklendiği için olası veri kaybı

önlenmiş olur. HDFS, NameNode ve DataNode süreçlerinden oluşmaktadır (Enacore, 2016).

NameNode ana (master) veri bloklarının yaratılmasından, dağıtılmasından, gerektiği zaman silinmesinden ve veri erişiminden sorumludur. Veri dosyaları ile ilgili tüm bilgiler burada yönetilir ve her veri kümesinde sadece bir adet NameNode bulunur (Enacore, 2016).

DataNode ise veri bloklarını saklamakla görevlidir. Kendi diskindeki verileri sakladığı gibi diğer disklerdeki verilerin yedeklerini de saklar. Bir veri kümesinde birden çok DataNode bulunabilir (Enacore, 2016).

1.7.2. MapReduce

MapReduce HDFS üzerinde saklanan büyük boyutlu verileri işleyebilmek için kullanılır. Verileri filtrelemek için geliştirilen Map fonksiyonu ve verilerden çıktı elde etmek için kullanılan Reduce fonksiyonlarından oluşur. Map ve Reduce işlemlerinden sonra Hadoop, Map ve Reduce'a ait işleri farklı sunucular üzerine dağıtarak aynı anda paralel olarak işlenmesini ve bu süreç sonunda verilerin tekrar bir araya getirilmesini sağlar. Bu sayede işlemlerin ağı meşgul etmeden daha hızlı şekilde yapılması sağlanarak aynı anda birden çok işlemin gerçekleştirilmesine olanak tanınır (İhs, 2016).

MapReduce, JobTracker ve TaskTracker süreçlerinden oluşur. JobTracker MapReduce programının kümeler üzerinde doğru dağıtılıp çalıştırılmasını ve oluşacak bir sorunda bu sürecin sonlandırılıp tekrar başlatılmasını sağlar. TaskTracker ise JobTracker'dan iş talep eder ve işi sunucularda çalıştırarak sonlandırır. Böylece tamamlanan iş parçacıklarının çıktıları birleştirilir ve HDFS üzerinde bir dosya olarak yazılır (Enacore, 2016).

MapReduce işlemini Özkan'ın (2013) verdiği bir örnek üzerinde açıklayalım. Son yüzyıldaki en yüksek hava sıcaklığı bulunmak istensin. Dünya üzerinde hava sıcaklıklarını kaydeden çok fazla sensör bulunduğu için bu verileri tek bir veritabanında saklamak imkansızdır. Veriler birden fazla sunucuda saklandığından üzerinde işlem yapmak istendiğinde map ve reduce fonksiyonları kullanılmalıdır (Özkan, 2013).

Şekil 1.2.'de sensörlerden gelen verilerin neler olduğu ve ilgili sistemlerde nasıl saklandığı gösterilmektedir. Örneğin dördüncü satırda gerçek değeri 01.01.1950 olan tarih verisi 19500101 olarak, beşinci satırda ise gerçek değeri 03:00 olan saat verisi 0300 olarak saklanırken yirmi altıncı satırda gerçek değeri -12.8 derece olan sıcaklık verisi 10 ile çarpılarak -0128 olarak saklanmıştır (Özkan, 2013).

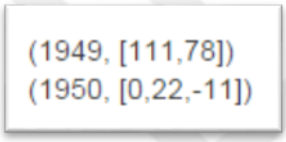
```
0057
332130   # USAF weather station identifier
99999    # WBAN weather station identifier
19500101 # observation date
0300     # observation time
4
+51317   # latitude (degrees x 1000)
+028783  # longitude (degrees x 1000)
FM-12
+0171    # elevation (meters)
99999
V020
320      # wind direction (degrees)
1        # quality code
N
0072
1
00450    # sky ceiling height (meters)
1        # quality code
C
N
010000   # visibility distance (meters)
1        # quality code
N
9
-0128    # air temperature (degrees Celsius x 10)
1        # quality code
-0139    # dew point temperature (degrees Celsius x 10)
1        # quality code
10268    # atmospheric pressure (hectopascals x 10)
1        # quality code
```

Şekil 1.2. Sensör Verileri

Verilen örnekte Map aşamasında ilgilendiğimiz yıl ve sıcaklık değerleri için anahtar/değer çiftleri oluşturulur. Map işleminden sonra verilerin formatı Şekil 1.3.'teki gibi olacaktır (Özkan, 2013).

Şekil 1.3. Map İşleminden Sonra Veri Formatı

Map işleminden sonra bu verilere bazı ön işlemler uygulanabilir. Örneğin Şekil 1.4.'te bu veri çiftleri yıl bazında gruplandırılarak Reduce işlemi için hazır hale getirilmiştir (Özkan, 2013).

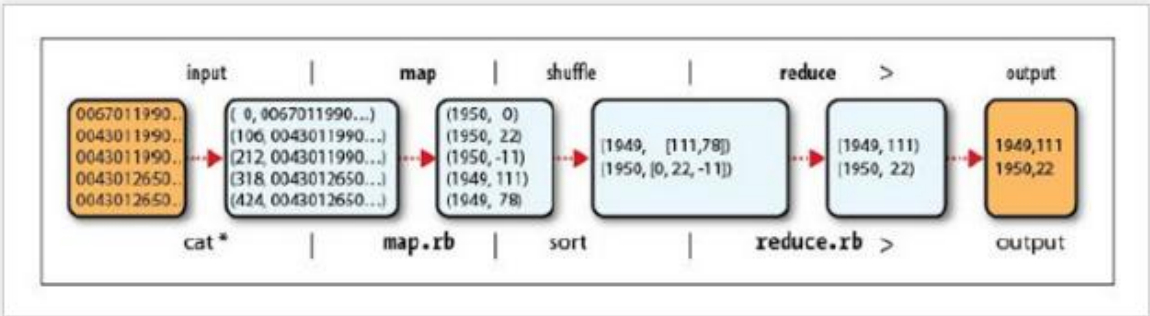


Şekil 1.4. Map İşleminden Sonra Gruplanan Veri Formatı

Reduce sürecinde ise bu veriler istenen şekilde işlenir ve sonuçlandırılır. Reduce aşaması, istenilen asıl iş mantığının gerçekleştirildiği aşamadır. Reduce aşamasında Şekil 1.5.'teki gibi her bir satır için yıl bazında en yüksek değerler bulunur (Özkan, 2013).

Şekil 1.5. Reduce İşleminden Sonra Veri Formatı

MapReduce uygulamasında bu örnek için gerçekleştirilen adımlar özet olarak Şekil 1.6.'da verilmiştir (Özkan, 2013).



Şekil 1.6. MapReduce Adımları

1.7.3. Betik Diller

1995 yılında Brendan Eich tarafından geliştirilen bu dil daha çok web sayfa içeriklerinin dinamik olması için kullanılan bir dildi. Html etiketleri arasında tanımlanır ve uygulamaları denetlemek için kullanılırdı. Büyük veri teknolojilerinin gelişmesi ile birlikte veri içi filtrelemeler, düzenlemeler yapmak için de kullanılmaya başlanmıştır. Büyük veri teknolojisine ait en sık kullanılan betik dil ise Hive'dır.

Hive SQL'e benzer yapısı ile Hadoop üzerinde Java programlama diline ihtiyaç duymadan sorgulama yapmayı sağlar. Hive kullanılarak belirli formatlarda (csv, arff, vb.) kaydedilmiş dosyalar SQL'de olduğu gibi tablo olarak tanıtılabilir. Hive bu bilgiyi saklayarak yazılan SQL benzeri kodları arka planda MapReduce kodlarına çevirerek çalıştırır. Bu sayede SELECT, JOIN, UPDATE gibi komutları çalıştırıp paralel işlemciler üzerinde aynı anda büyük veriyi işleyip analiz etme imkanı tanır (İlter, 2012).

1.7.4. Makine Öğrenmesi

Otomatik modelleme olarak da adlandırılan bu yaklaşım, veriyi mümkün olan en iyi uyumu yakalayabilmek için birçok farklı modelle sınar. Bu yöntemin avantajı, hızlı hareket eden verilerde ilişkileri en iyi açıklayan modelleri hızlı bir şekilde yaratabilmesidir. Dezavantajı ise verdiği sonuçların yorumlanması ve açıklanmasının

klasik yöntemlere göre daha zor olmasıdır. Yine de büyük veri teknolojilerinin hızı ve hacmi, makine öğrenmesi kullanımını önemli hale getirmektedir (Davenport, 2014). Makine öğrenmesi yöntemleri veri madenciliği algoritmalarında kullanılan yöntemlerin temelini oluşturur. Büyük veri teknolojilerinde de veri madenciliği yöntemleri makine öğrenmesi başlığı altında yer almaktadır.

1.7.4.1. Makine Öğrenmesi Kütüphaneleri

Makine öğrenmesi algoritmalarını (Random Forest, Lojistik Regresyon vs.), bu algoritmalara ait kodları ve kodların çalışması için gerekli olan dll, jar dosyası gibi dosya tiplerini içinde barındırır. En yaygın kullanılan makine öğrenmesi algoritmaları Mahout ve Scala'dır.

Mahout, veri hazırlama, model oluşturma, oluşturulan modelden bilgiye ulaşma gibi özelliklere sahiptir. Sıklıkla Sınıflandırma ve Kümeleme için kullanılır. Mahout içinde en sık kullanılan sınıflandırma algoritmaları Lojistik Regresyon, Bayesci ve Random Forest, kümeleme algoritmaları ise kMeans, Canopy ve MinHash'dir.

Scala, nesne yönelimli ve fonksiyonel programlama dillerini içerdiği için programlama dili olarak da görülmektedir. Kendi derleyicisine sahiptir, bu yüzden kolaylıkla Java kodlarını derleme ve çalıştırma özelliğine sahiptir. Daha çok mühendislik alanlarında büyük veri ile ilgilenen kişiler tarafından tercih edilmektedir. Java'nın sunmuş olduğu tüm kütüphaneleri ve özellikleri kullanabildiğinden Java ortamında üretilen tüm projeleri Scala ortamında da üretmek mümkündür.

1.7.5. Veri Madenciliği Yöntemleri

Veri madenciliği, büyük veri yığınları arasından bilgiye ulaşmayı, yeni bilgi keşfetmeyi, bu bilgileri kullanarak gelecek için tahminler yapmayı amaçlar.

Veri madenciliği istatistiksel yöntemler serisi olarak görülmekle birlikte klasik istatistiksel yöntemlerden farklılık göstermektedir. Veri madenciliğinde amaç, yeni ve keşfedilmemiş olanı çıkarmanın yanında görsel olarak anlaşılır modeller yaratmaktır. Bu yüzden veri madenciliğinde esas istenen insan yönetiminde, insan-bilgisayar etkileşimi kullanarak analiz yapmaktır. Veri madenciliği klasik istatistik analiz yaklaşımlarından farklı olarak makine öğrenmesi, veritabanı yönetimi ve yüksek performanslı işlem gibi alanları da içerir.

1.7.5.1. Sınıflama

En popüler veri madenciliği yöntemlerindendir. Sınıflama yöntemlerinde genel amaç merak edilen bir nesnenin özelliklerini inceleyerek o nesneyi daha önceden tanımlanmış bir sınıfa dahil etmektir. Bu yüzden sınıfın özellikleri en ince ayrıntısına kadar belirlenmiş olmalıdır. Örneğin, göğüs kanseri hastalarının verileri (yaşayıp yaşamağı, öldü ise teşhisten sonra kaç yıl yaşadığı, yaşı, cinsiyeti vs.) incelenerek herhangi bir göğüs kanseri hastanın hastalığı ile ilgili tahminlerde bulunmak için kullanılabilir.

1.7.5.2. Kümeleme

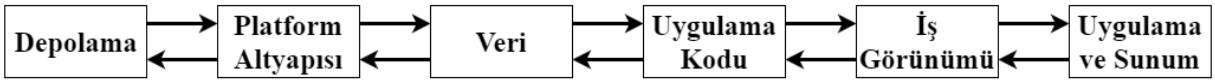
Belirli terim, veri ya da özelliklere göre gruplar oluşturulur. Bu gruplarda en sık yer alan verilerden faydalanılarak bir benzerlik ölçütü oluşturulur ve bu ölçüte göre kümeleme yapılır. Kümeleme algoritmaları sağlık alanında daha çok hastalık teşhisi koymak, hastalıklara ait risk faktörleri belirlemek ve hastalık grupları oluşturmak için kullanılmaktadır.

1.7.5.3. Birliktelik

Birliktelik kuralları aslında birbiriyle ilişkili olabilecek değişkenler arasındaki bağıntıların gösterilmesi ve bu ilişkinin büyüklüğünün saptanabilmesi için kullanılır. Örneğin, bir bankada hesabı olan müşterilerin özel emeklilik sigortasına da sahip olduğuna dair bir bağıntı bulunmuşsa, maaş hesabı olan diğer müşterilere de özel emeklilik sigortası yaptırmaları için öneri de bulunulabilir.

1.7.6. Büyük Veride Teknoloji Katmanları

Tüm stratejik trendler gibi büyük veri de kendini alışılmış sistemlerden ayıran özelleşmiş nitelikler sunmaktadır. Büyük veri teknoloji katmanları sırasıyla Depolama, Platform Altyapısı, Veri, Uygulama Kodu, İşletme Görünümü (modeller, görünümler, küpler) ve Uygulamalardan (görselleştirme, iş zekası, iş analitiği) oluşmaktadır (Davenport, 2014).



Şekil 1.7. Büyük Veri Teknoloji Katmanları

1.7.6.1. Depolama

Disk teknolojileri ticarileşerek daha verimli hale geldikçe büyük ve farklı miktarlardaki veriyi disk üzerinde depolamak da maliyet açısından daha verimli hale gelmiştir. Hadoop ortamlarında depolama genellikle ticari sunuculara bağlı birden fazla disk kullanılarak daha hızlı ve ekonomik olarak yapılabilmektedir (Davenport, 2014).

1.7.6.2. Platform Altyapısı

Büyük veri platformu genellikle büyük verinin yüksek performanslı işlenmesini sağlayan fonksiyonların bütünüdür. Büyük veri platformları genellikle bir Hadoop (veya benzeri açık kaynak kodlu proje) temeline sahiptir (Davenport, 2014).

Hadoop gibi açık kaynaklı teknolojiler fiilen büyük veri işleme platformlarına dönüşmüştür. Alışılmış veri ambarlarıyla sınırlı olmayan kurumlar sadece tek bir Hadoop platformunu, karmaşık iş yüklerini ayırmak ve yapılandırılmamış büyük veri hacimlerini analize hazır yapılandırılmış veriye çevirmek için kullanabilmektedirler (Davenport, 2014).

1.7.6.3. Veri

Büyük verinin gelişmesi de uygulamaları gibi kapsamlı ve karmaşık olmuştur. Birkaç örnek saymak gerekirse büyük veri, insan genomu dizileri, kanserli hücre davranışları, sosyal medya etkileşimleri ve yaşamsal sağlık sinyalleri olarak karşımıza çıkabilir. Teknoloji katmanları arasındaki veri katmanı, verinin ayrı bir yönetim ve idareyi hak eden bir değer olduğunu göstermektedir (Davenport, 2014).

1.7.6.4. Uygulama Kodu

Büyük veri kullanıldığı iş uygulamasına göre nasıl değişiyorsa, veriyi manipüle etmek ve işlemek için kullanılan kodlar da değişmektedir. Hadoop, MapReduce adlı işleme altyapısını sadece veriyi diskler arasında dağıtmak için değil, aynı zamanda bu veriye karmaşık hesaplama talimatları uygulamak için de kullanmaktadır. Platformun yüksek performans becerileriyle uyumlu olarak, MapReduce talimatları büyük veri platformundaki farklı düğümler arasında paralel olarak işlenmekte, sonra da yeni bir veri yapısı veya veri seti sağlamak için hızla birleştirilmektedir (Davenport, 2014).

Apache Pig ve Hive, Hadoop kullanarak uygulama kodlarında MapReduce fonksiyonlarının yürütülebilmesi için üst seviye bir dil sağlayan iki açık kaynaklı betik dildir. Pig veri okumak, filtrelemek, dönüştürmek, birleştirmek ve yazmak gibi işlemleri tarif etmek için bir betik dil sunmaktadır. Bu betik dil Java'dan üst seviye bir dildir (Pig Latin adlı Pig dili Java'ya dönüştürülüyor) ve programlamada daha fazla üretkenliği, esnekliğe olanak tanımaktadır. Hive aynı işlemleri daha fazla toplu işlem (batch) yönelimiyle gerçekleştirmektedir ve veriyi SQL sorgularına uygun ilişkisel formata çevirebilmektedir. Bu Hive'ı SQL sorgu diline hakim analistler için işlevsel hale getirmektedir (Davenport, 2014).

1.7.6.5. İş Görünümü

İş görünümü katmanı büyük veriyi ilave analizlere hazır hale getirmektedir. Büyük veri uygulamasına göre, MapReduce veya özel kod kullanılarak gerçekleştirilen ek işlemlerle, istatistiksel bir model, düz dosya, ilişkisel tablo veya veri küpü gibi bir ara veri yapısı inşa edilmektedir. Sonuçta ortaya çıkan bu yapı, ilave analiz veya SQL tabanlı sorgu araçlarıyla sorgulama yapmak için yaratılmış olur. SQL'in yıllardır kullanılması ve birçok kişinin SQL sorgusu yaratmayı biliyor olması nedeniyle birçok sağlayıcı "SQL ile Hadoop" denen yaklaşımlara kaymaktadır. Bu işletme görünümü, bir kuruluşta hali hazırda mevcut olan araçlar ve bilişim çalışanları sayesinde büyük verinin daha kullanılır hale gelmesini sağlamaktadır (Davenport, 2014).

1.7.6.6. Uygulama ve Sunum

Bu katmanda büyük veri ön işleme adımlarından sonra oluşan sonuçlar analiz edilerek iş kullanıcıları veya diğer sistemler tarafından otomatik kararlar vermek için görüntülenir. Birçok büyük veri tüketicisi büyük verinin görsel olarak gösterilmesini tercih etmektedir. Özelleştirilmiş iş zekası teknolojileri sayesinde veri görselleştirme araçları bilginin duyusal ve grafiksel bir şekilde gösterilmesine olanak tanımaktadır (Davenport, 2014).

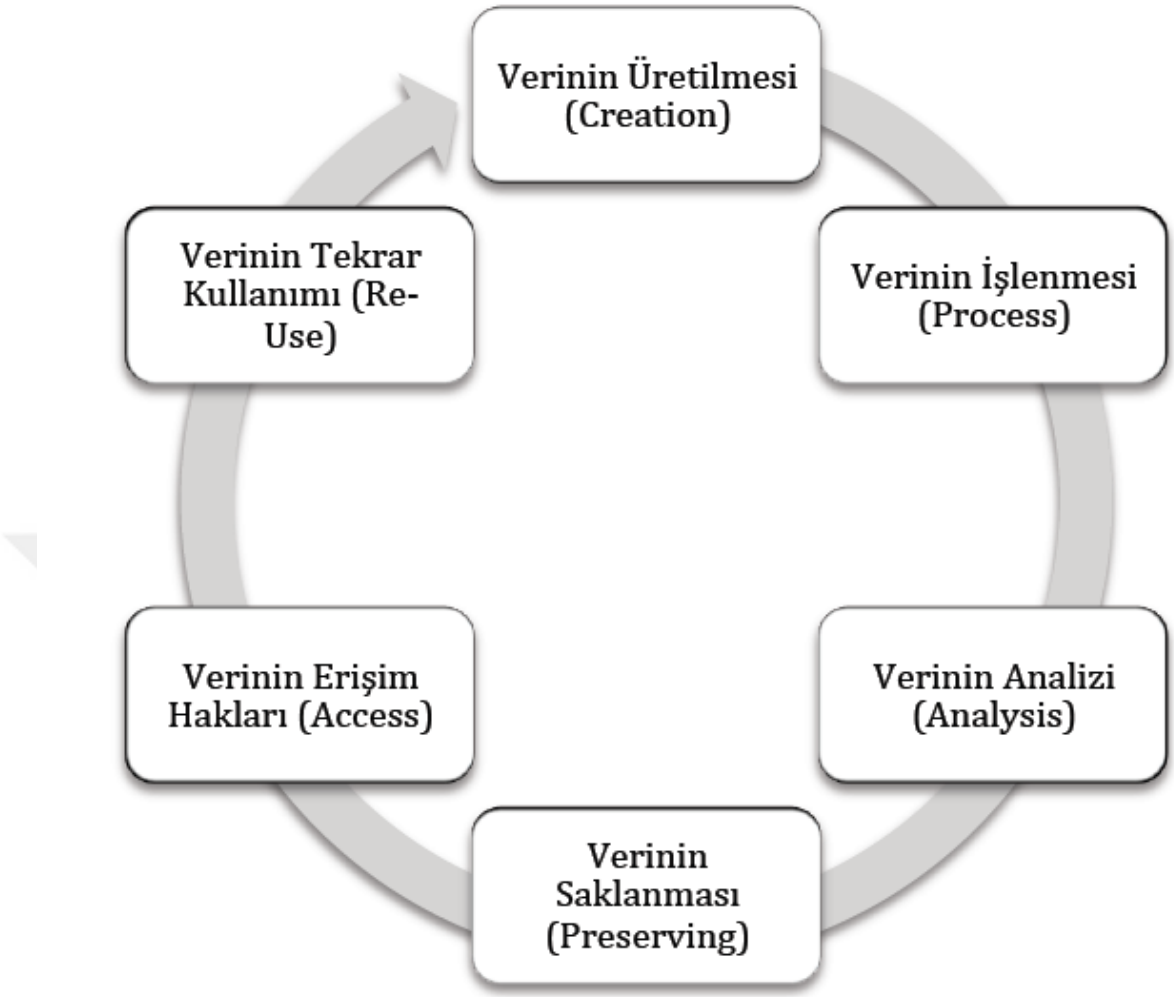
1.8. Büyük Veri Yaşam Döngüleri

1.8.1. Büyük Veri Yaşam Döngüsü Modelleri

Her kavram, değer nasıl gün geçtikçe değişiyorsa büyük veri kavramı da değişmektedir, gelişmektedir ve bu gelişimler dahilinde ihtiyaç bazlı çeşitli yaşam döngülerine sahiptir (Şeker, 2014).

1.8.1.1. Verinin Yaşam Döngüsü (Data Life Cycle)

Aşağıdaki yaşam döngüsü Essex Üniversitesi tarafından oluşturulmuştur. Verinin üretilmesiyle başlayan, işlenmesine, analiz edilmesine, saklanmasına, güvenlik ile ilgili erişim haklarının belirlenmesine kadar devam edip, verinin tekrar kullanılması ile en başa dönen bir süreçtir (Corti ve ark., 2014)



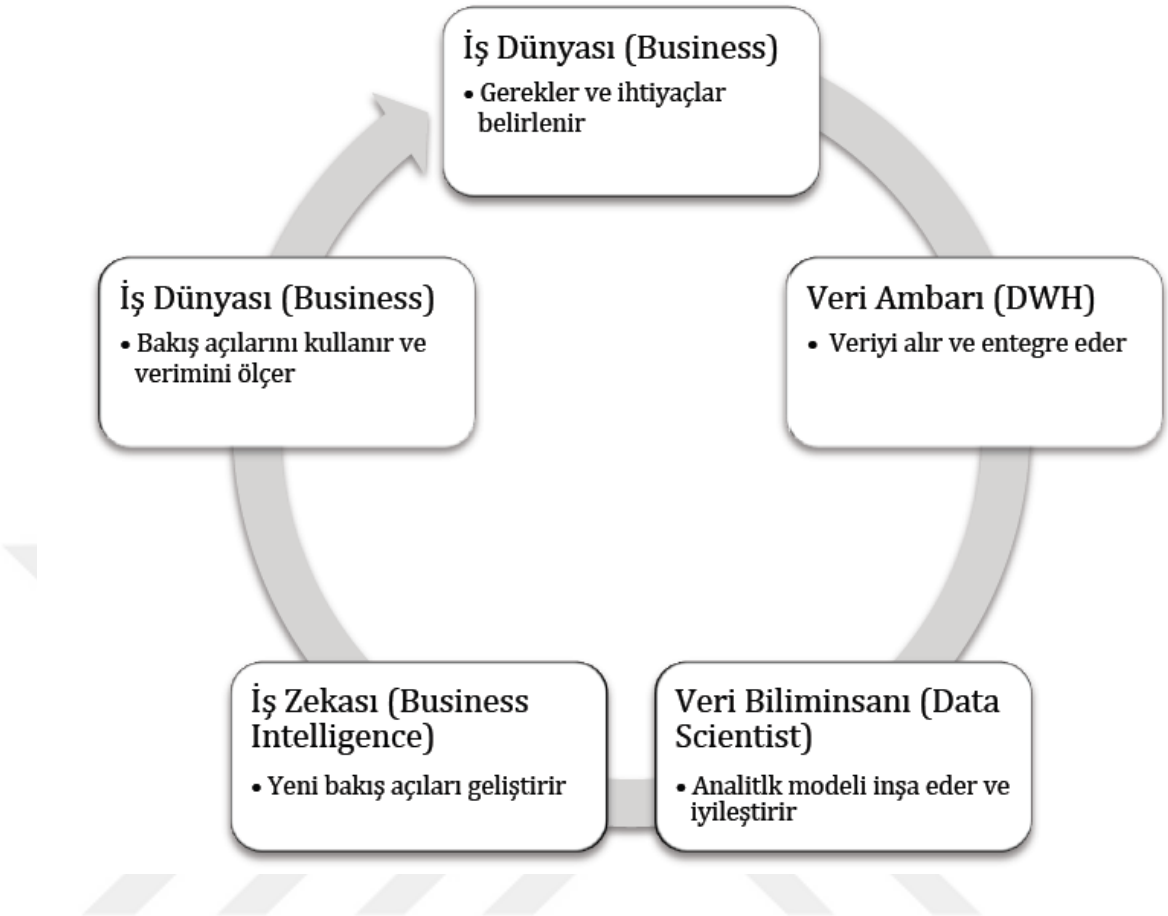
Şekil 1.8. Verinin Yaşam Döngüsü (Data Life Cycle)

Verinin üretilmesi aşamasında veriye neden ihtiyaç duyulduğunun, verinin nasıl üretileceğinin ve verinin hangi kaynaklardan geleceğinin belirlenmesi gerekir. Verinin işlenmesi aşamasında gereksiz verilerin temizlenmesi, eksik verilerin toplanması, verilerin analiz edilebilecek formatlara dönüştürülmesi gibi bazı ön işlemler gereklidir. Bu işlemlerden sonra veri analizi başlamaktadır. Veri analizi aşamasında sadece istenen sonuçlar elde edilmez aynı zamanda sürpriz sonuçlar ya da yeni fikirler ortaya çıkabilir. Örneğin elimizde Facebook'a ait veriler olsun. Bu veriler ile sadece kullanıcılarının yüzde kaçının hangi ülkeden olduğu gibi basit bilgilere değil aynı zamanda kaç tane sahte hesap olduğu, bu sahte hesapları kimlerin ne için açtığına kadar birçok yeni ve değişik bilgiye de ulaşabiliriz. Verinin saklanması aşamasında ise verinin nasıl saklanacağına, saklama için hangi formatın

uygun olduğuna karar verilmesi gerekmektedir. Veriye erişim hakları en önemli aşamalardan biridir. Veriye kimlerin erişebileceğinin, erişim haklarının kimler arasında nasıl paylaştırılacağının iyi belirlenmesi ve bu erişim haklarının yasal çerçevede belirlendiğinden emin olunması gerekmektedir. Son adım olan verinin tekrar kullanılması aşamasında ise gerekli tüm aşamalardan geçen verinin tekrar kullanılması sağlanmaktadır. Kullanılan veri ya da en azında metodoloji başka fikirler, analizler için tekrar kullanılabilir, bu yüzden verinin tekrar kullanıma hazır halde tutulması büyük veri yaşam döngüsü için önemlidir (Şeker, 2015b).

1.8.1.2. Büyük Veri Yaşam Döngüsü

Aşağıdaki model EMC firması tarafından hazırlanmıştır (Schmarzo, 2012). Bu modelde veri ile iş dünyası süreçlerinin nasıl kesiştiği anlatılmaktadır. İlk olarak iş dünyasına ait ihtiyaçlar belirlenmektedir. Daha sonra bu ihtiyaçlara dair veriler derlenir, toplanır ve veri ambarı süreci başlar. Bu süreçte toplanan veri alınır ve sisteme entegre edilir. Bu aşamadan sonra veri bilim insanı veri ambarlarında toplanan verileri alır ve analitik modeller (veri madenciliği modelleri vs.) yardımıyla bu veriyi analiz eder, raporlar. Bir sonraki aşamada artık iş zekası devreye girer ve oluşan raporlar doğrultusunda iş dünyasını şekillendirebilecek yeni fikirler ortaya çıkar. Bu fikirler doğrultusunda yeni ihtiyaçlar belirir ve döngü yeniden başlar (Şeker, 2015a).



Şekil 1.9. EMC Firması Tarafından Geliştirilen Büyük Veri Yaşam Döngüsü

1.8.1.3. Analiz Yaşam Döngüsü

"The Analytics Life Cycle" olarak adlandırılan bu model SAS tarafından geliştirilmiştir ve dünya çapında yaygın olarak kabul görmektedir.

İş dünyası aşaması bir üst düzey yönetici olarak düşünülebilir. Bu kişinin doğru kararlar alıp, yeni fikirler ortaya koyması için doğru analizler sonucu raporlanmış kesin bilgilere ihtiyacı vardır (Şeker, 2015a).

Şekil 1.10. Analiz Yaşam Döngüsü

Veri bilim insanı (data scientist) verinin araştırılması, görsellerle daha anlaşılır hale gelmesi ile uğraşan ve bunun için hazır araçları kullanan kişidir. Veri madencisi yani istatistikçi rolündeki kişi ise veri ile ilgili analizler yaparak veriden yeni değerler, bulgular çıkarmaya çalışır. Veri sınıflandırılması, kategorize edilmesi, kümelere ayrılması ve tahmin modelleri elde edilmesi gibi veri madenciliği uygulamaları ile uğraşır. İstatistik bilgisinin iyi olması gerekmektedir, çünkü sonuçların doğruluğu, yorumlanması gibi konularda bu bilgiye ihtiyaç vardır. Döngüdeki son kişi ise BT sistem yöneticisidir. Bu kişi kurulan modelin sisteme entegre edilmesi, sorunsuz çalışması, eğer model ya da süreç ile ilgili bir problem var ise bu problemin en kısa sürede en sorunsuz yoldan çözülmesi ile uğraşır. Birden çok veri tabanı arasındaki entegrasyon, network üzerinden güvenli veri paylaşımı gibi kritik görevlere sahiptir (Şeker, 2015a).

Bu döngüyü daha detaylı ele alacak olursak, ilk olarak gereksinimlerin detaylıca tanımlanması gerekmektedir. Daha sonra belirlenen problem dahilinde verinin hazırlanması, alınacağı kaynakların belirlenmesi, eğer gerekirse verinin farklı formatlardan dönüşümlerinin yapılarak tek bir formatta toplanması gerekir. Diğer aşama ise hazırlanan verinin veri madencisi ve veri bilim insanı tarafından işlenmesi, analiz edilmesi, modelin oluşturulması aşamasıdır. Daha sonraki aşamalarda da BT sistem yöneticisi tarafından model doğrulaması yapılır. Bu aşamadan sonra süreç tekrar yöneticiye döner ve yönetici çıkan sonuçları kontrol eder, sorun varsa döngü tekrarlanır sonuçlar yinelenir yoksa sonuçlar ışığında karar verme süreci başlar. Her şart altında döngü tekrar başlatılarak verinin benzer ya da farklı problemler için tekrar kullanılması sağlanır (Şeker, 2015a).

1.9. Büyük Veriye Uygun Sektörler (Kullanım Alanları)

Veriler açısından dezavantajlı olan kuruluşlar, fazla veriye sahip olmayanlar veya verileri iyi yapılandırılmamış olan kuruluşlardı, bu nedenle analitik açıdan zayıftılar. Bu gibi kuruluşlar arasında sağlık hizmeti veren şirketler, büyük veri açısından ne kadar kritik öneme sahip olsa da, elektronik tıbbi kayıtların yaygın olmaması ve hastalarla ilgili çok sayıda metin tabanlı not olması (elektronik tıbbi kayıt verisinin yüzde ellisi yapılandırılmamış metinlerden oluşuyor) nedeniyle dezavantajlıydılar.

Çizelge 1.3. Büyük Veri Kullanım Alanları

Sektörlerin Tarihsel Veri ve Analitik Kullanımı		
Veri dezavantajı olanlar	Düşük başarılılar	Üstün başarılılar
Sağlık hizmetleri kuruluşları	Geleneksel bankalar	Tüketici ürünleri
B2B firmalar	Telekom	Sigorta
Endüstriyel ürünler	Medya ve eğlence	Online
	Perakende	Seyahat ve taşımacılık
	Elektrik sağlayıcıları	Kredi kartları

1.9.1. Sağlık Hizmetleri

Sağlık hizmetleri sektöründe hastane ve polikliniklerin uygulamaya koyduğu elektronik hasta kayıt sistemlerinden gelen veriler yapılandırılmış formatta olmasına rağmen, doktor ve hemşirelerinden notlarından oluşan oldukça hacimli miktarda metin verisi ise yapılandırılmış formattadır ama bu veriler analiz edildiği takdirde çok büyük değer içermektedir. Bu metin verileri doğal dil işlemcisi teknolojileriyle gitgide daha fazla kayıt altına alınıp sınıflandırılabilmektedir. Sigorta şirketlerinin elinde muazzam miktarda tıbbi dava verisi vardır, ama bunlar sağlık hizmetleri sağlayıcılarının verileriyle birleştirilmemiştir. Tüm bu veriler birleştirilip sınıflandırılarak analiz edilebilse, hastaların durumu hakkında çok daha fazla bilgi sahibi olunabilir. CAT taramaları ve MRI'lardan gelen görüntü verileri olağanüstü veri kaynaklarından biridir ama bu veriler sadece doktorlar tarafından incelenmekte ama sistematik bir şekilde analiz edilmemektedir. İnsan genomu verisi (kişi başı yaklaşık 2 terabayt miktarında) hızla birçok hastadan toplanabilecek kadar ucuzdur (birkaç yıl içinde hasta başına 1000 dolara mal olacak). Buna ek olarak uzaktan sağlık (tele tıp ve uzaktan izleme) ve kişisel veri (kişisel izleme) cihazlarından alınabilecek devasa miktarda veri bulunmaktadır. Doktor ve hastanelerin tüm hastaların kilo, tansiyon, kalp atışı, fiziksel aktivite ve hatta zihinsel durumuyla ilgili verileri toplayabildiği bir ortamda toplanabilecek verinin büyüklüğü, çeşitliliği ve bu verilerden elde edilebilecek değerler göz ardı edilemeyecek kadar değerlidir (Davenport, 2014).

1.9.2. Sağlıkta (Tıp Alanında) Büyük Veri

Sağlık alanında daha klasik veritabanı yöntemleriyle saklanamayacak çeşitte ve yönetilemeyecek büyüklükte veri üretilmektedir ve üretilen verinin hızı çok fazladır. Asıl veri hızını artıran günlük düzenli ölçüm verileri(kan şekeri, kan basıncı) ve EKG verileri gibi verilerdir. Zamanla kullanımı yaygınlaşmaya başlayan üç boyutlu görüntülerin, genomik ve biyometrik sensör verileri gibi büyük veri türlerinin bu hızı üstel olarak artırması beklenmektedir (Beyan, 2016).

1.9.2.1. Sağlıkta Büyük Veri Analitiklerinin Kullanım Alanları

Büyük veri teknolojilerinin sağlık alanına uygulanması çok büyük avantajlar sağlamaktadır:

-Hastalıklara ilişkin tanılar daha erken evrelerde konabilir ve tedavi şansı artmaktadır.

-Kişilerin sağlık durumlarının kontrolü sağlanabilmekte ve toplum sağlığı merkezleri daha aktif kullanılabilmektedir.

-Hasta ve hastalık geçmişlerine ait veriler kullanılarak hangi tedavinin hangi hastalık ya da hasta için daha uygun olduğu belirlenebilmektedir.

-Maliyet analizi daha kolay, hızlı ve doğru yapılabilmektedir.

-Klinik deneylerin ve hasta kayıtlarının analiz edilmesi ile ürünler pazara verilmeden önce endikasyonları daha iyi ve hızlı tanımlanabilir ve yan etkileri daha net anlaşılabilmektedir.

-Klinik çalışmalarda deney tasarımının daha iyi ve hasta toplama yönteminin daha kontrollü gerçekleşmesi sağlanarak deneysel hatalar azaltılabilmekte ve ürünün pazara sürülme süreci hızlandırılabilmektedir.

-Geleceğe yönelik hastalık tahminleri yapılabilmekte ve bu sayede bulaşıcı ya da salgın hastalıklara ilişkin koruyucu tedbirler (aşı, ilaç) daha erken sağlanabilmektedir.

-Klinik, operasyonel ve genomik veriler birleştirilerek hastalık riski ve en uygun tedavi yöntemleri belirlenebilmektedir.

-Reçete ve diğer hizmetlerin kötü kullanımını önleyici tedbirler geliştirilebilmektedir (Beyan, 2016).

1.9.2.2. Sağlıkta Büyük Veri Metodolojisi

Sağlıkta büyük veri metodolojisi dört adımdan oluşmaktadır:

1. Kavramsal çalışma: Farklı disiplinlerden gelen uzmanlar probleme kavramsal açıklama getirmektedirler. Bu sayede yapılan açıklamaların problem üzerindeki artıları ve eksileri daha net şekilde karşılaştırılabilmektedir (Beyan, 2016).

2. Öneri geliştirme: Probleme yönelik çözümler geliştirilen aşamadır. Bu aşamada problem nedir, hangi aşamalardan oluşur, büyük veri metodolojisi hangi aşamada kullanılmalıdır gibi sorulara cevaplar aranmaktadır. Önceden yapılmış benzer çalışmalar incelenmekte ve bu çalışmalar ışığında problem altyapısı oluşturulmaktadır (Beyan, 2016).

3. Metodoloji: Bu aşamada problem detaylandırılarak parçalara bölünmektedir. Probleme dair verilerin hangi kaynaklardan hangi tekniklerle toplanacağı belirlenmekte ve gerekli ise verilerin hangi formatta tutulacağı, gerekirse dönüşüm teknikleri belirlenmektedir. Veri toplandıktan sonra büyük veri teknikleri uygulanmakta ve gerekli tahmin modelleri oluşturulmaktadır (Beyan, 2016).

4. Test ve değerlendirme: Son aşamada oluşturulan modeller gerekli analiz yöntemleri kullanılarak test edilmekte ve elde edilen bulgular değerlendirilmektedir (Beyan, 2016).

1.9.2.3. Sağlıkta (Tıp Alanında) Büyük Veri Örnekleri

Tıp bilimi büyük verinin insan sağlığı yararına kullanılması açısından büyük önem teşkil etmektedir. Geçmiş yıllardan bugüne kadar toplanan tüm sağlık verilerine veritabanları erişimleri sayesinde kolayca erişilmekte ve büyük veri teknolojileri ile büyük boyuttaki veriler analiz edilerek hem zamandan tasarruf sağlanmakta hem de yeni bilgilere ulaşılmaktadır (Çınar, 2014a).

Tıp alanında büyük veri ile yapılan güncel ve önemli çalışmalardan biri de Manchester Üniversitesi'nde gerçekleştirilmiştir. Bu çalışma sayesinde nörodejeneratif hastalıklara neden olduğu düşünülen önemli bir gen saptanmış ve bu gen en yüksek risk grubunda yer alan kişileri saptamada önemli bir biyogösterge olarak kabul edilmiştir (Çınar, 2014a).

Boston Üniversitesi'nde gerçekleştirilen bir diğer çalışmada ise artık ölümcül salgınların önüne geçilebileceği öngörülmektedir. Günümüzde internet ve mobil telefon kullanımının büyük bir hızla arttığının vurgulandığı bu çalışmada internet ve mobil telefonlardan elde edilen devasa verilerin HealthMaps tarafından işlendiği ve salgının olası seyir yönü, büyüklüğü gibi çıkarımların yapıldığı görülmüştür (Çınar, 2014c).

1.10. Çalışmanın Amacı

Veri hacmi ve değişken sayısının çok büyük olduğu durumlarda, kullanıcı bilgisayarları, tek sanal sunucuya sahip bilgisayarlar gibi klasik sistemlerde veri madenciliği programlarıyla yapılması mümkün olmayan analizlerin büyük veri teknolojileri yardımıyla yapılabildiğini ve sonuçların klasik yöntemlere göre daha iyi olduğunu göstermektir.

2. GEREÇ VE YÖNTEM

Health Facts, Amerika'daki hastanelere ait verileri ayrıntılı olarak kaydeden bir veri tabanıdır. Bu veri tabanı, gönüllü kurumlara ait elektronik medikal kayıtlar, demografik özellikler, hastane içi işlemler, laboratuvar bulguları, eczane verileri, hastane içi ölüm oranları gibi birçok veriyi içermektedir. Çalışmamızda kullanılan veri seti, Health Fact veri tabanında (Cerner Corporation, Kansas City, MO) yer alan verilerden alınmıştır. Bu veri seti 1999-2008 yılları arasında Amerika'nın orta (18 hastane), kuzey doğu (58 hastane), güney (28 hastane) ve batı (16 hastane) bölgelerinde yer alan toplam 130 hastaneye başvuran diyabet hastalarını içermektedir. Bu bölgelere ait hastanelerin birçoğunun yatak kapasitesi 100 ile 500 arasında değişirken yalnızca 38 hastanenin yatak kapasitesi 100'den düşük ve 14 hastanenin yatak kapasitesi ise 500'den büyüktür. Veri setinde hasta ve hastane sonuçları ile ilgili 50 değişken bulunmaktadır. Çalışmamıza bu veri setinden, daha önce diyabet tanısı ile hastaneye başvuran, hastanede 1-14 gün arasında yatan, laboratuvar testi ve ilaç tedavisi uygulanan 101766 kişi dahil edilmiştir (Bren, 2014).

Çalışmamızda büyük veri teknolojilerinden Mahout ve Scala kullanılarak "Yeni İlaç Reçete Edilmesi" değişkenine etki eden faktörlerin bulunması amaçlanmıştır. Birçok gerçek veri setinde olduğu gibi veri setimiz de eksik (incomplete), gereksiz (redundant) ve gürültülü (noisy) veri içermekteydi. Bu nedenle eksik veri yüzdesi büyük, bazı kategorileri çok az veri içeren ve bağımlı değişkene etkisi olmadığı düşünülen değişkenler çalışma dışı bırakılarak 49 bağımsız değişkenden 10 tanesi çalışmaya dahil edilmiştir. Bu değişkenler, ırk (Kafkas, Afrika Kökenli Amerikalı, Asyalı, İspanyol, Diğer), cinsiyet (Kadın, Erkek, Bilinmiyor), yaş (0-9, 10-19, 20-29, 30-39, 40-49, 50-59, 60-69, 70-79, 80-89, 90-100), hastaneye yatış şekli (acil, çok acil, hastanın isteğiyle, yeni doğan, travma, yetersiz bilgi, diğer), hastanede geçirilen süre, laboratuvar işlem sayısı, işlem sayısı (laboratuvar işlemleri hariç), ilaç tedavisi sayısı, sistemde kaydedilen tanı sayısı ve tekrar başvuru (yok, <30 gün, ≥30 gün)'dur. Veri setimizde yer alan değişkenler, tipleri, kategorileri ve eksik veri yüzdeleri Çizelge 2.1.'de verilmiştir.

Çizelge 2.1. Veri Setinde Yer Alan Değişkenler, Tipleri ve Kategorileri

Değişken Adı	Tipi	Değerleri ve Açıklamaları
Yeni İlaç Reçete Edilmesi	Nominal	Başvuru sırasında yeni bir diyabet ilacı reçete edilmesi: Kategoriler: Evet, Hayır
Bağımsız Değişkenler		
Irk	Nominal	Kategoriler: Kafkas, Afrika Kökenli Amerikalı, Asyalı, İspanyol, Diğer
Cinsiyet	Nominal	Kategoriler: Kadın, Erkek, Bilinmiyor
Yaş	Nominal	Kategoriler: 0-9, 10-19, …,90-100
Hastaneye Yatış Şekli	Nominal	Kategoriler: Acil, Çok Acil, Hasta İsteğiyle, Yeni doğan, Travma, Yetersiz bilgi, Diğer
Hastanede Geçirilen Süre	Nümerik	Hastaneye yatış ile taburcu arasında geçen süre
Laboratuvar İşlem Sayısı	Nümerik	Başvuru sırasında yapılan laboratuvar test sayısı
İşlem Sayısı	Nümerik	Başvuru sırasında yapılan işlem sayısı(laboratuvar işlemleri hariç)
İlaç Tedavisi Sayısı	Nümerik	Tedavi sırasında verilen toplam ilaç sayısı
Tanı Sayısı	Nümerik	Sistemde kayıtlı olan tanı sayısı
Tekrar Başvuru	Nominal	Kategoriler: "<30"eğer hasta 30 günden önce tekrar başvurduysa, "≥30" eğer hasta 30 günden sonra tekrar başvurduysa, "Hayır" tekrar başvuruya ait kayıt yoksa.

Çalışmamızda, veriler beş ayrı yüksek işlemcili sunucuya dağıtılmış ve eş zamanlı işlenerek sonuçların gösterilmesi için Hadoop platformundan (kütüphanesi) faydalanılmıştır. Bu kütüphane üzerinde veriye Map ve Reduce işlemleri uygulanarak verilerin sunucular üzerinde dağıtılması, yedeklenmesi, filtrelenmesi, işlenmesi, takibi ve sonuçlarının tek bir sunucu üzerinde gösterilmesi sağlanmıştır. Hadoop platformu üzerinde Hive dili kullanılarak sorgulama ve analiz işlemleri yapılmış ve eksik veriler tespit edilerek nicel değişkenler için ortalama değeriyle, kategorik değişkenler için ise o değişken için en sık görülen değerle değiştirilmiştir.

Veri setimiz %80 eğitim, %20 test verisi olmak üzere rasgele olarak iki ayrı veri setine bölünmüştür. Bu işlemlerden sonra veriler büyük veri analiz teknolojileri olan Mahout ve Scala programlarına aktarılmış, Random Forest ve Çok Katmanlı Algılayıcı algoritmaları yardımıyla 10 bağımsız değişken kullanılarak yeni ilaç reçete edilme tahmini yapılmıştır.

Random Forest Yöntemi

Random Forest yöntemi birden çok karar ağacının bir araya gelmesiyle oluşmaktadır. Breiman tarafından geliştirilen bu algoritmada, tek bir karar ağacı oluşturmak yerine birden çok sayıda çok değişkenli ağacın kararları birleştirilerek bir ağaç oluşturulur. Bu yöntemde çok değişkenli karar ağaçları oluşturmak için CART algoritmasından yararlanılır. Ağacın her seviyesi için hesaplamalar yapılarak nitelikler belirlenir, ardından bu nitelikler birleştirilerek karar ağacı için öznitelik seçimi yapılır ve bu işlem ağacın her seviyesi için tekrarlanır. CART algoritması bilgi kazancını hesaplayarak ağacın dallara hangi değişkenden başlanarak ayrılacağına karar verir. Ağaç oluşturulduktan sonra veriler tek tek bu ağaç modeline göre analiz edilerek sınıflama ve kümeleme yapılır.

Çok Katmanlı Algılayıcı

Çok Katmanlı Algılayıcı yöntemi Rumelhart ve arkadaşları tarafından geliştirilmiştir ve özellikle son yıllarda tahmin ve modelleme için sıklıkla kullanılmaya başlanmıştır. Model bir giriş, veri yapısına bağlı olarak bir veya daha fazla ara katma ve bir çıkış katmanından oluşmaktadır. Bilgi akışı giriş katmanında diğer katmanlara doğru olduğu için bu modele ileri beslemeli sinir ağı modeli de denilmektedir.

Giriş katmanı model tahmininde bulunurken yararlanacağımız bağımsız değişkenlerin bulunduğu katmandır. Ara katman sayısı istatistikçi tarafından belirlenir. Ara katman sayısını belirlerken veri setine hakim olmak, verileri iyi

bilmek çok önemlidir. Çıkış katmanında ise tahmin edilecek bağımsız değişken bulunur. Çok Katmanlı Algılayıcı yönteminde her bir girdi için o girdiye karşılık bir çıktı üretilir. Veri seti giriş katmanında bulunur, ara katmanlarda işlenir ve sonuçlar çıkış katmanında elde edilir.

Çizelge 2.2. Sınıflama Matrisi

Modelin Sınıf Tahmini		Gerçek Sınıf		
		Pozitif	Negatif	Toplam
	Pozitif	Doğru Pozitif Sayısı(TP)	Yanlış Pozitif Sayısı(FP)	TP+FP
	Negatif	Yanlış Negatif Sayısı(FN)	Doğru Negatif Sayısı(TN)	FN+TN
	Toplam	TP+FN	FP+TN	N

Toplam Örnek Sayısı = N = TP+FP+FN+TN

Doğru Sınıflama Oranı (Accuracy) = (TP + TN) /N

Doğru Pozitif Oranı (Duyarlılık) = TP/(TP+FN)

Doğru Negatif Oranı (Seçicilik) = TN/(TN+FP)

F-Ölçütü = 2 x Duyarlılık x Kesinlik / (Duyarlılık + Kesinlik)

Mahout ve Scala teknolojilerinde ayrı ayrı Random Forest ve Çok katmanlı Algılayıcı yöntemleri uygulanmış ve performansları Doğru Sınıflama Oranı ve F-Ölçütü kullanılarak karşılaştırılmıştır. Ayrıca her bir tahminleme yöntemi için Mahout ve Scala programlarının performansları da karşılaştırılmıştır.

3. BULGULAR

Mahout büyük veri teknolojisi ortamında uygulanan Random Forest ve Çok Katmanlı algılayıcı yöntemleri, Mahout Random Forest ve Mahout Çok Katmanlı Algılayıcı başlıkları altında incelenirken, Scala büyük veri teknolojisi ortamında uygulanan yöntemler Scala Random Forest ve Scala Çok Katmanlı Algılayıcı başlıkları altında incelenecektir.

Çalışmanın veri setini oluşturan bağımlı değişkenin düzeylerinde açıklayıcıyı değişkenlere ilişkin tanımlayıcı istatistikler Çizelge 3.1.'de ve Çizelge 3.2.'de verilmiştir.

Çizelge 3.1. Veri Setinde Yer Alan Nitel Değişkenlere Ait Tanımlayıcı İstatistikler

		Yeni İlaç Reçete Edilmesi		
		Yok (n=23403)	Var (n=78363)	Genel (n=101766)
Irk, n(%)	Kafkas	18050 (23.0)	60320 (77.0)	78370 (77.3)
	Afrika Kökenli Amerikalı	4413 (23.0)	14797 (77.0)	19210 (18.8)
	Asyalı	287 (19.1)	1217 (80.9)	1504 (1.4)
	İspanyol	167 (26.0)	476 (74.0)	643 (0.5)
	Diğer	487 (23.9)	1552 (76.1)	2039 (2.0)
Cinsiyet, n(%)	Kadın	12920 (23.6)	41788 (76.4)	54708 (53.7)
	Erkek	10480 (22.3)	36575 (77.7)	47055 (46.2)
	Bilinmiyor	1(33.3)	2(66.7)	3(0.1)
Yaş, n(%)	0-9	29 (18.0)	132 (82.0)	161 (0.2)
	10-19	92 (13.3)	599 (86.7)	691 (0.7)
	20-29	343 (20.7)	1314 (79.3)	1657 (1.6)
	30-39	927 (24.6)	2848 (75.4)	3775 (3.7)
	40-49	2281 (23.6)	7402 (76.4)	9683 (9.5)

	50-59	3856 (22.3)	13400 (77.7)	17256 (17.0)
	60-69	4873 (21.7)	17610 (78.3)	22483 (22.1)
	70-79	5878 (22.5)	20190 (77.5)	26068 (25.6)
	80-89	4284 (24.9)	12915 (75.1)	17199 (16.9)
	90-100	850 (30.4)	1943 (69.6)	2793 (2.7)
Hastaneye Yatış Şekli, n(%)	Acil	12725 (23.6)	41265 (76.4)	53990 (53.0)
	Çok Acil	3958 (21.4)	14522 (78.6)	18480 (18.2)
	Hasta İsteğiyle	4263 (22.6)	14606 (77.4)	18869 (18.4)
	Yeni Doğan	3(30.0)	7(70.0)	10 (0.1)
	Travma	7 (33.3)	14 (66.7)	21 (0.1)
	Yetersiz Bilgi	896 (18.7)	3889 (81.3)	4785 (4.7)
	Diğer	49 (15.3)	271 (84.7)	320 (0.3)
	NULL	1510 (28.5)	3781 (71.5)	5291 (5.2)
Tekrar Başvuru, n(%)	<30	2247 (19.8)	9110 (80.2)	11357 (11.2)
	≥30	7228 (20.3)	28317 (79.7)	35545 (34.9)
	Hayır	13931 (25.4)	40934 (74.6)	54864 (53.9)

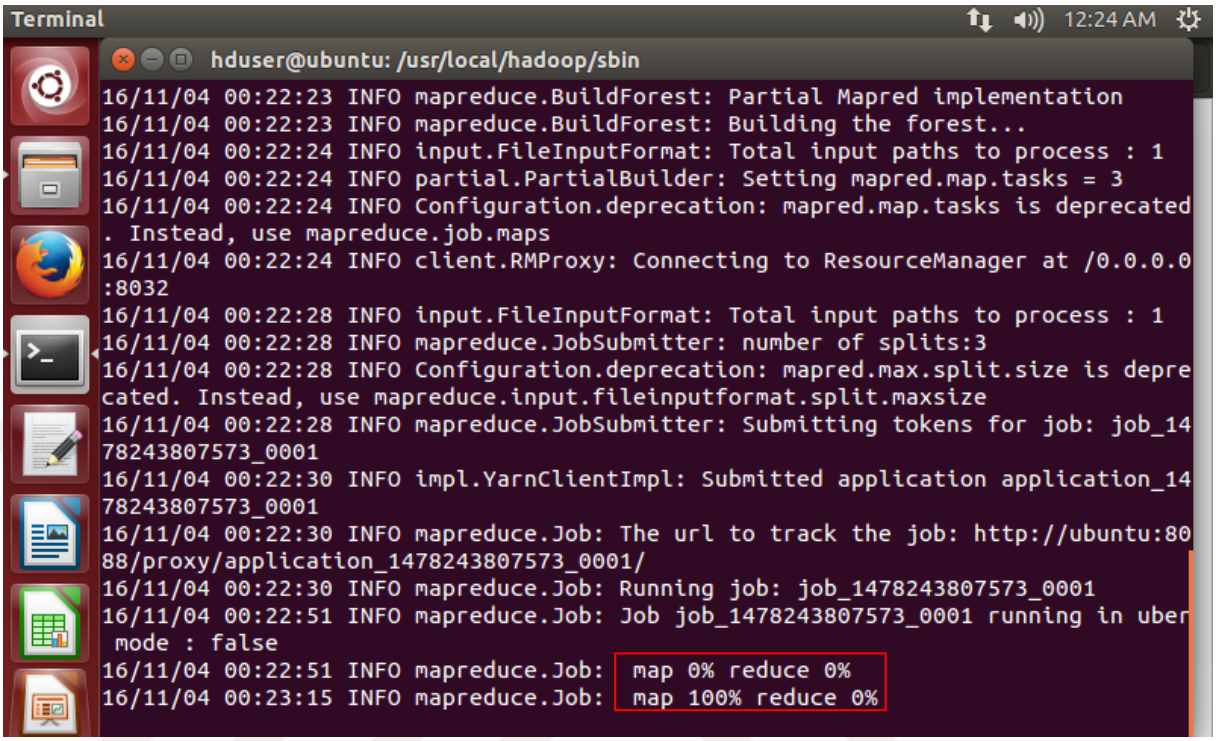
Çizelge 3.2. Veri Setinde Yer Alan Nicel Değişkenlere Ait Tanımlayıcı İstatistikler

	Yeni İlaç Reçete Edilmesi					
	Yok (n=23403)		Var (n=78363)		Genel (n=101766)	
	Ort.±SS	Ortanca (Min-Maks)	Ort.±SS	Ortanca (Min-Maks)	Ort.±SS	Ortanca (Min-Maks)
Hastanede Geçirilen Süre(gün)	4.1 ± 2.9	3.0 (1.0 – 14.0)	4.5 ± 3.0	4.0 (1.0 – 14.0)	4.4 ± 3.0	4.0 (1.0 – 14.0)
Laboratuvar İşlem Sayısı	41.9 ± 19.1	43.0 (1.0 – 129.0)	43.5 ± 19.8	45.0 (1.0 – 132.0)	43.1 ± 19.7	44.0 (1.0 – 132.0)
İşlem Sayısı	1.4 ± 1.7	1.0 (0.0 – 6.0)	1.3 ± 1.7	1.0 (0.0 – 6.0)	1.3 ± 1.7	1.0 (0.0 – 6.0)
İlaç Tedavisi Sayısı	13.2 ± 7.0	12.0 (1.0 – 69.0)	16.8 ± 8.2	15.0 (1.0 – 81.0)	16.0 ± 8.1	15.0 (1.0 – 81.0)
Tanı Sayısı	7.3 ± 1.9	8.0 (1.0 – 16.0)	7.4 ± 1.9	9.0 (1.0 – 16.0)	7.4 ± 1.9	8.0 (1.0 – 16.0)

3.1. Mahout Random Forest

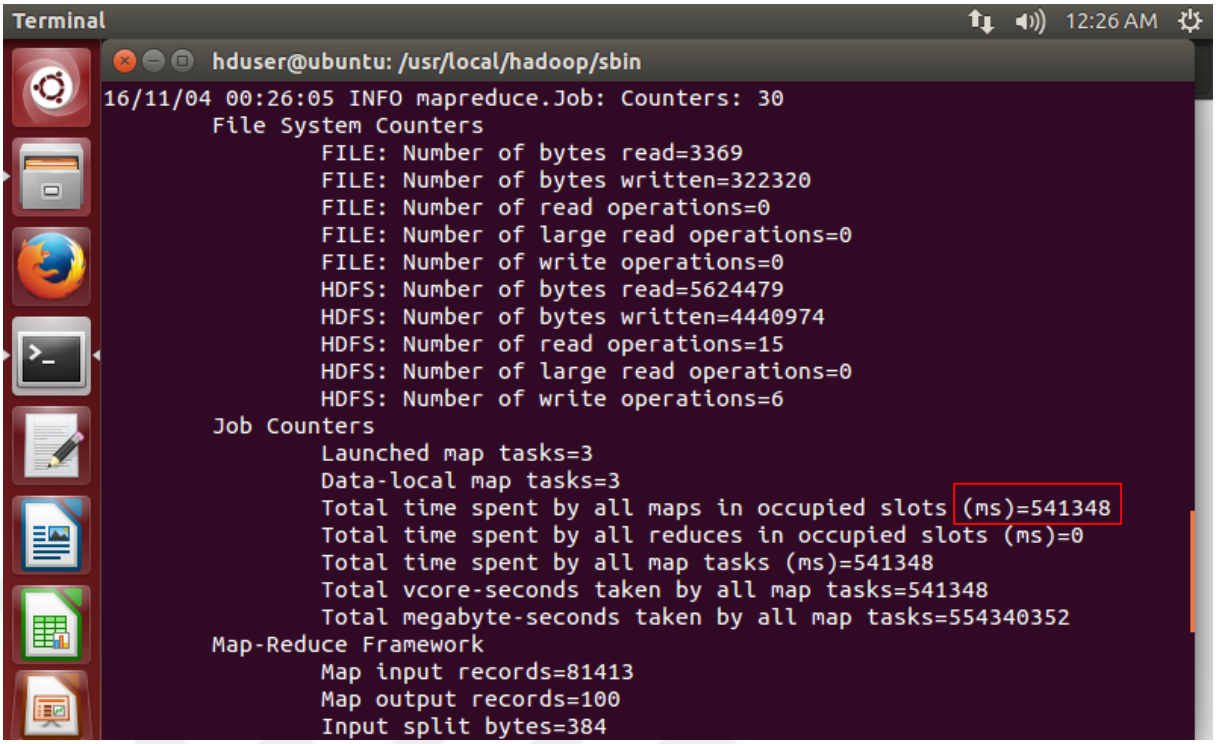
Mahout Random Forest yönteminin ilk aşamasında eğitim verisi Mahout ortamında çalıştırılmış, veri en uygun modeli oluşturana kadar çeşitli sayılara bölünerek en uygun model oluşturulmaya çalışılmıştır. Aynı zamanda verileri filtreleyip anahtar/değer çiftleri oluşturmak için map ve veriyi indirgemek için reduce işlemleri uygulanmıştır. Şekil 3.1.'de eğitim verisinde Mahout Random Forest yönteminin veriyi bölme sayısı, map ve reduce aşamalarına ait örnek yüzdeler gösterilmiştir. Bu analiz sürecinde alınan örnek şekilde veri üçe bölünmüş (number

of splits:3), map ve reduce aşamalarına ait yüzdeler map 0% reduce 0% (başlangıç aşaması) ve map 100% reduce 0% olarak gösterilmiştir.



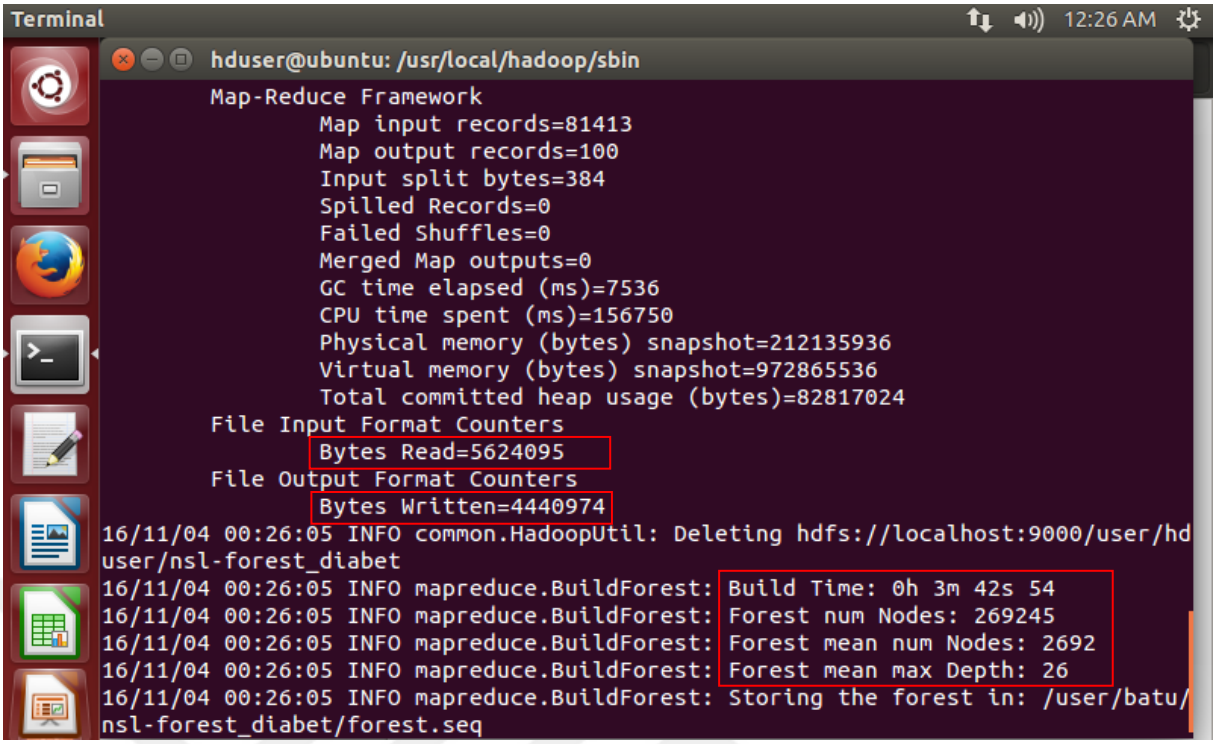
Şekil 3.1. Mahout Random Forest İçin Map ve Reduce İşlemleri

Mahout Random Forest eğitim modelini oluştururken okunan, yazılan operasyon ve bayt miktarları ve işlem süreleri Şekil 3.2.'de gösterilmiştir. Bu modelde tüm donanımı kullanarak yapılan map ve reduce işlemleri için geçen süre yaklaşık dokuz dakikadır (541348 ms). Bu veriyi 5 ayrı sunucu üzerine dağıtıp eş zamanlı yaptığımız düşünülürse, normal özelliklerdeki bir bilgisayarda bu işlem yaklaşık bir saat sürmektedir. Okunan ve yazılan bayt miktarları karşılaştırılacak olursa, dosyadan okuma yapılırken okunan ve yazılan bayt miktarlarının HDFS üzerinde işlenirken okunan miktarlara göre çok düşük olduğu görülmektedir (FILE Number of bytes read=3369, HDFS Number of bytes read=5624479).



Şekil 3.2. Mahout Random Forest Eğitim Modelini Oluştururken Okunan, Yazılan Operasyon ve Bayt Miktarları ve İşlem Süreleri

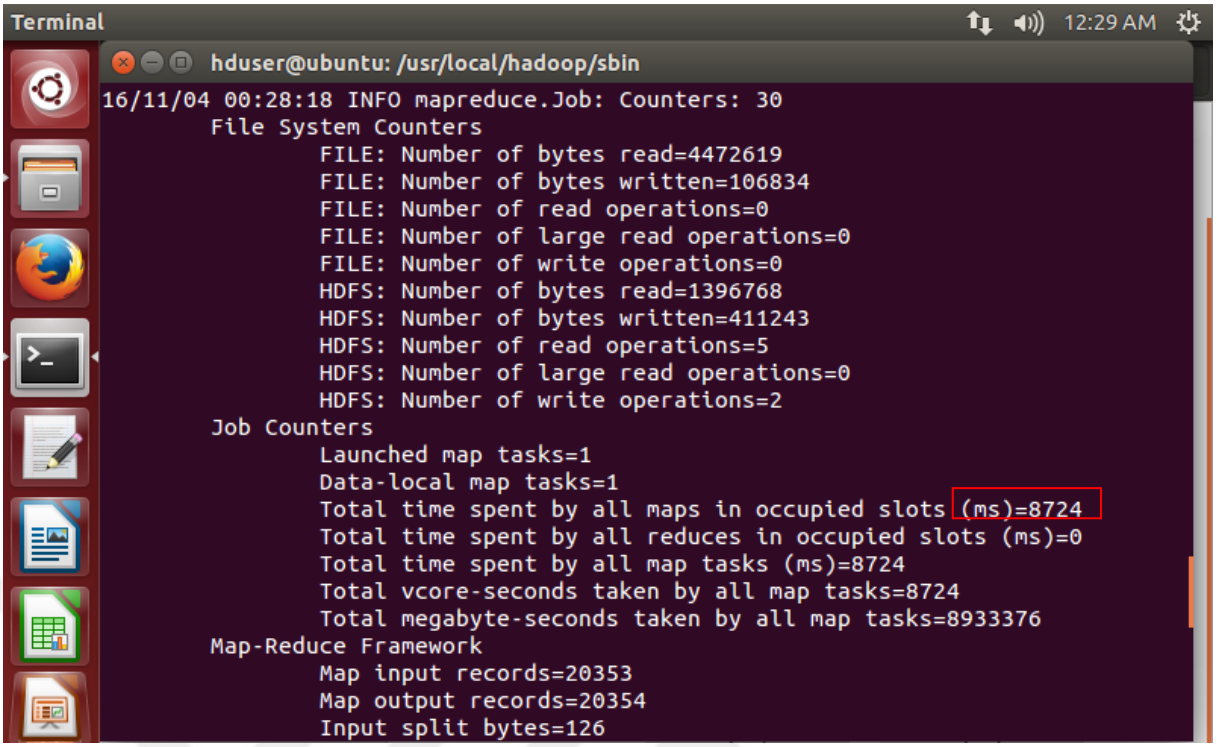
Şekil 3.3.'te Mahout Random Forest yönteminde eğitim modelini oluşturmak için harcanan süre, bayt miktarları ve oluşturulan ağacın düğüm ve derinlik sayıları gösterilmiştir. Bu modelin okuma işlemi esnasında 5624095 bayt veri gerekirken, işlenip yazılması aşamasında ise 4440974 bayt veri gerekmektedir. Bu aşamada bile reduce işlemi sayesinde yaklaşık %25'lik bir kazanç sağlanmıştır. Eğitim modeli 3 dakika 42 sn'de oluşturulmuştur. Bu işlem günümüzde kullanılan standart donanıma sahip beş ayrı bilgisayara dağıtılıp işlenebilseydi yaklaşık 30 dakika sürecekti. Mahout teknolojisi Random Forest yöntemindeki karar ağacına ait düğüm ve derinlik sayılarını en ideal sonucu, en az işlem gücüyle (donanım kullanımı) en az sürede verecek şekilde seçerek modeli o düğüm ve derinlik sayısına göre oluşturmaktadır. Bizim modelimizde en ideal düğüm sayısı (Forest num Nodes) 269245, ortalama düğüm sayısı (Forest mean num Nodes) 2692, derinlik (Forest mean max Depth) ise 26 olarak bulunmuştur.



Şekil 3.3. Mahout Random Forest Eğitim Modeli İçin Harcanan Süre, Bayt Miktarları ve Oluşturulan Ağacın Düğüm ve Derinlik Sayıları

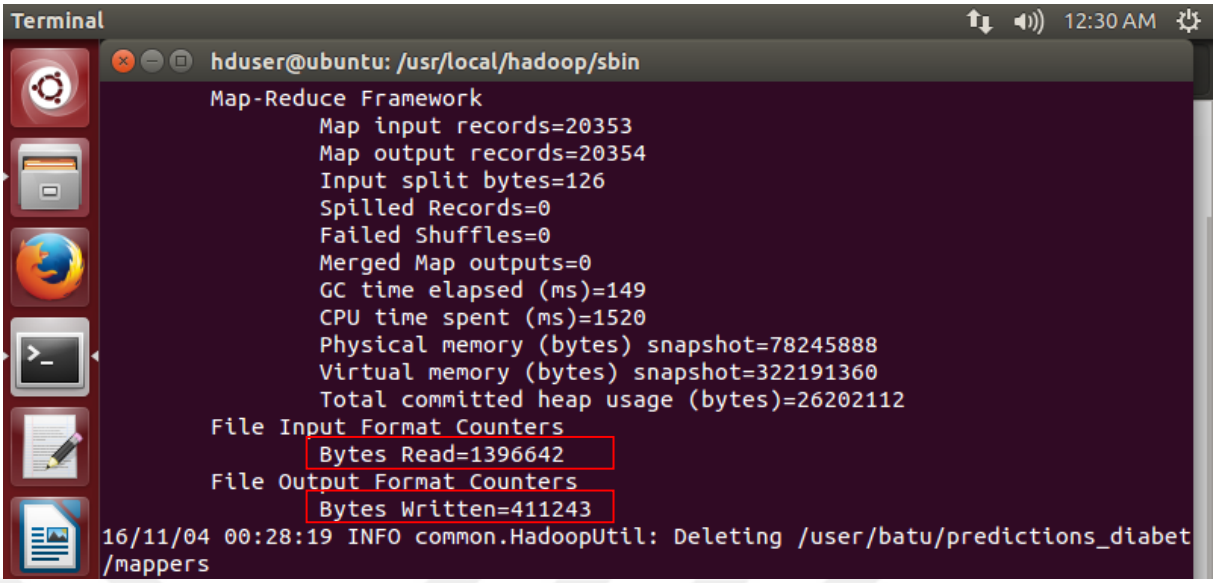
Mahout Random Forest yardımıyla eğitim verisi kullanılarak geliştirilen model test verisi kullanılarak Doğru Sınıflama Oranı ve F- Ölçütü değerleri elde edilmiştir.

Mahout Random Forest eğitim modelini oluştururken okunan, yazılan operasyon ve bayt miktarları ve işlem süreleri Şekil 3.4.'de gösterilmiştir. Bu modelde tüm donanımı kullanarak yapılan map ve reduce işlemleri için geçen süre yaklaşık 9 saniyedir (8724 ms). Okunan ve yazılan bayt miktarları karşılaştırılacak olursa, dosyadan okuma yapılırken okunan ve yazılan bayt miktarlarının HDFS üzerinde işlenirken okunan miktarlara göre yüksek olduğu görülmektedir (FILE Number of bytes read=4472619, HDFS Number of bytes read=1396768). Bunun nedeni model eğitim aşamasında oluşturulduğu için HDFS üzerinde yapılacak işlem sayısının azalmış olmasıdır.



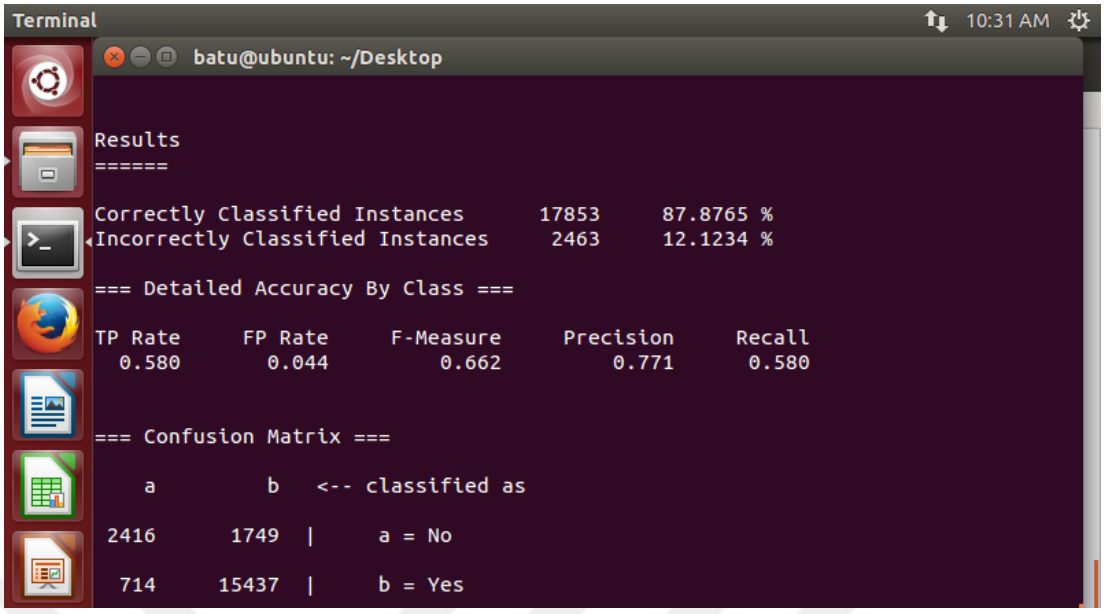
Şekil 3.4. Mahout Random Forest Test Modeli İçin Harcanan Süre ve Diğer Operasyonlar

Test modeli için Mahout Random Forest yönteminde gereken bayt miktarları Şekil 3.5.'te gösterilmiştir. Veriyi okurken 1396642 ve yazarken 411243 bayt gerekmektedir. Veriyi yazarken gereken bayt miktarının üçte birine düşmüş olması Mahout teknolojisinin reduce işleminde ne kadar başarılı olduğunu göstermektedir. Bu da test verisi sonuçları için harcanan süre miktarını önemli ölçüde düşürmektedir.



Şekil 3.5. Mahout Random Forest Test Modeli İçin Gereken Bayt Miktarları

Şekil 3.6.'da Mahout teknolojisi kullanılarak yapılan 26 derinliğe sahip Random Forest modeli sonuçları gösterilmiştir. Test modeli sonucu Doğru Sınıflama Oranı %87.8 ve F-Ölçütü değeri 0.662 olarak bulunmuştur. Aşırı uyum içermeyen bir model olduğu ve nicel değişkenleri de içerdiği için bu sınıflama oranı klasik veri madenciliği programları (WEKA, R, RapidMiner vs.) ile yapılan Random Forest analizlerine göre tercih edilebilir.



Şekil 3.6. Mahout Random Forest Test Modeli Sonuçları

Model sonucunda elde edilen sınıflama matrisi Çizelge 3.3.'te verilmiştir.

Çizelge 3.3. Mahout Random Forest Sınıflama Matrisi

Modelin sınıf tahmini	Gerçek Sınıf		
Yeni İlaç reçete edilmesi durumu	Edilmeyenler	Edilenler	Toplam
Edilmeyenler	2416	714	3130
Edilenler	1749	15437	17186
Toplam	4165	16151	20316

Toplam Örnek Sayısı = N = 20316

Doğru Sınıflama Oranı (Accuracy) = (TP + TN) /N = (2416+15437)/20316=0.878

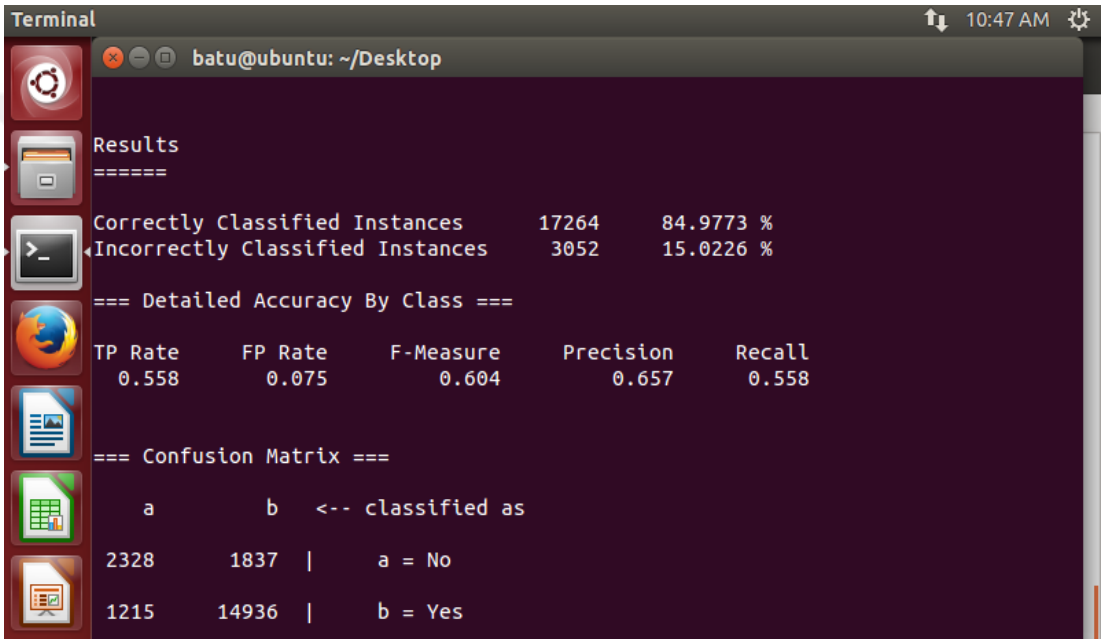
Doğru Pozitif Oranı (Duyarlılık) = TP/(TP+FN) = 2416/(2416+1749)=0.580

Doğru Negatif Oranı (Seçicilik) = TN/(TN+FP) = 15437/(15437+714)=0.956

F-Ölçütü = 2 x Duyarlılık x Kesinlik(Precision) / (Duyarlılık + Kesinlik) =2x0.580x0.771/(0.580+0.771)=0.662

3.2. Scala Random Forest

Şekil 3.7.'de Scala kullanılarak yapılan ve Mahout ile benzer parametreleri içermesi açısından 26 derinlik içeren Random Forest modeli sonuçları gösterilmiştir. Test modeli sonucu Doğru Sınıflama Oranı %84.9 ve F-Ölçütü değeri 0.604 olarak bulunmuştur. Bu oranlar gayet yüksek olsa da Mahout teknolojisi ile oluşturulan Random Forest modeline göre daha düşük çıkmıştır (Doğru Sınıflama Oranı: %87.8 - %84.9, F-Ölçütü: 0.662 – 0.604).



Şekil 3.7. Scala Random Forest Test Modeli Sonuçları

Model sonucunda elde edilen sınıflama matrisi Çizelge 3.4.'te verilmiştir.

Çizelge 3.4. Scala Random Forest Sınıflama Matrisi

Modelin sınıf tahmini	Gerçek Sınıf		
Yeni İlaç reçete edilmesi durumu	Edilmeyenler	Edilenler	Toplam
Edilmeyenler	2328	1215	3543
Edilenler	1837	14936	16773
Toplam	4165	16151	20316

Toplam Örnek Sayısı = N = 20316

Doğru Sınıflama Oranı (Accuracy) = (TP + TN) /N = (2328+14936)/20316 = 0.849

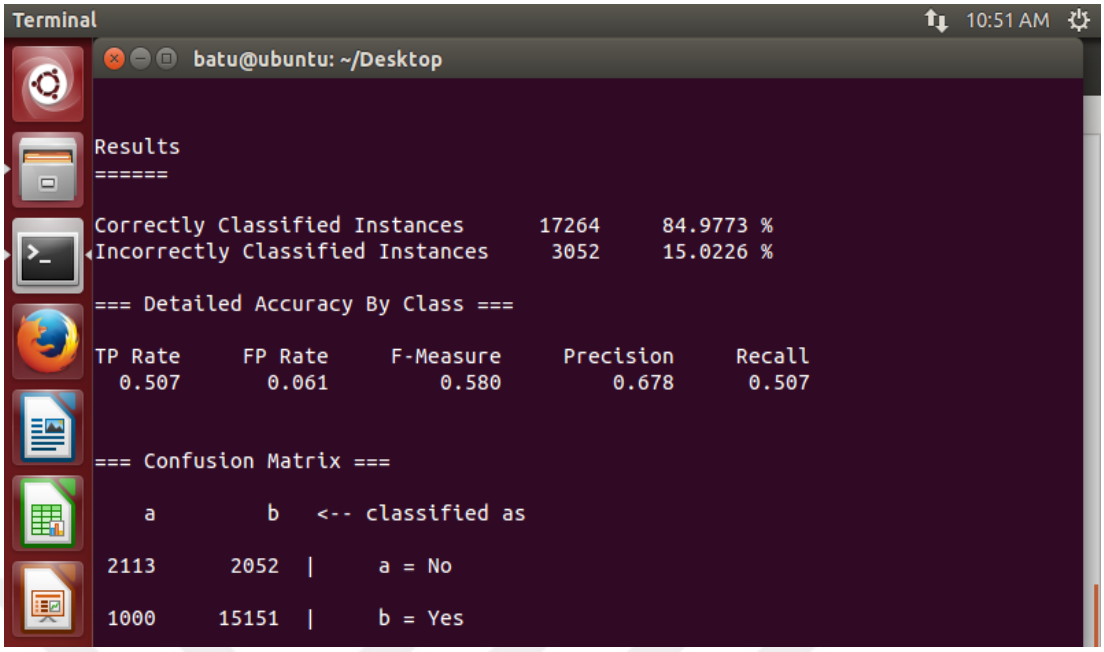
Doğru Pozitif Oranı (Duyarlılık) = TP/(TP+FN) = 2328/(2328+1837) = 0.558

Doğru Negatif Oranı (Seçicilik) = TN/(TN+FP) = 14936/(14936+1215) = 0.925

F-Ölçütü = 2 x Duyarlılık x Kesinlik(Precision) / (Duyarlılık + Kesinlik) =2x0.558x0.657/(0.558+0.657) = 0.604

3.3. Mahout Çok Katmanlı Algılayıcı

Şekil 3.15.'te Mahout teknolojisi kullanılarak yapılan Çok Katmanlı Algılayıcı analizi sonuçları gösterilmiştir. Test modeli sonucu Doğru Sınıflama Oranı %84.9 ve F-Ölçütü değeri 0.580 olarak bulunmuştur. Aşırı uyum içermeyen bir model olduğu ve nicel değişkenleri de içerdiği için bu sınıflama oranı klasik veri madenciliği programları (WEKA, R, RapidMiner vs.) ile yapılan Çok Katmanlı Algılayıcı analizlerine göre tercih edilebilir.



Şekil 3.8. Mahout Çok Katmanlı Algılayıcı Test Verisi Sonuçları

Model sonucunda elde edilen sınıflama matrisi Çizelge 3.5.'te verilmiştir.

Çizelge 3.5. Mahout Çok Katmanlı Algılayıcı Sınıflama Matrisi

Modelin sınıf tahmini	Gerçek Sınıf		
Yeni İlaç reçete edilmesi durumu	Edilmeyenler	Edilenler	Toplam
Edilmeyenler	2113	1000	3113
Edilenler	2052	15151	17203
Toplam	4165	16151	20316

Toplam Örnek Sayısı = N = 20316

Doğru Sınıflama Oranı (Accuracy) = (TP + TN) /N = (2113+15151)/20316=0.849

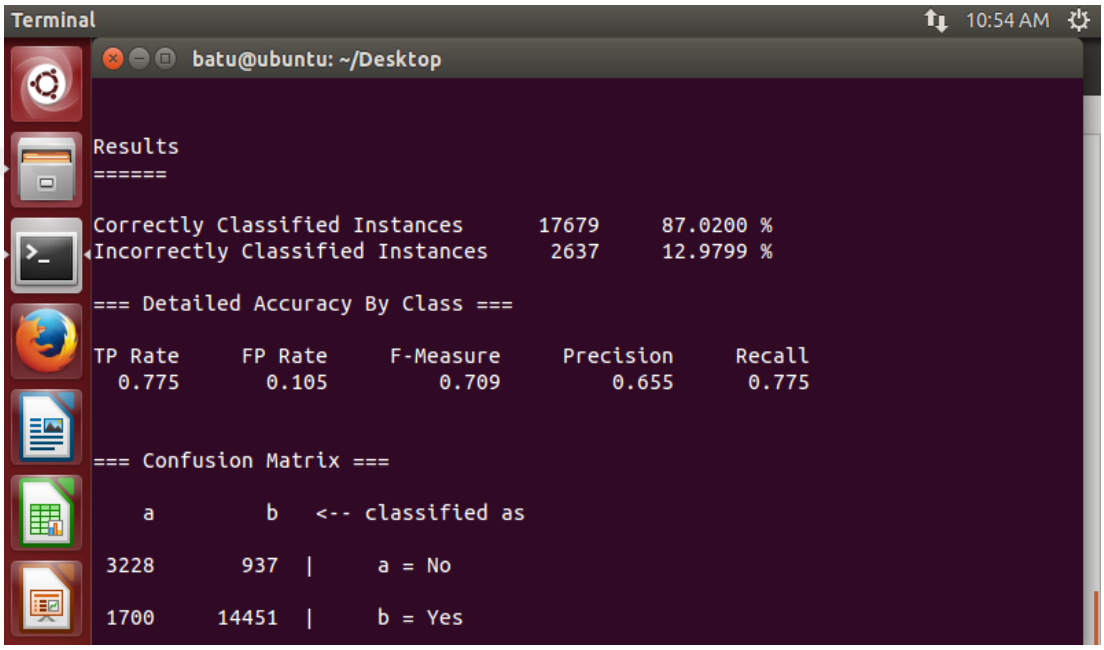
Doğru Pozitif Oranı (Duyarlılık) = TP/(TP+FN) = 2113/(2113+2052)=0.507

Doğru Negatif Oranı (Seçicilik) = TN/(TN+FP) = 15151/(15151+1000)=0.939

F-Ölçütü = 2 x Duyarlılık x Kesinlik(Precision) / (Duyarlılık + Kesinlik) =2x0.507x0.678/(0.507+0.678)=0.580

3.4. Scala Çok Katmanlı Algılayıcı

Şekil 3.9.'da Scala teknolojisi kullanılarak yapılan Çok Katmanlı Algılayıcı analizi sonuçları gösterilmiştir. Test modeli sonucu Doğru Sınıflama Oranı %87.0 ve F-Ölçütü değeri 0.709 olarak bulunmuştur. Bu oranlar Mahout teknoloji ile yapılan Çok Katmanlı Algılayıcı analizi sonuçları ile karşılaştırıldığında daha yüksek çıkmıştır (Doğru Sınıflama Oranı: %87.0 - %84.9, F-Ölçütü: 0.709 – 0.580).



Şekil 3.9. Scala Çok Katmanlı Algılayıcı Test Verisi Sonuçları

Model sonucunda elde edilen sınıflama matrisi Çizelge 3.6.'da verilmiştir.

Çizelge 3.6. Scala Çok Katmanlı Algılayıcı Sınıflama Matrisi

Modelin sınıf tahmini	Gerçek Sınıf		
Yeni İlaç reçete edilmesi durumu	Edilmeyenler	Edilenler	Toplam
Edilmeyenler	3228	1700	4928
Edilenler	937	14451	15388
Toplam	4165	16151	20316

Toplam Örnek Sayısı = N = 20316

Doğru Sınıflama Oranı (Accuracy) = (TP + TN) /N = (3228+14451)/20316=0.870

Doğru Pozitif Oranı (Duyarlılık) = TP/(TP+FN) = 3228/(3228+937)=0.775

Doğru Negatif Oranı (Seçicilik) = TN/(TN+FP) = 14451/(14451+1700)=0.895

F-Ölçütü = 2 x Duyarlılık x Kesinlik(Precision) / (Duyarlılık + Kesinlik) =2x0.775x0.655/(0.775+0.655)=0.709

3.5. Mahout Sınıflama Yöntemleri Performans Karşılaştırması

Mahout teknolojisi kullanılarak yapılan Random Forest ve Çok Katmanlı Algılayıcı yöntemlerine ait doğru sınıflama oranları sırasıyla %87.8 ve %84.9 olarak bulunmuştur. Bu oranlara göre Mahout teknolojisi kullanılarak elde edilen en yüksek doğru sınıflama oranı Random Forest yöntemine aittir.

Çizelge 3.7. Mahout Sınıflama Yöntemlerinin Performans Karşılaştırılması

	Random Forest	Çok Katmanlı Algılayıcı
Doğru Sınıflama Oranı	0.878	0.849
F-ölçütü	0.662	0.580

3.6. Scala Sınıflama Yöntemleri Performans Karşılaştırması

Scala teknolojisi kullanılarak yapılan Random Forest ve Çok Katmanlı Algılayıcı yöntemlerine ait Doğru Sınıflama Oranları sırasıyla %84.9 ve %87.0 olarak bulunmuştur. Bu oranlara göre Scala teknolojisi kullanılarak elde edilen en yüksek doğru sınıflama oranı Çok Katmanlı Algılayıcı yöntemine aittir.

Çizelge 3.8. Scala Sınıflama Yöntemlerinin Performans Karşılaştırılması

	Random Forest	Çok Katmanlı Algılayıcı
Doğru Sınıflama Oranı	0.849	0.870
F-Ölçütü	0.604	0.709

3.7. Sınıflama Yöntemlerinin Teknolojilere Göre Performans Karşılaştırması

Scala ve Mahout teknolojileri kullanılarak yapılan Random Forest ve Çok Katmanlı Algılayıcı yöntemlerine ait doğru sınıflama oranları ve F-ölçütü değerleri Çizelge 3.9.'da verilmiştir. Random Forest ve Çok Katmanlı Algılayıcı yöntemlerine göre Mahout ve Scala teknolojilerinin Doğru Sınıflama Oranları sırasıyla %87.8 ve %84.9 ve %84.9 ve %87.0 olarak bulunmuştur. Bu oranlara göre Mahout teknolojisi kullanılarak elde edilen Random Forest doğru sınıflama oranının diğer veri madenciliği yöntemlerine göre daha yüksek olduğu görülmektedir.

Çizelge 3.9. Mahout ve Scala Teknolojilerinin Sınıflama Yöntemlerine Göre Performanslarının Karşılaştırılması

	Random Forest	Çok Katmanlı Algılayıcı
Mahout Doğru Sınıflama Oranı	0.878	0.849
Scala Doğru Sınıflama Oranı	0.849	0.870
Mahout F-Ölçütü	0.662	0.580
Scala F-Ölçütü	0.604	0.709

4. TARTIŞMA

Demirtaş ve ark. (2015)'a göre büyük veri; "işletme, devlet ve organizasyonların, çoğunluğu yapılandırılmamış olan ve sonu gelmez bir şekilde birikmeye devam eden, geleneksel ilişki bazlı veritabanı teknikleri yardımıyla çözülemeyecek kadar yapısallıktan uzak, çok büyük, çok ham ve üstel bir şekilde büyümekte olan veri setlerini bütünleştirerek istatistik ve veri madenciliği teknikleriyle gizli kalmış bilgileri ve sürpriz korelasyonları ortaya çıkarması" olarak tanımlanır. Literatürde Türkiye'de sağlık alanında büyük veri üzerinde veri madenciliği yöntemleri kullanılarak yapılmış bir çalışmaya rastlanmamıştır. Çalışmamızda büyük verinin sağlık alanındaki uygulama alanlarına değinilmiş ve veri madenciliği yöntemleri kullanılarak nasıl analiz edilebileceği gösterilmiştir.

Yurtdışı çalışmalarda ise; Hay ve ark. (1990) küresel bulaşıcı hastalık takibi için büyük veri kullanmaya yönelik fırsatları tartışmıştır. Harita üzerinde gerçek zamanlı risk izleme sağlayan bir sistem geliştirerek makine öğrenme ve kitle kaynak kullanımının hastalık takibine yönelik sürekli güncellenen bir atlas geliştirmekte yeni olanaklar sağladığına işaret etmişlerdir. Epidemiyolojik bilgiyle harmanlanmış çevrim içi sosyal medyanın kamu sağlığı gözetimini kolaylaştıran değerli ve yeni bir veri kaynağı olduğuna inanmıştır.

Hastalık takibine yönelik sosyal medya kullanımı, Young ve ark. (1991) tarafından ortaya koyulmuş, araştırmacılar çalışmalarında 553186016 tweet mesajı ve 9800 adet HIV riskiyle ilgili anahtar kelime (örn. cinsel davranışlar ve uyuşturucu kullanımı) çıkartmışlardır. Yaygınlık analizine dayalı olarak HIV'le ilgili tweet mesajları ve HIV vakaları arasında anlamlı ve pozitif bir ilişki (p=0,010) olduğunu bulmuşlardır; bu da sosyal medyanın (örn. Twitter, Facebook) önemini ve küresel hastalık olayları üstündeki potansiyel etkisini ortaya koymaktadır.

Sosyo-demografik ve mediko-yönetimsel faktörlerin dağılımı ve etkisini araştırmak için, Lamarche-Vadel ve ark. (1992) MD ve ICD hastalarının bağımsız

ilişkilendirmesini analiz etmiştir. MD, ICD10 koduyla belirlenirken, UCD, bir ölüm tescilinden çıkartılmıştır. MD ve UCD farklı olaylar olsa, bu olayların bağımsız olduğu tespit edilir. Sağlık sigortası verileri kullanılarak, 2008'den 2009'a kadarki dönemde 421460 vefat eden hastadan bilgi çıkartılmıştır. Sonuçlara göre hastane içi ölümlerin %8,5'i ve hastane dışı ölümlerin %19,5'i bağımsız olaylardır ve bağımsız ölüm, yaşlı hastalarda daha yaygındır. Sonuçlar, büyük ölçekte veri analizinin tıbbi olayların ilişkilendirilmesini etkili biçimde analiz edilmede kullanılabileceğini göstermektedir.

Nambisan ve ark. (1993) sosyal medyada yayınlanan mesajların bunalımı elemek ve potansiyel olarak belirlemek için kullanılabileceğini belirtmiştir. Analizleri, depresif bozukluklar ve tekrarlayan düşünceler/ruminasyon davranışları arasındaki ilişkilendirmeye yönelik önceki araştırmalara dayanmaktadır. Büyük veri analiz araçları, mesajlardaki ya da Twitter'da yayınlanan tweet mesajlarındaki gizli davranışsal ve duygusal kalıpları çıkartarak önemli bir rol oynamaktadır. Bu tweet mesajlarında, belki de daha önceden gizlenmiş bir semptom olan hastalıkla ilgili bir davranış kalıbını tespit edebiliriz. Yazarlar gelecek araştırmaların, gizli duygu ve hisleri daha fazla anlamak için bunalımlı kullanıcıların sohbetlerini daha derinlemesine inceleyebileceğini öngörmektedir. Buna ek olarak Dabek ve Caban (1994) anksiyete, davranış bozukluğu, bunalım ve travma sonrası stres bozukluğu gibi psikolojik şartların gelişme ihtimalini tahmin edebilen bir sinirsel ağ modeli sunmuştur. Araştırmacılar ayrıca modellerinin 89840 hastadan oluşan bir veri setine karşı etkinliğini analiz etmiştir ve sonuçlar tüm şartlarda %82,35'lik bir tahmin oranı sağlanabildiğini göstermektedir.

Çalışmamızdaki veriler hastaneye diyabet şüphesi ile başvuran 101766 kişiyi içermektedir. Büyük veri teknolojileri yardımıyla veri madenciliği tekniklerinden Random Forest ve Çok Katmanlı Algılayıcı kullanılarak bu veriler üzerinde Yeni İlaç Reçete Edilme tahmini yapılmaya çalışılmıştır. Mahout teknolojisi kullanılarak yapılan Random Forest ve Çok Katmanlı Algılayıcı yöntemlerine ait Doğru Sınıflama Oranları sırasıyla %87.8 ve %84.9 olarak bulunmuştur. Bu sonuçlar büyük

veri teknolojilerinin veri madenciliği yöntemlerinde yüksek sonuçlar verdiğini ve kullanılabilir olduğunu göstermektedir.

5. SONUÇ VE ÖNERİLER

Günden güne artan veri boyutu nedeniyle büyük veri önümüzdeki yıllarda da artmaya devam edecektir ve her bir veri bilimcisi her yıl çok daha fazla miktarda veri yönetmek zorunda kalacaktır. Bu veri çok daha çeşitli, büyük, hızlı olacaktır ve geleceğin en heyecan verici fırsatlarından biri hâline gelecektir. Bu sebeple büyük veri bilimsel veri araştırması ve iş uygulamaları için yeni uç sınır olmaktadır.

Hem bilgisayımsal kaynakların hem de bu kaynaklardan faydalanan algoritmaların artan kabiliyeti sayesinde büyük verinin analizi ancak ve ancak yakın zamanda mümkün hale gelmiştir. Bu araç ve teknikleri sağlık alanına yönelik kullanan araştırmalar kritik önemdedir çünkü bu alan, yeni tekniklerin tüm düzeylerde gerçek dünya kararları vermek için uygulanabilmesinden önce çok sayıda sınama ve teyit gerektirmektedir. Bilgisayımsal gücün büyük veriyi etkin algoritmalar (ve elbette donanımsal gelişmeler) aracılığıyla ele alabilme yeteneğini elde etmesi, veri madenciliğinin sağlık alanındaki (geleneksel veya diğer türlü) verinin büyük hacmi, hızı, çeşitliliği, doğruluğu ve değeriyle baş etmesini sağlamaktadır.

Büyük verinin kullanımı, daha fazla sınama durumuna veya araştırmaya yönelik daha fazla özelliğe imkan vererek sağlık alanında avantajlar sağlamaktadır; bu da çalışmaların daha çabuk geçerlilik kazanmasını ve sadece küçük miktarda örnekler olumlu anlamda var olduğu zaman eğitime yönelik yeterli örneğin elde edilmesini beraberinde getirmektedir. Tıp bilişimine giden yolda, tıbbi verinin her türlü düzeyinde elde edilen büyük veriden kesinlikle yararlanılmakta ve olabildiğince çok tıbbi soruyu en iyi şekilde analiz etme, çıkarsama ve cevaplama yolları bulunmaktadır. Bahsedildiği üzere herhangi bir tıbbi sorunun, ister insan ölçeğindeki biyoloji düzeyinde, ister klinik düzeyde, isterse de evren düzeyinde olsun, cevaplanmasındaki genel amaç hastaların sağlığını geliştirmektir. İnsan bedeni (şu an hepsi erişilebilir olmasa da), birçok düzeyi içeren bileşik ve karmaşık bir sistemdir.

Dolayısıyla tüm bu seviyelerle ilgili sürekli uzayan tıbbi sorular listesini cevaplamak için tüm düzeylerdeki veriler üzerinde araştırma yapmak gerekmektedir.

Son zamanlarda yapılan çalışmalar dört veri düzeyinin hepsini içeren sorulara moleküler, doku, hasta ve evren düzeylerinden verilerin analizi yoluyla cevap aramaktadır. Büyük veri doğru kullanıldığı takdirde klinik sorulara yanıt bulmak amacıyla tüm bu veri düzeylerini kullanacak kapsama sahiptir ama bu amaç sadece klinik değil, tüm düzeylerdeki soruların yanıtlamasından faydalanabilir. Belki de gelecek araştırmalar, korelasyon ve bağlantı bulmak için her düzeydeki verileri kullanmaya odaklanarak doktorlara daha fazla tanı, tedavi ve hastalara yardım etme yolu sunabilir. Sağlık alanında gelecekte yapılacak her çalışma, insan mevcudiyetinin her düzeyindeki veriyi kullanan, dönüşümsel bir yaklaşıma sahip olmalıdır. Bu teknolojilerin hepsi gelecek vaat etmekte, gelecek çalışmalar için yol haritası çizmekte ve sağlık alanındaki her erişilebilir düzeyden veriyi kullanmanın önemini göstermektedir. Bu sunulan teknolojilerin her biri geliştirilebilir ve sonuçların farklı evrenlerde aynı olup olmadığını görmek için büyük hacimli ve çeşitlilikli veri setleri (potansiyel olarak diğer düzeylerdeki veriler) kullanılarak sınanabilir. Bu teknolojiler, veri madenciliği ve sağlık alanına yönelik büyük veri analizi yoluyla erişilebilecek imkanlara kısa bir bakış niteliği taşımaktadır. Bilgisayımsal güç arttıkça, daha etkin ve doğru yöntemler geliştirilecektir. Bu da her hastadan toplanan verinin çok büyük veri olduğu moleküler veri alt düzeyi (örn. atomik ölçek) gibi yeni insan mevcudiyeti düzeylerinin analize hazır hâle gelmesini sağlayabilecektir.

ÖZET

Sağlık Alanında Büyük Veri ve Veri Madenciliği Yöntemlerinin Kullanımı

Büyük veri, kuruluşların (kamu kurumları, hastane vs.) yapılandırılmış ve yapılandırılmamış veri setlerini birleştirerek veri madenciliği ve istatistiksel yöntemler kullanıp keşfedilmemiş bilgileri ve öngörülmeyen korelasyonları ortaya çıkarmasıdır. Veri depolama maliyetlerinin düşmesi ve büyük hacimli verilerle uğraşırken iş yükünü birden çok sanal sunucuya dağıtmaya imkan tanıyan NoSQL veritabanlarının, Hadoop ve benzeri veritabanı tasarımlarının ortaya çıkması ile büyük veri akımı daha hızlı yayılmaya başlamıştır. Yeni teknolojilerin gelişmesi üzerine daha az maliyet ile daha hızlı ve değişik türden verilerin analiz edilmesine ve yeni bulguların ortaya çıkmasına imkan sağladığı için kısa sürede tüm dünyada kabul görmüş ve kullanılmaya başlanmıştır. Son zamanlarda özellikle sağlık alanında bilinen klasik yöntemlerle saklanamayan ve analiz edilemeyen video görüntüleri, hekim notları gibi bilgilerin de büyük veri ile analiz edilmesiyle beraber yeni ve faydalı birçok bilginin elde edilmesine olanak sağlamaktadır. Bu çalışma hastaneye diyabet şüphesi ile başvuran 101766 kişiyi içermektedir. Bu modelde yeni ilaç reçete edilme tahmini için ırk, cinsiyet, yaş, hastane yatış şekli, hastanede geçirilen süre, laboratuvar işlem sayısı, işlem sayısı, ilaç tedavi sayısı, tanı sayısı ve tekrar başvuru olmak üzere toplam 10 değişken bulunmaktadır. Mahout ve Scala yardımıyla Random Forest ve Çok Katmanlı Algılayıcı veri madenciliği yöntemleri kullanıldı. Mahout için sırasıyla bu yöntemlerin Doğru Sınıflama Oranları %87.8 ve %84.9 iken Scala'da bu oranlar %84.9 ve %87.0 olarak bulunmuştur.

Anahtar Kelimeler: Büyük Veri, Veri Madenciliği, Sınıflama, Tahmin

SUMMARY

Big Data and Data Mining Methodologies in Medicine

Big data is that businesses, states, hospitals and organizations integrate different digital data sets and use information that has remained hidden and surprise correlations through methods of statistics and data mining. With lower data storage costs and the emergence of NoSQL databases, Hadoop and similar database designs allowing the distribution of workload to multiple virtual servers when dealing with high-volume data, the flow of big data has been spreading even faster. Since newly developed technologies allow for analyzing different types of data for lower costs and in a faster way and emergence of new findings, big data has been accepted by all the world in a short notice and started to be used. Recently, especially the analysis of information such as videos, physician notes which cannot be stored and analyzed with classical methods through big data has facilitated obtaining several pieces of new and useful information. This study consists of 101766 individuals who consulted the hospital with the suspicion of new diabetes medication. In this model, there are 10 variables for the prediction of diabetes: race, gender, age, admission type, time in hospital, number of laboratory procedures, number of procedures, number of medications, number of diagnoses and readmitted. With the help of Mahout and Scala, data mining methods of Random Forest and Multilayer Perceptron were used. True classification rates of these methods were found to be %87.8 and %84.9 for Mahout whereas these rates were found to be %84.9 and %87.0 for Scala.

Keywords: Big Data, Data Mining, Classification, Prediction

KAYNAKLAR

ALTUNIŞIK R (2015). Büyük Veri: Fırsatlar Kaynağı mı Yoksa Yeni Sorunlar Yumağı mı?. *YILDIZ Social Science Review* **1**: 45-76.

ARSLAN E (2012). Veri Madenciliği Yöntemleri. Erişim Adresi: [https://emraharslanbm.wordpress.com/2012/07/17/veri-madenciligi-yontemleri/]. Erişim Tarihi: 20/8/2016.

BEYAN T (2016). Sağlıkta dönüşüm ve büyük veri uygulamaları. Erişim Adresi: [https://saglikvebilisim.wordpress.com/2016/02/18/saglikta-donusum-ve-buyuk-veri-uygulamalari/]. Erişim Tarihi: 10/6/2016.

BREN D (2014). Diabetes 130-US hospitals for years 1999-2008 Data Set. Erişim Adresi: [https://archive.ics.uci.edu/ml/datasets/Diabetes+130-US+hospitals+for+years+1999-2008]. Erişim Tarihi: 24/7/2016.

CHANDRAMAN T (2015). Learning Apache Mahout. Birmingham: Packt Publishing.

CORTI L, EYNDEN VV, BISHOP L, WOOLLARD M (2014). Managing and Sharing Research Data, A Guide to Good Practice. Essex: SAGE.

COURT D (2015). Getting big impact from big data. McKinsey Quarterly.

ÇINAR SN (2014a). Büyük veri madenciliği Alzheimer keşfine doğru yol alıyor. Erişim Adresi: [http://bilgicagi.com/buyuk-veri-madenciligi-alzheimer-kesfine-dogru-yol-aliyor/]. Erişim Tarihi: 10/6/2016.

ÇINAR SN (2014b). Büyük veri "sağlığımızı" nasıl kurtarır?. Erişim Adresi: [http://bilgicagi.com/buyuk-veri-sagligimizi-nasil-kurtarir/]. Erişim Tarihi: 10/6/2016.

ÇINAR SN (2014c). Büyük veri, ebola salgınını önler mi?. Erişim Adresi: [http://bilgicagi.com/buyuk-veri-ebola-salginini-onler-mi/]. Erişim Tarihi: 10/6/2016.

DABEK F, CABAN JJ (2015). Brain informatics and health. In: Guo Y, Friston K, Aldo F, Hill S, Peng H, eds. A Neural Network Based Model for Predicting Psychological Conditions. Springer International Publishing: 252–61.

DAVENPORT T, BARTH P, BEAN R (2012). How 'big data' is different. *MIT Sloan Management Review* **54(1)**: 42-46.

DAVENPORT T (2014). Big Data @ Work. Harvard Business School Publishing.

DEAN J, GHEMAWAT S (2004). MapReduce: Simplied Data Processing on Large Clusters. Google Inc.

DEMİRTAŞ B, ARGAN M (2015). Büyük Veri ve Pazarlamadaki Dönüşüm: Kuramsal Bir Yaklaşım. *Pazarlama ve Pazarlama Araştırmaları Dergisi* Sayı: **15**

DUMBILL E (2013). Making sense of big data. Big Data: 1-2.

ENACORE (2016). Yapısal Olmayan Verinin Analizi ve Big Data. Erişim Adresi: [http://www.enacore.com.tr/tr/dokumantasyon/19-yapisal-olmayan-verinin-analizi-ve-big-data]. Erişim Tarihi: 13/6/2016.

GIACOMELLI P (2013). Apache Mahout Cookbook. Birmingham: Packt Publishing.

GOBBLE MAM (2013). Big data. The next big thing in innovation. Research-Technology Management. January-February: 64-66.

GÖKSU C (2014). Datawarehouse Türkiye. Erişim Adresi: [http://datawarehouse.gen.tr/big-data-nedir-geleneksel-veri-yonetimine-etkisi-ne-olur/]. Erişim Tarihi: 19/8/2016.

GUTERMEN J (2009). Big Data. O'Reilly Media, Inc. 27.

HAY SI, GEORGE DB, MOYES CL (2013). Big data opportunities for global infectious disease surveillance. *PLoS Med.*: **10(4)**: e1001413.

HOLDEN K, ANDY K, PATRICK W, MATEİ Z (2015). Learning Spark. Cambridge:O'Reilly.

HOY B (2014). Big data: An introduction for librarians. *Medical Reference Services Quarterly* **33(3)**: 320-326.

IAN HW, EIBE F, MARK AH (2011). Data Mining Practical Machine Learning Tools and Techniques. Oxford: MK.

IŞIKLI B (2009). Veri Madenciliği (Data Mining) Nedir ve Nerelerde Kullanılır-1. Erişim Adresi: [https://burakisikli.wordpress.com/2009/02/15/veri-madenciligidata-mining-nedir-ve-nerelerde-kullanilir-1/]. Erişim Tarihi: 25/8/2016.

İHS (2016). Hadoop ve Veri Ambarı. Erişim Adresi: [http://kurumsal.ihs.com.tr/hadoop-veri-ambari/]. Erişim Tarihi: 19/7/2016.

İLTER H (2012). Apache Hive. Erişim Adresi: [http://devveri.com/hadoop/apache-hive]. Erişim Tarihi: 27/8/2016.

JACOBS A (2009). The Pathologies of Big Data. *Data – ACM Queue* **7(6)**: 1-12.

KARAU H, KONWINSKI A, WENDELL P, ZAHARIA M (2015). Learning Spark Lightning-Fast Data Analysis. Cambridge:O'Reilly.

KELION L (2013). Q&A: NSA's Prism internet surveliance scheme. Erişim Adresi: [http://www.bbc.com/news/technology-23051248]. Erişim Tarihi: 6/6/2016.

LAMARCHE-VADEL A, PAVILLON G, AOUBA A (2008). Automated comparison of last hospital main diagnosis and underlying cause of death ICD10 codes. *BMC Med Inform Decis Mak.***14(1)** :44.

LCIA (2011). Big Data: Big Opportunities to Create Business Values.

LOHR S (2012). The age of big data. *New York Times* **11**.

MANYIKA R, MICHAEL C, BRAD B, JACQUES B, RICHARD D, CHARLES R, ANGELA HB (2011). Big data: The next frontier for innovation, competition, and productivity. McKinsey Global Institute.

MEDE AA (2014). Büyük Veri'ye ilişkin doğru bilinen yanlışlar. Erişim Adresi: [http://www.cio.com.tr/teknoloji/buyuk-veriye-iliskin-dogru-bilinen-yanlislar/]. Erişim Tarihi: 6/6/2016.

NAMBISAN P, LUO Z, KAPOOR A (2015). Social media, big data, and public health informatics: ruminating behavior of depression revealed through twitter. 48th Hawaii International Conference on IEEE System Sciences (HICSS): 2906–13.

NARAYANAN V (2014). Using big-data analytics to manage data deluge and unlock real-time business insights. *Journal of Equipment Lease Financing*, **32**: 1-7.

NICK P (2015). Machine Learning with Spark. Birmingham: Packt Publishing.

ÖZKAN E (2013). Büyük veri (Big Data) ve Map Reduce. Erişim Adresi: [http://www.erencanozkan.com/2013/08/buyuk-veri-big-data-ve-map-reduce.html]. Erişim Tarihi: 5/6/2016.

PARTNERS N (2012). Big Data Executive Survey: Creating a Big Data Environment to Accelerate Business Value. NewVantage Partners.

SAĞIROĞLU S, SİNAÇ D (2013). Big data: A review. 2013 International Conference on Collaboration Technologies and Systems (CTS): 42-47.

SCHAEFFER DM, OLSON PC (2014). Big data options for small and medium enterprises. *Review of Business Information Systems* **18 (1)**: 41-46.

SCHMARZO B (2012). New Roles in the Big Data World. EMC, Infocus.

SNOAD L (2011). Is data safe in brands' hands?. Erişim Adresi: [http://www.marketingweek.co.uk/analysis/data-strategy/is-data-safe-in-brands-hands/3032564.article]. Erişim Tarihi: 5/6/2016.

ŞEKER SE (2014). Sosyal Ağlarda Akan Veri Madenciliği. *YBS Ansiklopedi* **1**: 21-25.

ŞEKER SE (2015a). Büyük Veri ve Büyük Veri Yaşam Döngüleri. *YBSAnsiklopedi* **2**: 10-17.

ŞEKER SE (2015b). Sosyal Ağlarda Veri Madenciliği (Data Mining on Social Networks). *YBS Ansiklopedi* **2**: 30-39.

YOUNG SD, RIVERS C, LEWIS B (2014). Methods of using real-time social media technologies for detection and remote monitoring of HIV outcomes. *Prev Med*. **63**: 112–117.

YÖNTEM HE (2013). Büyük Veri ile Analitik Uygulamalar. IBM Connected 2013.

ÖZGEÇMİŞ

1- Bireysel Bilgiler

Adı	: Batuhan
Soyadı	: BAKIRARAR
Doğum yeri ve tarihi	: Afyonkarahisar, 06.03.1987
Uyruğu	: T.C.
Medeni durumu	: Evli
Askerlik durumu	: Yaptı
E-posta	: batuhan_bakirarar@hotmail.com
Telefon	: 541 667 64 27

2- Eğitimi

Yüksek Lisans	: Ankara Üniversitesi, Tıp Fakültesi, Biyoistatistik AD, Mezuniyet yılı: -
Lisans	: İstanbul Kültür Üniversitesi, Mühendislik Fakültesi, Bilgisayar Mühendisliği, Mezuniyet yılı: 2011
Lise	: Antalya Çağlayan Lisesi, Fen Bilimleri, Hazırlık + 3 yıl, Mezuniyet yılı: 2005
Yabancı Dil	: İngilizce – KPDS : 76,25

3- Mesleki Deneyimi

2012-2013 : Kütahya Porselen Bilgi İşlem Merkezi, Bilgisayar Mühendisi

2013-2015 : Talya Bilişim, Bilgisayar Mühendisi

2015- : Ankara Üniversitesi, Tıp Fakültesi, Biyoistatistik AD, Araştırma Görevlisi

4- Bilimsel Etkinlikleri

Ulusal ve Uluslararası Kongrelerde Sunulan Bildiriler ve Posterler

1. **Bakırarar B,** Kar İ, Yavuz Y, Köse SK, "Veri Madenciliği'ne Genel Bakış Ve Çok Katmanlı Algılayıcı Yönteminin WEKA Programında İncelenmesi: Sağlık Alanında Bir Uygulama", XVII. Ulusal Biyoistatistik Kongresi, Girne, Kıbrıs, 05-09 Kasım 2015
2. Kar İ, Bakırarar B, Köse SK, "İki Sürekli Değişken Arasındaki Uyum Araştırmalarında En Uygun Kesim Noktasının Tespiti: ROC Analizi mi, Tek Değişkenli Lojistik Regresyon Analizi mi?", XVII. Ulusal Biyoistatistik Kongresi, Girne,Kıbrıs, 05-09 Kasım 2015
3. **Bakırarar B,** Kar İ, Yavuz Y, "Büyük Veriye Genel Bakış: Sağlık Alanında Örnek Uygulama", XVIII. Ulusal Biyoistatistik Kongresi ve I. Uluslararası Biyoistatistik Kongresi, Beldibi, Antalya, 26-29 Ekim 2016
4. Kar İ, Bakırarar B, Köse SK, "Veri Madenciliği Sınıflandırıcılarının Performanslarının Karşılaştırılması", XVIII. Ulusal Biyoistatistik Kongresi ve I. Uluslararası Biyoistatistik Kongresi, Beldibi, Antalya, 26-29 Ekim 2016

Seminerler

1. "Sağlık Alanında Büyük Veri Uygulamaları", Ankara Üniversitesi, Ekim 2016