

**T.C.
BAHÇEŞEHİR ÜNİVERSİTESİ**

FİNANSAL VERİLER İLE VERİ ANALİZİ

Yüksek Lisans Tezi

ROJDA KILIÇ

İSTANBUL, 2021

**T.C.
BAHÇEŞEHİR ÜNİVERSİTESİ**

**LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ
BİLGİ TEKNOLOJİLERİ YÜKSEK LİSANS PROGRAMI**

FİNANSAL VERİLER İLE VERİ ANALİZİ

Yüksek Lisans Tezi

ROJDA KILIÇ

Tez Danışmanı: DR. ÖĞR. ÜYESİ ÖZGE YÜCEL KASAP

İSTANBUL, 2021



BAHÇEŞEHİR ÜNİVERSİTESİ
LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ

15/06/2021

YÜKSEK LİSANS TEZ ONAY FORMU

Program Adı:	Bilgi Teknolojileri Yüksek Lisans Programı (Türkçe, Tezli)
Öğrencinin Adı Soyadı:	Rojda Kılıç
Tezin Adı:	Finansal Veriler ile Veri Analizi
Tez Savunma Tarihi:	15.06.2021

Bu tezin Yüksek Lisans tezi olarak gerekli şartları yerine getirmiş olduğu Lisansüstü Eğitim Enstitüsü tarafından onaylanmıştır.

Doç. Dr. Burak KÜNTAY
Enstitü Müdürü

Bu Tez tarafımızca okunmuş, nitelik ve içerik açısından bir Yüksek Lisans tezi olarak yeterli görülmüş ve kabul edilmiştir

	Ünvanı, Adı Soyadı	İmza
Tez Danışmanı:	Dr. Öğr. Üyesi Özge Yücel Kasap	
2. Üye :	Dr. Öğr. Üyesi Tamer Uçar	
3. Üye :	Dr. Öğr. Üyesi Can Razbonyalı	



EN DEĞERLİ VARLIĞIM; AİLEME
İTHAF EDİYORUM.

TEŐEKKÖR

Tezimi hazırlarken bana her daim destek olan danışman hocam Dr.Öğr. Üyesi Özge Yücel Kasap'a, bu süreç boyunca beni hep motive eden aileme ve arkadaşlarıma, yüksek lisansa başlama konusundaki desteklerinden dolayı Aktif Yatırım Bankası ailesine teşekkür ederim.

İstanbul, 2021

Rojda Kılıç



ÖZET

FİNANSAL VERİLER İLE VERİ ANALİZİ

Rojda Kılıç

Bilgi Teknolojileri

Tez Danışmanı: Dr. Öğr. Üyesi Özge Yücel Kasap

Haziran 2021, 56 sayfa

Finans dünyasında hareketlilik gün geçtikçe artmaktadır. Hareketlilik ve büyüme beraberinde risk faktörünü de getirmektedir. Kredi ürünü karşılıklı güven içerisinde sunulsa da banka ve kredi veren kurumların her kredi ile birlikte aldıkları riski de göz önüne almaları ve öngörüler elde edebilmenin yollarını bulmalındırlar.

Veri analizi, mevcut durum ve konumu analiz ederek potansiyelleri keşfetmek konusunda en büyük destekçidir. Her geçen gün büyüyen veri kaynakları makine öğrenme algoritmaları ile buluşarak bilgi üretimini sağlamaktadır. Bilgi vasıtasıyla da rotalar belirlenmekte ve gidilecek yol şekillenmektedir. Bu tez çalışmasında Alman kredi veri seti ile finansal veriler ile veri analizi konusuna değinilmiş ve ilgili veri seti Random Forest ile Naive Bayes sınıflandırma algoritmaları ile oluşturulmuş modellerde çalıştırılmıştır. İlgili veri setinde daha önce kullanılmamış olan bu algoritmalar ile birlikte iyi performans sonucuna sahip modeller oluşturulmuştur.

Anahtar Kelimeler: Kredi Risk Analizi, Veri Analizi, Sınıflandırma Algoritmaları

ABSTRACT

DATA ANALYSIS WITH FINANCIAL DATA

Rojda Kılıç

Information Technologies

Thesis Supervisor: Assist Prof. Özge Yücel Kasap

June 2021, 56 pages

Mobility in the financial world is increasing day by day. Mobility and growth also bring risk factors. Even though the loan product is offered with mutual trust, banks and lending institutions should consider the risk they take with each loan and find ways to obtain predictions.

Data analysis is the biggest supporter in discovering potentials by analyzing the current situation and location. Data sources, which are growing day by day, meet with machine learning algorithms and provide information production. Routes are determined by means of information and the path to be taken is formed. In this thesis, the German loan data set and financial data and data analysis were discussed and the related data set was run in models created with Random Forest and Naive Bayes classification algorithms. With these algorithms, which have not been used before in the relevant data set, models with good performance results have been created.

Keywords: Credit Risk Analysis, Data Analysis, Classification Algorithms

İÇİNDEKİLER

TABLolar.....	x
ŞEKİLLER.....	xi
KISALTMALAR	xiii
SEMBOLLER.....	xiv
1. GİRİŞ.....	1
2. LİTERATÜR TARAMASI.....	3
3. VERİ ANALİZİ.....	8
3.1. VERİ ANALİZİ TANIMI	8
3.2. VERİ ANALİZİ ÖNEMİ	9
3.3. VERİ ANALİZİ SÜRECİ	9
3.4. VERİ ANALİZİ TÜRLERİ	11
3.4.1. Betimsel Analiz	11
3.4.2. Teşhis Edici Analiz.....	11
3.4.3. Öngörücü Analiz.....	12
3.4.4. Çıkarımsal Analiz	12
3.4.5. Kural Koyucu Analiz.....	12
3.4.6. Metin Çözümlemesi Analizi.....	13
3.4.7. İstatiksel Analiz.....	13
4. MAKİNE ÖĞRENİMİ.....	14
4.1. MAKİNE ÖĞRENİMİ TANIMI	14
4.2. MAKİNE ÖĞRENİMİ TARİHÇESİ	15
4.3. MAKİNE ÖĞRENİMİ METODLARI.....	16
4.3.1. Denetimli Makine Öğrenimi.....	16
4.3.2. Sınıflandırma	17
4.3.3. Regresyon	17
4.3.4. Denetimsiz Makine Öğrenimi.....	17
4.3.4.1 Kümeleme	18
4.3.4.2 İlişkilendirme	19
4.3.5. Yarı Denetimli Öğrenim	19
4.3.6. Takviyeli Öğrenim	20

4.4.	MAKİNE ÖĞRENİMİNDE KULLANILAN ALGORİTMALAR	21
4.4.1.	Naive Bayes Algoritması.....	21
4.4.2.	Destek Vektör Makineleri Algoritması	22
4.4.3.	Lojistik Regresyon	22
4.4.4.	Karar Ağaçları.....	23
4.4.5.	Random Forest Algoritması	25
4.4.6.	K-Means Algoritması.....	25
4.4.7.	Yapay Sinir Ağları.....	26
5.	PYTHON PROGRAMLAMA DİLİ.....	28
5.1.	PYTHON PROGRAMLAMA DİLİ VE TARİHÇESİ	28
5.2.	PYTHON PROGRAMLAMA DİLİ İLE VERİ ANALİZİ.....	28
6.	FİNANS SEKTÖRÜ VE VERİ ANALİZİ	30
6.1.	FİNANS SEKTÖRÜNDE VERİ ANALİZİ ÖNEMİ	30
6.2.	FİNANS SEKTÖRÜNDE KREDİ TANIMI	32
7.	YÖNTEM	33
7.1.	GENEL BAKIŞ	33
7.2.	KULLANILAN VERİ ANALİZİ ARAÇLARI.....	33
7.3.	KULLANILAN MAKİNE ÖĞRENİMİ ALGORİTMALARI.....	34
7.4.	PROBLEMİN TANIMI VE ÇALIŞMANIN AMACI.....	34
7.5.	VERİNİN HAZIRLANMASI	35
7.5.1.	Verinin Çalışma Ortamına Dahil Edilmesi.....	36
7.5.2.	Veri Setinin Özellikleri	36
7.5.3.	Eksik Veri İnceleme Kısmı	38
7.5.4.	Öznitelik Oluşturma	39
7.5.5.	Özniteliklerin Değerlendirilmesi ve Verinin Tanınması	40
7.5.6.	Verileri Dönüştürme.....	43
7.6.	MODEL KURMA VE DEĞERLENDİRME	44
7.6.1.	Model 1: Random Forest Algoritması Kullanımı	44
7.6.2.	Model 2: Naive Bayes Algoritması Kullanımı.....	47
8.	MODEL UYGULAMALARI DEĞERLER VE KARŞILAŞTIRMA.....	50
9.	TARTIŞMA VE SONUÇ	55
	KAYNAKÇA.....	57

EKLER.....	Error! Bookmark not defined.
EK 1 : Python ile Veriyi Alma ve Hazırlama Kodları	Error! Bookmark not defined.
EK 2 : Python ile Veri Dönüştürme Kodları.....	Error! Bookmark not defined.
EK 3 : Model 1 Hazırlık ve Değerlendirme Python Kodları	Error! Bookmark not defined.
EK 4 : Model 1 Karmaşıklık Matrisi, Ölçü Değerleri Hesaplaması ve Grafik Oluşturulması Çalışmalarına Ait Python Kodları.....	Error! Bookmark not defined.
EK 5 : Model 2 Hazırlık Aşaması, Tahminleme Yapılması ve Matris Değerleri ile Grafik Oluşturulması Çalışmalarına Ait Python Kodları	Error! Bookmark not defined.



TABLÖLAR

Tablo 7.1: Veri setine ait özellikler listesi	36
Tablo 8.1: Model metrik değeri sonuçlarının karşılaştırılması	54



ŞEKİLLER

Şekil 3.1: Veri bilimi dünyası ve veri analizi	8
Şekil 3.2: Veri analizi adımları.....	11
Şekil 4.1: Derin öğrenme, makine öğrenmesi ve yapay zeka arasındaki ilişki	14
Şekil 4.2: Makine öğrenimi ve yapay zeka tarihçesi	16
Şekil 4.3: Denetimsiz öğrenim çalışma yapısı	19
Şekil 4.4: Makine öğrenimi metodları ve kullanılabilecek durumlar ilişkisi.....	20
Şekil 4.5: Bayes formülü.....	21
Şekil 4.6: Karar ağacı örnek çalışması.....	24
Şekil 4.7: Öklit uzaklığı formülü	26
Şekil 5.1: Python programlama dili ile veri analizi arasındaki ilişki.....	29
Şekil 7.1: Verinin ilk hali	36
Şekil 7.2: Veri seti geneli ve öznitelik özelinde eksik veri kontrolü sonucu	38
Şekil 7.3: Verinin yeni öznitelik oluşturulması sonrası hali	40
Şekil 7.4: Risk, meslek ve medeni durum ile cinsiyet öznitelikleri değer ile sayı dağılımları.....	41
Şekil 7.5: Risk özniteliği değişkenine göre iyi kredi ile kötü kredi değerlendirmesi grafiği	42
Şekil 7.6: Medeni durum ile cinsiyet özniteliği değişkeni ve müşteri dağılımı grafiği ..	42
Şekil 7.7: Meslek özniteliği değişkeni ve müşteri dağılımı grafiği	43
Şekil 7.8: Verinin son halinin ilk 5 satırı	43
Şekil 7.9: Karmaşıklık matrisi yapısı.....	46
Şekil 7.10: Model 1 karmaşıklık matrisi değerleri	46
Şekil 7.11: Model 2 karmaşıklık matrisi değerleri	48
Şekil 7.12: Veri seti korelasyon değerlendirmesi	49
Şekil 8.1: Model 1 uygulaması sonrası elde edilen karmaşıklık matrisi grafiği	51
Şekil 8.2: Model 1 uygulaması sonrası elde edilen metrik değerler sonuçları.....	51
Şekil 8.3: Model 1 uygulaması sonrası elde edilen AUC-ROC eğrisi.....	52
Şekil 8.4: Model 2 eğitim verileri ile uygulaması sonrası elde edilen başarı oranı değeri	52

Şekil 8.5: Model 2 uygulaması sonrası elde edilen karmaşıklık matrisi ve metrik değerler sonuçları.....	53
Şekil 8.6: Model 2 uygulaması sonrası elde edilen AUC-ROC eğrisi.....	53



KISALTMALAR

AUC	:	Area Under the Curve
BDDK	:	Bankacılık Düzenleme ve Denetleme Kurulu
DVM	:	Destek Vektör Makineleri
KNN	:	K-Nearest Neighbor
LR	:	Lojistik Regresyon
MH	:	Meyilli Hızlandırma
ROA	:	Rastgele Orman Algoritması
ROC	:	Receiver Operating Characteristic
UCI	:	University of California, Irvine
WEKA	:	Waikato Environment for Knowledge Analysis
XGBoost	:	Extreme Gradient Boosting

SEMBOLLER

Yüzde

:

%



1. GİRİŞ

Günümüzde giderek büyüyen veri havuzu mevcuttur. Kurum ve kişiler için bilginin kaynağı olan bu havuz, veri analizi dünyasının da gözdesidir. İşlem akışları, hareketler ve davranışlar sektörler özelinde dikkate alınmaktadır. Finans, sağlık ve eğitim gibi birçok sektörde veri analizi önemini giderek artırmaktadır. Çalışmamızın çıkış noktası finans sektöründe veri analizinin önemi ve ihtiyacıdır. Para akışının söz konusu olduğu süreçlerde riskler de peşinden gelir ve riskin olduğu noktada doğru konumlanma ihtiyacı da artar. Veri analizinin tercih sebebi olma sebeplerinin başında bizlere kullanabileceğimiz bilgiyi verebiliyor olması gelmektedir. Mevcut durumların analizleri ile birlikte tahmin analizleri yapılarak risk başta olmak üzere belli başlı konularda atılacak adımlar planlanabilmektedir.

Yatırım ve tasarruf kararlarının birbirinden farklı birimler tarafından alındığı ülke ekonomileri söz konusu olduğunda, yatırımlara odaklanıp, fon ve destekleri buluşturarak işlemler yapan, finansal araçların oluşturduğu yapı, finansal sistem olarak isimlendirilmektedir. Bu sistemi fon ve parasal getiriye yatırımlar ile buluşturan araçlar ile araçlar oluşturmaktadır. Krediler, hazine bonosu ile mevduat ürünleri bu araçlar arasında sayılabilir. (Barışık ve diğ. 2010)

Sistemin sürekliliği ve dalgalanmalar dolayısıyla risk faktörü daha fazla öne çıkmaktadır. Kredi dünyası da bu payda da payına düşeni almaktadır. Kredi, karşılıklı güvene dayalı olarak sözleşmeler ile birlikte kişi ile kurum arasında başlayan bir süreçtir. Burada bankaların ve kredi veren kurumların aldıkları riski öğrenebilme arzuları vardır. Risk tolere edilebilir seviye üzerindeyse bu doğrultuda hareket edilecektir. Mevcut müşteriye analiz etme, olası risklere dair öğretilmiş modellere sahip olmak yeni kredi süreçlerinde en büyük destekçileri olmaktadır. Veri analizi bu konuda bankaların ilk çaldıkları kapıdır. Giderek artan veri analizi çalışmaları ile kredi riskleri ve potansiyeli kavrama konularına dair çalışmalar devam etmektedir.

Müşteriye ait bilgileri ve bilgiyi elde edebileceğimiz özniteliklerin yer aldığı veri seti toplanır. Toplanan veriler sırasıyla veri analizi aşamalarından geçirilir. Eksik verilerin bulunması, ihtiyaç dahilinde yeni özniteliklerin oluşturulması, verinin temizlenmesi ve algoritma seçimi olmak üzere adımlar gerçekleştirilir. Veri, model kullanımına uygun hale getirilir ve seçili algoritma ile model oluşturularak veri seti çalıştırılır. Sonrasında metrik değerler elde edilir ve bu değerler yorumlanır. Temelde bu şekilde akışa sahip olan veri analizi süreci sonucunda bir performans değeri ile tamamlanır. Burada seçilen algoritmanın veri seti için doğru olup olmadığı sonucu da elde edilmiş olacaktır.

Amaç bilgiye ulaşmaktır, burada başvuru olan veri analizi teknikleri olmakta ve yardımcı olarak da makine öğrenimi teknikleri ile birlikte algoritmaları gelmektedir. Havuzdan ihtiyaç duyulan bilgi elde edilerek hedeflere ilerlenilmektedir. Burada başvuru olan bilimsel yol veri analizi artan talep ve çalışmalar dolayısıyla kullanıcıya kolaylık sağlayan platformlarda da çalışmaları sürdürebilme imkanı sunmaktadır.

Tezin birinci bölümünde finans sektöründe veri analizi, makine öğrenimi ile kredi risk analizi başlıklarına dair ve Alman kredi veri seti üzerine yapılan çalışmalar incelenmiş, elde edilen sonuçlar ile çıkarımlara yer verilmiştir. İkinci bölümünde veri analizi tanımına, önemine, süreç ile türlerine değinilmiştir. Üçüncü bölümde makine öğrenimi tanımı yapılmış ve tarihçesinden bahsedilmiştir. Bununla beraber makine öğrenimi metodları ile makine öğrenimi algoritmaları detaylı olarak anlatılmıştır. Dördüncü bölümde Python programlama dili ve tarihçesi aktarılmış ve Python ile veri analizi başlığı ele alınmıştır. Beşinci bölümde finans sektörü ve veri analizi başlığı ele alınmıştır. Veri analizi ihtiyacı sebepleri maddeler halinde belirtilmiştir. Altıncı bölümde yöntemle yer verilmiştir. Burada veri seti tanıtılmış, kullanılan programlama dili ile yazılım belirtilmiştir. Verinin geçirildiği süreçler ve detaylarına yer verilmiştir. Modeller oluşturulmuş ve çalıştırılmıştır. Yedinci bölümde sonuçlar değerlendirilmiş ve en başarılı olan model değerlendirme sonucunda belirlenmiştir.

2. LİTERATÜR TARAMASI

Literatür kısmında, çalışmanın temelinde ele aldığı konu ile ilişkili çalışmaların yanında, çalışmada kullanılan veri setinin kullanıldığı çalışmalara da yer verilmiştir. Temelde finans sektöründe veri analizi önemi ve örnek çalışmalar ile birlikte sonrasında kredi risk ve inceleme çalışmalarında veri analizi kullanımına dair çalışmalara yer verilmiştir. İrdelenen ve dikkate alınan bu konular sonrasında kullanılacak olan veri seti ile daha öncesinde yapılmış çalışmalara da yakından bakılacaktır.

Demirtürk, çalışmasında kredi kavramı ve kredi analiz ihtiyacının neden ortaya çıktığı üzerinde durmuştur. Kredi karşılıklı sözleşme ve güven doğrultusunda sunulan bir destektir. Ancak kredi veren kurum için doğru konumlanmak ve zarara uğramamak da önemlidir. Veri analizine burada başvurulmaktadır. Kredi alacak kurum ve kişiler için mevcut durum, geçmiş ve gelecekteki durumların değerlendirilmesi, bütün faktörlerin göz önünde bulundurulması gerekmektedir. Burada elde bulunan veri havuzu veri analizi algoritmaları da ele alınarak değerlendirildiğinde ön görüler ortaya çıkarılacak ve faaliyetler bu doğrultuda şekillenebilecektir. Neden ve sonuçlar belirtilerek konu ele alınmıştır. (Demirtürk 2011)

Yıldız , çalışmasında Türk Ticari Bankacılık sektöründe karlılığı veri analizi bakımından ele almıştır. Finans ve ekonomi alanlarında giderek artan bir veri havuzu vardır. Veri bakımından zengin olan bu alanlarda oluşabilecek durum çeşitliliği de fazlasıyla mevcuttur. Bu olayların öngörülebilir olması veya istenilen akışta ilerlenebilmek adına veri analizine başvurulmaktadır. Dünya’da finansal riskler giderek artmaktadır. Piyasaların değişken olması ve bu hareketlilik, karlılık hedefiyle hareket eden kurumlar için tehlike arz ettiğinden karlılık hedefine ulaşma yolunda veri analizi üzerine değinilmiş ve ele alınmıştır. (Yıldız 2018)

Yıldız , çalışmasında sektörde ülke kredi notları ile makroekonomik değişkenler ilişkisini iredelerken veri analizine başvurmuştur. Ülkelere ait gelirler, kredi notları verilirken önem arz eden diğer kalemlerin de ele alınarak bu kalemlerden elde edilen

verilerin analizleri sonucu elde edilen sıralamalar üzerinde durulmuştur. Bu sıralamalar ile birlikte yorumlamaların yer aldığı çalışmada, ilişkilerin irdelenmesi sağlanmıştır. Buradaki çalışmada da görüldüğü üzere elimizdeki veri havuzu ile birlikte elde edebileceğimiz bilgiler bizlere farklı alan ve koşullarda değerlendirme yapabilme imkanı sunabilmektedir. Bu şekilde çıkarımlar ile değerlendirmeler yapılmıştır. (Yıldız 2014)

Kalaycı, çalışmasında BDDK'nın yayınladığı rapor doğrultusunda küçük ve orta ölçekli işletme duumundaki müşterilerin kullandıkları kredi oranlarının arttığına ve gün geçtikçe de artacağına değinmiştir. Bu neden dolayısıyla da müşteri riskinin çok daha iyi tahmin edilebilir ve öngörülebilir olmasının önemini belirterek analizlerin kredi riskleri ve skor analizleri için makine öğrenimine başvurulmasını konu edinmiştir. Elde yer alan kalemlere uygun veriler ile makine öğrenimi algoritmalarından harmanlanmış şekilde sistem oluşturarak konuyu ele almıştır. ROA ve MH yöntemleri ile harmanlanmış sistem oluşturmuş ve bu sistemin algoritmaların tek başlarına elde ettikleri performansdan daha fazla olduğu görülmüştür. Bu çalışma ile eldeki kredi verileri ile makine öğrenmesini bir araya getirmiş ve çalışmayı tamamlamıştır. (Kalaycı 2018)

Çizer, çalışmasında kredi üzerine tanımlamaları yapmış ve kredi vererek alınan riske değinmiştir. Buradan bankalar için bu kredi risklerini minimuma indirmenin ne derece önemli olduğu sonucu elde edilmiştir. Son dönemlerde bu neden dolayısıyla ortaya çıkan sonucun kredilendirme faaliyetlerinde veri madenciliği, makine öğrenimi algoritmalarından yapay sinir ağları, sınıflandırma ve regresyon teknikleri ile bulanık mantık kullanımının artışı gözlemlenmektedir. Bu teknik ve algoritmalar ile birlikte potansiyel risklerin tahmin edilebilmesi ve karlılık ile kaynak verimliliği açısından kazanca geçilebilmesi daha olasıdır. Çalışmanın devamında eldeki kredi verisi ile birlikte DVM ve karar ağaçları ile analiz, bulanık mantık ile genetik algoritma analizi yapılmıştır. Yapay sinir ağları analizi de yapılarak modeller sonucunda elde edilen değerler karşılaştırılmış ve kullanılan algoritmalar başarılarına göre sıralanmıştır. (Çizer 2018)

Can, çalışmasında bankaların mevcut kaynaklarını doğru kullanabilmek ve bu hareketli piyasada doğru konumlanabilmek açısından aldıkları kararlardan notlar paylaşmıştır. Risk ve tahmin analizlerinin önemli olduğu bu alanda, verinizi tanıyarak ona uygun modelleri kurmanın önemi artmıştır. Çok çeşitli ve karmaşık veri havuzuna sahip olunması durumu göz önüne alındığında değerlendirme kalemleri ile kriterlere doğru cevap verilebiliyor olması gerekmektedir. Makine öğrenmesi algoritmaları ve metodları kurum ve kişilerin karlılıklarını düşünüp riskleri nasıl savurabileceklerini kendilerine sorduklarında karşısına çıkmakta ve seçilmektedir. Bu çerçevede dahilinde mevcut veri setimiz ile XGBoost sınıflandırıcısında model uygulanmış ve çalıştırılmıştır. Daha öncesinde yapılan XGBoost çalışmaları arasındaki en yüksek başarı oranına sahip olunmuştur. %75,60 başarı oranı ile doğru seçim olduğunu gösterdiğini de belirtmiştir. (Can 2020)

Yavuz, çalışmasında teknoloji ile birlikte gündemimize dahil olan çeviklik dolayısıyla oluşan talep ve arz artış oranına değinmiştir. Rekabet ortamını beraberinde getiren bu süreç ile birlikte şirketlerin konumlanmasından ve risklerin varlığından bahsetmiştir. Kredi risk değerlendirmesinin tüm risk yönetimi konsuundaki yeri ve öneminin göz ardı edilemeyeceği gerçeğiyle yüzleştirdikten sonra finansal endüstri dinamikliği ile birlikte, kredi ile kredi kartı taleplerinin ve başvurularının çokluğunu belirtmiştir. Artışın getireceği risk faktörü ve bu rekabet ortamından başarı ile çıkabilme hedefinde en önemli destekçi veri analizidir. Analiz beraberinde bilgiyi, bilgi de öngörü ve davranış modellerini getirmektedir. Bu mantık ve yol ile birlikte UCI Repository üzerinden elde edilmiş olan Tayvan kredi veri seti kullanılmış ve bu veri setine tahmine dayalı analitik yöntemler uygulanarak analiz değerlendirmeleri tamamlanmıştır. (Yavuz 2020)

Çetiner, çalışmasında kredi risk analizinin önemi ama bir o kadar da zor bir veri madenciliği problemi olduğundan bahsetmiştir. Global kredi hacmi artmakta ancak bir o kadar da ekonomik dalgalar yaşanmaktadır. Gelecekteki kayıpları azaltmak için eski ve mevcut duurmun iyi analiz edilmesi, kabul edilebilir risk doğrultusunda belirlenmiş tolerans değeri ile karşılaştırarak adım atılması gerekmektedir. Çalışma içeriisinde seçili algoritmalar tek çalıştırılması sonrası elde edilen performans değerlerine karşın az da olsa iyileştirilmiş bir performans değeri elde edebilmek adına kombine sınıflandırma

yöntemi oluşturulmuştur. Bu yöntem de algoritmalarından tek tek elde edilen başarıya karşın daha başarılı sonuçlar elde edebilme imkanı mevcuttur. Mevcut veri setimize DVM ile LR teknikleri uygulanmıştır. WEKA ve GeneXproTools yazılımları ile yapılan çalışmalar sonucunda elde edilen performans değerleri karşılaştırılmış ve iyi sonuç veren sınıflandırma teknikleri ile hibrid sınıflandırma tekniği oluşturularak daha iyi bir performans sonucu elde edilmiştir. (Çetiner 2011)

Ada Fu, Wang, Xu Yu ve Zhou, çalışmalarında veriler ile veriler üzerindeki değişiklikleri ile özellikle sınıflandırma tekniklerine yönelmiş, buradaki değişimleri gözlemlemişlerdir. Kullanılan teknik ile bir sonuç elde edilir. Ancak sonucu iyileştirmek için oynamalar da yapılabilir. Veri üzerinde yapılacak düzenlemeler ile birlikte metodun çıktısı üzerinde de değişimler olabilmektedir. Bunların gözlemlenmesi ve bu doğrultuda seçimlerin de değiştirilebiliyor olması gerekmektedir. Veri ile çalışırken sonucunda elde etmek isteyeceğimiz en başarılı ve performansı en iyi olanı seçebilmektir. Doğru bilgiyi elde edebilmek bu konudaki hedeflerin başında yer almaktadır. Çalışma bu doğrultuda yapılmış ve incelemeler gerçekleştirilmiştir. (Ada Fu, Wang, Xu Yu ve Zhou 2003)

Griffith, O'Dea ve O'Riordan, çalışmalarında sınıflandırma problemlerini çözmek için özellik seçimi ile sinir ağıları tekniklerini birleştirerek yeni bir yaklaşım oluşturmuşlardır. Mevcut veri setimiz oluşturulan problem çözme kombininde çalıştırılmış ve ortaya sunulan bu yeni çözüm örneğinin başarılı olup olamayacağı konusunda fikir verebilmiştir. Sinir ağı seçili özellikler ile eğitildikten sonra sınıflandırmak için kullanılmıştır. Burada yapılan bu yenilikçi yaklaşım ile veri madenciliği ve makine öğrenmesi algoritmalarının ana hatları gözden geçirilmiştir. Oluşturulan bu kombin yaklaşım ile birlikte mevcut veri setimiz ile de ön sonuçlar elde edilmiştir. Çalışmanın amacı ana hatların yanında alternatif çözümler sunabilmenin mümkün olduğunu göstermektir. (Griffith, O'Dea ve O'Riordan 2013)

Ekin, Hammer, Kogan ve Winter, çalışmalarında mevcut veri setimizi yeni yaklaşımlarında kullanmışlardır. Sınıflandırma metodları üzerine çalışmaları yapmışlardır. Eğitim kümesi oluşturmuş ve Öklid uzayında bir dizi nokta ile bu küme arasında işlemler yapmışlardır. Alt kümeler elde ederek mevcut noktaların dağılımı

incelenmiştir. Sorgu noktası ile eğitim kümesindeki mesafeler takip edilmiştir. KNN algoritması ile birlikte yeni yöntemler ile elde edilen sonuçlar kapsamlı bir şekilde hesaplanmışlardır. Bu hesaplar değerlendirilmiş ve mesafeye dayalı yöntemlerin performanslarının diğer sınıflandırma yöntemlerine göre çok daha iyi olduğunu öne sürmüşlerdir. Doğru parametre seçimi ile uygulaması kolay olduğunu öne sürdükleri bu yöntemleri ile hızlı sonuç elde edebildiklerini de ayrıca çıkarım olarak elde etmişlerdir. (Ekin, Hammer, Kogan ve Winter 2016)

Gunopulos ve Ratanamahatana, çalışmalarında karar ağaçları ile özellik seçimi tekniklerini birleştirerek Naive Bayesian sınıflandırıcıyı ölçeklendirmek üzere işlemler yapmışlardır. Burada mevcut veri setimizi elde ettikleri yaklaşımı test etmek adına kullanmışlardır. NB bazı alanlarda çok iyi çalışabiliyorken bazılarında aynı başarıyı göstermemektedir. Karar ağaçları ilişkili özelliklerin yer aldığı veri analizi çalışmalarında daha iyi performans göstermektedir. Bu çalışma içerisinde seçici bir Byes sınıflandırıcı ile oluşum sağlanmış ve çalışmalar yapılmıştır. Veri setimiz ile yapılan çalışma sonrasında da görünen bu seçici sınıflandırıcının NB'nin yanında karar ağacı tekniğinden de daha iyi sonuç verdiği görülmektedir. Performansı artırmak için yapılacak çalışmaların ve yeni yaklaşımların en büyük katkılarından biri de daha iyi sonuçlar veren modeller elde edebiliyor olmaktır. (Gunopulos ve Ratanamahatana 2002)

Dehuri, Jagadev, Mohapatra ve Pandey, çalışmalarında makine öğrenmesi ile kredi risk analizine değinmişlerdir. Bankaların kredi verme konusundaki büyük hacmi ve konumu ile risk ilişkisinin tekrar ele alındığı bu çalışmada mevcut veri setimiz ile birlikte tasarlanan akış ilişkilendirilmiştir. Veri analizinin yapıyor olma sebeplerinin en başında potansiyeli fark etmek ve tehlikeleri öncesinde sezip engelleyebilir olabilmek. Burada engelleme riskli müşteriye kabul etmemek, çalışmaların ve modellerin sonucunda elde edileceği gibi limitlerin üzerinde işlemler yapmamaktır. Bunun yanında bu çalışmalar bizlere öngörü elde edebilme imkanı vermektedirler. Karar ağacı, KNN ve DVM kullanılarak veri seti üzerinde çalışmalar yapılmış ve mdoeller oluşturulmuştur. Amaç burada daa çalışmalar sonrasında bize en iyi sonucu verebilecek olan tekniği ve algoritmayı seçebilmektir. Analiz sonucundaki hedef bizi ulaşmak istediğimiz noktalara getirebilecek olan seçeneği bulmaktır. (Dehuri, Jagadev, Mohapatra ve Pandey 2017)

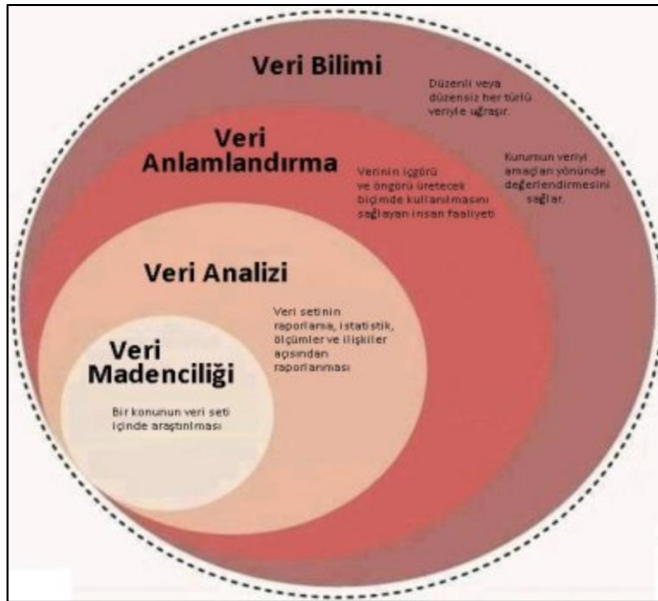
3. VERİ ANALİZİ

3.1. VERİ ANALİZİ TANIMI

Veri, bilginin hammaddesi, ham halidir. Veri Analizi ile ham veri kullanılmak üzere bir ortamda toplanır, temizlenir ve seçilecek algoritma ile oluşturacak modellerde işlenerek ihtiyacımız olacak ya da kullanılabilecek bilgiler elde edilir. Sonuç elde etme ve karar alabilmek adına bu süreçlerde destek olan modelleme aşamalarıdır. (Özdemir 2010) Elde olan ve çalışma yapılan veri ile ilgili anlam üretebilmek, veri setinde temsilin ne olduğuna dair çıkarımlarda bulunabilmek hedefiyle yapılan bir sınıflandırma, yorum elde edebilme süreci olan veri analizi bununla beraber mevcut uygulama süreç ve sorunlarını tanımlamak adına başvurulmuş, etkili bir yoldur.

Kurumlar, pazarlama stratejilerini belirlerken ve kritik karar alırken veri analizinden faydalanmaktadır. Bilimsel olan bu süreç sonucunda elde edilecek sonuçlar ile strateji belirlemek konusunda birçok işletme harekete geçmiştir.

Şekil 3.1: Veri bilimi dünyası ve veri analizi



Kaynak: ugurozmen.com-Pazarlamada Dijital Dönüşüm

3.2. VERİ ANALİZİ ÖNEMİ

Veri analizi, mevcutta kullanılacak olan veriler için en uygun yöntemlerin seçilerek yapılması durumunda kurumların stratejik karar verme sürecinde yapılabilecek birçok hatayı engelleyebilme imkanı vermektedir. Bankacılık, finans, perakende, sağlık gibi birçok sektör veri analizlerini müşteri memnuniyetini ölçmek ve artırmak amacıyla da kullanmaktadır. İleriye dönük yararlı olabilecek doğru kararlar almak gerektiği durumda başvurulan yöntem veri analizidir. Veriyi bu doğrultuda kullanmak ve irdellemek yoluyla oluşan bu süreç temelde bu kazanç amacıyla devam ettirilmektedir.

Firmalar veri analizi ile ellerinde yer alan bilgileri etkili bir şekilde değerlendirerek gelecekteki stratejilerini daha doğru belirleyebilir hale gelmektedir. Büyüme stratejileri, satış konusundaki davranışlar ile yatırım gibi birçok konuda firmalar ileriye dair fikirler edinebilmek adına veriye başvurmaktadır.

Veri analizinin sağladığı faydaların en önemlilerinden biri müşteri açısından memnuniyet oluşumunu artırmaya yönelik uygulamalara dair olan katkısı sayılmaktadır. Veri analizi sayesinde müşteriler tarafından iletilen geri bildirimler daha doğru değerlendirilebilmekte ve bu sayede müşteri memnuniyetinin etkili bir biçimde sağlandığı hizmetler ve ürünler sunulabilmektedir. Müşteriler açısından memnuniyeti sağlayabilen kurumlar kazançlı ve yüksek prestij sahibi olabilmektedir.

3.3. VERİ ANALİZİ SÜRECİ

Veri analizi sürecinin temel amacı ham veri üzerinden anlam çıkarabilmektir. Bu amaçla veri üzerinde araştırma yapacak kişinin, okuduklarını, gözlemlediklerini ve katılımcılar vasıtasıyla elde ettiği verileri yorumlamalı, temizleyebilmeli ve modeller ile çalışabilecek hale getirebilmelidir. Veriden anlam çıkarmak için araştırma sürdüren kişi, devamlı olarak somut veri seti ile soyut kavramlar arasında bağlantı kurabilmekte, tanım ve yorumlar arasında gidip gelerek tümevarımsal ve tümdengelimsel fikir yürütme süreçlerini birlikte kullanır. (Corbin ve Strauss 1990; Glaser 1967; Maxwell 1996)

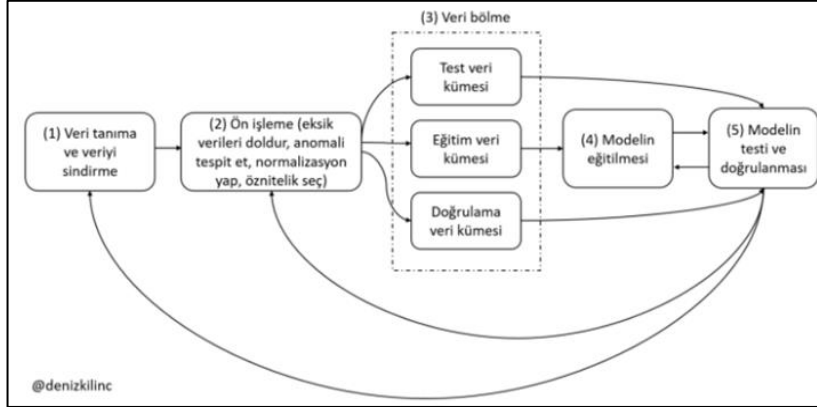
Veri analizi süreci adımları beş temel başlık altında sıralanabilir. Aslında bu veri analizinin nasıl yapıldığının algoritmasıdır. Bu algoritma, analizi yapacak olanlara yol göstermektedir.

- I. Problemin belirlenmesi
- II. Belirlenen problem için toplanacak veriye altyapı oluşturulması
- III. Verinin toplanması
- IV. Analizin yapılması ve verinin işlenmesi
- V. Optimizasyon ve tasarım aşaması

Bu adımlar ile veri analizinin yol haritası çerçevelenmiş olur. Bu çerçeve doğrultusunda ilerlemeler kaydedilir. Veri analizinin yaşam döngüsü genele bakıldığında da dört temel başlığa atanmış ve bu yol hattı bu şekilde dile getirilmiştir.

- a) Verinin içe alınması; veri toplama aşamasıdır.
- b) Verinin kullanıma hazır hale getirilmesi; hazırlık ve ön işleme aşamasıdır.
- c) Veri modelleme; analiz aşamasıdır. Modelleme algoritmaları oluşturulur ve analizler yapılır.
- d) Karar alma; elde edilen bilginin kullanılma aşamasıdır. Veri içeri alınıp, işlemlerden geçtikten sonra analiz edilerek bilgi elde edilir. Bu bilginin kullanılması ve yorumlanması gerekmektedir.

Şekil 3.2: Veri analizi adımları



Kaynak: <https://medium.com/deniz-kilinc-veri-analizi>

3.4. VERİ ANALİZİ TÜRLERİ

Veri analizi türleri yedi temel başlık altında ele alınabilir. Bilgiyi elde etmek için girilen bu yolda bizleri istediğimiz sonuca ulaştırma konusunda destekleyici olmaktadır.

3.4.1. Betimsel Analiz

Geçmiş verileri özetler ve görselleştirir. Mevcutta olan olaylar analiz edilir ve gözler önüne serilir. Durumu anlamak ve canlandırabilmek açısından önemlidir. Burada durum tanımlanmış olur. Ortalama ile standart sapma ve yüzde ile sıklık şeklinde verileri elde edebilmek için ya tam veriye ya da özetlenmiş sayısal veri örneğine dayanmaktadır. Bir e-ticaret firması örneği düşünüldüğünde kampanya döneminde hangi ürünlerin ne derece satıldığı ve tercih sebepleri gibi konular betimlendiğinde bu doğrultuda mevcut durum ve ilişki gözler önüne serilmiş olmaktadır.

3.4.2. Teşhis Edici Analiz

Yaşanılan durumun ve geçmiş verilerin ne olduğunu değil nasıl olduğunu incelemeye yönelik yapılan analizdir. Bu analiz ile neden sonuç ilişkileri kurulur ve bu doğrultuda çıkarımlar elde edilmiş olur.

Burada hedef istatistiksel analizi ileriye taşıyarak bir şeyin neden olduğunu cevaplamak içindir. Neden olduğu sorusuna cevap verildiğinde ve böylelikle davranış kalıpları ortaya çıkarılabildiğinde bu doğrultuda hareket edilebilir. Verilerin davranış kalıplarını tanımlamak için yararlıdır.

3.4.3.Öngörücü Analiz

Öngörücü analiz, geçmiş verilerden elde edilmiş istatistiksel modellere ve tanımlanmış eğilimlere göre olası sonuçları elde etmeyi sağlar. Öngörücü analiz yapabilmek adına strateji geliştirmek için model oluşturabilmek ve doğrulama yapabilmek gereklidir. Bu analiz türünde genellikle makine öğrenimine başvurulur.

Mevcut veya geçmiş verilere dayanarak gelecekteki sonuçlar hakkında tahminlerde bulunmak için kullanılır. Doğruluğu, ne kadar ayrıntılı bilgiye sahip olunduğuna göre değişmektedir. Bir trendin, modelin veya olayın neden gerçekleştiği anlaşılabilir ise, karşılaşılabilecek olan durumlara dair bilinçli bir projeksiyon geliştirilebilir ve oluşması muhtemel potansiyel sorunların kontrolden çıkması önlenabilir. Bu sayede girişimler kontrol altına alınabilir ve etkili kampanyalara imza atılabilir. Kampanyalar, oluşma ihtimali olan potansiyeller doğrultusunda oluşturulabilir.

3.4.4.Çıkarımsal Analiz

Bir tam veri kümesi kullanılarak örnek analizler yapılır. Aynı veri kümesindeki farklı örnekler yorumlanarak farklı sonuçlar elde edilebiliyor olmaktadır. Küçük miktarda veri kullanarak, daha büyük miktardaki gruplar hakkında yorum yapabilmek ya da kararlar alabilmek için kullanılır. Küçük veri grubu davranışları üzerinden çıkarımlar ile birlikte büyük veri grubu üzerine yorumlar yapılabilir.

3.4.5.Kural Koyucu Analiz

Kural koyucu analizler temelde makine öğrenimi analizleri ile yapay sinir ağlarını içermektedir. Bu analiz türü için diğer üç tür analize dayalı sağlam bir temele ihtiyaç

vardır. mevcut bir problem ya da karar sürecinde hangi davranışın gerçekleştirilebileceğini belirlemek için önceki tüm analizler tarafından elde edilen sonuçları birleştiren tekniktir. En etkili veri analizi yöntemlerinden biri olarak, pratik iş stratejileri geliştirmek için eğilimleri kullanır. Veri konusunda odaklı olan çoğu kurum ve kişiler , öngörülü analiz ve teşhis analizinin performansını artırmak için kural koyucu analizi tercih eder. Mevcut durumlara ve sorunlara dayanan verileri analiz ederek kararlar alması bu tercihin sebebidir.

3.4.6. Metin Çözümlemesi Analizi

Metin analizine veri madenciliği denir. Ham veriler üzerinden iş öngörülerini elde etmeye yaramaktadır. Metin çözümlemesi analizi çalışmalarını sosyal medya ortamındaki yorumlar kullanılarak yapılan çalışmalarda görebilmekteyiz. Burada mevcut metinler üzerinden incelemeler yapılır ve genele yaygınlaştırılabilir kullanım ortaklıkları üzerine çıkarımlar yapılabilir.

3.4.7. İstatiksel Analiz

İstatistiksel analiz, bir veri kümesinin veya bir veri örneğinin analizi için verilerin toplanmasını, analizini, yorumlanmasını, sunumunu ve modellenmesini içermektedir. Aslında ham verinin geçtiği akışlar ve sürecin isimlendirilmesidir. İstatiksel analiz, betimsel analiz ile çıkarımsal analizi kavramaktadır. Burada sürecin dağılımlarına ve aşamalarına göre alt kırılımlar oluşmaktadır.

4. MAKİNE ÖĞRENİMİ

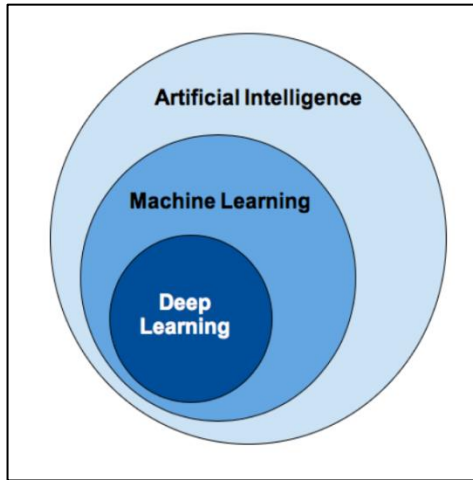
4.1. MAKİNE ÖĞRENİMİ TANIMI

Makine öğrenimi, verileri ve deneyimleri kullanarak insan beynine benzer, öğrenme yeteneğine sahip donanımda bilgisayar oluşumunu sağlayan süreçtir. Elde edilen bilgiler ile donatma süreci olarak ele alınır. (Altunışık 2015)

Algoritma tabanlı bilgisayarların yaptığı işlemlerde hata payı yoktur ve işlemler adımlar halinde belli aşamalarla tamamlanır. Mevcut örnek veriler doğrultusunda bilgisayarların aldığı temel kararlar ve durumlar vardır. Bir girdiye dayalı çıktı sürecidir. Burada karar verme sürecindeki insanlarda olduğu gibi hatalar oluşabilir. Bu akışa sahip ve sonuçları elde ettiğimiz bir süreçtir. (Altunaydın ve Çelik 2018)

Makine öğrenimi yapısal yetenek olarak düşünüldüğünde öğrenebilen ve veriler üzerinden tahmin elde edebilen algoritmaların çalışma ve oluşumlarını araştıran bir yapıdır. Makine öğreniminin temel amacı, kendilerini geliştirmek için eğitebilecek modeller oluşturmaktır. Karmaşık kalıpları algılar ve önceki sorunları kullanarak yeni sorunlara çözümler bulur.

Şekil 4.1: Derin öğrenme, makine öğrenmesi ve yapay zeka arasındaki ilişki



Kaynak: Erbey, A.,Yapay Zekaya Giriş, <https://medium.com/@alierbey/yapay-zekaya-giris>

4.2. MAKİNE ÖĞRENİMİ TARİHÇESİ

1940'lı yıllarda nöronların elektriksel çökmeleri üzerine çalışmalar ve incelemeler yapılmış, bununla beraber insanın karar alma mekanizması incelemeleri hızlanmış ve artmıştır. Bu doğrultuda 1950'li yıllarla birlikte özellikle yapay zeka üzerine araştırmalar artmıştır.

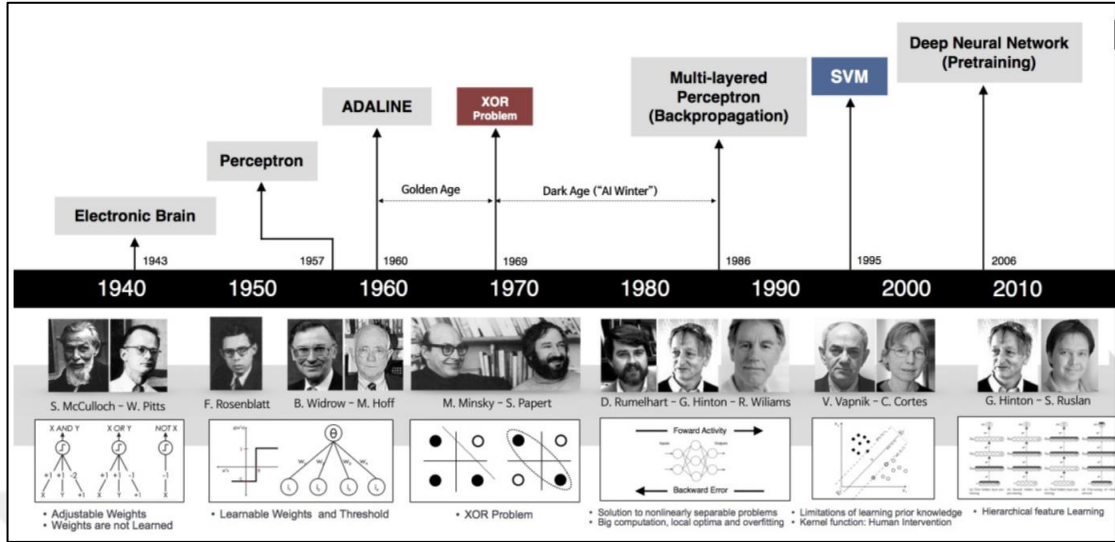
Alan Turing tarafından geliştirilen ve bir makinenin bir insanı taklit etme yeteneğini test etmek için geliştirilen Turing Testi ile 'makine zekası' kavramı yerini almıştır. 1956 yılında Marvin tarafından düzenlenen yaz okulunda 'yapay zeka' kavramı yerini almıştır. (Altunaydın ve Çelik 2018)

Makine öğrenmesi, 1959 yılında bilgisayar biliminin yapay zekada sayısal öğrenme ve model tanıma çalışmalarından geliştirilmiş bir alt dalıdır. 1959 yılında Arthur Samuel, bilgisayarların programlamadan bağımsız olarak öğrenme yetilerini kazanabileceklerini söylemiş ve bir dama oyunu geliştirmiştir. Bu oyunu diğerlerinden ayıran özelliği geçmişteki hamlelerinden ders çıkararak kendini geliştirmesi olmuştur. Makine öğrenmesinin temeli olan bu hareket ile deneyim ve veriler kullanılarak öğrenen sistemler oluşturma dönemi başlamıştır. Bu çalışma ile birlikte 'makine öğrenimi' kavramı ile devam edilmiştir. (Topal 2017)

Makine öğrenmesini ilk dile getirenlerden ve hayatımıza dahil edenlerden en çok anılan isim psikiyatr Frank Rosenbalt'dır. İnsan sinir sistemi üzerine Cornell Üniversitesi'nde çalışmaktadır. Buradan yola çıkarak alfabede yer alan harfleri algılayabilen perceptron olarak adlandırdığı makineyi hayata geçirmiştir. (Fradkov 2020)

1959 yılında olan bu ve benzeri psikoloji dünyası bilim insanları hareketleri makine öğrenimini hayatlarımıza dahil etmiştir. Derin Öğrenme, Makine Öğrenmesi ve hepsini kapsayan Yapay Zeka dünyasını oluşturan bu sürecin temelinde insan zihni ve öğrenme mekanizması yer almaktadır. Bu doğrultuda zaman içerisinde çok farklı çalışmalar yapılmış ve yöntemler geliştirilmiştir.

Şekil 4.2: Makine öğrenimi ve yapay zeka tarihçesi



Kaynak: <https://bilisim.io/2017/11/17/yapay-zekanin-kisa-bir-tarihcesi/>

4.3. MAKİNE ÖĞRENİMİ METODLARI

Makine öğrenimi, girdi verileri ile analizi gerçekleştiren modelleri kullanarak, belirli bir aralıkta tahmin ve değer üretebilerek çıktı elde edilmesini sağlamaktadır. Bu algoritmalara yeni veriler gönderilirken, performansı iyileştirmek ve zamanla zeka geliştirmek için operasyonları öğrenir ve optimize ederler. Bu süreç ve akış sırasında kullanılan metodlar vardır. Bu metodlar dört başlı altında ele alınmaktadır.

4.3.1. Denetimli Makine Öğrenimi

Denetimli makine öğreniminin makine örnek olarak öğretilir. Burada istenen giriş ve çıkışları içeren, çalışılan veri kümesi ile algoritma sağlanır. Bu giriş ve çıkışlara ulaşmak için yöntem bulunur. Algoritma verideki kalıpları tanımlar, gözlemler ve bunun sonucunda öğrenir. Öğrendikleri ile tahminlerde bulunur ve yüksek düzeyde doğruluk/performans elde edilinceye kadar düzeltmeler ile işlemler sağlanır.

Denetimli makine öğrenimi tahmin modelleri geliştirmek için sınıflandırma ve regresyon tekniklerini kullanır. Tahmin ; geçmiş ve şimdiki verilere dayanarak gelecek

ile ilgili çıkarımlar yapma sürecidir ve genellikle eğilimleri analiz etmek için kullanılır. (Altunaydın ve Çelik 2018)

4.3.2. Sınıflandırma

Sınıflandırma, verilerin veri setinde tanımlanan kategorilere özel özelliklerine göre dağıtılmasıdır. Bu teknikte, makine öğrenmesi programı gözlemlediği değerlerden bir sonuç çıkarmalıdır ve yeni gözlemlerin hangi kategoriye ait olduğunu belirlemelidir. Burada yapılan, girdi verilerini kategorilere ayırmaktır. Veriler etiketlenebilir, kategorilere veya belirli gruplara ayrılabiliriyorsa sınıflandırma tekniği kullanılabilir. (Bulut 2019)

Sınıflandırma yapmak için kullanılabilecek olan yaygın algoritmalar arasında ; Karar Ağaçları, KNN, Naive Bayes, Lojistik Regresyon sayılabilir.

4.3.3. Regresyon

Regresyon, verilerin mevcut özelliklerine göre diğer özelliklerini tahmin etmek veya sonuç elde etmektir. Makine öğrenmesi programı, değişkenler arasındaki ilişkileri tahmin etmelidir. Regresyon analizi, bir bağımlı değişkene ya da bir dizi başka değişkene odaklanır. Bu tahmin için faydalıdır. Veri aralığıyla çalışıldığında yanıtın niteliği geçen süre gibi gerçek bir sayı olduğunda regresyon tekniği kullanılabilir. (Bulut 2019)

Regresyon yapmak için kullanılabilecek olan yaygın algoritmalar arasında ; Doğrusal Model, Doğrusal Olmayan Model, Regülasyon, Kademeli Regresyon, Sinir Ağları sayılabilir.

4.3.4. Denetimsiz Makine Öğrenimi

Denetimsiz makine öğrenimi çıktı verisi yoktur ve öğrenme işlemi verilerin aralarındaki ilişki ve bağlantılar ile gerçekleşmektedir. Eğitim veri seti yoktur. Denetimsiz

öğrenmenin sağladığı avantaj olarak istatistiklere dayanan tahminlerin, etiketlenmemiş örnekler arasındaki kalıpları, korelasyonları ve ilişkileri tanımlamak için kullanılmasıdır. (Panjeh 2018)

Verilerle ilgili birçok özellik bilinmesine rağmen, verilerin grup ve kategorilerine ilişkin özelliklerinin belirlenemediği durumlar mevcuttur. Bu tip durumlarda özelliklerin bir alt gruplarının seçilmesi ya da özellikler birleştirilerek yeni özellikler elde edilmesi işlemleri yapılabilir.

Denetimsiz makine öğrenimi tahmin modelleri geliştirmek için kümeleme ve ilişkilendirme tekniklerini kullanır.

4.3.4.1 Kümeleme

Verilerdeki doğal gruplamalar bilinmez. Buradaki çalışma ile birbirine benzer veri grupları bulunur. Kümeleme, çok değişkenli obir analiz yöntemidir ve birbirine benzer değişkenlere sahip, grup sayısı tam olarak bilinmeyen benzer durumda olan alt kümelere ayrılabilmeyi sağlamaktadır.

Kümeleme, birim ve bu birimlere ait değişkenlerin sınıflamaları hakkında kesin bilginin bulunmadığı bir veri setinde alınan n tane birimin, p tane değişkene ilişkin gözlem sonuçları ile ilgilenmektedir. Homojen nesneler toparlanarak heterojen olan gruplar elde edilir ve birimler hiyerarşik bir şekilde düzenlenmektedir. Sınıflandırma yapmak gözlem sonuçlarının çok az bir kayıpla bir araya toplanmasını sağlamaktadır. Bu doğrultuda çıkarımlar ve tahminlemeler de kolaylıkla yapılabilir.

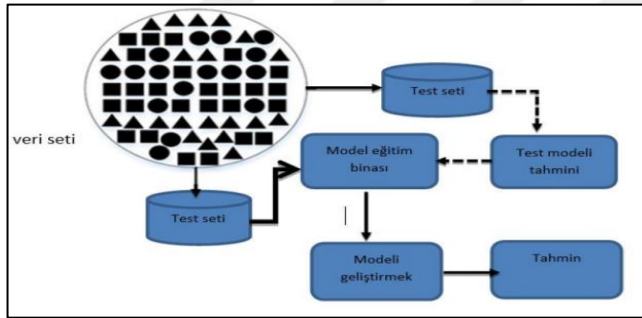
Kümeleme algoritmaları verileri işler ve verilerde olan doğal kümeleri bulur. Algoritmanın kaç kümeyi tanımlaması gerektiği kısmı da değiştirilebilir.

4.3.4.2 İlişkilendirme

Aynı veri grubundaki, veriler ele alındığında aralarındaki ilişkiler incelenir ve aynı olan ilişki ile bağlantılar belirlenir. Genellikle birlikte ortaya çıkan öge kümeleri tanımlanmış olur.

Verinin büyük bölümünü açıklayan kurallar keşfedilebilir. Örnek olarak verilebilecek olan; bir markette kullanıcıların aldıkları ürünler incelendiğinde X ürününü alanlar Y ürününü almaya eğilimlidir. T ürünü ile Z ürünü beraber alınmaktadır gibi mevcut üzerinden aralarındaki ilişkiler çıkarılarak sonuçlandırmalar yapılabilmektedir. İlişkilendirme burada aslında bu veri setinden farklı yollar çıkarabilmeyi ve etki edilen alanı büyütebilmeyi sağlamaktadır.

Şekil 4.3: Denetimsiz öğrenim çalışma yapısı



Kaynak: Panjeh, M. (2018) Bilgisayar Ağlarında Makine Öğrenimi Algoritmalarını Kullanarak İzinsiz Giriş Tespiti. Yüksek Lisans Tezi

4.3.5. Yarı Denetimli Öğrenim

Yarı denetimli öğrenim, bir öğrenme türüdür. Etiketlenmiş ve etiketlenmemiş veriler ile eş zamanlı çalışır. Bu teknik iki aşamalı kümeleme ve sınıflandırma içermektedir.

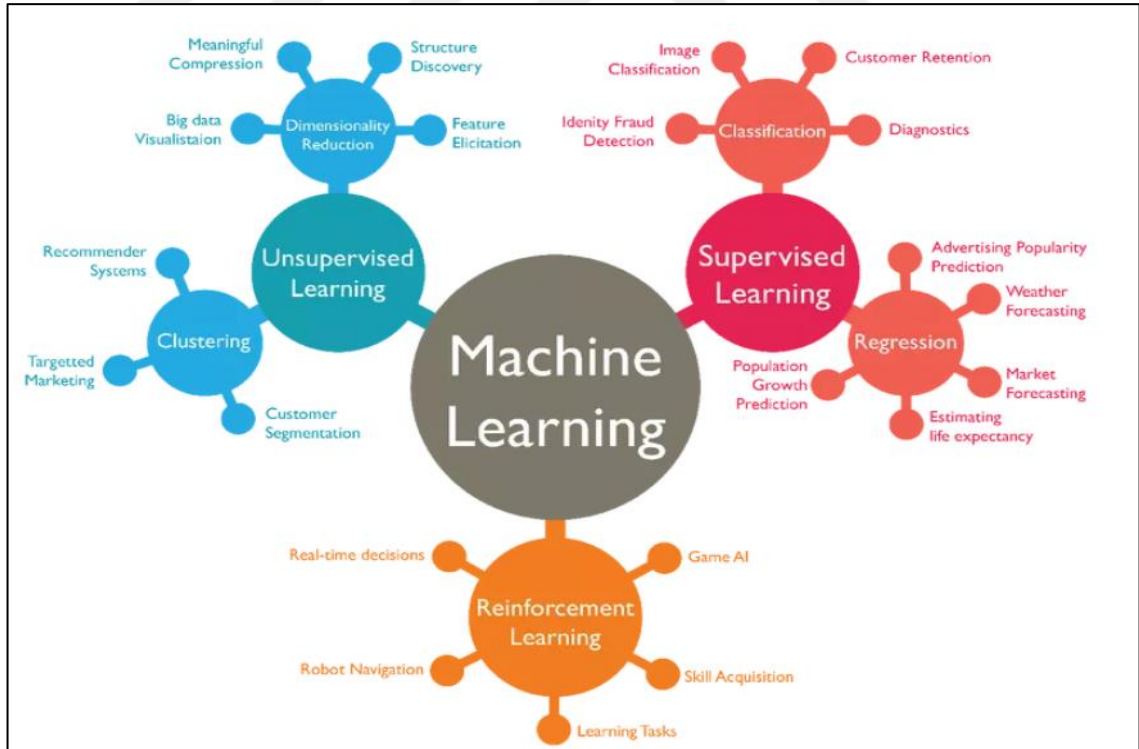
Yarı gözetimli öğrenme, eğitim için etiketlenmemiş verilerin kullanımını da sağlayan denetlenen öğrenme görevleri ve tekniklerinin bir sınıfıdır. Genellikle etiketlenmemiş veri sayısı etiketli verilerden daha fazla olabilmektedir. Yarı denetimli öğrenme, denetimsiz öğrenme ve denetimli öğrenme arasında yer almaktadır. Hem etiketli hem de

etiketsiz olan veriler ile aynı zamanda çalışabiliyor olmasından dolayı bu şekildedir. Etiketlenmemiş veriler az miktarda da olsa etiketlenmiş veriler ile birlikte kullanıldığı durumlarda öğrenim doğruluğunda fark edilebilir bir iyileşme oluşturmaktadır.

4.3.6. Takviyeli Öğrenim

Takviyeli öğrenim, bir makine öğrenimi algoritmasına bir dizi eylem, parametre ve son değer içeren, düzenli öğrenme süreçlerine odaklanmaktadır. Kuralları belirleyerek, makine öğrenimi algoritması farklı seçenekleri ve olasılıkları keşfetmeye çalışır, hangisinin en uygun olduğunu belirlemek için her bir sonucu izleyerek ve değerlendirmektedir. Bu teknik ile makine denemesi ve hata öğretilir. Uygun olanı seçtiğinde ödüllendirilirken uygun olmayanda cezalandırma sistemi vardır. Geçmiş deneyimlerden öğrenir ve en iyi olacak cevaba adaptasyon sağlanır.

Şekil 4.4: Makine öğrenimi metodları ve kullanılabilecek durumlar ilişkisi



Kaynak: <https://medium.com/halil-ibrahim-safak/makine-ogrenmesi>

4.4. MAKİNE ÖĞRENİMİNDE KULLANILAN ALGORİTMALAR

Makine öğrenimi metodları çerçevesinde bilgiye ulaşım sürecinde kullanılabilecek çok fazla algoritmalar mevcuttur. Bu algoritmalar ile modeller oluşturulur ve hazırlanmış veriler ile birlikte modeller çalıştırılır.

Doğru makine öğrenme algoritmasını seçmek çeşitli faktörlere bağlıdır. Bunlara veri boyutu, kalite ve çeşitlilik ile istenilen cevap temel olarak söylenebilir. Dikkat edilecek diğer hususlar doğruluk, eğitim süresi, parametreler, veri noktaları ve daha fazlasını içermektedir.

4.4.1. Naive Bayes Algoritması

Denetimli öğrenme metodunun altında yer alan sınıflandırma tekniği algoritmalarındandır. Bayes teoremine dayanmaktadır ve her değeri diğer değerden bağımsız olacak şekilde sınıflandırmaktadır. Bayes teoremi, bir rassal değişken için olasılık dağılımı içinde koşullu olasılıklar ile marjinal olasılıklar arasındaki ilişkiyi gösterir. Şekil 4.5’de teoremin temelinde yer alan formül açıklayıcı bir şekilde yer almaktadır. Olasılık kullanarak ve belirli bir özellik dizisine istinaden bir sınıfı tahmin etmenizi sağlar. Yani ilgili olayın oluşumunu temel alır ve bir olayın oluşma olasılığını hesaplar. İyi performans gösterdiği için sıkça tercih edilen bir algoritmadır. (Albayrak ve diğ. 2016)

Şekil 4.5: Bayes formülü

$$P(A|B) = \frac{P(B|A) P(A)}{P(B)}$$

The diagram illustrates the components of Bayes' theorem with arrows pointing from text labels to the corresponding terms in the formula:

- $P(A|B)$: THE PROBABILITY OF "A" BEING TRUE GIVEN THAT "B" IS TRUE
- $P(B|A)$: THE PROBABILITY OF "B" BEING TRUE GIVEN THAT "A" IS TRUE
- $P(A)$: THE PROBABILITY OF "A" BEING TRUE
- $P(B)$: THE PROBABILITY OF "B" BEING TRUE

Kaynak: <https://medium.com/ekrem-hatipoglu/machine-learning-classification-naive-bayesi>

Tembel bir öğrenme algoritması olan Naive Bayes dengesiz veri kümesinde çalışabilmektedir. Algoritmanın çalışma şekli bir eleman için her durumun olasılığını hesaplamak ve olasılık değeri en yüksek olana göre sınıflandırmak şeklindedir.

4.4.2. Destek Vektör Makineleri Algoritması

Denetimli öğrenme metodunun altında yer alan sınıflandırma tekniği algoritmalarındandır. Destek Vektör Makineleri, düzlem üzerindeki noktaların bir doğru veya hiper düzlem ile ayrıştırılması ve sınıflandırılmasıdır. Yüksek boyutlu uzaylarda ve boyut sayısının örneklem sayısından fazla olduğu durumlarda etkilidirler.

DVM, dış bükey optimizasyona dayalıdır ve yapısal risk minimizasyonu prensibine uygun olarak çalışır. Veriye ait herhangi bir birleşik dağılım fonksiyonu bilgisine ihtiyaç duymaz ve dağılımdan bağımsızdır. DVM, örüntü tanıma ve sınıflandırma problemlerinin çözümü için Vapnik tarafından geliştirilmiştir. DVM'in temelleri istatistiksel öğrenme teorisine diğer bir ifadeyle Vapnik-Chervonenkis teorisine dayanmaktadır. (Ayhan ve Erdoğan 2014)

Destek Vektör Makineleri'nin en önemli avantajı, sınıflandırma problemini kareli optimizasyon problemine dönüştürerek çözüyor olmasıdır. Bundan dolayı problemin çözümüne ilişkin öğrenme aşamasında işlem sayısı azalmakta ve diğer algoritmalara göre daha hızlı çözüme ulaşılabilmektedir. Teknik anlamda öne çıkan bu özelliğinden dolayı, özellikle büyük hacimli veri setlerinde büyük avantaj sağlamaktadır. Ayrıca optimizasyon temelli olduğu için sınıflandırma performansı, hesaplama karmaşıklığı ve kullanılabilirlik açısından diğer tekniklere göre daha başarılı olabilmektedir. (Ayhan ve Erdoğan 2014)

4.4.3. Lojistik Regresyon

Denetimli öğrenme metodunun altında yer alan regresyon tekniği algoritmalarındandır. Lojistik regresyon, bir sonucu belirleyen bir veya daha fazla bağımsız değişken bulunan bir veri kümesini analiz etmek için kullanılan bir yöntemdir. Yalnızca iki olası sonuç

vardır. Lojistik regresyon, y değişkenin olma olasılığı ile veri setindeki x öznitelikleri arasındaki ilişkiyi anlamamıza yardımcı olmaktadır. Veri setimizde eğitim, medeni durum ve yaşanan semt gibi özniteliklerin olduğunu varsayalım. Maaşın 20.000 TL'den büyük olma olasılığını incelemek istediğimizde bize bahsettiğimiz bu öznitelikler arasındaki ilişkiyi anlamaya yardımcı olmaktadır. Eğitim ile maaş kıyası gibi öznitelik ile aralarındaki bu ilişki üzerine çalışmamızı sağlamaktadır.

Algoritma için sağlanan önceki verilere dayanarak meydana gelen bir olay için olasılığının tahmin edilmesine odaklanır. İkili bağımlı bir değişkeni kullanır. Verilere sürekli devam eden S şeklinde bir eğri yerleştirir. Lojistik regresyonda amaç, ikili yapıdaki bağımlı değişken ile bağımsız değişkenler arasındaki ilişkiyi en uygun şekilde açıklayan modelin bulunmasıdır.

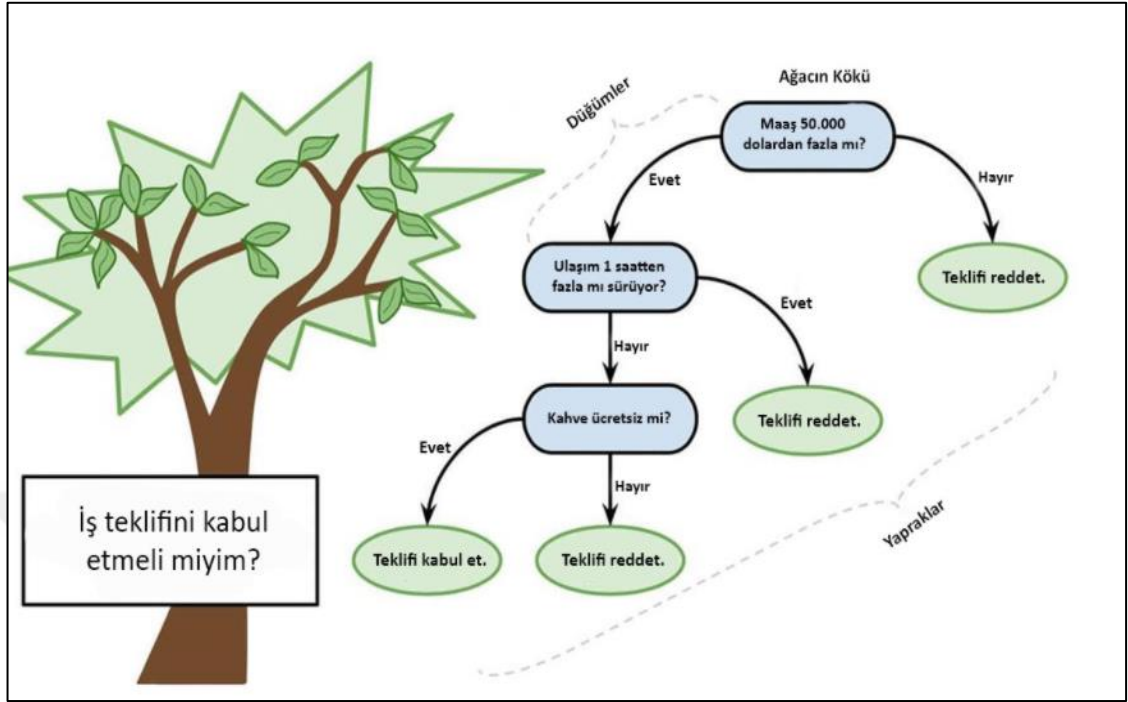
4.4.4. Karar Ağaçları

Denetimli öğrenme metodunun altında yer alan sınıflandırma/regresyon tekniği algoritmalarındandır. Karmaşık veri setlerinde kullanılabilir. Veri ön işleme gerek duymaz.

Karar ağaçları, özellik ve hedefe göre karar düğümleri ve yaprak düğümlerinden oluşan ağaç yapısı formunda bir model oluşturan bir yöntemidir. Karar için olası olan her sonucu göstermek amacıyla dallanma yöntemini kullanır ve akış şemasına benzer bir ağaç yapısıdır. Ağaç üzerindeki her düğüm belirli bir değişken üzerinden uygulanan bir testi temsil eder ve her dal bu testin sonucu olarak yer almaktadır.

Karar ağacında bölüm kriteri bilgi kazanımı değeridir. En yüksek bilgi kazanımına sahip özellik ağaca eklenir. Bu işlem veri setindeki tüm özellikler için uygulanarak ağaç oluşturulur. (Ayhan ve Erdoğan 2014)

Şekil 4.6: Karar ağacı örnek çalışması



Kaynak: <https://medium.com/deep-learning-turkiye/karar-ağaçları-makine-öğrenmesi-serisi>

Şekil 4.6’da görüldüğü gibi örnek bir çalışma üzerinden karar ağaçları yapısı gösterilmiştir. Burada veri setini olabildiğince açıklayabilen bir kök seçimi olmalıdır. Burada kök değer bulunmak istendiğinde her düğüm için Gini ya da Entropi değerlerine bakılmaktadır. Gini, alt küme saflık değeri anlamına gelmektedir ve her sınıf için hesaplanır. Çıkan sonuçların karelerinin toplamı 1 değerinden çıkartılır. 0 ile 1 arasında bir sonuç alınır ve 0’a ne kadar yakınsa ayrım o derece iyidir anlamına gelmektedir. Burada da her düğüm için Gini değerine bakılmış ve bu doğrultuda maaş düğümü kök olarak seçilmiştir. Kök hücreden sonra gelecek düğümler seçilirken de Gini değeri üzerinden işlemler yapılmaktadır. Entropi, bir grupta bozukluğu hedef değere göre ölçmek temel fikri ile \log_2 tabanında işlem yapar. Entropi dengeli bir ağaç çıkarmak hedefindeyken, Gini frekansı fazla olan bir sınıfı ayırtmaya meyillidir.

Bu örnekte de ilgili hesaplamalar sonrası veri seti Şekil 4.6’da görüldüğü hale gelebilmiştir.

4.4.5. Random Forest Algoritması

Denetimli öğrenme metodunun altında yer alan sınıflandırma/regresyon tekniği algoritmalarındandır.

Birden fazla algoritmayı birleştiren bir topluluk öğrenme yöntemidir. Birden fazla karar ağacı oluşturulur, daha doğru ve istikrarlı bir tahmin elde etmek için birleştirme işlemi yapılır. Her bir sınıflandırıcı zayıf ancak başka sınıflandırıcılar ile birleştiğinde mükemmel sonuçlar üretebilir. Bundan dolayı da en çok tercih edilen algoritmalarından biridir.

Algoritma bir ağaç benzeri bir grafik veya karar modeli ile başlar ve en üste bir girdi girilir. Daha sonra ağaçtan aşağı doğru hareket eder, veriler belirli değişkenlere bağlı olarak daha küçük kümeler bölünür. Bu algoritma ile varyans yani overfitting azalır. Bunun yanında bootstrap yöntemiyle oluşturulan alt veri kümelerinde outlier bulunma olasılığı da düşürülmüş olur. Özniteliklerin ne kadar önemli olduğunu veren algoritma sayesinde x sayıda öznitelik verip en faydalı y tanesini seçmesini isteyebilir ve bu bilgi istenilen başka bir modelde kullanılabilir.

4.4.6. K-Means Algoritması

Denetimsiz öğrenme metodunun altında yer alan kümeleme tekniği algoritmalarındandır. Tanımlanmış kategori veya grup içermeyen verileri kategorilere ayırmak için kullanılır. K değeri küme sayısını belirler ve bu değeri parametre olarak alınmaktadır. K değişkeni ile temsil edilen grupların sayısı ile veri içindeki grupları bularak çalışmasını sürdürür. K değeri belirlendikten sonra algoritmada rastgele K tane merkez noktası seçer. Sonrasında verilen özelliklere dayanarak her bir veri noktasını K gruplarından birine atamak için tekrarlar. Verilerin hepsi ile rastgele olacak şekilde belirlenmiş merkez noktaları arasındaki uzaklığın hesaplanarak, verinin en yakın olduğu merkez noktaya göre bir kümeye atanmasıdır. Sonrasında her küme için tekrar bir merkez noktası seçilir ve yeni seçilmiş olan merkez noktalarına göre kümeleme işlemi yapılır. Bu durum sistemi kararlı hale getirinceye kadar devam etmektedir.

Her kümenin içinde yer alan veri noktaları homojendir ancak diğer kümelerdeki veri noktalarına karşı heterojen olma özelliği taşırlar. K-Means'in atama mekanizması her verinin sadece bir kümeye ait olabilmesine izin vermektedir. Çalışma yönteminde, Şekil 4.7'de görülen Öklit uzaklığı formülü temel alınarak kümeleme yapılmaktadır.

Şekil 4.7: Öklit uzaklığı formülü

$$p = (p_1, p_2, \dots, p_n) \text{ ve } q = (q_1, q_2, \dots, q_n)$$
$$\sqrt{\sum_{i=1}^n (p_i - q_i)^2} = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \dots + (p_n - q_n)^2}$$

Kaynak: Sarıman, G. Veri Madenciliğinde Kümeleme Teknikleri Üzerine Bir Çalışma: K-Means ve K-Medoids Kümeleme Algoritmalarının Karşılaştırılması

4.4.7. Yapay Sinir Ağları

Takviyeli öğrenme metodunun altında yer alan algoritmalarındandır. Yapay nöronlardan oluşmakta ve insan beyninin basit bir modelini gerçeklemektedirler. Ancak yapıları daha basittir. Optimum ağırlık değerlerinin bulunması doğruluk oranlarını artırır. Eğitim veri setinden faydalanarak yapısında bulunan nöronlar arasındaki bağlantıların ağırlıklarını bulmaya çalışır. (Ayhan ve Erdoğan, 2014)

Yapay sinir ağı, örnek ve deneyim yoluyla da öğrenirler. Doğrusal olmayan ilişkileri yüksek olan verileri modelleme konusunda faydalıdır. Yapay sinir ağı örneklerle ilgili bilgiler toplamakta, genellemeler yapmakta ve daha sonra hiç görmediği örnekler ile karşılaştırıldığında daha öncesinde öğrenmiş olduğu bilgileri kullanarak, mevcut örnekler hakkında karar verebilir olmaktadır. Bu öğrenme ve genelleme özelliklerine sahip olmaları dolayısıyla birçok alan ve sektörde çözülmesi zor olan durum ve problemler için kullanılmaktadır.

Yapay sinir ağılarının temeli olarak varsayılan hesaplama modeli hareketleri 1940'ların başında araştırmalarına başlayan W.S. McCulloch ve W.A. Pitts'in, 1943 yılında yayınlamış oldukları makale ile atılmıştır. 1954 yılında B.G. Farley ile W.A. Clark tarafından bir ağ içerisindeyken uyarılara tepki verebilen, uyarılara karşı adapte olabilen model oluşturulabilmiştir. 1960 yılı ile birlikte ilk sinirsel bilgisayar ortaya çıkmıştır. 1963 yılında basit modellerin sahip oldukları eksiklikler fark edilebilmiş, buna rağmen başarılı sonuçların alınabilmesi 1970 ve 1980'lerde termodinamikteki teorik yapıların doğrusal olmayan ağların geliştirilmesinde kullanılmasına kadar gecikmiştir. 1985 yapay sinir ağları tanınırlığının arttığı, yoğun araştırma çalışmalarının başlamış olduğu yıl olarak bilinmektedir.

5. PYTHON PROGRAMLAMA DİLİ

5.1. PYTHON PROGRAMLAMA DİLİ VE TARİHÇESİ

Python, nesne yönelimli, yorumlamalı, birimsel ve etkileşimli yüksek seviyeli bir programlama dilidir. Modüler yapısı, sınıf dizgesini ve her türlü veri alanı girişini destekler. Birden fazla çeşit platformda çalışabilir (Unix, Linux, Mac, Windows, Amiga, Symbian). Python ile sistem programlama, kullanıcı arabirimi programlama, ağ programlama, web programlama, uygulama ve veri tabanı yazılımı programlama gibi birçok alanda yazılım geliştirilebilir. (Yazgan 2017)

1980 yılında ABC programlama diline alternatif olarak geliştirilmiş ve 2000 yılında yayınlanmaya başlanmıştır. Eklemeler ve güncellemeler ile versiyonları ile kullanıcılara sunulmaya devam edilmiştir. 1991–1996 yılları arasında başında Amsterdam’da Guido Van rossum tarafından geliştirilmeye başlanmıştır. Adını Guido van Rossum’un çok sevdiği, Monty Python adlı altı kişilik bir İngiliz komedi grubunun Monty Python’s Flying Circus adlı gösterisinden almıştır. Python Yazılım Vakfı çevresinde toplanan gönüllülerin çalışmalarıyla gelişimler sürdürülmektedir. Kolay öğrenilebilir, kolayca okunabilir bir programlama dilidir. Çapraz platform desteğine sahiptir, birçok farklı platformda kendisine yer bulmaktadır.

5.2. PYTHON PROGRAMLAMA DİLİ İLE VERİ ANALİZİ

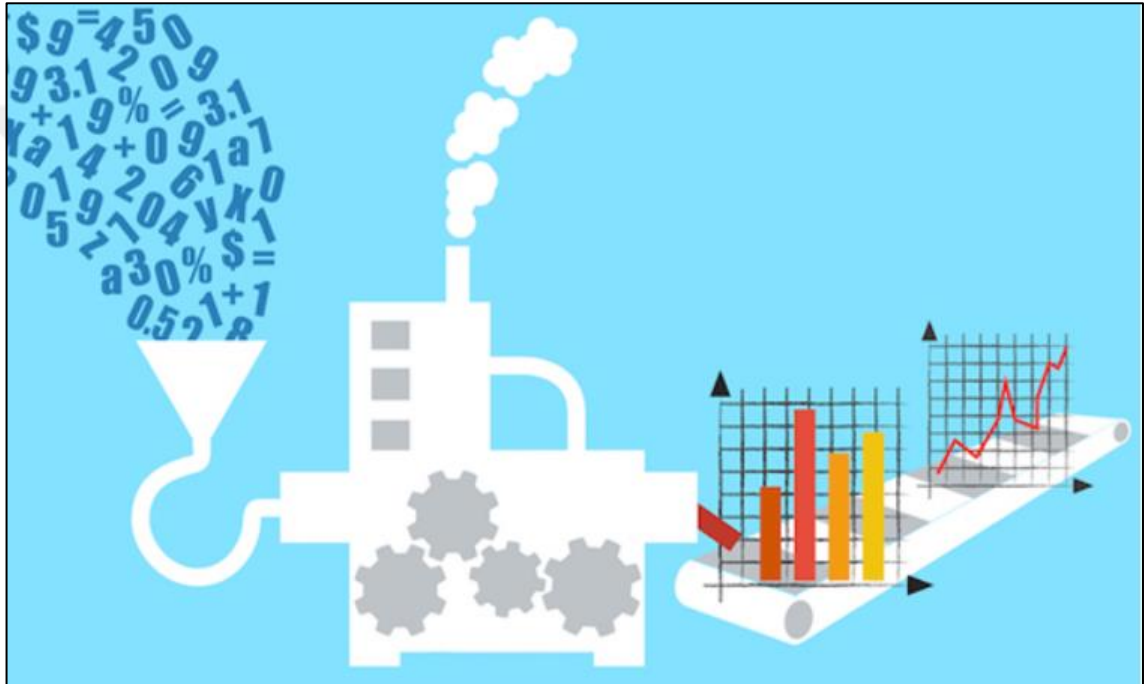
Veri bilimi, karmaşık problemleri çözmek için hem yapılandırılmış hem de yapılandırılmamış veriyi, işe yarar bilgiye dönüştürmeye yarar. Bu süreçte, bilimsel problem çözme teknikleri, matematik, istatistik ve yazılım geliştirme disiplinleri birlikte kullanılır.

Kolay öğrenilebilmesi, kolay okunması, bakımının kolay olması ve hazır birçok veri bilimi kütüphanesi bulundurması özellikleri dolayısıyla kullanımı artan ve bugün veri analizi denilince akla ilk gelen dillerden biri Python’dır. Popülerliği ve kullanımı

arttıkça, var olan kütüphaneleri daha fazla geliřmekte ve ayrıca yeni kütüphaneler de eklenmektedir.

Python programlama dili, verinin bilgiye dönüşmesi aşamasında ona bu başkalařımları geçirirken bütün yetenekleri ile destek olmaktadır. Bilgi üretimin gerçekteřtięi fabrika kullanılacak IDE iken üretimi saęlayacak olan python programlama dili olacaktır.

řekil 5.1: Python programlama dili ile veri analizi arasındaki iliřki



Kaynak: <https://medium.com/@mbektas/python-ile-kesifsel-veri-analizi>

6. FİNANS SEKTÖRÜ VE VERİ ANALİZİ

6.1. FİNANS SEKTÖRÜNDE VERİ ANALİZİ ÖNEMİ

Finansal sistem, yatırım ve tasarruf kararlarının farklı birimler tarafından verildiği ekonomilerde fonları buluşturarak getirisi en yüksek sektörlerden başlamak üzere yatırımlara kanalize edilmesine aracılık yapan finansal araçların oluşturduğu yapıdır. Finansal sistemde fonların yatırımlara aktarımını sağlayan araçlar ve araçlar mevcuttur. Nakit para, mevduat, krediler, hazine bonusu ve tahviller, varlığa dayalı menkul kıymetler vb. finansman araçları arasında yer alır. Sistemdeki finansal araçlar ise; bankalar, menkul kıymet borsaları, yatırım ortaklıkları gibi kurum ve kuruluşlardan oluşur. Finansal sistemi, banka ile piyasa temelli olmak üzere ikiye ayırarak tanımlama yapılabilmektedir. Bankalar, finansal sistem içerisinde önemli derecede yüksek bir paya sahiptir. Bu duruma banka temelli olan finansal sistem, sermaye piyasası kurumları yüksek pay sahibi olduğu durumuna da piyasa temelli finansal sistem adı verilmektedir. Banka temelli finansal sistem içerisinde, mevduat, mevduat sigortası, kredi şeklinde araçlar yoğun bir şekilde kullanılmaktadır. Piyasa temelli finansal sistem içerisinde türev ürünler yoğun bir şekilde yer alabilmektedir. (Barışık ve diğ. 2010)

Finans sektörü bu parçaları göz önüne alındığında veri bakımından da en zengin sektörlerden biridir. Bu veri birikimi büyük bir kaynaktır. Verinin analizi ve elde edilecek bilgilerin önemi de sektörde yer alan firmalar tarafından dikkate alınmaktadır. Belli başlıklar altında toparlayarak aslında verinin hangi amaçlarla kullanıldığı da listelenebilmektedir. Temelde amaç, yeni potansiyelleri fark etmek ve daha fazla getiri elde etmektir. Müşterilere tavsiyelerden, iş stratejilerine kadar birçok alanda finans dünyasının sorunlarını çözmekte kullanılabilir.

Risk Yönetimi

- Risk yönetimi şirketin güvenliği, güvenilirliği ve stratejik kararlarından sorumlu finansal kuruluşlar için son derece önemli bir alandır. Bunun nedeni şirket risklerinin müşteriler, rakipler ve yatırımcılar olmak üzere her yönden

gelmesidir. Risklerin boyutu, kayıp türü ve önemi değişebilmektedir. Veri bilimi ve makine öğrenmesi riskleri oluşumuna göre sınıflandırmak ve potansiyel zararı öngörmek konusunda önemli bir araçtır . Çok sayıda müşteri verisi, finansal borç ödeme ve sigorta sonuçları konusunda eğitim alan algoritmalar yalnızca risk puanlama modellerini arttırmakla kalmaz ve aynı zamanda maliyet verimliliği ile sürdürülebilirliği de arttırabilmektedir.

Müşteri Analizi

- Mevcut müşteri davranışları incelenerek ürün reklamları bu şekilde güncellenebilmektedir. Bununla beraber hareketleri doğru yorumlayarak potansiyele sahip ürün fikirleri de elde edilebilmektedir. Tahmin analizleri ile müşteri davranışları doğrultusunda öngörüler elde edilmektedir.

Sahtekarlık Vaka Tespitleri

- Sahtecilik tespiti, finans sektöründeki en önemli konu başlıklarından biridir. Finansal kurumların çoğu, müşterilerinin verilerini ve fonlarını güvende tutmak için mücadele etmektedir. Sistemdeki bir boşluk, bilgisayar korsanlarına yol açmakta ve bu milyonlarca dolandırıcılığa sebep olmaktadır. Kurumlar açısından bu konu hem itibar hem de getiri kaybına neden olmaktadır. Gerçek zamanlı izleme ve sahtekarlık tespiti, herhangi bir kuruluşun siber güvenliğini artırır. Sisteme her zamanki faaliyetlerle birlikte giren ve çıkan verileri sürekli izleyerek, veri bilimi yazılımı ve platformlarının anomali geldiği anda tutarlı bir şekilde tespit edilmesi mümkündür. Bu, dolandırıcılık oluşumunu azaltmakta ve ayrıca hasar kontrolüne yol açmaktadır. Bu çalışmaların yanında siber saldırıları durdurmak için sistemdeki değişiklikleri ve manipülasyonları tespit edecek şekilde tahmin modelleri de yapılmaktadır. (Barışık, Darıcı ve Öztürk, 2010)

Bu başlıklar temel olarak finans sektörü için verinin kullanımının gerçekleştirildiği noktalardır. Veri, burada büyük bir kaynak ve geleceğe dair adımlar atılırken de önemli bir destekçidir. Kurumların stratejileri bu verilerin analizleri doğrultusunda şekillenmekte ve kaynak olarak kullanılmaktadır.

6.2. FİNANS SEKTÖRÜNDE KREDİ TANIMI

Kredi temel olarak karşılıklı güvene dayanan ve gelecekte geri ödeme sözünü içeren bir işlem olarak tanımlanmaktadır. Borçlanan kimsenin borcunu ödeyemeyecek yada ödemek istemeyecek olması ihtimali, bu karşılıklı güven ilişkisinin bir borç anlaşması ve bir teminat ile garanti altına alınmasını gerektirmektedir. (Erim ve Türk 2005) Bu doğrultuda sözleşmeler hazırlanmakta ve koşullar belirlenmektedir. Kredi ancak geri ödeyeceği konusunda güven duyulan kişi ve kuruluşlara verilir. Kredi talebinde bulunan bir kişi veya kurumun güvenilir olup olmadığı, çeşitli yöntemlerle hakkında yapılacak istihbarat ve mali tahlil ile anlaşılır. Bununla birlikte, müşterinin aldığı krediyi kısmen ya da tamamen geri ödeyememesi her zaman mümkündür.

Geri ödeyememe ile birlikte batık krediler ortaya çıkmaktadır. Burada birden fazla türde durum oluşabilmektedir. Bireylerin ihtiyaçları doğrultusunda başvurdukları krediler ile büyük bir kredi kullanım havuzu oluşmuştur. Bu havuzda oluşan veriler ile mevcut durumlar analiz edilerek sonuçlar elde edilmektedir. Burada bankalar ile kredi kuruluşlarının kendilerini korumaları ve oluşabilecek risklere göre doğru konumlanabilmeleri için bu konuda incelemeleri okumaları iyi bir şekilde yapmaları gerekmektedir.

7. YÖNTEM

7.1. GENEL BAKIŞ

Giderek artan kredi havuzu ile birlikte risk analizi önemi de ele alınarak bu alanda çalışma yapılmıştır. Bu kapsamda UCI Repository üzerinden kullanıma açık ve güvenlik sıkıntısı oluşturmayacak Alman kredi veri setine ulaşılmıştır. Elde edilen veri seti incelenmiştir. Bu kapsamda ve veri seti de göz önüne alınarak ilgili çalışmalara bakılarak literatür taraması gerçekleştirilmiştir. Bu çalışma ile birlikte çalışılacak makine öğrenimi algoritmalarına karar verilmiştir.

Çalışma kapsamında 1000 tane müşteriye ait 20 özniteliğe sahip Alman kredi veri seti özellikleri ile birlikte incelenmiştir. 1000 satırlı veri seti ele alındıktan sonra baz alınacak özniteliğe karar vermek adına öznitelikler ve değerleri incelenmiştir. Burada kredi riski üzerine gidileceği için bu bağlamda hangi kredi geçmişi özniteliği üzerinden çalışılmıştır. Kredi geçmişi özniteliği kategorik değerlere sahiptir. Burada alabileceği değerler değerlendirerek ya da üzerinden değerlerini alabilen Risk özniteliği oluşturulmuştur. Risk özniteliği temel alınarak çalışma yapılmıştır. Veri setimizi Risk özniteliği baz alınarak ilgili modellerde çalıştırılması sağlanmıştır. Kategorik değerler nümerik değere çevrilmiştir. Çalışmada makine öğrenimi denetimli öğrenme başlığı altında yer alan sınıflandırma algoritmaları kullanılmıştır. Risk özniteliği ile birlikte çalıştırılan modellerin performansları değerlendirilmiştir.

7.2. KULLANILAN VERİ ANALİZİ ARAÇLARI

Çalışmada UCI Machine Learning Repository platformundan veri seti elde edilmiştir. UCI Repository, makine öğrenimi ile veri analizi çalışması yapmak isteyen ve akademik anlamda çalışmalarını sürdüren bireylere destek olmak amaçlı veri setlerinin olduğu platformdur.

Veri seti ile ilgili analizler yapılırken Python programlama dili kullanılmıştır. Python programlama dili birçok alanda sağladığı kolaylıklar dolayısıyla seçilen nesne yönelimli, yorumlamalı, birimsel ve etkileşimli yüksek seviyeli bir programlama dilidir.

Python ile daha kolay çalışabilmek adına Anaconda kurulumu yapılmıştır. Anaconda, veri bilimi ve benzeri bilimsel uygulamalar için python kullanacak olanlara uygun hazırlanmış tümleşik bir python dağıtımıdır. Anaconda kurulumu ile birlikte python ve kullanacağımız JupyterLab ile Jupyter Notebook kurulumları tamamlanmıştır. Jupyter Notebook web tabanlı etkileşimli bir hesaplama ortamıdır. JupyterLab, yeni nesil bir kullanıcı arayüzüdür. IDE benzeri bir deneyim sunarak birden çok not defteri veya dosya ile çalışma imkanı sunar. Çalışma JupyterLab üzerinden yapılmıştır.

7.3. KULLANILAN MAKİNE ÖĞRENİMİ ALGORİTMALARI

Mevcut veri seti ile ilgili çalışmalar incelendiğinde ve veri setimizde göz önüne alındığında Naive Bayes ile Random Forest algoritmaları kullanılmıştır. İncelenen çalışmalarda bu veri seti ile daha öncesinde Random Forest algoritması çalıştırılmadığı göz önüne alındığında, verinin bu algorithmada iyi bir sonuç elde ettirebileceği düşünülerek seçilmiştir. Bununla beraber Naive Bayes algoritması da incelenmiş çalışmalarda doğrudan kullanılmadığı için seçilmiştir.

Buradaki amaç veri setimizin nasıl sonuçlar verebildiğini gözlemlemek adına daha öncesinde çalışılmamış olan algoritmalar ile modeller kurarak sonuçlar elde etmektir. Denetimli öğrenim başlığı altında yer alan iki farklı sınıflandırma algoritması kullanılarak ve mevcutta kategorik ile nümerik olan veriler kullanıma uygun hale getirilerek bu algoritmalar ile modeller oluşturulmuş ve çalışmalar tamamlanmıştır.

7.4. PROBLEMİN TANIMI VE ÇALIŞMANIN AMACI

Uluslararası piyasa ve ülke gündemleri incelendiğinde mevcut para akışı ve finans dünyasında kredi kavramı önemli bir yere sahiptir. Akış artıp, havuz büyüdükçe ve bununla beraber risk kavramı da kurum ve kişilerin gündemine daha fazla geldikçe

analiz etme ihtiyacı da doğmaktadır. Veri analizinin yoğun bir şekilde yapıldığı ve makine öğrenimi algoritmaları kullanılarak yaklaşımların ele alındığı en önemli sektörlerden biri finans sektörüdür. Atılacak adımlar ve doğru konumlanabilme ihtiyacı ile mevcutun analizini mümkün olan en iyi performansa sahip sistemler ile yapmak ve gerekli bilgilere erişerek öngörülerini ortaya çıkarmak en önemli hedef haline gelmiştir. Problemin çıkış noktası budur. Riski ortaya dökebilecek daha doğru bir tanımla veriyi doğru analiz edebilecek algoritmayı seçerek çalışmayı tamamlamaktır.

Çalışmanın amacı elimizde mevcut olan kredi verisini kullanarak daha önce kullanılmamış olan algoritmalar ile çalıştırarak daha iyi performansa sahip olabilecek olan çalışmayı yapmaktır. Model kurulumu ve veriyi tanıyarak doğru çıkarımı yapmayı sağlayacak algoritmayı seçmek temel amaçlardandır.

7.5. VERİNİN HAZIRLANMASI

UCI üzerinden elde edilen veri seti, model oluşturuluncaya kadar belli aşamalardan geçmektedir. Keşif Veri Analizi olarak temel başlık altında yer alan bu kısımda alt kırılımlar mevcuttur. Asıl amaç, veri bulunmuş ve algoritmaya karar verildikten sonra o algoritmaya uygun hale veriyi getirebilmek ve modeli sağlıklı bir şekilde kurabilmektir. Bu amaç doğrultusunda ve ön hazırlık olarak da nitelendirilebilecek bu sürecin temel başlıkları adımlar üzerinden de ele alınmaktadır. Çalışma bu akışta devam etmiştir. Adımlar gerçekleştirilirken python kodları ile birlikte JupyterLab üzerinden çalışma yapılmıştır.

Veri Ön İşleme olarak da anılan bu süreçte ele alınabilecek adımlar aşağıdaki gibidir. Burada analiz edecek kişi ihtiyacı doğrultusunda istediği adımları gerçekleştirebilir.

- I. Eksik veri doldurma
- II. Yeni öznitelikler oluşturma
- III. Kategorik değerleri dönüştürme
- IV. Veriyi ölçeklendirme ve normalize etme
- V. Aykırı verileri tespit etme

7.5.1. Verinin Çalışma Ortamına Dahil Edilmesi

Python ile yapılacak veri analizi için kullanım açısından daha efektif olan JupyterLab içerisinde çalışma sayfası oluşturulmuştur. Sonrasında aşağıda paylaşılan kodlar ile birlikte ilgili veri linkinden veriler çekilmiş ve toparlanmıştır. Bu çalışmanın gerçekleşmesini sağlayan python kodlarına EK 1 ile ulaşılabilir.

Şekil 7.1’de, veri setinden öznitelik ve değerleri ile belli sayıda kullanıcı bilgisinin olduğu kesit yer almaktadır.

Şekil 7.1: Verinin ilk hali

	Status of existing checking account	Duration in month	Credit history	Purpose	Credit amount	Savings account/bonds	Present employment since	Installment rate in percentage of disposable income	Personal status and sex	Other debtors / guarantors	Present residence since	Property	Age in years	Other installment plans	Housing	Number of existing credits at this bank	Job	Number of people being liable to provide maintenance for	Telephone	Foreign worker
0	A11	6	A34	A43	1169	A65	A75	4	A93	A101	4	A121	67	A143	A152	2	A173	1	A192	A201
1	A12	48	A32	A43	5951	A61	A73	2	A92	A101	2	A121	22	A143	A152	1	A173	1	A191	A201
2	A14	12	A34	A48	2096	A61	A74	2	A93	A101	3	A121	49	A143	A152	1	A172	2	A191	A201
3	A11	42	A32	A42	7882	A61	A74	2	A93	A103	4	A122	45	A143	A153	1	A173	2	A191	A201
4	A11	24	A33	A40	4870	A61	A73	3	A93	A101	4	A124	53	A143	A153	2	A173	2	A191	A201
...
995	A14	12	A32	A42	1736	A61	A74	3	A92	A101	4	A121	31	A143	A152	1	A172	1	A191	A201
996	A11	30	A32	A41	3857	A61	A73	4	A91	A101	4	A122	40	A143	A152	1	A174	1	A192	A201
997	A14	12	A32	A43	804	A61	A75	4	A93	A101	4	A123	38	A143	A152	1	A173	1	A191	A201
998	A11	45	A32	A43	1845	A61	A73	4	A93	A101	4	A124	23	A143	A153	1	A173	1	A192	A201
999	A12	45	A34	A41	4576	A62	A71	3	A93	A101	4	A123	27	A143	A152	1	A173	1	A191	A201

Veri seti içerisinde kategorik ve nümerik değerler yer almaktadır. Başlangıç hali yukarıda yer alan şekilde görünmektedir.

7.5.2. Veri Setinin Özellikleri

Veri setinde tanımlanmış 20 öznitelik ve 1000 adet kullanıcıya ait bilgiler yer almaktadır. Sınıflandırmaya uygun olan bu veri setimiz içerisinde kategorik ve nümerik olmak üzere çoklu değer çeşidi mevcuttur.

Tablo 7.1’de, veri setine ait özniteliklerin listesi yer almaktadır.

Tablo 7.1: Veri setine ait özellikler listesi

Öznitelikler(Attributes)
Status of existing checking account (Mevcut hesap durumu)

Duration in month (Süre (Ay cinsinden))
Credit history (Kredi geçmişi)
Purpose (Amaç)
Credit amount (Kredi tutarı)
Savings account/bonds (Mevcut hesap/bono tutarı)
Present employment since (Çalışma süresi)
Installment rate in percentage of disposable income (Gelire göre olası taksit tutarı)
Personal status and sex (Medeni durum ve cinsiyet)
Other debtors / guarantors (Diğer borçlu ve garantörler)
Present residence since (Mevcut ikametgah süresi)
Property (Mülk)
Age in years (Yaş)
Other installment plans (Diğer taksit planları)
Housing (Yerleşim)
Number of existing credits at this bank (Bu bankada yer alan mevcut kredi)
Job (Meslek)
Number of people being liable to provide maintenance for (Bakmakla yükümlü olduğu kişi sayısı)
Telephone (Telefon)
Foreign worker (Yabancı işçi mi?)

Burada özniteliklerden ele alacağımız ve ihtiyacımız doğrultusunda yeni bir öznitelik oluşturmak için kullanacağımız öznitelik ‘Kredi Geçmişi’ olacaktır. Bireylerin kredi geçmişi onların davranışları ve ileriye dönük olarak da nasıl hareket edebilir sorusuna cevap niteliğindedir. Burada özniteliğimiz kategoriktir ve her biri için belli karşılıkları vardır. Ele aldığımız ve risk özniteliğinde değerlendirileceği için değerler ayrıca incelenmelidir.

Kredi geçmişi özniteliğinin alabildiği değerler:

- A. A30: Kredi alınmadı/Tüm krediler usulüne uygun ödendi
- B. A31: Bu bankadaki tüm krediler usulüne uygun ödendi
- C. A32: Mevcut krediler şu ana kadar usulüne uygun geri ödendi
- D. A33: Geçmişte ödeme gecikmesi mevcut
- E. A34: Kritik hesap/mevcut diğer krediler (bu bankada değil)

7.5.3. Eksik Veri İnceleme Kısım

Bu aşamada amaç veri içerisinde boş değer kontrolünü sağlamaktır. Eğer göz ardı edilemeyecek kadar boş değerler mevcut ise bu durumda farklı aksiyonlar alınmaktadır. Eksik verinin modelde bize çıktısı göz önüne alındığında mümkün olduğunca yapılması gereken bir çalışmadır. Veri setimiz üzerinde aşağıda yer alan python kodları ile birlikte inceleme yapılmış ve boş değere rastlanılmamıştır.

Eksik verinin toplamda ve öznitelik bazlı olarak bulunması için kullanılan python kodu aşağıdaki gibidir.

```
df.isnull().sum().sum()
df.isnull().sum()
```

Şekil 7.2’de, veri setinde toplamda ve öznitelik bakımından boş değer sayısı gösterilmektedir.

Şekil 7.2: Veri seti geneli ve öznitelik özelinde eksik veri kontrolü sonucu

df.isnull().sum().sum()	
0	
df.isnull().sum()	
Status of existing checking account	0
Duration in month	0
Purpose	0
Credit amount	0
Savings account/bonds	0
Present employment since	0
Installment rate in percentage of disposable income	0
Personal status and sex	0
Other debtors / guarantors	0
Present residence since	0
Property	0
Age in years	0
Other installment plans	0
Housing	0
Number of existing credits at this bank	0
Job	0
Number of people being liable to provide maintenance for	0
Telephone	0
foreign worker	0
Risk	0
dtype: int64	

7.5.4.Öznitelik Oluşturma

Mevcut veriyi inceleyerek riski teşhis etmek ve değerlendirme yapabilmek adına mevcut kredi geçmişi özniteliği üzerinden yeni bir öznitelik oluşturulmuştur. Burada risk teşkil eden iki değer mevcuttur. Geçmişte ödeme gecikmesi mevcut müşteri ile kritik hesaba sahip müşteriler risk teşkil etmektedir. Aldığımız krediler ve içinde bulunduğumuz bu akışta önemli bir yere sahip olan kredi geçmişimiz bizlere birer kredi puanı kazandırır. Mevcutta gecikme yapan ya da kritik hesap olarak işaretlenmiş müşterilere tekrar kredi vermek riski göze almak demektir. Kredi geçmişi özniteliği ile birlikte bu doğrultuda aşağıda yer alan python kodu ile birlikte nümerik değer almış olan Risk özniteliği oluşturulmuştur. Sonrasında kredi geçmişi bilgisinin yer aldığı öznitelik kısmı çıkarılmıştır.

Risk özniteliği oluşturulması için kullanılan python kodu aşağıdaki gibidir.

```
conditions = [  
    (df['Credit history'] == 'A33'),  
    (df['Credit history'] == 'A34'),  
    (df['Credit history'] == 'A32'),  
    (df['Credit history'] == 'A31'),  
    (df['Credit history'] == 'A30'),  
    (df['Credit history'] == "")  
]  
values = [1, 1, 0, 0, 0, 0]  
df['Risk'] = np.select(conditions, values)  
df.drop(columns=['Data', 'Credit history'])  
print(df)
```

Burada geçmişte ödeme gecikmesi olan müşteriler ile kritik hesaba sahip müşterilere risk barındırdıkları için 1 yani True; diğer üç seçeneğe sahip olan düzenli ödeme yapan müşteriler için 0 yani False olarak değerler atanmıştır. Bu işlem sonrasında veri setimiz tekrar 20 özniteliğe sahip olarak ancak biri bu sefer risk olarak düzenlenmiştir.

Şekil 7.3, yeni öznitelik oluşturulması sonrası veri setinin son haline ait kesiti sunmaktadır.

Şekil 7.3: Verinin yeni öznitelik oluşturulması sonrası hali

	Status of existing checking account	Duration in month	Purpose	Credit amount	Savings account/bonds	Present employment since	Installment rate in percentage of disposable income	Personal status and sex	Other debtors / guarantors	Present residence since	Property	Age in years	Other installment plans	Housing	Number of existing credits at this bank	Job	Number of people being liable to provide maintenance for	Telephone	Foreign worker	Risk
0	A11	6	A43	1169	A65	A75	4	A93	A101	4	A121	67	A143	A152	2	A173	1	A192	A201	1
1	A12	48	A43	5951	A61	A73	2	A92	A101	2	A121	22	A143	A152	1	A173	1	A191	A201	0
2	A14	12	A46	2096	A61	A74	2	A93	A101	3	A121	49	A143	A152	1	A172	2	A191	A201	1
3	A11	42	A42	7882	A61	A74	2	A93	A103	4	A122	45	A143	A153	1	A173	2	A191	A201	0
4	A11	24	A40	4870	A61	A73	3	A93	A101	4	A124	53	A143	A153	2	A173	2	A191	A201	1
...
995	A14	12	A42	1736	A61	A74	3	A92	A101	4	A121	31	A143	A152	1	A172	1	A191	A201	0
996	A11	30	A41	3857	A61	A73	4	A91	A101	4	A122	40	A143	A152	1	A174	1	A192	A201	0
997	A14	12	A43	804	A61	A75	4	A93	A101	4	A123	38	A143	A152	1	A173	1	A191	A201	0
998	A11	45	A43	1845	A61	A73	4	A93	A101	4	A124	23	A143	A153	1	A173	1	A192	A201	0
999	A12	45	A41	4576	A62	A71	3	A93	A101	4	A123	27	A143	A152	1	A173	1	A191	A201	1

7.5.5.Özniteliklerin Değerlendirilmesi ve Verinin Tanınması

Bu aşamada öznitelikler değerlendirilir ve aralarında ilişki kurulması gerektiği düşünülenler bu doğrultuda değerlendirilir. Burada göz önünde canlandırmak adına öne çıkan öznitelikler kullanılarak grafikler oluşturulabilir ve bunlar üzerinden veri daha iyi tanınabilir. Bu doğrultuda çalışmamızda risk ile birlikte medeni durum ile cinsiyet ile birlikte meslek öznitelikleri için aşağıda yer alan python kodları yardımıyla müşteri dağılımları bulunmuş ve sonrasında grafikler oluşturulmuştur.

Risk, meslek ve medeni duurm ile cinsiyet özniteliklerinin değerlerinin sayısal dağılımlarını veren python kodu aşağıdaki gibidir.

```
df['Risk'].value_counts()
df['Job'].value_counts()
df['Personal status and sex '].value_counts()
```


Şekil 7.4: Risk, meslek ve medeni durum ile cinsiyet öznitelikleri değer ile sayı dağılımları

df['Risk'].value_counts()	
0	619
1	381
Name: Risk, dtype: int64	
df['Job'].value_counts()	
A173	630
A172	200
A174	148
A171	22
Name: Job, dtype: int64	
df['Personal status and sex '].value_counts()	
A93	548
A92	310
A94	92
A91	50
Name: Personal status and sex , dtype: int64	

Burada Risk özniteliğimizin değerlerine göre sayısal dağılımları mevcuttur. 619 müşteri için 0 değerine sahipken 381 müşteri için 1 olarak yer almaktadır.

Meslek özniteliğinin alabildiği değerler ve sayı karşılıkları:

- A. A171: İşsiz/vasıfsız-yerleşik olmayan, 22 müşteri
- B. A172: Vasıfsız-yerleşik, 148 müşteri
- C. A173: Vasıflı işçi/ofis çalışanı, 200 müşteri
- D. A174: Yönetim/serbest meslek/kalifiye işçi/memur, 630 müşteri

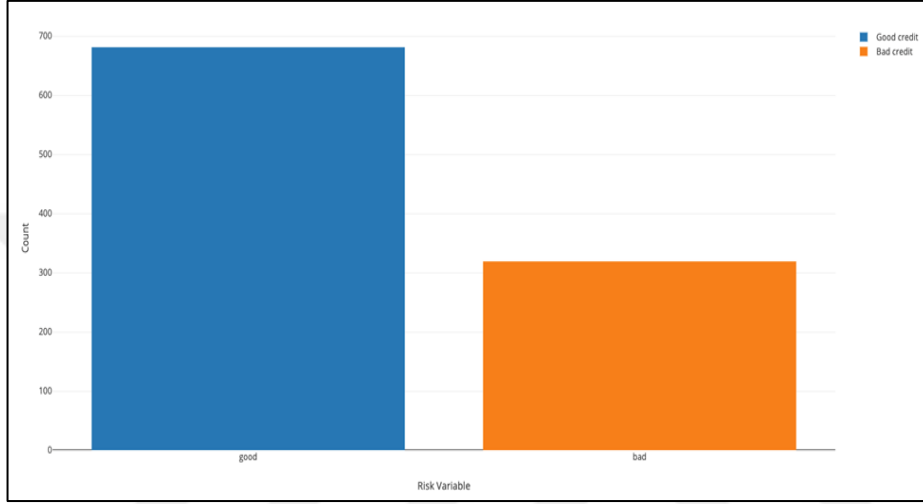
Medeni durum ile cinsiyet özniteliğinin alabildiği değerler ve ve sayı karşılıkları:

- A. A91: Erkek- boşanmış, 50 müşteri
- B. A92: Kadın-boşanmış/evli, 310 müşteri
- C. A93: Erkek-bekar, 548 müşteri
- D. A94: Erkek-evli, 92 müşteri
- E. A95:Kadın-bekar, 0 müşteri

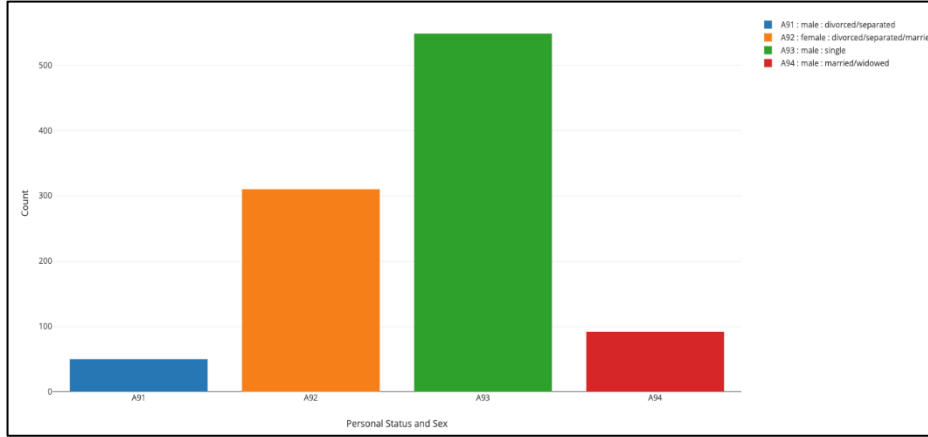
Şekil 7.4’de yer alan değerler alınarak, öznitelik değerlerine göre dağılımları gösteren grafikler oluşturulmuştur. Şekil 7.5, Risk özniteliğine ait müşteri ve değerlere göre dağılım grafiğini göstermektedir. Şekil 7.6, medeni durum ile cinsiyet özniteliğine ait dağılımı aldığı değerlere ve müşteri sayısına göre grafik şeklinde göstermektedir. Şekil

7.7, meslek ile müşteri dağılımını aldığı değerler üzerinden dağılım grafiği ile göstermektedir.

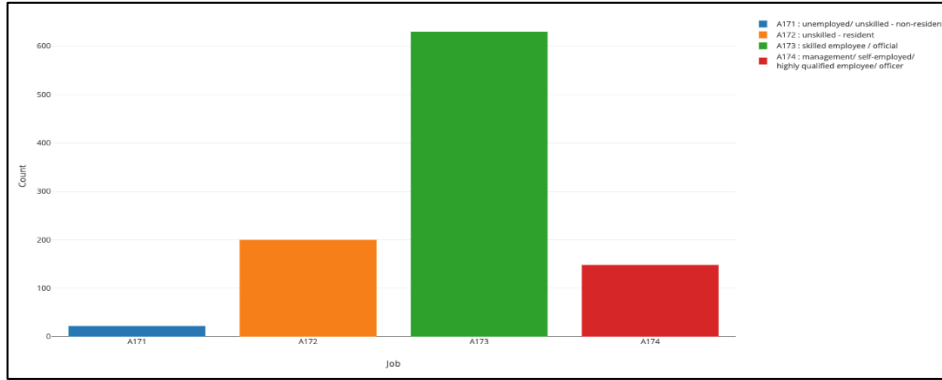
Şekil 7.5: Risk özniteliği değişkenine göre iyi kredi ile kötü kredi değerlendirmesi grafiği



Şekil 7.6: Medeni durum ile cinsiyet özniteliği değişkeni ve müşteri dağılımı grafiği



Şekil 7.7: Meslek özniteliği değişkeni ve müşteri dağılımı grafiği



7.5.6. Verileri Dönüştürme

Kategorik veriler ile çalışmak ve hesaplama yapmak bilgisayarın bu değerleri anlaması açısından zorluklar içermektedir. Makine öğrenmesi algoritmaları ile oluşturulacak olan modellerin doğru çalışabilmesi için kategorik verilerin sayısal karşılıkları yani bir diğer deyiş ile temsillerine dönüştürülmesi gerekmektedir. Bunu yapabilmek adına Sklearn kütüphanesi altında yer alan OneHotEncoder (Pandas.get_dummies) sınıfları kullanılabilir. Veri seti üzerinde de çalışma kapsamında bu sınıf kullanılmıştır. Bu yaklaşımda kategorik türde öznitelige ait tüm değerler yeni birer öznitelik haline getirilir. Her örnek asıl olarak hangi yeni öznitelige sahipse onun değeri “1”, diğer yeni özniteliklerin değerleri ise “0” olur. Burada yapılan bir diğer çalışma ise veri artık modele uygun olacağı için kendimize baz aldığımız ve kilit öznitelik olarak belirlediğimiz risk özniteliğini de verimizden ayırmak olacaktır. Bütün bu çalışmayı sağlayan python kodları EK: 2’ de yer almaktadır.

Şekil 7.8: Verinin son halinin ilk 5 satırı

	Duration in month	Credit amount	Installment rate in percentage of disposable income	Present residence since	Age in years	Number of existing credits at this bank	Number of people being liable to provide maintenance for	Checking_Status_A11	Checking_Status_A12	Checking_Status_A13	...	Housing_A152	Housing_A153	Property_A171	Property_A172	Property_A173	Property_A174	Telephone_A191
0	6	1169	4	4	67	2	1	1	0	0	...	1	0	0	0	1	0	0
1	48	5951	2	2	22	1	1	0	1	0	...	1	0	0	0	1	0	1
2	12	2096	2	3	49	1	2	0	0	0	...	1	0	0	1	0	0	1
3	42	7882	2	4	45	1	2	1	0	0	...	0	1	0	0	1	0	1
4	24	4870	3	4	53	2	2	1	0	0	...	0	1	0	0	1	0	1

7.6. MODEL KURMA VE DEĞERLENDİRME

Burada ilk olarak Random Forest algoritması kullanılarak model oluşturulmuş, sonrasında da Naive Bayes algoritması ile oluşturulacak model çalışması yapılmıştır. Bu veri seti üzerinde bu algoritmaların çalıştığı bir araştırmaya rastlanılmadığı için ilgili Random Forest ve Naive Bayes algoritmaları tercih edilmiştir.

7.6.1. Model 1: Random Forest Algoritması Kullanımı

Modele hazırlık aşamasında kullanılan bütün python kodları EK: 3 'de yer almaktadır. Burada yapılan ilk işlem test ve eğitim verilerini oluşturmak olmuştur. %80'i eğitim ve %20'si test olmak üzere dağıtım gerçekleştirilmiştir. Eğitim veri seti üzerinde model denemesi yapıldığı için veri analizi çalışmaları ve anlatımlarında belirtildiği üzere en az %60 ve en fazla %90 olacak şekilde ayrılması gerekmektedir. Bu yorum ve çalışmalar doğrultusunda modeller için eğitim datası oluşturulurken %80'lik kısmı, test için de %20'lik kısmı alınmıştır.

Eğitim ve test veri setleri dağılımı yapan python kodu aşağıdaki gibidir.

```
X_train,X_test,Y_train,Y_test = train_test_split(df1,Risk,test_size=0.20,random_state = 30)
```

Test_size kısmında görünen değer %20'nin teste ait olacağını göstermektedir. Random_state kısmında sabit bir değer olması da veri her bölündüğünde farklı yerlerden ve değerlerden bölünmesi engellenir, sabit bir değer verilmesi sayesinde aynı test verileriyle test edilmesi sağlanmaktadır.

Burada ekstra bir işlem olarak elde edilen X_train ve X_test veri kümelerine StandartScaler sınıfının verdiği yetkinlikler ile değişkenler, ortalaması 0 ve standart sapması 1 olan bir dağılıma çevrilmektedir.

```
sc=StandardScaler()
```

```
X_train_std=sc.fit_transform(X_train)
X_test_std=sc.transform(X_test)
```

Random Forest modeli için gerekli hazırlıklar yapıldıktan sonra tahminleme yapılması işlemi için aşağıda kodu yer alan işlem ile birlikte Y_test_pred veri kümesi elde edilir. Modelin oluşturulması aşamasında kullanılan python kodlarına EK:3 ile ulaşılabilir.

Şekil 7.10’da Model 1 çalıştırılması sonrası elde edilen karmaşıklık matrisi yer almaktadır. Random Forest algoritması verilen veri setini küçük parçalara ayırarak karar ağaçları oluşturur. Tahmin işlemi sırasında da karar ağaçlarındaki tahminlerin ortalamasını alır. Bu şekilde çalıştıktan sonra da aşağıda yer alan kodda da görüldüğü üzere Y_test_pred veri seti elde edilir ve burada karmaşıklık matrisinde yer alan değerleri bizlere verir.

Tahminleme işlemini gerçekleştiren ve karmaşıklık matrisini oluşturan python kodu aşağıdaki gibidir.

```
Y_test_pred = rf_random.predict(X_test_std)
confusion_matrix = confusion_matrix(Y_test, Y_test_pred)
confusion_matrix
```

Karmaşıklık matrisi değerleri elde edildiği durumda da modelin çalıştırılması ve sonucu elde edilmiş olmaktadır.

Şekil 7.9’da bir karmaşıklık matrisinde hangi alanların yer aldığı ve bu alanların nelere karşılık geldiği görülmektedir.

Burada değerlerin, her bir alanın geldiği anlamlar aşağıdaki şekildedir:

- A. TP: Algoritma çıktısı pozitif, asıl durum da pozitif
- B. FP: Algoritma çıktısı pozitif ancak asıl durum negatif
- C. FN: Algoritma çıktısı negatif ancak asıl durum pozitif

D. TN: Algoritma çıktısı negatif, asıl durum da negatif

Şekil 7.9: Karmaşıklık matrisi yapısı

		Var Olan Durum	
		Pozitif	Negatif
Tahmin	Pozitif	TP	FP
	Negatif	FN	TN

Şekil 7.10'daki değerlere Şekil 7.9'da yer alan bilgiler eşliğinde bakılacak olursa, başlangıçta %20'lik kısmı test datası olarak ayrılmıştır. Bu datalar üzerinde Risk özniteliği bazlı model çalıştırılması sonrası TP değeri 102 olarak çıkmıştır. Bu asıl durumda evet olan 102 müşteri verisi için algoritma da evet dönmüştür. TN değeri de 56'dır. Bu da asıl durumda negatif olan 56 müşteri verisi için algoritma da negatif dönmüştür. TP ve TN değerinin önemli olma sebebi algoritmanın doğruluk değeri bulunurken kullanılıyor olmasıdır. Doğruluk değeri TP ve TN değerlerinin toplamalarının teste verilmiş olan toplam veriye bölünmesi demektir. Burada $102+56/200$ işlemi bize doğruluk değerini vermektedir. Bu da 0.79 olarak karşımıza çıkar ve gayet iyi bir sonuçtur.

Şekil 7.10: Model 1 karmaşıklık matrisi değerleri

		Var Olan Durum	
		Pozitif	Negatif
Tahmin	Pozitif	102	28
	Negatif	14	56

7.6.2. Model 2: Naive Bayes Algoritması Kullanımı

Naive Bayes algoritması ile birlikte yeni bir model oluşturmak için gerekli olan python kodları EK:4'de yer almaktadır. Naive Bayes 18.yy'da Thomas Bayes'in Bayes Teoremi temel alınarak geliştirilmiştir.

Model oluşturulmasına başlanmadan önce verinin korelasyon hesabı yapılmıştır. Korelasyon iki değişken arasındaki doğrusal ilişkinin derecesini ve yönünü açıklayan bir kavramdır. Şekil 7.12'de veri setimize ait korelasyon grafiği yer almaktadır. İlişki hesaplamalarında veya ilişki analizlerinde bu kavram sıklıkla kullanılmaktadır. Yapılan analiz ve hesaplamaların nihai amacı ele alınan değişkenler arasındaki ilişkinin gücünü ve yönünü ortaya koymak için standart bir değer elde etmektir. Aşağıda yer alan python kodları çalıştırılarak veri seti korelasyon grafiği elde edilmiştir.

```
plt.figure(figsize=(14,12))
sns.heatmap(df1.astype(float).corr(),linewidths=0.1,vmax=1.0,
            square=True, linecolor='white', annot=True)
plt.show()
```

Naive Bayes modeli için GaussianNB sınıfı eklenmiştir. Bu sınıfta yer alan fit() fonksiyonu ile birlikte Naive Bayes verdiğimiz veri kümeleri üzerinde uygulanmış olacaktır. X_train ve Y_train olarak oluşturduğumuz eğitim datalarımız ile fit fonksiyonu yardımıyla model oluşturulmuş ve training score değeri elde edilmiştir. Bu işlemlere ait python kodları EK:5'de yer almaktadır. Daha sonrasında tahminleme işlemi de model.predict()fonksiyonuna X_test veri kümesinin verilmesi ile yapılmıştır ve Y_pred, tahminleme sonrası veri kümesi elde edilmiştir.

Şekil 7.11'de yer alan Model 2 karmaşıklık matrisi değerlerine, Şekil 7.9'da yer alan bilgiler eşliğinde bakılacak olursa, başlangıçta %20'lik kısmı test datası olarak ayrılmıştır. Bu datalar üzerinde Risk özneteliği bazlı model çalıştırılması sonrası TP değeri 82 olarak çıkmıştır. Bu asıl durumda evet olan 82 müşteri verisi için algoritma da evet dönmüştür. TN değeri de 49'dur. Bu da asıl durumda negatif olan 49 müşteri verisi

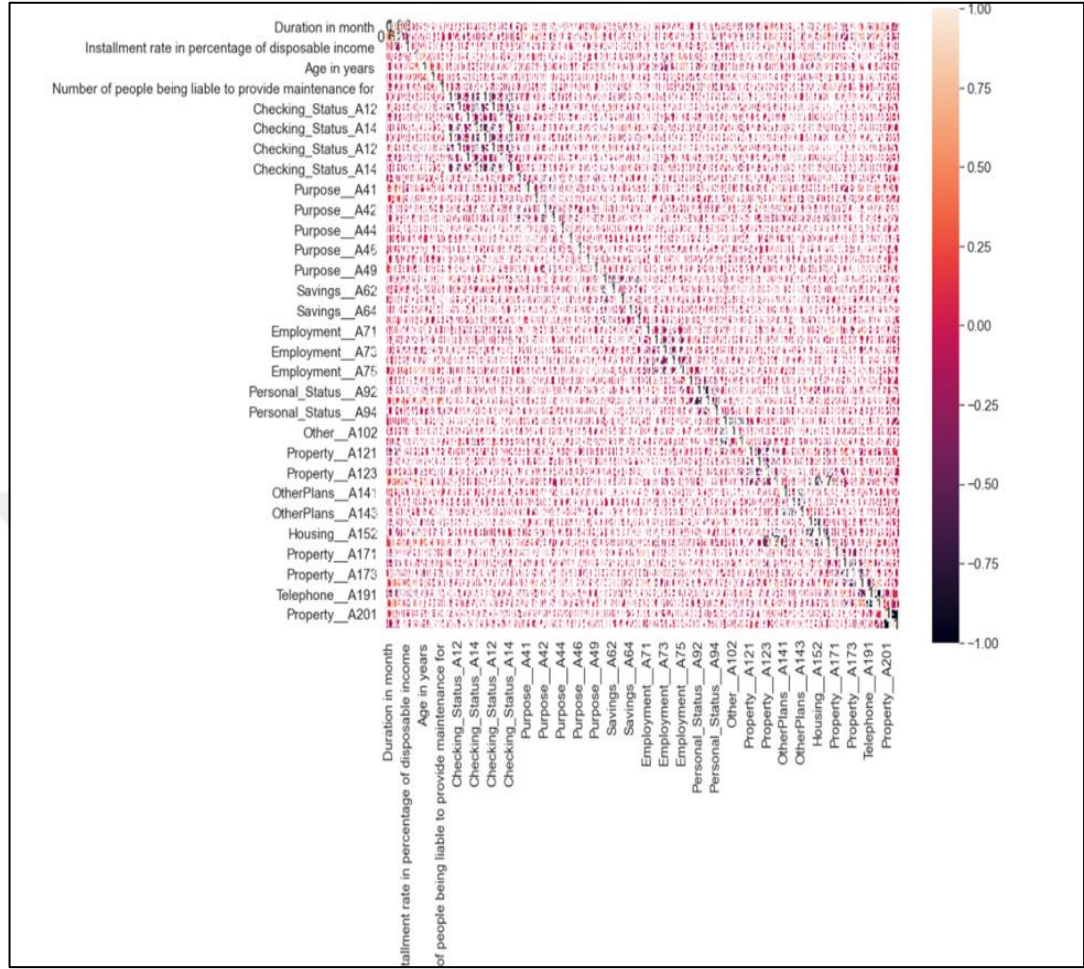
için algoritma da negatif dönmüştür. Doğruluk değeri TP ve TN değerlerinin toplamalarının teste verilmiş olan toplam veriye bölünmesi demektir. Burada $82+49/200$ işlemi bize doğruluk değerini vermektedir. Bu da 0.655 olarak karşımıza çıkmaktadır. Ortalama bir sonuçtur ve Model 1 performansının gerisinde kalmaktadır.

Şekil 7.11: Model 2 karmaşıklık matrisi değerleri

		Var Olan Durum	
		Pozitif	Negatif
Tahmin	Pozitif	82	48
	Negatif	21	49

Algoritma çalışırken ona verilen veri setindeki her bir eleman için her durum olasılığını hesaplar ve olasılık değeri en yüksek olana göre sınıflandırır. Test kümesinde yer alan bir değer eğitimi kümesinde gözlemlenemeyen bir değeri varsa olasılık değeri 0 kabul edilir ve tahmin yapılmaz. İlgili sınıflar eklendikten sonra kullanılan tahminleme fonksiyonları aldıkları veri setleri üzerinde algoritmalara ait akışları gerçekleştirerek sonuçları vermektedir.

Şekil 7.12: Veri seti korelasyon değerlendirmesi



8. MODEL UYGULAMALARI DEĞERLER VE KARŞILAŞTIRMA

Veri analizinde verinin toplanması, çalışılabilir hale getirilmesi, veriye uygun olabilecek algoritmanın seçilmesi, modelin kurulması ve çalıştırılması aşamaları sonrası elde ettiğimiz sonuçlar doğrultusunda değerlendirme yapmamız gerekmektedir. Burada modeli çalıştırarak elde ettiğimiz sonuçların birer anlamı vardır. Kimi durumda sonuçlar üzerinde iyileştirilmeler olması adına değişimler ve çalışmalar yapılabilmektedir. Kredi veri setini bu aşamalardan geçirerek iki farklı algoritma ile ayrı ayrı çalıştırılması sağlanmıştır. Yapılan çalışma sonrası iki farklı model ve onlara ait değerler elde edilmiştir. Burada karmaşıklık matrisi yanında ROC Curve değerinin olduğu grafikler de yer almaktadır. Değerlendirme kısmı ve sonuçlara geçilmeden önce karşılaştırmada kullanılacak olan terimlere değinilmiştir.

Confusion yani Karmaşıklık matrisi, veri setindeki var olan durum ile sınıflandırma modelinin doğru ve yanlış tahminlerinin sayısını tablo olarak gösterir. Makine öğrenimi modellerinde tahmin edilen ve gerçek sınıf etiketlerini göstermek amaçlı kullanılır.

Recall yani duyarlılık değeri, pozitif olarak tahmin etmemiz gereken işlemlerin ne kadarını pozitif tahmin ettiğimizi gösterir.

Precision yani kesinlik değeri pozitif tahminlenmiş değerlerin gerçekten kaç tanesinin pozitif olduğunu gösterir.

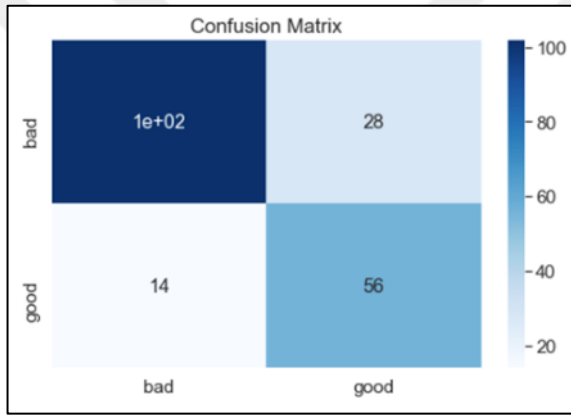
Accuracy yani doğruluk değeri modelde doğru tahmin edilen alanların toplam veri kümesine oranını hesaplamaktadır. Tek başına bir gösterge de olsa diğer metriklere de bakılması gerekmektedir. Kesin olarak doğruluğu göstermemektedir.

F1 Score değeri, kesinlik ile duyarlılık değerlerinin harmonik ortalamasını göstermektedir. Uç durumların göz ardı edilmemesi adına harmonik ortalama alınmaktadır.

ROC Curve yani ROC Eğrileri, makine öğrenmesinde performans ölçümü için önemlidir. Sınıflandırma model performans değerlendirmesinde AUC-ROC eğrisinden yararlanılır. AUC , ROC eğrisi altında kalan alanı göstermektedir. Kapsanan alan ne kadar büyükse makine öğrenimi modeli sınıfları ayırt etme konusunda daha iyidir anlamına gelmektedir.

Şekil 8.1’de Random Forest algortiması ile oluşturulmuş Model 1’e ait elde edilen karmaşıklık matrisinin grafiği yer almaktadır.

Şekil 8.1: Model 1 uygulaması sonrası elde edilen karmaşıklık matrisi grafiği



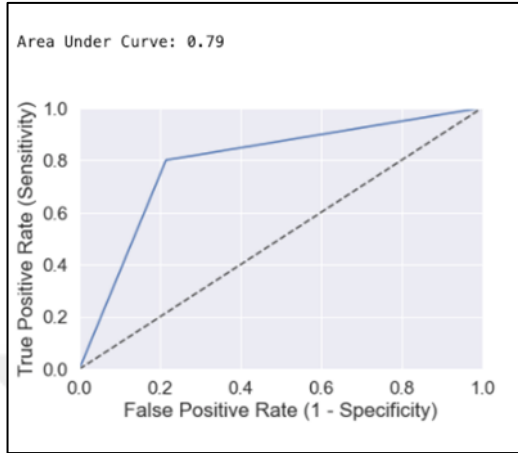
Şekil 8.2, Model 1’e ait metrik değerlerin elde edilen sonuçlarını göstermektedir. Bu değerler üzerinden performans karşılaştırmaları yapılmaktadır. Model 1’in başarı oranı 0.79 olarak tespit edilmiştir.

Şekil 8.2: Model 1 uygulaması sonrası elde edilen metrik değerler sonuçları

```
Sensitivity_recall : 0.8793103448275862  
Specificity: 0.6666666666666666  
Precision: 0.7846153846153846  
Accuracy: 0.79
```

Şekil 8.3, Model 1'e ait AUC-ROC eğrisi grafiğini göstermektedir. Burada doğruluk; kapsanan alanın büyüklüğüne göre model performans değerlendirmesi yapılmaktadır.

Şekil 8.3: Model 1 uygulaması sonrası elde edilen AUC-ROC eğrisi



Model 1, Random Forest algoritması kullanılarak oluşturulmuş ve veri seti bu modelde çalıştırıldıktan sonra ilgili metrik ve değerlerin yer aldığı, yukarıda paylaşılan sonuçlar elde edilmiştir. Burada doğruluk değeri gayet iyi bir sonuç gözükmemektedir. Bununla beraber diğer metrikleri değerlendirdiğimizde de aksi veya yanlış bir durum gözükmemektedir.

Şekil 8.4, Naive Bayes algoritması kullanılarak oluşturulmuş Model 2'nin çalıştırılması sonrası eğitim verilerinin çalıştırılması ile elde edilen sonucu göstermektedir. Modelin eğitim verileri ile verdiği başarı oranı %71 olarak tespit edilmiştir.

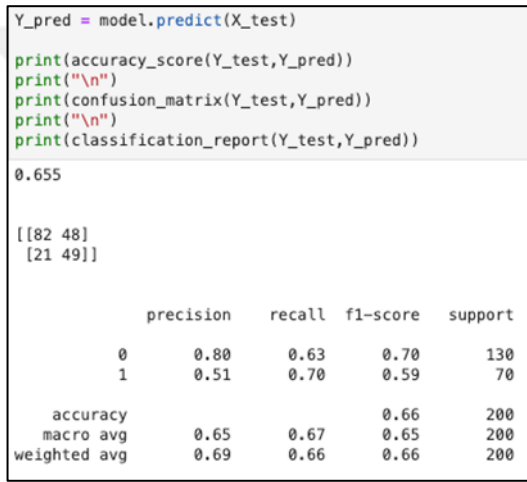
Şekil 8.4: Model 2 eğitim verileri ile uygulaması sonrası elde edilen başarı oranı değeri

```
# Printing the Training Score
print("Training score data: ")
print(model.score(X_train, Y_train))

Training score data:
0.71375
```

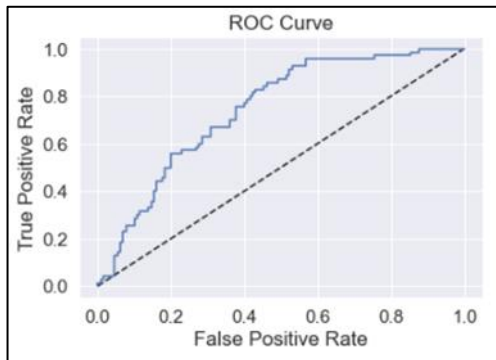
Şekil 8.5, Model 2 çalıştırılması sonrası elde edilen karmaşıklık matrisi değerleri ile karşılaştırmada kullanılacak olan metrik değer sonuçlarını göstermektedir. Şekil üzerinde işlemi yapmak için kullanılan python kodları da görülmektedir. Şekilden de anlaşılacağı gibi model “0” sınıfına ait olmayan şeyleri “0” olarak etiketlemekten kaçınmaya özen göstermektedir. Öte yandan, duyarlılık nispeten daha düşük çıkmıştır, bu da sınıflandırıcının dikkatli olması nedeniyle bir grup “0” sınıfı veriyi karıştırdığı anlamına gelmektedir. F1 değeri de bu durumu yansıtmaktadır.

Şekil 8.5: Model 2 uygulaması sonrası elde edilen karmaşıklık matrisi ve metrik değerler sonuçları



Şekil 8.6, Model 2’ye ait AUC-ROC eğrisi grafiğini içermektedir. Burada kapsanan alan Model 1’e ait olan ile kıyaslanacaktır. Bu eğri, performans değerlendirmesi yapılırken kullanılmaktadır.

Şekil 8.6: Model 2 uygulaması sonrası elde edilen AUC-ROC eğrisi



Model 2, Naive Bayes sınıflandırma algoritması kullanılarak oluşturulmuş ve veri seti burada çalıştırılmıştır. Model çalıştırılması sonrası elde edilen metrik değerler ile AUC-ROC eğrisi yukarıdaki şekillerde yer almaktadır. Burada doğruluk değerinin 0.655 olduğu görülmektedir. Bu değer Model 1’de elde ettiğimiz değerden çok daha düşüktür. Duyarlılık değerlerine bakıldığında da Model 2, Model 1’in altında kalmaktadır.

Tablo 8.1, iki modele ait metrik değerlerin sonuçlarını karşılaştırmalı olarak göstermektedir.

Tablo 8.1: Model metrik değer sonuçlarının karşılaştırılması

	Model 1	Model 2
Doğruluk	0.79	0.655
Duyarlılık	0.88	0.63
Kesinlik	0.78	0.80

Tablo 8.1’de görüldüğü üzere Random Forest algoritması kullanarak oluşturduğumuz model Naive Bayes ile oluşturulmuş modele göre daha iyi sonuçlar vermiştir. Naive Bayes algoritmasının başarısız olmasında etkili olabilecek nokta verilerin aralarında her şekilde bağımlılık olabileceğidir. Kesinlik değeri çok küçük önde de gözükse Model 1 ile aralarında çok fazla fark yoktur. Tüm bu değerler ve incelemeler göz önüne alındığında Model 1’in veri setimiz için daha iyi olduğu ve sınıflandırmayı daha iyi yaptığı söylenebilir. Random Forest algoritması kullanılarak daha önce çalışma yapılmama sebeplerinden biri bu çalışmadan yeterli derecede iyi sonucun alınamayacağı fikridir. Ancak mevcut çalışma sonucumuz gösteriyorki veri seti için gayet yüksek bir başarı ile sonuçlanmıştır. Bu çalışma sayesinde iki farklı sınıflandırma algoritması ile de başarılı sonuçlar alınabildiği ancak Random Forest’in daha da önde olduğu görülmektedir.

9. TARTIŞMA VE SONUÇ

Etkileşim ve popülasyonun artması ile birlikte ulusal ve uluslararası boyutta sektör ve piyasalar göz önüne alındığında tahminin üzerinde bir veri havuzu oluşmaktadır. Bu veri havuzu petrol gibi düşünülebilir. Büyük bir kaynak ve zenginlik anlamına gelmektedir. Olası veriler düşünüldüğünde bu ilk başta kulağa anlamsız gelebilse de mevcut veri havuzu ile elde edilebilecek bilgilerin var oluşu gerçeği bu konuda kurum ve kişileri hareketlendirmektedir.

Kişi ve kurumların elinde yer alan bu devasa veri havuzunda beklenen bilgiyi elde etmeye götüren yol veri analizidir. Veri analizi, makine öğrenmesinin sağladığı teknik ve algoritmalar ile, eldeki mevcut yöntemlerden oluşturulan yeni yaklaşımlar ve ele alış şekilleri ile birlikte giderek büyüyen ve çeşitlenen bir alandır. Sağlık, eğitim, finans gibi birçok alanda başvurulmuş ve giderek de sahiplenilerek iş akışlarına dahil edilen veri analizi, bilgi edebilme imkanı sunarak öngörülerini beraberinde getirebildiği için vazgeçilmezdir.

Her sektörde oluşabilecek riskler vardır ve riskler belli oranda tolere edilebilir. Kişi ve kurumlar olası riskleri farkedebilmek ve bu doğrultuda adım atıp konumlanabilmek için de veri analizine başvurur. Makine öğrenimi teknikleri her türlü veride sonuç elde edebilme konusunda destekleyici algoritmalara sahiptir. Bu teknikler doğrultusunda mevcut durum analizleri ile birlikte elde edilen bilgiler ile tahmin analizleri gerçekleştirilir ve olası risk hesaplamaları yapılır.

Çalışmanın çıkış noktası da finans sektöründe veri analizi önemi ve ihtiyacı noktasıdır. Kredi akışı ve artan veri havuzu hacmi ile birlikte bankalar için risk almak daha tehlikeli ve tolere edilebilir düzeyde kalmasını sağlamak önem arz etmektedir. Bu ihtiyaç ile birlikte mevcut veriler için uygun algoritmalar seçilir ve modeller oluşturularak süreç tamamlanır. Burada veri için uygun model ve makine öğrenimi algoritmasını seçmek önemlidir. Çalışmada kredi veri seti incelenmiş ve literatür taraması sonrası da daha önce kullanılmamış olan algoritmalar üzerinde durulmuştur. Veri seti, kategorik ve nümerik değer alan özneliklerden oluşmaktadır. Burada ilgili çalışmalar sonrası da

sınıflandırma tekniği kullanılarak analiz yapılması uygun görülmüştür. İncelenmiş çalışmalarda yeni yaklaşımların ele alınmasının yanında, Karar Ağaçları ve DVM tekniklerinin çoğunlukla kullanıldığı görülmüştür. İyi sonuç elde edilebileceği öngörüsü ve daha önce bu algoritmaların çalıştırılmadığı da göz önüne alındığında Random Forest ve Naive Bayes algoritmaları ile kurulan modellerde veri setinin çalışması yapılmıştır.

Çalışmada JupyterLab arayüzü ile Python programlama dili kullanılmıştır. Veri mevcut veri analizi süreçleri göz önüne alınarak ilgili aşamalardan geçirilmiştir. Kredi geçmişi özniteliği üzerinden ilgili kategorik değerler değerlendirilerek Risk özniteliği oluşturulmuştur. Geç ödenmiş ve düzenli olmayan kredi ödemelerine sahip müşteriler ile birlikte kritik hesaba sahip müşteriler risk taşıdıklarından, bu değerlendirmeye ile birlikte şartlar belirlenmiş ve öznitelik değerleri elde edilmiştir. Kategorik veriler, modellerin kullanabileceği şekle One-Hot Encoding ile getirilmiştir. Veri çalışmaya hazır hale getirildikten sonra modeller oluşturulmuştur. Bu aşamada test ve eğitim veri setleri de hazırlanmış ve sırasıyla modellerde çalıştırılmıştır.

Model çalışma sonuçları metrik değerlendirme kalemleri sonuçları karşılaştırılmıştır. Veri seti ile daha önce kullanılmamış olan Random Forest algoritması 0.79 doğruluk değeri ile birlikte gayet başarılı bir sonuca ulaştırmıştır. Ancak Naive Bayes algoritması ile oluşturulmuş model sonrasında doğruluk değeri 0.655 olarak elde edilmiştir. Diğer metrik değerler ile birlikte AUC-ROC eğrileri de kıyaslandığında Random Forest algoritması modeli performans açısından en iyi sonucu vermiştir. Bu doğrultuda finansal veri seti ile veri analizi işlemleri gerçekleştirilmiş ve başarılı model oluşturulmuştur.

KAYNAKÇA

Kitaplar

Glaser, B. G. & Strauss, A. L. (1967). The Discovery of Grounded Theory: Strategies for Qualitative Research. Chicago: Adline Publishing Company.

Maxwell, J. A. (1996). Qualitative Research Design: An Interactive Approach. California: SAGE Publications.

Strauss, A. & Corbin, J. (1990). Basics of Qualitative Research: Grounded Theory Procedures and Techniques. New Delhi: SAGE Publications.

Şeker, Ş. E. (2013). İş Zekası ve Veri Madenciliği. Cinius Yayınları.

Sürekli Yayınlar

Akpınar, H. (2000). Veri Tabanlarında Bilgi Keşfi ve Veri Madenciliği. İstanbul Üniversitesi İşletme Fakültesi Dergisi, C:29.

Ada Fu, S., Wang, K., Xu Yu, J. & C., Zhou (2003). Mining Changes of Classification by Correspondence Tracing. Proceedings of the 2003 SIAM International Conference on Data Mining (SDM). S:1-12.

Albayrak A., Görmez G., Kaynar O. & Yıldız A. (2016), Makine Öğrenmesi Yöntemleri ile Duygu Analizi. International Artificial Intelligence and Data Processing Symposium (IDAP'16). S:234-241.

Altunaydın, S. S. & Çelik Ö. (2018). A Research on Machine Learning Methods and Its Applications. Journal of Educational Technology & Online Learning, August : DOI: 10.31681.

Altunışık, R. (2015). Büyük Veri: Fırsatlar Kaynağı mı Yoksa Yeni Sorunlar Yumağı mı?. Yildiz Social Science Review, 1(1), 48.

Ayhan S. & Erdoğan Ş. (2014). Destek Vektör Makineleriyle Sınıflandırma Problemlerinin Çözümü İçin Çekirdek Fonksiyonu Seçimi. Eskişehir Osmangazi Üniversitesi İİBF Dergisi, 9(1), S:175- 198.

Barışık S., Darıcı H. & Öztürk N. (2010). Gelişmekte Olan Piyasalarda Finansal Derinleşme ve Büyüme İlişkisi: Panel Veri Analizi. ZKÜ Sosyal Bilimler Dergisi, C:6.

Bulut, F. (2019). Bankacılık Sektöründe Yeni Özniteliklerle ve Makine Öğrenmesi Yöntemleriyle Müşteri İlişkileri Yönetiminin Zenginleştirilmesi. Avrupa Bilim ve Teknoloji Dergisi, S:16.

- Dehuri, S., Jagadev, A. K., Mohapatra, S.K. & Pandey, T. N. (2017). Credit Risk Analysis using Machine Learning Classifiers. International Conference on Energy, Communication, Data Analytics and Soft Computing, S: 1848-1854.
- Ekin, O., Hammer, P. L., Kogan, A. & Winter, P. (2016). Distance-Based lassification Methods. Information Systems and Operational Research, S:337-354.
- Erim, N. & Türk A. (2005). Finansal Gelişme ve İktisadi Büyüme. Kocaeli Üniversitesi Sosyal Bilimler Enstitüsü Dergisi, C:10.
- Fradkov, A. L. (2020). Early History of Machine Learning. Science Direct, IFAC PapersOnLine 53-2.
- Griffith, J. & O'Dea, P. & O'Riordan, C. (2013). Combining Feature Selection and Neural Networks for Solving Classification Problems. National University of Ireland, Galway. S:1-9.
- Gunopulos, D. & Ratanamahatana, C. (2002). Scaling up the Naive Bayesian Classifier: Using Decision Trees for Feature Selection. University of California, S: 790-800.
- Özdemir, M. (2010). Nitel Veri Analizi: Sosyal Bilimlerde Yöntembilim Sorunsalı Üzerine Bir Çalışma. Eskişehir Osmangazi Üniversitesi Sosyal Bilimler Fakültesi, C:11.
- Topal, Ç. (2017). Alan Turing'in Toplum bilimsel Düşünü: Toplumsal Bir Düş Olarak Yapay Zeka. DTCF Dergisi, 57(2), 1350-1352.
- Yazgan K. (2017). Akrabalık Matrisinin Tersinin Henderson Metodu İle Hesaplanması ve Hesaplama Python Programlama Dili Kullanımı. Tekirdağ Ziraat Fakültesi Dergisi, C:14.

Diğer Yayınlar

Amasyalı, M. F. (2008). Yeni Makine Öğrenmesi Metotları ve İlaç Tasarımına Uygulamaları. *Doktora Tezi*. Yıldız Teknik Üniversitesi, İstanbul.

Can, Ö.Y. (2020). Makine Öğrenmesi Teknikleri Kullanılarak Kredi Risk Analizi. *Yüksek Lisans Tezi*. İstanbul Aydın Üniversitesi, İstanbul.

Çetiner, E. (2008). Classifier Performances for Credit Risk Analysis. *Yüksek Lisans Tezi*. Işık Üniversitesi, İstanbul.

Çizer, E. (2018). Makine Öğrenmesi Teknikleriyle Kredi Risk Analizi. *Yüksek Lisans Tezi*. Marmara Üniversitesi, İstanbul.

Demirtürk B., (2011). Bankalarda Kredilendirme Sürecinde İstihbarat, Nitel ve Nicel Veri Analizi ve Derecelendirme. *Yüksek Lisans Tezi*. Marmara Üniversitesi, İstanbul.

Kalaycı, S. (2018) Makine Öğrenmesi Yöntemleri ile Kredi Risk Analizi. *Yüksek Lisans Tezi*. İstanbul Teknik Üniversitesi, İstanbul.

Panjeh, M. (2018). Bilgisayar Ağlarında Makine Öğrenimi Algoritmalarını Kullanarak İzinsiz Giriş Tespiti. *Yüksek Lisans Tezi*. Atatürk Üniversitesi, Erzurum.

Yavuz, D. (2020). Using Predictive Analytics Methods for Credit Risk Analysis. *Yüksek Lisans Tezi*. Yeditepe Üniversitesi, İstanbul.

Yıldız, Ö.D. (2018). Panel Veri Analizi ile Türk Ticari Bankacılık Sektöründe Karlılık. *Yüksek Lisans Tezi*. İstanbul Ticaret Üniversitesi, İstanbul.

Yıldız, Ü. (2014). Kredi Değerlendirme Kuruluşları ve Ülke Kredi Notlarının Makroekonomik Belirleyicileri: Panel Veri Analizi. *Yüksek Lisans Tezi*. Anadolu Üniversitesi, Eskişehir.