



**T.C. İSTANBUL T CARET
 NİVERSİTESİ**

FEN BİLİMLERİ ENSTİT S 

**MAKİNE  ĞRENMESİNDEKİ SINIFLANDIRMA Y NTEMLERİNİN
KAR ILA TIRILMASI VE E-T CARET  ZERİNDE BİR UYGULAMA**

Rabia  ZDEMİR

Danışman

Prof. Dr. M nevver TURANLI

**Y KSEK LİSANS TEZİ
İSTATİSTİK ANABİLİM DALI
İSTANBUL - 2021**

KABUL VE ONAY SAYFASI

Rabia ÖZDEMİR tarafından hazırlanan "Makine Öğrenmesindeki Sınıflandırma Yöntemlerinin Karşılaştırılması Ve E-Ticaret Üzerinde Bir Uygulama" adlı tez çalışması 05/03/2021 tarihinde aşağıdaki jüri üyeleri önünde başarı ile savunularak, İstanbul Ticaret Üniversitesi Fen Bilimleri Enstitüsü **İstatistik Anabilim Dalı'nda YÜKSEK LİSANS TEZİ** olarak kabul edilmiştir.

Danışman	Prof. Dr. Münevver TURANLI İstanbul Ticaret Üniversitesi
Jüri Üyesi	Prof. Dr. Dilek ALTAŞ Marmara Üniversitesi
Jüri Üyesi	Doç. Dr. Seda BAĞDATLI KALKAN İstanbul Ticaret Üniversitesi

Onay Tarihi: 15.03.2021

İstanbul Ticaret Üniversitesi, Fen Bilimleri Enstitüsünün 15.03.2021 tarih ve 2021/308 numaralı Yönetim Kurulu Kararının 1. maddesi gereğince, ders yüklerini ve tez yükümlülüğünü yerine getirdiği belirlenen “Rabia ÖZDEMİR” (TC:29645371276) adlı öğrencinin mezun olmasına oy birliği ile karar verilmiştir.

Prof. Dr. Necip ŞİMŞEK
Enstitü Müdürü

AKADEMİK VE ETİK KURALLARA UYGUNLUK BEYANI

İstanbul Ticaret Üniversitesi, Fen Bilimleri Enstitüsü, tez yazım kurallarına uygun olarak hazırladığım bu tez çalışmada,

- tez içindeki bütün bilgi ve belgeleri akademik kurallar çerçevesinde elde ettiğimi,
- görsel, işitsel ve yazılı tüm bilgi ve sonuçları bilimsel ahlak kurallarına uygun olarak sunduğumu,
- başkalarının eserlerinden yararlanılması durumunda ilgili eserlere bilimsel normlara uygun olarak atıfta bulunduğumu,
- atıfta bulunduğum eserlerin tümünü kaynak olarak gösterdiğimi,
- kullanılan verilerde herhangi bir tahrifat yapmadığımı,
- ve bu tezin herhangi bir bölümünü bu üniversitede veya başka bir üniversitede başka bir tez çalışması olarak sunmadığımı

beyan ederim.

15.03.2021

Rabia ÖZDEMİR

İÇİNDEKİLER

İÇİNDEKİLER	i
ÖZET	iv
ABSTRACT	v
TEŞEKKÜRLER	vi
ŞEKİLLER DİZİNİ.....	vii
ÇİZELGELER DİZİNİ	viii
SİMGELER VE KISALTMALAR DİZİNİ	ix
1. GİRİŞ	1
2. LİTERATÜR ÖZETİ.....	5
3. E-TİCARET.....	8
3.1. E-Ticaret Kavramı	8
3.2. E-Ticaretin Tarihsel Gelişimi	12
3.3. E-Ticaret Türleri.....	15
3.3.1. İşletmeden işletmeye yapılan e-ticaret (Business to business-B2B).....	16
3.3.2. İşletmeden tüketiciye yapılan e-ticaret (Business to consumer-B2C)	17
3.3.3. İşletmeler ile kamu idaresi arasında yapılan e-ticaret (Business to government-B2G)	18
3.3.4. Tüketiciden tüketiciye yapılan e-ticaret (Consumer to consumer-C2C)....	18
3.3.5. Tüketiciler ile kamu idaresi arasında yapılan e-ticaret (Consumer to government-C2G)	18
3.3.6. Devletler arasında yapılan e-ticaret (Goverment to government-G2G).....	19
3.4. E-Ticaretin Avantaj ve Dezavantajları	19
4. MAKİNE ÖĞRENMESİ	22
4.1. Makine Öğrenmesi Kavramı	22
4.2. Makine Öğrenmesinin Tarihsel Gelişimi	25
4.3. Makine Öğrenmesi Yöntemleri	27
4.3.1. Denetimli öğrenme.....	27
4.3.2. Denetimsiz öğrenme.....	29
4.3.3. Yarı denetimli öğrenme.....	30
4.3.4. Takviyeli öğrenme	31

4.4.	Makine Öğrenmesi Uygulama Süreci	31
4.4.1.	Problemin tanımlanması.....	32
4.4.2.	Verilerin toplanması süreci	33
4.4.3.	Verilerin hazırlanması süreci	33
4.4.4.	Eğitim ve test verisinin belirlenmesi.....	34
4.4.5.	Makine öğrenmesi yönteminin uygulanması/model oluşturma	35
4.4.6.	Modelin yorumlanması ve doğruluğunun ölçülmesi	35
5.	MAKİNE ÖĞRENMESİ SINIFLANDIRMA YÖNTEMLERİ.....	36
5.1.	Lojistik Regresyon	36
5.1.1.	Lojistik Regresyon'un kısa tarihçesi	37
5.1.2.	Lojit dönüşüm ve Lojistik Regresyon modeli	38
5.1.3.	İkili (Binary) Lojistik Regresyon Analizi	41
5.1.4.	Çoklu Sınıflayıcı (Multinomial) Lojistik Regresyon Analizi	41
5.1.5.	Çoklu Sıralayıcı (Multiordinal) Lojistik Regresyon Analizi	42
5.2.	Naive Bayes.....	42
5.3.	Destek Vektör Makineleri	46
5.3.1.	Destek Vektör Makineleri'nin kısa tarihçesi.....	47
5.3.2.	Doğrusal Destek Vektör Makineleri	48
5.3.3.	Doğrusal Olmayan Destek Vektör Makineleri.....	52
5.4.	Rastgele Ormanlar	54
5.4.1.	Rastgele Ormanlar'ın kısa tarihçesi	56
5.4.2.	Rastgele Orman algoritması.....	56
5.5.	Sınıflandırma Model Performanslarının Değerlendirme Ölçütleri.....	58
5.5.1.	Holdout yöntemi	58
5.5.2.	Karışıklık matrisi.....	59
5.5.3.	Doğruluk oranı (Accuracy rate)	60
5.5.4.	Hata oranı (Error rate).....	61
5.5.5.	Kesinlik (Precision).....	61
5.5.6.	Duyarlılık/Hassasiyet (Sensitivity/Recall rate) oranı.....	62
5.5.7.	F1 ölçütü (F1-measure).....	62
5.5.8.	Alıcı işlem karakteristik (Receiver operating characteristic-ROC) eğrileri	63
5.5.9.	Eğri altındaki alan (AUC-area under curve)	64

6. UYGULAMA	65
6.1. Problemin Tanımlanması	65
6.2. Verinin Toplanması	66
6.3. Veri Ön-işleme/Hazırlama.....	68
6.4. Eğitim ve Test Verisinin Ayrılması.....	70
6.5. Modelin Oluşturması, Test Edilmesi ve Değerlendirilmesi	70
7. SONUÇ ve ÖNERİLER	75
KAYNAKÇA	77
ÖZGEÇMİŞ	82

ÖZET

Yüksek Lisans Tezi

MAKİNE ÖĞRENMESİ SINIFLANDIRMA YÖNTEMLERİNİN KARŞILAŞTIRILMASI VE E-TİCARET ÜZERİNDE BİR UYGULAMA

Rabia ÖZDEMİR

İstanbul Ticaret Üniversitesi
Fen Bilimleri Enstitüsü
İstatistik Anabilim Dalı

Danışman: Prof. Dr. Münevver TURANLI
2021, 82 sayfa

Bilgi teknolojilerinin gelişmesi ile internet kavramının insanların hayatına girmesi dünyadaki iş akışını değiştiren bir buluş olmuştur. Teknolojik gelişmeler ve bununla birlikte yaygınlaşan internet sayesinde büyük boyutlarda veri akışı oluşmaktadır. Bu veri akışının daha kontrol edilebilir ve akıllı bir şekilde yönetilmesi için zamanla yapay zeka, makine öğrenmesi gibi kavramlar ortaya çıkmıştır.

Günlük faaliyetlerin internet ortamına taşınması ile birlikte ticaret kavramı da elektronik ortama taşınmış ve e-ticaret kavramı oluşmuştur. E-ticaret kavramı zamanla küresel boyutta işlemlerin yapıldığı platformlar oluşturmuş ve birçok insan tarafından kullanılmaya başlanmıştır. E-ticaret ile gerçekleştirilen işlemlerin her aşamasında büyük oranda veri kaynağı oluşmaktadır. Bu verilerin doğru bir şekilde işlenmesi ve doğru kararlar alınabilmesi amacı ile makine öğrenmesi yöntemleri kullanılmaktadır. Makine öğrenmesi e-ticaretin en çok kullanıldığı alanlardandır.

Tez çalışmasında, makine öğrenmesindeki kavramlardan, tarihsel gelişimlerden ve makine öğrenmesindeki sınıflandırma yöntemlerinin nasıl çalıştıkları anlatılmaya çalışılmıştır.

E-ticaret platformlarında satın alma yapan müşteriler, tüm ürün, hizmet veya sistem sağlayıcıları tarafından büyük önem taşıyan internet kullanıcılarıdır. Bu nedenle e-ticaret platformları üzerinden satın alma işlemi yapan müşterilerin tespit edilmesi büyük önem taşımaktadır. İşletmeden tüketiciye yapılan e-ticaret ile ürün satmakta olan incikcincik.co adlı e-ticaret web sitesi verileri kullanılarak, makine öğrenmesi Lojistik Regresyon, Naïve Bayes, Destek Vektör Makineleri ve Rastgele Ormanlar sınıflandırma yöntemleri ile satın alan ve almayan müşterilere yönelik tahmin edilmiş ve hangi sınıflandırma yönteminin daha başarılı tahmin sonuçları olduğuna dair karşılaştırmalar yapılmıştır.

Anahtar Kelimeler: Destek vektör makineleri, e-ticaret, lojistik regresyon, makine öğrenmesi, naïve bayes, rastgele ormanlar, sınıflandırma yöntemleri.

ABSTRACT

M.Sc. Thesis

COMPARISON OF MACHINE LEARNING CLASSIFICATION METHODS AND AN APPLICATION ON E-COMMERCE

Rabia ÖZDEMİR

**Istanbul Commerce University
Graduate School of Applied and Natural Sciences
Department of Statistics**

**Supervisor: Prof. Dr. Münevver TURANLI
2021, 82 pages**

With the development of information technologies, the introduction of the concept of the internet into people's lives has been an invention that changes the workflow in the world. Thanks to technological developments and the widespread internet, large data flows occurs. In order to, concepts such as artificial intelligence and machine learning have emerging to manage this data flow in a more controllable and intelligent way.

Along with the movement of daily activities to the internet environment, the concept of trade was moved to electronic environment and the concept of e-commerce was formed. The concept of e-commerce has created platforms where transactions are made on a global scale over time and is used by many people. The transactions carried out with e-commerce also provide a large amount of data sources with every stage. Machine learning methods are used in order to process this data correctly and to make right decisions. Machine learning is one of the most used areas of e-commerce.

In the thesis, concepts, current sub-methods, historical developments and how machine learning classification algorithms work were explained within the scope of machine learning

Customers who make purchases on e-commerce platforms are internet users who are of great importance by all product, service or system providers. For this reason, it is of great importance to identify customers who make purchases through e-commerce platfoms. Using e-commerce web site data named incikcincik.co, which is selling products through e-commerce from business to consumer, with machine learning Logistics Regression, Naïve Bayes, Support Vector Machines and Random Forests classification methods were estimated for customers who has been purchased and non purchased. Comparisons have been made as to which classification method has more successful estimation results.

Keywords: Classification methods, e-commerce, logistic regression, machine learning, naïve bayes, random forests, support vector machines.

TEŞEKKÜRLER

Lisans ve yüksek lisans öğrenimim ve tez çalışmam boyuncaengin bilgileri ile bana yol gösteren, her türlü desteğini esirgemeyen, öğrencisi olmaktan onur duyduğum, saygı değer ve sevgi dolu kişiliği ile her zaman örnek olan kıymetli danışman hocam sayın Prof. Dr. Münevver TURANLI'ya tüm katkılarından dolayı teşekkür ve saygılarımı sunarım.

Eğitim hayatım boyunca desteklerini esirgemeyen bölüm hocalarım Prof. Dr. Ünal Halit ÖZDEN, Prof. Dr. Özlem BAŞAR ve Doç. Dr. Seda BAĞDATLI KALKAN'a teşekkürlerimi sunarım.

Tez çalışmam boyunca bana her konuda destek veren ve yardımlarını esirgemeyen kadim dostum Burçin KARAKAŞ'a teşekkürü borç bilirim.

Son olarak her zaman yanımda olan, eğitim hayatım boyunca ve her konuda beni destekleyen aileme ve yakın arkadaşlarıma teşekkürlerimi sunarım.

Rabia ÖZDEMİR
İSTANBUL, 2021

ŞEKİLLER DİZİNİ

Şekil 3.1. Dünya çapında dijital alıcılar	12
Şekil 3.2. Küresel B2B e-ticaret brüt mal hacmi 2013-2019	17
Şekil 4.1. Yapay zeka, makine öğrenmesi, derin öğrenme algoritmalarının karşılaştırılması.....	24
Şekil 4.2. Makine öğrenmesi yöntemleri	27
Şekil 4.3. Denetimli öğrenme.....	28
Şekil 4.4. Denetimsiz Öğrenme	30
Şekil 4.5. Makine öğrenmesi uygulama süreci aşamaları	32
Şekil 4.6. Girdi, model, çıktı uzayı gösterimi	35
Şekil 5.1. Lojistik Regresyon fonksiyon grafiği	40
Şekil 5.2. Veri seti içerisindeki birden fazla hiper düzlemin gösterimi	49
Şekil 5.3. Maksimal hiper düzlemi kenar boşluğu destek vektörleri gösterimi	50
Şekil 5.4. Veri setine eklenen yeni gözlemin hiper düzlem üzerinde etkisi	51
Şekil 5.5. Doğrusal olmayan SVM sınıflandırması	54
Şekil 5.6. ROC uzayı	63
Şekil 6.1. Veri setinin görünümü	68
Şekil 6.2. Korelasyon matrisi	69
Şekil 6.3. Sınıflandırma yöntemlerinin ROC eğrileri.	73

ÇİZELGELER DİZİNİ

Çizelge 3.1. Bölgelere göre dünyadaki internet kullanımı 2020 3. çeyreği tahminlerine göre	10
Çizelge 3.2. Dünya çapında toplam internet kullanımı 2000-2020	11
Çizelge 3.3. Türkiye'de İnternet Kullanımı 2000-2020	11
Çizelge 5.1. İki sınıflı karışıklık matrisi.....	59
Çizelge 6.1. Araştırma kapsamındaki değişkenler ve açıklamaları	67
Çizelge 6.2. Sınıflama yöntemlerinin karşılaştırılması	71
Çizelge 6.3. AUC değerleri	74

SİMGELER VE KISALTMALAR DİZİNİ

ACC	Accuracy
AUC	Area Under Curve
B2B	Business to Business
B2C	Business to Consumer
B2G	Business to Government
C2C	Costumer to Costumer
C2G	Consumer to Government
ÇD	Çapraz Doğrulama
DN	Doğru negatif
DP	Doğru Pozitif
EDI	Electronic Data Interchange
ERR	Error rate
E-ticaret	Elektronik Ticaret
G2G	Government to Government
GBT	Gradient Boosting Tree
IP	İnternet Protokolü
LR	Lojistik Regresyon
NB	Naive Bayes
OOB	Out-of-Bag
RF	Rastgele Ormanlar
ROC	Receiver Operating Characteristic
SVM	Destek Vektör Makineleri
USD	Amerikan Doları
VPN	Sanal Özel Ağ
YN	Yanlış Negatif
YP	Yanlış pozitif
YSA	Yapay Sinir Ağları

1. GİRİŞ

Bilgi sistemlerinin teknolojik gelişmeleri ile devamında getirdiği buluşlardan birisi de dünyadaki akışı ve işleyişini değiştiren internettir. İnternetin buluşu sonrasında dünya çapında büyük bir değişim yaşandı ve bu değişim her gün gelişen teknoloji ile birlikte sürekli olarak devamlılığını sürdürmektedir. İnternete ilişkin altyapının düzenlenmesi ile internete erişmek oldukça kısa sürede sağlanmıştır. İnternetin kullanımının artması ve yaygınlaşması birçok sektörde yeni alanların doğmasına sebep olmuş ve kişi ve kuruluşların gerçekleştirdikleri faaliyetleri sürekli gelişen çağa ayak uydurmak, varlıklarını sürdürebilmek, ulaşılabilirlikleri ve hacimlerini genişletmek için var olmaları gereken bir ortam olma zorunluluğu haline gelmiştir. İnternet ilk zamanlarda iletişim için kullanılmış olsa da ticaretin zamanla elektronik ortama taşınması konusunda da oldukça büyük bir role sahiptir.

Elektronik ticaret (e-ticaret) kısaca internet üzerinden mal ve hizmetleri içeren alım ve satım işlemlerinin yapılması anlamına gelmektedir. E-ticaret yolu ile birçok kişi ve her boyuttaki işletmeler internet erişimi olan birçok kişiye ulaşım sağlayabilmekte; bununla birlikte yeni pazarları, satış yapmak için yeni iş fırsatlarını yakalamayı, yeni hizmetler, yeni iş modelleri ile çalışma kapsamlarını genişletebilme ve küresel pazarlarda yer edinebilme gibi fırsatları elde etmektedir. Pazarlama alanı iş dünyasında çok önemli bir etkiye sahiptir ve e-ticaret ile küresel pazarlamada herkesin bu rekabetin içine girebilmesini sağlayabilmektedir. E-ticaret reklam, pazarlama, ürün, hizmet, kurumlar arasında yapılan satışlar, bankacılık işlemleri gibi birçok farklı alanda yaygın olarak kullanılmaktadır.

E-ticaretin bu kadar büyük hacimlere erişebilmesinin en büyük nedeni yeni ve genellikle uzak pazarlara veya küresel boyuttaki sektörlerle erişim sağlayabilmesi ve her gün yeni fırsatlar sunmasıdır. Gelişen her yeni teknoloji ülkelerin ekonomilerini çeşitlendirebilmesi için fırsatlar sunabilmektedir. Ekonomik faaliyetlerde dijital ortamlarda yapılan işlemlerin gelişmesi ile birlikte dijital ekonomi kavramları gelişmiş ve e-ticaret ülkelerin ekonomik göstergeleri içerisinde yer alan bir unsur haline

gelmiştir. Teknolojiyi kullanan ülkeler e-ticaret kapsamında ekonomilerini oldukça büyütmiştir. Türkiye de bu gelişen küresel ekonomide yer almakta, e-ticaret alanında girişimler yapmakta ve bu alanda hızla büyümeye devam etmektedir.

Günlük yaşamın her alanında var olan bilgisayar, tablet, cep telefonu gibi teknolojik aletler e-ticaretin kullanımı için stratejik bir önem kazandırmıştır. İnsanların iş yapma şekillerini değiştiren e-ticaret, bilginin evrimi, ulaşımı ve teknolojik gelişmeler ile ülkeler arasındaki sınırları kalkmakta ve bütün dünyadaki internet kullanıcıları tek bir ortamda buluşturabilmektedir. Bu durum birçok fiziksel açıdan zorluğu ortadan kaldırmış ve bununla birlikte çözüm bulunması gereken birçok yeni alanların ortaya çıkmasına neden olmuştur. Her teknolojik alet kullanan ve internette işlem gerçekleştiren kişi, erişim gerçekleştirdiği andan itibaren veri akışı sağlamaktadır. Hayatımızın bir parçası olan veri akışı anlık ve geleneksel yapılan işlemlerden elde edilemeyecek, kolayca hesaplanıp, işlenemeyecek ve saklanamayacak boyutta büyük ölçektedir. Bu büyük hacimdeki veriler hızlı bir şekilde artmakta ve farklı yapıda, türde verilerden oluşmaktadır.

Veri sayısı arttığı durumda mevcut durumu değerlendirmek, analiz etmek ve geleceğe yönelik tahmin yapabilmek oldukça zorlaşmaktadır. Belirli sayıda veri olsa bile yapılan yorumlar ve tahminler yanıltıcı olabilmektedir. Teknolojik gelişmeler ile makinelerin öğrenmesi ve düşünmesine yönelik işlemlerin yapılabilmesi dair fikirler ortaya atılmış ve uygulamalar yapılmıştır. Oluşturulan sistemler, mevcuttaki verileri matematiksel ve istatistiksel işlemleri kullanarak çeşitli modelleme ve analiz yöntemleri geliştirmiştir. Makine öğrenmesi, oluşturulan programlamalar ile otomatik olarak öğrenen ve bir çok alanın geliştirilmesine dair işlemler yapabilen yapay zekanın bir uygulamasıdır. Makine öğrenmesi ile verilere erişilebilir, verileri özetlenebilir, mevcut durum değerlendirmesi yapılabilir, makine öğrenmesi yöntemleri kullanarak birçok problemin yanıtını bulabilir, geleceğe dair tahmin yapılarak iş planları oluşturulabilir.

Büyük boyuttaki verileri analiz edebilmek için birçok makine öğrenmesi yöntemi bulunmaktadır. Makine öğrenmesi öğrenme şekline göre; denetimli, denetimsiz, yarı

denetimli, pekiştirmeli öğrenme olarak ayrılabilir. Veri tipleri yapılarına göre yöntem ve analizler farklılık göstermektedir. Verilerden doğru sonuçlar elde edilebilmesi için bir sınıflandırma işlemi ile ayrılması gerekmektedir. Sınıflandırma işlemlerinin doğru bir şekilde gerçekleştirilmesi için veri setindeki veriler eğitim ve test olmak üzere ayrılmakta ve eğitim verisi makineye tanıtılıp, bu veri ile bir model oluşturması beklenmektedir. Daha sonrasında test verisi bu model üzerinde denenerek doğru tahmin yapması beklenmektedir. Sınıflandırmaya işlemi denetimli makine öğrenmesi olarak isimlendirilmektedir. Makine öğrenmesi yöntemleri ile ne kadar çok sayıda veri ile çalışılır ve eğitilirse oluşturulan modelin doğru tahminler yapması durumu da artmaktadır.

E-ticaret sektöründe yapılan işlemler sürekli olarak gelişime ihtiyaç duymaktadır. Bu gelişim mevcut verileri kullanarak ve doğru analizler yaparak bu analizler sonucunda ne yapılması gerektiği nasıl bir hedef çizileceğine dair yön vermektedir. Bu yön vermelerin sağlanabilmesi için makine öğrenmesi algoritmaları kullanılmaktadır. Makine öğrenmesini kullanmak profesyonel sürecin gelişmesini kolaylaştırmaktadır. Çözüm bulunması gereken konular ile ilgili süreçlerde çeşitli algoritma ve yöntemler kullanılarak doğru sonuçlara ulaşılmaktadır.

Tez çalışmasının amacı e-ticaret platformlarında müşterilerin satın alma ve almamalarına ilişkin doğru sınıflandırma yapan makine öğrenmesi yönteminin bulunmasıdır. Bu doğrultuda öncelikle e-ticaret kavramının açıklanmasına yer verilmiştir. E-ticaretin var olmasının ve yaygınlaşmasının en büyük sebebi internetin yaygınlaşması olduğu için dünya ve Türkiye çapında internet kullanımları gösterilmiştir. E-ticaretin ne anlama geldiği, e-ticaret ile gerçekleştirilen faaliyetlere yönelik bilgiler, e-ticaretin kazandırdığı avantajlar ve bunlarla birlikte mevcuttaki dezavantajlarından bahsedilmiştir. E-ticaretin varlığı internet tarihine bağlı olmakla birlikte son yıllarda insanlar tarafından bilinmekte ve kullanılmaktadır, bu sebepten e-ticaretin tarihsel gelişimine dair süreçten bahsedilmiştir.

Makine öğrenmesi, bilginin gücünü temsil eden verinin doğru bir şekilde analiz edilmesi ve bu doğrultuda belirtilen amaca yönelik çalışmaların yapılmasına olanak sağlamaktadır. Makine öğrenmesinin ne anlama geldiği ve öğrenme şekline göre makine öğrenmesi yöntemlerinin farklılaşması durumunda makine öğrenmesi yöntemlerinden bahsedilmiştir. Makine öğreniminin fikir olarak çıkmasından itibaren günümüze kadar olan tarihsel süreci incelenmiş ve bugünkü konumu itibari ile kullanılan güncel algoritmalar çalışma kapsamında kullanılmıştır. Makine öğrenmesinin uygulanabilmesi için hangi aşamaların ve adımların izlenmesi gerektiğine dair bilgiler yer almaktadır.

Makine öğrenmesi sınıflandırma yöntemleri birden çok kategorili sınıflandırma işlemleri yapabilmektedir. E-ticaret yolu ile alışveriş yapan müşterilerin satın alıp almamalarına dair doğru tahmin yapılmaya çalışıldığı için ikili sınıflandırma yöntemleri kullanılmıştır ve bu yönde açıklamalar yapılmıştır. Bu sınıflandırma işlemlerinin gerçekleştirilmesi için denetimli makine öğrenmesi Lojistik Regresyon (LR), Naïve Bayes (NB), Destek Vektör Makineleri (Support Vector Machines, SVM) ve Rastgele Ormanlar (Random Forest, RF) yöntemleri kullanılmıştır. Her yöntemin nasıl çalıştığına dair açıklamalar yapılmış ve bu yöntemler birbirleri ile karşılaştırılarak hangisinin bu çalışma kapsamında en iyi doğruluğu veya iyi tahmin yaptığına dair sonuçlara ulaşılmıştır. Karşılaştırılma yapılırken hangisinin başarılı sonuç verdiğini tespit edebilmek için dikkat edilmesi gereken ölçüm değerlerinden bahsedilmiştir. Bu alanda çalışma yapmak isteyen kişiler için hangi analizin tercih edilmesi gerektiğine dair fikir oluşturabilecek geleceğe yönelik önerilere yer verilmiştir.

2. LİTERATÜR ÖZETİ

Literatürde, e-ticaret kavramı altında, çevrimiçi alışveriş yapan kitlelere ait bilgilerin tespit edilmesi için demografik araştırmalar yapılmış ve satın alma alışkanlıklarına yönelik tespitler yapabilmek için de anket çalışmaları uygulanmıştır. Bu alanda yapılan çalışmalar betimleyici istatistikler ile özetlenmiş, T-Testi, Anova Testi gibi çıkarımsal istatistik yöntemleri kullanılmıştır. Günümüzde bir çok şirket e-ticaret alanındaki çalışmalarını makine öğrenmesi sınıflandırma yöntemleri ile yapmaktadır fakat satın alma tahminine yönelik işlemlerin yapılabilmesi için makine öğrenmesi sınıflandırma yöntemlerinin kullanılmasına yönelik literatürde spesifik bir çalışma bulunmamaktadır. E-ticaretin önemine yönelik literatürde yapılan çalışmalardan aşağıda bahsedilmektedir.

Makine öğrenmesi kullanım alanları oldukça geniş yelpazeye sahip olduğu için literatürde bir çok konu hakkında çalışmalar yer almaktadır. Sınıflandırma yöntemleri incelendiğinde en çok çalışma yapılan alanlar kredi risk değerlendirilmeleri, metin sınıflandırma, haritalama ve hastalık teşhisleri en çok kullanılan konular olarak dikkat çekmektedir. Literatürde bir çok konuya ilişkin sınıflandırma metodlarının birbirleri ile karşılaştırılması hakkında çeşitli çalışmalar bulunmaktadır. LR, NB, SVM ve RF yöntemlerini içeren, iki sınıflı tahmin etmeye yönelik yapılan bazı güncel ve kapsamlı sınıflandırma çalışmaları aşağıda açıklanmaya çalışılmıştır.

Türkiye’de, 2015 yılında, çevrimiçi alışverişe yönelik tutum ve davranış durumunu anlayabilmek için Buğra Çelik anket yolu ile veri toplayarak çalışmasını gerçekleştirmiştir. Türkiye’nin 2015 senesindeki mevcut durumu gözden geçirildiğinde araştırmaya katılan insanların yüzde 97,4’ü internetten alışveriş yaptıklarını belirtmiş ve internetten alışveriş yaparken bilgisayar, akıllı telefon veya tabletten en az bir tanesini kullandıklarını ve her üç adet internet kullanıcısı grubundan bir kullanıcının çevrimiçi satın alma eylemini gerçekleştirdiğini belirtmiştir. Yapılan bu çalışma sayesinde Türkiye’de e-ticaretin kayda değer ölçüde mühim olduğunu vurgulamıştır.

Bagui ve diğer ekip arkadaşları (2017), ağ trafiğinin sınıflandırılmasına yönelik, sanal özel ağların (VPN), İnternet Protokolü (IP) ağlar üzerinden kullanıcılar arasında en bilinen uzaktan erişim iletişim yöntemlerinden olduğunu belirtmiş ve bu sebepten dolayı zamanla alakalı hususları kullanarak sanal özel ağ (VPN) olan ve olmayan ağ trafiğini sınıflandırmak için en doğru sonuç veren yöntemi araştırmıştır. Faaliyet gösterdikleri çalışmada makine öğrenmesi SVM, LR, NB, K-En Yakın Komşu, RF, Gradient Boosting Tree (GBT) yöntemlerini test etmiş ve araştırma neticesinde RF ve GBT modellerinin en yüksek doğruluk oranına ulaştıkları ve daha başarılı performans öngörüsünde bulunduğunu saptamışlardır.

Mık (2018), işletmeden tüketiciye yapılan e-ticaret (Business to Consumer-B2C) türünde, çevrimiçi alışveriş pazarında kullanılması önemli olabilecek bir rehber oluşturmak için, tüketicilerin çevrimiçi alışveriş davranışlarını etkileyen faktörleri istatistiksel araçlarla analiz etmiştir. Anket yolu ile veri toplamış ve araştırma neticesinde iade prosedürünün, gerekli altyapıların olması ve alana özgü inovasyonların satın alma eğilimi olan kişilerin alışveriş yapmalarına yönelik davranışlarını pozitif şekilde etkilediği gözlemlenmiştir.

2019’da Sercan Türker, zararlı Android yazılımlarının tespit edilebilmesi üzere makine öğrenmesindeki sınıflandırma yöntemlerini kullanmış ve en başarılı tahmini yapan yöntemi araştırmıştır. Türker çalışmasında birbirinden farklı veri setlerinden yararlanmış ve öncelikle AMD veri seti ile uygun veri boyutunu belirlemek için Destek Vektör Makineleri, Rastgele Ormanlar ve Lojistik Regresyon yöntemleriyle farklı boyuttaki veri setlerinin sınıflandırma işlemi sonucundaki doğruluk yüzde oranlarını incelemiştir. Veri sayısının kısıtlı olduğu veri seti ile en yüksek doğruluk payının RF yönteminin sağladığını, veri miktarı arttıkça LR’nin daha yüksek doğruluk oranı verdiğini tespit etmiştir. Genel olarak 1.000 adetlik veri setinde tüm modeller ortalama %98,18 doğruluk oranı olduğunu tespit etmiş ve çalışmasının geri kalanında 1.000 adetten oluşan veri setini kullanmıştır. Daha sonrasında K-En Yakın Komşu (KNN), Karar Ağaçları, Lojistik Regresyon, Destek Vektör Makineleri, Rastgele Orman ve Çok Katmanlı Algılayıcı makine öğrenmesi yöntemlerini birbirinden farklı veri setlerini kullanarak

sınıflandırma işlemi yapmıştır. Türker yapmış olduğu çalışma sonucunda Destek Vektör Makineleri ve AdaBoost yöntemlerinin en yüksek doğruluk oranına sahip olduğunu tespit etmiştir.

2020 yılında Elif Çelik, telekomünikasyon sektöründeki mevcut müşterilerin satın aldıkları hizmetten vazgeçeceğini önceden tahmin etmek ve böylece müşterinin elde tutulması için çalışmalar yapılabilmesi amacı ile müşteri kayıp analizi üzerinde çalışma yapmıştır. Müşterilerin ayrılıp ayrılmayacağını makine öğrenmesi Lojistik Regresyon, K-En Yakın Komşu (KNN), Naive Bayes, Rastgele Ormanlar, Karar Ağaçları, Destek Vektör Makineleri ve Gradient Boosting Tree sınıflandırma yöntemlerini kullanarak en doğru tahmin yapan modellerin RF ve KNN yöntemleri olduğunu tespit etmiştir.

Alan (2020), Kaggle’da bulunan 32 farklı veri seti ile bağımlı değişkeni iki ve ikiden fazla kategorik olan veri setleri için sınıflandırma yöntemleri uygulamış ve en iyi doğruluk oranı veren yöntemleri tespit etmiştir. Alan makine öğrenmesindeki Naive Bayes, Rasgele Ormanlar, Karar Ağaçları, Lojistik Regresyon, Yapay Sinir Ağları (YSA), Destek Vektör Makineleri, AdaBoost, Gradyan Artırma sınıflandırma yöntemlerini kullanmıştır. 20 veri setinde houldout yönteminin çapraz doğrulama (ÇD) yöntemine göre daha iyi sonuç verdiğini gözlemlemiş ÇD kullanılabileceği durumlarda Hold-out yöntemine gerek duyulmadığını fakat ÇD kullanıldığında sınıflandırma çok fazla maliyet gerektiriyorsa ve veri sayısının yetersiz olduğu durumlarda Hold-out yönteminin kullanılabileceği tespit etmiştir. İki sınıflı veri setlerinde RF ve YSA başarılı sonuçlar vermiş olsa da, Alan sınıflandırma sürecinde, sınıflandırma için veri setinin dengeli olup olmama durumu, kullanılan test yönteminin hangisi olduğu büyük önem taşıdığını belirtmiştir.

Tez çalışması kapsamında yöntemlere ve kavramlara ilişkin yapılan ilk çalışmalar, nasıl ve kimler tarafından ortaya çıktıkları ilgili bölümlere ait başlıklar altında incelenmiş ve aktarılmaya çalışılmıştır.

3. E-TİCARET

Tez çalışmasının bu bölümünde e-ticaret kavramı, e-ticaretin tarihsel gelişimi ve e-ticaret türleri detaylı olarak anlatılmaya çalışılmıştır.

3.1. E-Ticaret Kavramı

Yerel bölgelerde alım-satım, üretim-tüketim yaparak başlayan ticaret sektörü, denizcilik alanında yapılan buluşlar sonrasında yeni deniz yollarının ve farklı yerlerin keşfedilmesi ile mevcut ağından daha geniş bir alana ve hacme ulaşmaya başlamıştır. Bu gelişme, kaynaklarda ticaretin tarihte ilk geniş kapsamlarına ulaşmaya başlaması olarak geçmektedir.

Bilgi teknolojilerinin sürekli gelişmesi ve sonrasında internetin bulunuşu hayatımızda önemli bir yer edinmektedir. İnternetin bu denli hızlı ve köklü yaygınlaşması, toplumsal yaşam kalitesini arttırmış ve insanların günlük hayatta gerçekleştirdikleri çoğu yaşamsal faaliyetini (seyahat, reklam, tanıtım, ürün, hizmet, üretim vb.) internet ortamına taşınmasına neden olmuştur. Ticaret de zamanla internet ortamına taşınan bu kavramlar arasına girmiştir ve böylece “Elektronik Ticaret” adını almıştır. Bilgisayar bilimlerinin bulunuşu ve gelişmesi sayesinde ticaretin elektronik ortama taşınması sağlanmış, bu sayede dünya çapında insanların ürün ve hizmetlere daha hızlı ulaştığı bir döneme geçiş yapılmıştır.

Günümüzde internet, e-ticaret sayesinde oldukça dikkat çekmekte, bu yeni alışveriş tarzı her geçen gün insanlar tarafından oldukça tercih edilebilir bir seçenek olmaya devam etmektedir. Büyük bir hızla her geçen gün gelişimler gösteren çevrimiçi alışveriş, işletmelere ve pazarlama faaliyetlerine de yeni boyutlar kazandırmaktadır (Turan, 2008). Dünya genelinde geleneksel satış yapan firmalar, gelişen sektöre ve internet ortamına ayak uydurmadıkları veya değişime direndikleri sürece yok olma ile karşı karşıya kalmaktadır. Faaliyetlerine e-ticareti ekleyen firmalar sadece bulundukları bölgelere değil, ülke veya dünya çapında birçok tüketiciye ulaşabilme fırsatı bulabilmektedir.

Birçok teknolojik ürünler; bilgisayarlar, akıllı telefonlar, tabletler ve yeni üretilen akıllı aletler ile e-ticaret yapılabilir. Kişiler, her sektörde faaliyet gösteren küçük, orta veya büyük çaplı işletmeler, kurum ve/veya kuruluşlar mal ve hizmetlerini doğrudan tüketiciye satabilir. Kurslar, e-kitaplar, e-dergiler, yazılım ile ilgili veya medya içerikli indirilebilir içerikler, toplu olarak satılan ürünler, danışmanlık hizmetleri (koçluk, mentorluk vb.) pazarlama hizmetleri, çevrimiçi abonelikler gibi birçok farklı alan e-ticareti kapsamaktadır. Kullanıcılar kendi taleplerine göre birden fazla seçeneğe aynı anda ulaşım sağlayabilir, seçenekleri birbirleri ile karşılaştırabilir ve kendileri için en uygun olduğunu düşündükleri seçeneğe hızlı ve özgür şekilde karar verebilir.

Mal ve hizmetlerin küresel boyutta, üretilmesinin ve dağıtılmasının bilgi ve iletişim teknolojileri aracılığıyla yapılmasını kapsayan ve sürekli gelişen işletme yaklaşımının bir parçası olan e-ticaret; temelde yepyeni bir işleyiş, değişim ve dönüşümü gerçekleştiren dinamik bir sistemi ifade etmektedir. Genel çerçeve itibari ile bakıldığında bu dönüşüm, kişiler ve her boyuttaki işletmelerin iş hedefleri doğrultusunda, süreçlerini elektronik ortama taşımalarını, faaliyetlerini maliyet tasarrufu yaparak daha etkin ve verimli bir şekilde dönüştürmesini kapsamaktadır. E-ticaret dönüşümü yeni iş yapış şekli ile gerekli kültürel değişimi gerçekleştirmek üzere izlenmesi gereken yolların takip edilmesini sağlamaktadır. Bu sayede e-ticaret, yapısı gereği müşteri odaklı oluşturulmuş ve günümüzün teknolojilerini kullanarak müşterilerin ihtiyaç ve beklentilerine en hızlı yanıt verebilecek şekilde süreç odaklı bir sanal sistem olarak tasarlanmış, bu sebepten işletmelerin iş süreçlerinin bu yapıya uydurulması neredeyse zorunlu hale gelmiştir (Kozak, 2007).

E-ticaretin bu kadar yaygın kullanılması veya tercih edilmesinin sebebi yukarıda anlatıldığı gibi internet kullanımının hızlı ve ulaşılması kolay olmasından kaynaklanmaktadır. İnternetin daha yaygın ve hızlı olabilmesi için çalışmalar sürekli olarak devam etmektedir. World Stats internet sitesi; farklı güvenilir kaynaklar aracılığı ile kapsamlı bir şekilde dünya çapında internet kullanımını 2000 yılını temel alarak göstermektedir.

Çizelge 3.1. Bölgelere göre dünyadaki internet kullanımı 2020 3. çeyreği tahminlerine göre (World Stats, 2021)

Bölgelere Göre Dünyadaki İnternet Kullanımı 2020 3. Çeyreği Tahminlerine Göre						
Dünyadaki Bölgeler	Nüfus (2020 tahmini)	Nüfus (% Dünya)	İnternet Kullanıcıları	İnternet kullanıcılarının nüfusa oranı (%)	Büyüme	Dünya Çağında İnternet (%)
Afrika	1.340.598.447	17.2 %	631.940.772	47.1 %	13,898%	12.8 %
Asya	4.294.516.659	55.1 %	2.555.636.255	59.5 %	2,136%	51.8 %
Avrupa	834.995.197	10.7 %	727.848.547	87.2 %	593,000%	14.8 %
Latin Amerika/Karayipler	654.287.232	8.4 %	467.817.332	71.5 %	2,489%	9.5 %
Orta Doğu	260.991.690	3.3 %	184.856.813	70.8 %	5,527%	3.7 %
Kuzey Amerika	368.869.647	4.7 %	332.908,87	90.3 %	208,000%	6.8 %
Avusturalya	42.690.838	0.5 %	28.917.600	67.7 %	279,000%	0.6 %
Toplam	7.796.949.710	100 %	4,929,926,187	63.2 %	1,266%	100%

Çizelge 3.1’de görüldüğü üzere dünya üzerindeki bölgelerin internet kullanımlarına bakıldığında; nüfusa göre en çok internet kullanımının Kuzey Amerika bölgesinde %90,3 oranında olduğu görülmektedir fakat Kuzey Amerika bölgesi dünya çapındaki internet kullanımının sadece %6,8’ini oluşturmaktadır. Nüfusu en çok internet kullanan ikinci bölge %87,2 oranı ile Avrupa bölgesidir. Avrupa bölgesi dünya çapındaki internet kullanımının %14,8’ini oluşturmaktadır. Asya bölgesindeki insanların %59,5’i internet kullanmakta ve %51,8 oranı ile bu bölge dünya çapında en çok internet kullanımının olduğu bölgedir. Afrika bölgesinde ise nüfusun %47,1’i internet kullanmakta ve bu nüfus dünya çapındaki internet kullanımının %12,8’ini oluşturmaktadır. Çizelge 3.2’de dünya çapında internet kullanıcıları gösterilmiştir.

Çizelge 3.2. Dünya çapında toplam internet kullanımı 2000-2020 (World Stats, 2021)

DÜNYA ÇAPINDA İNTERNET KULLANICILARI				
İnternet Kullanıcıları		Nüfus (Tahmini)		Büyüme
2020 Q1	2000 Q4	2020	2000	2000 - 2020
4.574.150.134	360.985.492	7.796.615.710	6.145.006.989	1267,00%

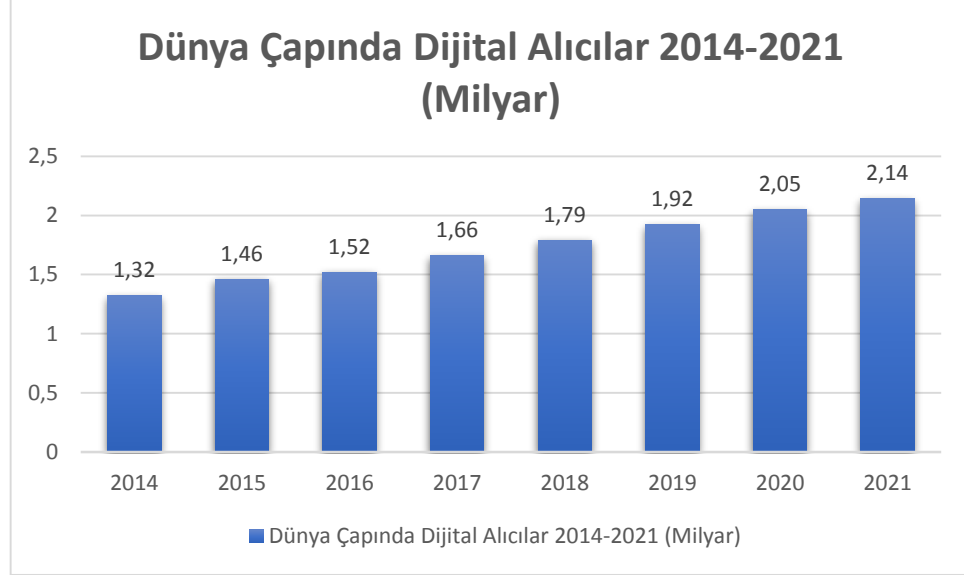
Çizelge 3.2’de görüldüğü gibi; 2000 yılında 6 milyarlık dünya nüfusunun sadece 360 milyonu internet kullanırken 2020 yılında bu kullanım alanı dünya çapında genişlemiş ve 8 milyara yaklaşan dünya nüfusunun 4.5 milyarı artık interneti hayatlarına dahil etmişlerdir. Bu 20 yıllık süreçte internet kullanımı %1267 oranında devasa bir büyüme göstermiştir.

Çizelge 3.3. Türkiye’de İnternet Kullanımı 2000-2020 (World Stats, 2021)

TÜRKİYE’DE İNTERNET KULLANIMI				
İnternet Kullanıcıları		Nüfus (Tahmini)		Büyüme
2020 Q1	2000 Q4	2020	2000	2000 - 2020
69.107.183	2.000.000	84.339.067	63.240.121	3455,00%

Türkiye, dünya çapında en çok internetin kullanıldığı ilk 20 ülke arasında 13. sırada yer almaktadır. Çizelge 3.3’de görüldüğü gibi 2000 yılında Türkiye nüfusunun %3,6’sı

internet kullanıyor iken, 2020’de bu oran nüfusun %81,9’u olarak büyük bir artış göstermiştir. Türkiye yaygınlaşan internet altyapısı ile 2000-2020 yılları içerisinde internet kullanımı %3455’lik bir büyüme ile gerçekleştirmiştir.



Şekil 3.1. Dünya çapında dijital alıcılar (Statista, 2021)

Diğer taraftan 2020’de internet üzerinden alışveriş yapan kişi sayısı Şekil 3.1’de de görüldüğü gibi dünyada 2.05 milyar insana ulaşmıştır. İnternet kullanıcılarının %41’i ve dünya nüfusunun %26’sı internet üzerinden satın alma yapmaktadır. Dünya çapında internet kullanımı ve satın alma işlemleri sürekli olarak doğrusal bir artış sergilemektedir.

3.2. E-Ticaretin Tarihsel Gelişimi

E-ticaretin kökeni, müşterilere yönelik olarak, müşteri davranışlarına göre gelişen, değişen, yeni, sanal bir dünyanın tarihi oluşturmaktadır. E-ticaret asıl varlığını 1995 sonrasında göstermeye başlamıştır fakat 1995 öncesinde bu fikrin ve e-ticaretin temelleri atılmıştır. Bu sebepten e-ticaret tarihini 1995 öncesi ve sonrası olarak ayırmak mümkündür.

1995 öncesinde yapılan çalışmalara bakıldığında; Jamie Bartlett (2015) The Dark Net adlı kitabında, John Markoff (2005) yayınladığı What the Dormouse Said: How the Sixties Counterculture Shaped the Personal Computer Industry kitabında 1970'den önce, çevrimiçi olarak yapılan ilk işlemin bir uyuşturucu anlaşması satışı olduğunu söylemektedir. Kaliforniya'daki Stanford Üniversitesi ve Massachusetts'teki (MIT-Massachusetts Institute of Technology) öğrenciler yapay zeka laboratuvarlarındaki, dünyanın öncüsü olan, ilk internet ağı Arpanet hesabını kullanarak marihuana satışı gerçekleştirmiştir. Daha sonrasında resmi olarak ilk bilinen ve yayınlanan bilgi; 1979 yılında Peter Champion ve Michael Aldrich adlı iki bilgi teknolojisi mühendisi bir bilgisayarı televizyona bağlayarak çevrimiçi alışverişin doğuşuna öncülük etmişlerdir (Coleman & Ganong, 2014). 1982'de France Telecom tarafından Minitel adında özel bir ağ oluşturulmuş ve kullanıcıların telefon numaralarını aramasına, seyahat rezervasyonlarını ayarlamasına, finansal işlemler yapmasına ve çevrimiçi alışveriş yapmasına olanak sağlanmıştır (Cornelius vd., 2002). Bir takım kaynaklara göre bu işlem ilk elektronik alışveriş olarak geçmektedir (Çelik B. , 2015). Bahsedilen bu ilk gelişmelerin ardından, e-ticaretin kronolojik ilerleyişi aşağıda sunulduğu gibidir (Uygur, 2010).

1960'lı yılların ortalarında nakliye ve bazı perakende firmaları elektronik veri değişimi (electronic data interchange-EDI) ile kağıtsız ofisler kurmaya çalışmıştır. EDI 1970'li yıllarda benimsenmeye başlanmıştır (Tian ve Stewart, 2008).

1979 yılında İngiltere'de Michael Aldrich gerçek zamanlı olarak, çok kullanıcıli işlemleri işlemek için bir bilgisayarı telefon hattı üzerinden değiştirilmiş bir ev televizyon setine bağlamıştır. Sistem 1980 yılından itibaren pazarlanan ve İngiltere, İrlanda ve İspanya'da satılan ağırlıklı olarak işletmeler arası e-ticaret sistemleri sunmuştur.

1982 yılında kullanılmış bilgisayarları satmak için Boston Bilgisayar Borsası ilk çevrimiçi platformu oluşturmuş ve bu ilk e-ticaret şirketi olmuştur. 1982 yılında sadece

Fransa’da kullanılan Minitel, France Telecom tarafından çevrimiçi sipariş için kullanıma sunulmuştur.

1984 yılında CompuServe tarafından Electronic Mall adında elektronik alışveriş platformunu internet erişimi olan herkese açmıştır. En büyük özelliği web tabanlı mağazaların olması, ürün demoları ve fotoğraflarının kullanıldığı ilk platform olmasıdır.

1986’da akademik amaçlar için kullanılması amaçlı NSFNET ağı sınırlı bir şekilde kullanılmaya başlanmıştır. ARPANET’ten sonra kurulan ilk ağ özelliğini taşımaktadır. 1990 yılında, Berner Lee tarafından NETcomputer kullanılarak ilk web sunucusu ve web tarayıcısı World Wide Web’de icat edilmiştir.

1994 yılında Netscape, Ekim ayında Mozilla kod ismiyle Navigator web tarayıcısını kullanıma sunmuştur. Diğer bir gelişme olarak Pizza Hut kurumsal web sitesinden siparişler alma işlemini uygulamaya koymuştur. Aynı yıl içerisinde ilk çevrimiçi banka faaliyete girmiş ve dergi abonelikleri, çevrimiçi çiçek satışı gibi e-ticaret girişimlerine başlanmıştır. Bisiklet ve otomobil gibi ürünler çevrimiçi alışverişe çıkarılmıştır. Senenin sonlarına doğru ise, e-ticarete güvenlik hususunun önemli bir nokta olduğu görülmüş ve SSL şifreleme ile güvenliği seviyesi yükseltilmiş, Netscape 1.0’in tanıtımı yapılmıştır.

1995 yılında Jeff Bezos, amazon.com’u kurmuştur. 24 saat kesintisiz yayın gerçekleştiren, tanıtım faaliyeti olmayan ilk internet radyo istasyonları oluşturulmuş, NetRadio ve Radyo HK ismiyle yayın yapmaya başlanmıştır. Dell ve Cisco ticari faaliyetler için internet kullanımına girişmiştir. eBay bilgisayar programcısı Pierre Omidyar tarafından AuctionWeb olarak kurulmuştur.

1995 yılında bu önemli gelişmelerden sonra, internetin yayılması ile birlikte e-ticaretin artık herkes tarafından bilinmeye ve kullanılmaya başlandığı bir döneme girilmiştir. Kullanılan platformlar üzerinden bir pazar gelişmeye başlamıştır. 1995 yılı sonrasında Hindistan ve Kore kendi platformlarını kuran ülkelerden olmuş, internet üzerinden reklam verme işlemleri başlamıştır. 1998 yılında e-ticaret üzerinden ödeme yapılması

için bir sistem kurulmuştur. 1999 yılında dünyanın en büyük platformlarından olan Alibaba kurulmuştur. 2000 yılında internet şirketlerinin hisselerinin borsada yerle bir olması ile büyük çaplı bir dot-com balonu denilen bir kriz yaşanmıştır. 2003 yılında amazon.com karını ilk kez açıklayan firma olmuştur. Sonraki süreçlerde birçok platform oluşturulmuş, ödeme sistemleri genişletilmiş, birçok alanın ve sektörün elektronik ortamda varlıklarını oluşturması kaçınılmaz bir gereklilik haline gelmiştir.

Türkiye’de de e-ticaret, çok hızlı büyüyen sektörlerden birisidir. İnternette alışveriş faaliyetlerinin Türkiye’de öncelikli olarak 1992 yılında, Merkez Bankası ile diğer bankaları ortak bir platformda buluşturan Elektronik Fon Transferi uygulamasıyla başladığı bilinmektedir (Eren, 2009).

1998 yılında ilk tohumları atılan hepsiburada.com (Erişim Tarihi: 01.01.2021) 2000 yılında faaliyete girmiştir. 2001 yılında gittigidiyor.com (Erişim Tarihi: 01.01.2021) Türkiye’nin ilk açık artırma pazarı olarak sektöre giriş yapmıştır. Bugün mağaza ağı ile yaygın olan Teknosa (Erişim Tarihi: 01.01.2021) 2000 yılında harekete geçmiş, 2005 yılında ise teknosa.com ile faaliyete girmiş ve Türkiye'nin her yerine hizmet vermeye başlamıştır. Bu süreci 2010 yılında kurulan Trendyol ve 2012 yılında kurulan n11.com ve birçok e-ticaret sitesi daha zamanla açılarak bu tarihsel süreçte hızla yeni atılımlar yapmıştır.

3.3. E-Ticaret Türleri

E-ticaret, elektronik ortamda alıcı ve satıcıların bir araya getirildiği platformlar üzerinden gerçekleştirilmektedir. E-ticaretin bu özelliği sayesinde elektronik ortamdaki ticaretin kapsamı ulusal düzeyden uluslararası düzeye çıkmıştır. Sadece katılımcıları dikkate alınarak e-ticareti altı gruba ayırmak mümkündür (Kalaycı, 2008).

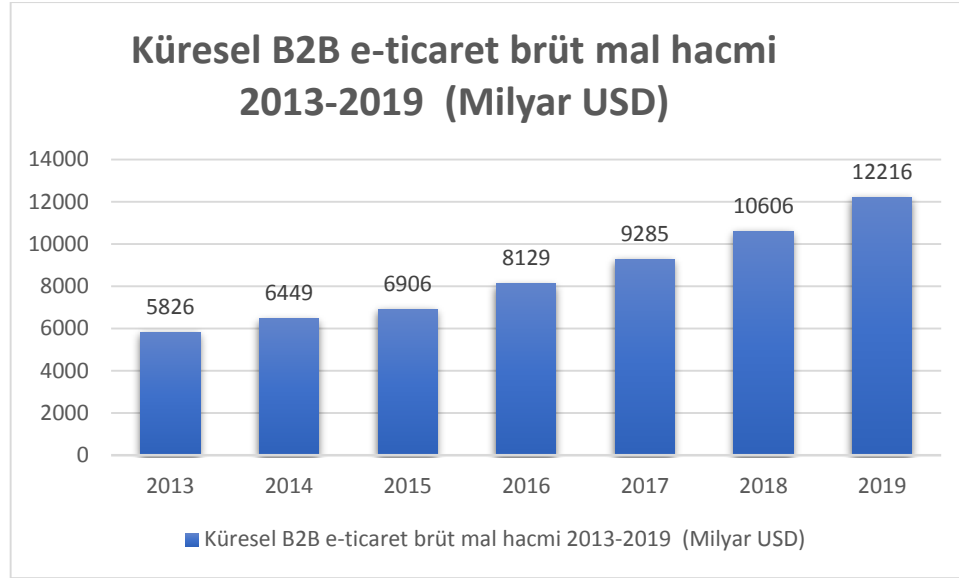
E-ticaretin kimler arasında yapıldığına bağlı olarak türleri belirlenmektedir. E-ticaret devlet, işletme ve tüketici olarak bu üç ana kavramlar ile birbirleri arasındaki ticaretler doğrultusunda aşağıda belirtildiği gibi türlere ayrılmaktadır:

- İşletmeden işletmeye yapılan e-ticaret (Business to Business-B2B)
- İşletmeden tüketiciye yapılan e-ticaret (Business to Consumer-B2C)
- İşletmeler ile kamu idaresi arasında yapılan e-ticaret (Business to Government B2G)
- Tüketiciden tüketiciye yapılan e-ticaret (Costumer to Costumer-C2C)
- Tüketiciler ile kamu idaresi arasında yapılan e-ticaret (Consumer to Government-C2G)
- Devletler arasında yapılan e-Ticaret (Goverment to Government-G2G)

Bunlardan en yaygın ve en çok kullanılan e-ticaret türleri; işletmeden işletmeye yapılan e-ticaret, B2B (Business to Business), işletmeden tüketiciye yapılan e-ticaret, B2C (Business to Consumer), tüketiciden tüketiciye yapılan e-ticaret, C2C (Consumer to Consumer)'dir.

3.3.1. İşletmeden işletmeye yapılan e-ticaret (Business to business-B2B)

B2B e-ticaret, otomasyon sistemi bulunan şirketlerin birlikte ticaret yaptığı diğer şirketlere (tedarikçi şirketler, bayiler, üretim işlemini yapan şirket, departmanlar, mağazalar vb.) entegre olması ile elde edilen ürünler, bilgi ve hizmetlerin şirketler arasında kullanımı, paylaşımı ve satışlarıdır. B2B e-ticaret, şirketlerin çağımız teknolojisini en iyi şekilde kullanarak sanal ağlar üzerinden tedarikçi şirketlere sipariş geçmesi veya tedarikçisi olarak görev yaptığı şirketlerden sipariş alması, faturalandırdığı meblağların tahsili ve kendisine fatura edilen tutarlara ödeme yapması gibi uygulamalardan oluşmaktadır. Daha önce yapılan araştırmalar doğrultusunda da yapılan e-ticaretler genellemesinde şirketler arası yapılan B2B e-ticaretin oranı, diğer e-ticaret türlerinden daha fazladır ve e-ticaret hacminin büyük kısmını oluşturmaktadır (Marangoz, 2011).



Şekil 3.2. Küresel B2B e-ticaret brüt mal hacmi 2013-2019 (Statista, 2021)

Şekil 3.2’de görüldüğü üzere dünya çapında işletmeden işletmeye yapılan e-ticaret hacmi 2019 yılında 12.2 trilyon Amerikan Doları (USD) olarak kayıtlara geçmiştir. İşletmeler arası e-ticaret, toplam e-ticaret kısmının büyük bir kısmını oluşturmaktadır. İşletmeden tüketiciye yapılan e-ticaret göz önünde olsa da işletmelerin kendi aralarında toplu olarak yapılan işlemlerin parasal boyutlarının büyük olmasından dolayı büyük hacimde yer kaplamaktadırlar.

3.3.2. İşletmeden tüketiciye yapılan e-ticaret (Business to consumer-B2C)

Şirketlerin doğrudan müşteriye e-ticaret platformları aracılığı ile satış yapma şeklidir. Şirketler ürünleri hakkındaki tüm bilgileri, ürün sözleşmelerini, ürün fiyat bilgilerini, kargolama, ödeme yöntemlerini internet kullanıcılarına kendileri belirtir. Tüm şart ve koşullar şirket tarafından belirlenmekte, müşteri ile birebir iletişim halinde olunmakta ve müşteriye doğrudan erişim sağlanmaktadır. E-ticaretin en önemli ve dikkat çeken türüdür ve satıcı ile müşteriye birleştirmek için Amazon gibi birçok dünya çapında e-ticaret platformları bulunmaktadır (bkz.: www.amazon.com).

Bu tez çalışmasında da, e-ticaretin dünya çapında ve yerel olarak en çok kullanılan işletmeden tüketiciye e-ticaret türü ele alınmaktadır.

3.3.3. İşletmeler ile kamu idaresi arasında yapılan e-ticaret (Business to government-B2G)

Şirketler ile kamu idaresi arasında yapılan e-ticaret, haberleşme ve diğer tüm etkinlikleri kapsayan e-ticaret şeklidir. B2G e-ticaret türünde firmalar satıcı, kamu kuruluşları ise alıcı konumundadır. Kamu ihalelerinin çevrimiçi olarak yayınlaması, şirketlerin de yine çevrimiçi olarak tekliflerini kamuya sunmaları; ya da şirketlerin e-devlet uygulamaları vasıtasıyla vergi ödemelerini gerçekleştirmeleri B2G e-ticareti kapsamaktadır. B2G e-ticaret, kamu idaresinin satın alma ya da şirketlerin ihaleye katılım noktasında gerçekleştirildiğinden dolayı, vergi ve benzeri ödemelerin gerçekleştirilmesi noktasında doğrudan e-ticaret türüne dahil olmaktadır (Toprak, 2014).

3.3.4. Tüketiciden tüketiciye yapılan e-ticaret (Consumer to consumer-C2C)

Tüketicilerin kendi aralarında bir takım e-ticaret platformları üzerinden birbirleri ile iletişime geçip, ürün alım ve satımı yaptıkları e-ticaret türüdür. Genellikle kullanılmış ürünlerin satışının yapıldığı bu e-ticaret yönteminin Türkiye’de Gittigidiyor öncülerindendir (bkz.: www.gittigidiyor.com).

3.3.5. Tüketiciler ile kamu idaresi arasında yapılan e-ticaret (Consumer to government-C2G)

Vatandaş ile devlet arasındaki ilişkilerini düzenlemek ve bu ilişkileri internet üzerinde yürütmek vatandaş ile devlet arasında yapılan e-ticaret şeklidir. Bunlara örnek olarak; sosyal güvenlik ödemelerinin, vergi borçları ve cezalarının (trafik gibi) internet yolu ile ödemelerinin yapılabilmesi; devletin, eğitim sisteminde, yaptığı sınavların sonuçlarını internet üzerinden kamuya açıklaması; çeşitli kimlik numaralarının internet yoluyla öğrenilmesi; devletten bilgi talep edilmesi örnek olarak gösterilebilir (Özbay ve Devrim, 2000).

3.3.6. Devletler arasında yapılan e-ticaret (Government to government-G2G)

Sektörel bilgileri içeren, yerel veya ulusal yönetimler arasında yapılan yazılı bilgilerin raporlaştırılarak paylaşılması işlemidir. Bu yöntem çok uluslu devletlerin daha çok kullandığı bir e-ticaret şeklidir. Bunun dışında örneğin; saklanan doğum ve ölüm kayıtlarının başka devletler ile paylaşılması devlet kurumları arasındaki bağlantıların sağlanması amaçlanmaktadır (Efendioğlu ve Sezgin, 2007).

3.4. E-Ticaretin Avantaj ve Dezavantajları

E-ticaretin tercih edilmesinin birçok sebebi bulunmaktadır. Bunlardan bazıları aşağıdaki gibi ifade edilmiştir:

- Geleneksel pazarlama ve satış yöntemleri ile belirli bir lokasyona ve kitleye erişim sağlanabilirken e-ticaret ile daha geniş kapsamlı, dünyanın farklı lokasyonlarında bulunan, farklı yaşlarda, farklı özelliklerde olan kitlelere ulaşmak mümkündür. Coğrafi sınırlamalar yoktur, internete erişim sağlayan herkese ulaşım sağlanabilir.
- Fiziksel bir mağazaya ihtiyaç duyulmadığından, e-ticaret işletmeler açısından işletmelerin büyük maliyetlerinden birisi olan mağaza veya dükkan gibi masraflarından tasarruf yapmalarını sağlar.
- 7/24 satış yapabilme imkanı sağlamaktadır. Mağaza saat sınırlamalarının olmaması, kullanıcılara istedikleri zamanda alışveriş yapma fırsatı tanımaktadır.
- Dilediğiniz kadar ürünü web sitesinde listeleyerek onları satışa çıkartabilme imkanı sağlamaktadır. Bir mağazaya sığamayacak kadar ürünü listelenebilir, e-ticaret platformları aracılığı ile satışa çıkartılabilir ve bununla birlikte yeni ve farklı ürün satışları gerçekleştirilebilir.
- Ürün veya hizmete ait özelliklerin detaylı olarak belirtilmesi kullanıcılar için detaylı araştırma yapabilme, fikir sahibi olma olanakları sağlamaktadır.
- Farklı mağaza, satıcı, üretici, bilgi sağlayıcı gibi kaynaklara kolay erişebilme ve karşılaştırmaların rahatlıkla yapılabilmesi, kullanıcıların farklılıkları görüp karar vermeleri açısından değişik opsiyonlar sağlaması e-ticaretin avantajlarından.

- İstenilen kategorideki ürüne kolayca erişilebilir ve ziyaretçiler ürün kategorisi sayfalarına göz atabilir ve ürün arama özelliğini kullanarak istedikleri ürüne hemen erişim sağlayabilir.
- Kullanıcılar veya müşteriler satın alma yaparken zaman açısından büyük kazanç sağlamaktadır.
- Mağaza fiyatına karşın çevrimiçi fiyatlar daha uygun olarak sunulmaktadır. İşletme maliyetleri daha az olduğu için çevrimiçi fiyatlar mağaza fiyatlarına göre daha uygun olmaktadır.
- Ürün satın almalarında müşteriye yormadan, müşterinin kapısına kadar siparişini teslim etme süreci de müşteriler için avantajlı ve e-ticaretin cazip gelen özelliklerindendir.
- Kampanya ve indirim gibi fırsatlar kullanıcılara e-posta veya sms (kısa mesaj) yolları ile haber verilmektedir.
- Makine öğrenmesi algoritmaları kullanılarak doğru zamanda doğru kitlelere ulaşım sağlanabilir, müşteri odaklı algoritmalar geliştirilebilir, müşterilerin tercih edeceği düşünülen alternatifler müşteriye tercih olarak sunulabilir.
- Web site veya diğer çevrimiçi platform alışverişlerinde kredi kartı veya sanal paralar kullanılarak ödeme kolaylıkları sağlanmaktadır. Ödeme esnasındaki güvenlik sistemleri de gün geçtikçe geliştirilmekte ve müşterilerin alışveriş yaptığı platformlara güvenerek alışveriş yapmalarını sağlamaktadır.

E-ticaretin dezavantajları ise aşağıda görüldüğü gibidir:

- Ürün satın alma süreçlerinde geleneksel alışveriş yöntemlerinde ürünü fiziki olarak görme, deneme, inceleme vb. imkanları tanınmamaktadır. E-ticaret platformlarında ürüne ait sadece görsel ve ürün açıklaması bulunmaktadır, örneğin; müşterilerin bir ses cihazı satın alırken ilgili ses cihazını dinleyerek test etme, deneme gibi bir seçenekleri yoktur.
- Satın alma işleminin yapılabilmesi için bir teknolojik cihaza ihtiyaç duyulmaktadır ve aynı zamanda internet bağlantısına erişim sağlanması gerekmektedir.

- Satın alma yapıldıktan sonra bir müddet ürünün müşteriye ulaşmasının beklenmesi gerekmektedir. Geleneksel satın alma sürecinde ise, ürün aynı zaman içerisinde satın alınır ve herhangi bir bekleme gerekmez.
- Ürünün lojistik maliyetlerinin bulunması; satın alınan ürünün kargolama hizmet bedelinin müşteri tarafından karşılanması geleneksel alışveriş yönteminde karşılaşılmayan bir durumdur.
- Ödeme seçeneği müşteri tercihinine göre değişebilmektedir. E-ticaret platformlarında yalnızca kredi kartı, banka kartı veya diğer sanal paralar ile satın alma yapılabilir ve bu nedenle nakit satın alma iskontolarından faydalanılmaz.
- E-ticaret web sitelerinin kayıt ettiği kişisel veriler, kimlik avcılığı, hesabın başka kişiler tarafından ele geçirilmesi gibi durumların yaşanabilmesi olasılığı güçlü güvenlik sistemleri olsa da mümkün olabilir ve bu nedenle kişisel bilgiler risk altındadır.
- Ödeme esnasında kullanılan kart bilgilerinin de kopyalanma ve siber saldırılara uğraması gibi durumlarla karşılaşılabilir.
- E-ticaret mağazasında canlı müşteri destek hizmetleri sınırlı olabilir. Müşteri hizmetleri günün belirli saatlerinde destek verebilir veya müşteri hizmetleri telefon numarasını arayarak müşteriye beklemede tutabilir. Geleneksel alışveriş yönteminde ise müşteri mağaza içerisinde istediği bir ürün hakkında satış temsilcileri tarafından anlık olarak destek alabilmektedir.

Farklı sektörler temel alınarak derin bir inceleme yapıldığında bu avantajlar ve dezavantajlar artabilmekte, azalabilmekte veya farklılıklar gösterebilmektedir. Gelişen teknoloji çağında e-ticaretin güvenilirliği ve karşılaşılan dezavantajların zamanla müşteri lehine giderilmesi için çalışmalar yapılmaya devam edilmektedir.

4. MAKİNE ÖĞRENMESİ

Tez çalışmasının bu bölümünde makine öğrenmesi kavramı, tarihsel gelişimi, yöntemleri ve makine öğrenmesi uygulama süreci anlatılmaya çalışılmıştır.

4.1. Makine Öğrenmesi Kavramı

1950’li yıllarda teknolojik buluşlar sebebi ile bilgi çağı denilen bir döneme girilmiştir. Bu teknolojik gelişmeler bilgi kavramı üzerine kurulmuştur. İnsanların hayatını kolaylaştırmak için geliştirilen bu sistemler zamanla insanları merkezine alarak insanlar için; insanlardan elde edilen bilgileri toplayarak, insan beynini, insan hareketlerini temel alan yapay zeka sistemleri kurulmaya başlanmıştır. Bilgisayar teknolojilerindeki her türlü bilgi, veri olarak geçmektedir ve insan en büyük veri kaynağıdır. Bilgilerin akıllıca kullanılabilmesi için elde edilen verileri makinelerle öğretmek faydalı sonuçlar elde edilmeye başlanmıştır.

Makine öğrenmesi insanların hayatının her alanında karşılaştıkları bir kavramdır. Hayatımıza hızlı bir şekilde giriş yapmış ve sürekli olarak gelişmektedir. Makine öğrenmesi güncel bir teknolojik gelişme olarak hayatımızda yer alsa da fikir ve düşünce olarak, geçmişte bugün bilinen makine öğrenmesi kavramının geleceğe yönelik tanımları yapılmıştır. Edmunda Berkeley, 1949 yılında, Giant Brains or Machine That Think (Büyük Beyinler Düşünen Makineler) adlı eserinde; büyük makineler hakkında bir çok haber olduğunu, beyine benzer kompleks yapılardan oluştuğundan bahsetmiştir. Makinelerin bilgiyi işleyebilir, hesaplayabilir ve sonuçlandırabilir yapılarda, aynı zamanda devasa hız ve becerilerinin olduğunu, makinelerin düşünebileceğini belirten tanımlamalarda bulunmuştur.

Her sektör ve alan için devamlı olarak makine öğrenmesi algoritmaları ve çözümlemeleri geliştirilmektedir. Örneğin; sağlık alanında hastalıklara tanı koyma ve teşhis etme, ilaçların keşfedilmesi ve üretilmesi, tıbbi görüntülemelerde teşhis etme, klinik denemeler, araştırmalar, yüz tanımlama, görüntü tanımlama ve okuma,

konuşmaların tanımlanması, konuşma esnasında kullanılan sözcüklerin yazılı olarak anlık çevrisinin yapılabilmesi, otomatik dil işleme süreçleri, çevirileri, trafik tahminleri, e-posta, istenmeyen (spam) filtrelemelerinin yapılması, kişisel sanal asistanlar, banka ve borsa takip edilmesi ve tahminlerinin yapılması gibi birçok alanda kullanılmaktadır. İnsan beynini temel alarak geliştirilen akıllı sistemler insan genleri, hücreleri gibi küçük bir kavramdan uzay ve evren gibi daha büyük kavramlara yönelik geliştirmeleri sürdürebilmektedir.

Yapay zekanın bir alt dalı olan makine öğrenmesi, öğrenme metotları ve bu metotların performansı ile ilgilenir. Bilgisayar sistemlerinin insanlar gibi öğrenme yeteneğinin olabilmesi için çeşitli algoritma, modelleme ve teknikleri geliştirmek amaçlı kullanılan bilimsel bir çalışma alanıdır (Alpaydın, 2010). Makinelerin öğrenebilmesi kavramı ilk bakışta zor olsa da yapay zekâ, matematik ve istatistik ile birlikte çalışarak makinelerin öğrenebilmesi olayı mümkün hale gelmiştir. Makine kavramı kelime anlamı olarak birçok insan için farklı anlamlar taşısa da, günümüzde yazılım ve donanımlardan meydana gelen bilgisayar sistemlerinde, bu yazılım kısmı “makine” olarak düşünülmektedir (Nilsson, 2010).

Yapay zeka ve makine öğrenmesi akıllı sistemler oluşturmak için kullanılan bilgisayar sistemleridir fakat ikisi de farklı terimlerdir. Yapay zeka, insan düşünme yeteneğini ve davranışını taklit edebilen akıllı makineler oluşturmak için daha büyük bir kavramı temsil eder, oysa makine öğrenimi, makinelerin açıkça programlanmadan verilerden öğrenmesini sağlayan bir yapay zeka uygulaması veya alt kümesidir (Ballouk, 2020). Yapay zekada bir amaca yönelik tahmin oluşturma veya karar verme makine öğrenmesi yöntemleri ile yapılmaktadır.

1950’li yıllarda yapay zeka kavramının ortaya çıkması ve gelişmesi, 1980’li yıllarda yapay zeka öğrenme işlemleri yapılmaya başlamıştır ve makine öğrenmesi kavramı gelişmiştir. Bu teknolojiler ile birlikte derin öğrenme kavramı da 2010’lu yıllarda açığa çıkmaktadır. Bu teknolojiler sık sık birbirleri ile iç içe olmasından kaynaklı olarak sıkça karıştırılmaktadır. Kısaca Şekil 4.1’deki gibi bu kavramlar ele alınmaktadır.



Şekil 4.1. Yapay zeka, makine öğrenmesi, derin öğrenme algoritmalarının karşılaştırılması

Makine öğrenmesi sadece bir veri tabanı sistemi olmayıp aynı zamanda yapay zekanın bir parçasıdır. Mevcut durumlar birbirinden farklılık gösterebilir, değişen durumlarda da kullanılması için bu sistemlerin öğrenebilme yeteneğinin olması gerekmektedir. Sistem bu değişiklikleri öğrenebilir ve uygulayabilirse, sistem tasarımcısı tüm olası durumlar için ayrı ayrı tahmin ve sonuç çıkartma ihtiyacı duymaz (Keskin A., 2018).

Makine öğrenmesi yapay zekanın temelini oluşturan matematik ve istatistik bilimlerinden yararlanarak modelleme ve analiz yapabilen öğrenme yöntemidir. Makine öğrenmesi çözümlerinin yapılabilmesi için oluşturulan bilgisayar sistemleri arka planda çalışan matematik ve istatistiksel formülasyonları kullanmaktadır.

Makine öğrenmesi algoritmaları, mevcut verileri veya geçmiş deneyimleri kullanarak, bir performans değerini optimize etmek amacıyla bilgisayarlarda yapılan programlamalar ile yazılımsal sistemlerden oluşur. Matematiksel modellerin oluşturulması için istatistik teorisini kullanır, çünkü asıl amaç istatistiksel yöntemlerde olduğu gibi örneklemden çıkarım yapmaktır. Makine öğrenmesinde bilgisayar bilimini eğitim aşamasında performans değerini optimize etmek için algoritma ya da modele gerekli veriyi depolama ve işleme imkanı sağlar (Alpaydın, 2016).

4.2. Makine Öğrenmesinin Tarihsel Gelişimi

Makine öğrenmesinin genel çerçevesi itibari ile aşağıdaki gibi tarihsel gelişimi gerçekleşmiştir (Marr, 2016).

1950 ve öncesi yıllarda; makine öğrenmesi kavramı ortaya çıkmadan önce makine öğrenmesi temelini oluşturan istatistiksel modellerin geliştirildiği dönemdir. Makine öğrenmesi birçok istatistiksel kavramlar, olasılık ve bayes teoremine bağlı olarak geliştirilmiştir.

1950 yılında, dünyanın akışını değiştiren İngiliz matematikçi Alan Turing makinelerin düşünebilmesine dair fikirlerini Computing Machinery and Intelligence adlı makalesinde bir zeka testinden bahsetmiştir ve Turning Testi olarak tarihte yerini, atılan ilk adım olarak almıştır.

1950 ve 1960 yılları arası, sinir ağlarının bulunması ve gelişimi olarak bilinmektedir. Dama oyunu ilk makine öğrenmesi uygulaması olarak tarihe geçmiş, derin sinir ağları temelleri atılmış ve Alpha adlı bir bilgisayarda uygulamalar yapılmıştır.

1960 ve 1970 yılları arasında, 1965'de Matematikçi Alexey Ivakhnenko, V.G. Lapa ve arkadaşları o ana kadar sadece teori ve fikir olan yöntemleri uygulayarak ilk derin öğrenme ağlarını kurmuştur. 1967'de En Yakın Komşu algoritması oluşturulmuş ve ilk olarak haritalama alanında rota oluşturmak için kullanılmıştır.

1970 ve 1980 yılları arasında, Fukushima bir yapay sinir ağı olan ve görsellerin kayıtlarını yapabilen Neocognitron'u yapmıştır. Bazı kişiler tarafından sinir ağlarının işe yaramayacağına dair tepkiler artmış, bazı yatırımcılar fonlarını kesmişlerdir. Bu dönem birçok kaynakta yapay zekanın kışı olarak da geçmektedir.

1980 ve 1990 yılları arasında, Gerald Dejong 1981'de, verileri analiz eden ve önemsiz verileri atarak izleyebileceği genel bir kural oluşturan Açıklama Tabanlı Öğrenme

(Explanation Based Learning-EBL) kavramını tanıtmıştır. 1985’de Terry Sejnowski, bir bebeğin yaptığı gibi kelimeleri telaffuz etmeyi öğrenen NetTalk’u icat etmiştir. Şekil tanıma, sinir ağları gibi daha önce bulunan modellerin bu on yıl içerisinde geliştirilmesine yönelik çalışmalar yapılmıştır.

1990-2000 yılları arasında, gözetimli ve gözetimsiz makine öğrenmesi kavramları daha belirgin hale gelmiştir. Bilgi odaklı bir yaklaşımdan, veriye dayalı bir yaklaşıma geçilmeye başlanmıştır. Bilim adamları, büyük miktarda veriyi analiz edebilmek, sonuçlar çıkarabilmek veya “öğrenmek” için bilgisayarlar programları oluşturmaya başlamıştır. 1993 yılında, Alman bilgisayar bilimcisi Schmidhuber, tekrarlayan sinir ağında 1.000’den fazla katman gerektiren bir “çok derin öğrenme” görevini çözmüştür. Sinir ağlarının karmaşıklığı ve becerisinde ileriye doğru büyük bir adım atılmıştır. 1997 yılında IBM’in geliştirdiği Deep Blue adlı bilgisayar, satrançta dünya şampiyonunu yenmiştir. Bu zaferden sonra makine öğrenmesi insanların ilgisini daha çok çekmeye başlamıştır. Vapnik daha önce çalışmaları olan Destek Vektör Makinelerini geliştirmiştir.

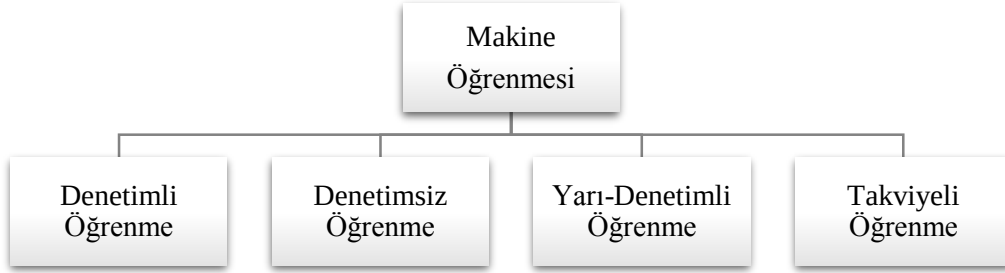
2000-2010 yılları arasında, 2006 yılında, resim ve videolardaki nesneleri, metinleri algılayıp ayırt edebilen derin öğrenme metotları Kinect adlı hareket algılama giriş cihazını geliştirilmiştir.

2010 yılından sonra, Google Brain ile nesneleri tanımlama ve sınıflama işlemi yapabilen bir sinir ağı geliştirmiştir. 2014 yılında Facebook, fotoğraflardaki bireyleri insan doğruluğu ile tanıyabilen ve doğrulayabilen bir yazılım algoritması olan DeepFace’i geliştirmiştir. Amazon kendi makine öğrenmesi platformunu açmıştır. Satrançtan daha karmaşık olarak bilinen eski bir Çin oyununu oynamak için Google DeepMind tarafından AlphaGo adlı bir algoritma geliştirilmiş ve girdiği oyunların birçoğunu kazanmıştır.

Tüm bu gelişmeler veri akışı arttıkça, veri akışını yönetebilmek, daha büyük boyutlara ulaşmak, mevcut verileri doğru makine öğrenmesi algoritmaları ile doğru bir şekilde kullanarak insan zekasına daha yakın yapay zeka sistemleri geliştirilecektir.

4.3. Makine Öğrenmesi Yöntemleri

Makine öğrenmesi yöntemleri mevcut veriler doğrultusunda doğru algoritmaları ve teknikleri kullanmak veya seçmek amaçlı geliştirilmiş yöntemlerdir. Veri boyutu, yapısı, çeşitliliği gibi farklı durumlar doğrultusunda ve elde edilmek istenen amaca yönelik farklı algoritmalar kullanılmaktadır. Makinelerin öğrenme şekillerine göre makine öğrenmesi yöntemleri Şekil 4.2’de görüldüğü gibidir. Bu yöntemler ve bu yöntemlere ilişkin kavramlar ilerleyen bölümlerde açıklanacaktır.



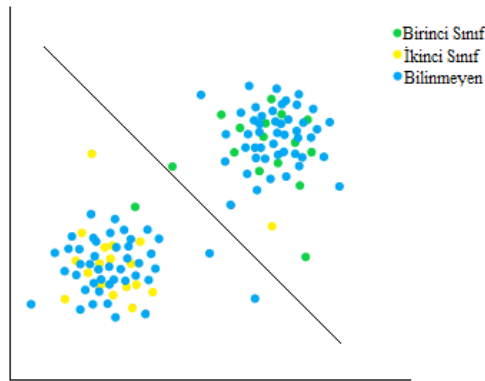
Şekil 4.2. Makine öğrenmesi yöntemleri

4.3.1. Denetimli öğrenme

Denetimli öğrenme girdi ve çıktı değerlerinin birlikte yer aldığı veri setleri için uygulanan bir öğrenme biçimidir. Hangi girdi değerlerinin hangi çıktı değerlerine karşılık geldiğinin bilinmesi durumu etiketli veri seti olarak adlandırılmaktadır. Denetimli öğrenme yönteminde algoritmalar girdi ve çıktı değerleri arasındaki ilişkiyi anlamlı bir şekilde ifade eden modeller çıkartmaya çalışır. Bu model yardımıyla yeni bir

örnek ve örneği oluşturan girdiler verildiğinde hangi çıktı değerinin oluşacağını tahmini yapılır. Üretilen model sonucunda elde edilen çıktılar, eğri uydurmak veya girdilerin sınıflarının tahmin edilmesi için kullanılabilir (Mohri vd., 2012).

Veri seti içerisinde her bir girdi tahmin edici x_i ($i=1,2,3,\dots,n$), için bir çıktı y_i değeri vardır. Veri seti eğitim ve test olarak iki kısma ayrılır. Veri setini eğitim ve test olarak ayırmak için bir takım yöntemler bulunmaktadır ve veri setinin boyutuna ve aynı zamanda en iyi sonucu vermesi durumuna göre bu ayırım değişiklik göstermektedir. Örneğin; veri seti %80 eğitim, %20 test veya %70 eğitim, %30 test verisi olarak ayrılabilir. Büyük kısım makineye öğretileceği için eğitim verisi olarak belirlenir. Ayrılan eğitim verisi içerisinde hem girdi (x_i) hem de çıktı (y_i) değerleri bulunmaktadır. Eğitim verisi, belirlenen denetimli öğrenme yöntemi ile makineye öğretilerek bir model çıkartılır ve makineye daha önceden tanıtılmayan test verisi üzerinde bu model uygulanır. Makinenin belirlenen yöntem ile daha önceden eğitim verisi ile öğrendiği girdi (x_i) ve buna karşılık gelen çıktı (y_i) değerlerini kullanarak oluşturduğu model ile, test verisi üzerinde her bir girdiye (x_i) karşılık gelecek çıktı (y_i) değerlerini bulması beklenir. Bu şekilde mevcut veriler ile oluşan model yeni veriler için en iyi tahmini bulmaya çalışır. Sınıflandırma ve regresyon, en çok bilinen denetimli makine öğrenmesi yöntemlerindendir.



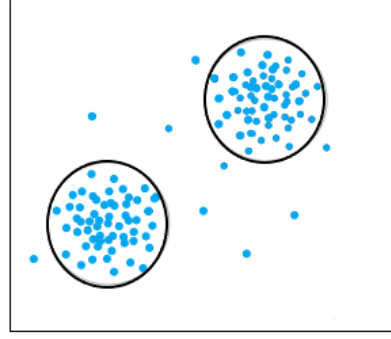
Şekil 4.3. Denetimli öğrenme

Şekil 4.3'te mevcutta yeşil ve sarı olarak gösterilen iki grup bulunmaktadır, mavi renk olarak gösterilen gözlemler hangi gruba ait olduğu bilinmeyen gözlemlerdir. Gözlemler arasında bazı örtüşmeler olduğu görülmektedir. Bu durumda kümeleme yapmak daha zor olmaktadır, amaç yeni gelen her bir gözlemin hangi gruba dahil olduğunu belirleyebilmektir.

4.3.2. Denetimsiz öğrenme

Denetimli öğrenmeden farklı olarak bu öğrenme yönteminde veri seti içerisinde her bir girdi tahmin edici x_i ($i=1,2,3,\dots,n$), için x_i 'lere karşılık gelen bir çıktı y_i değeri yoktur. Denetimsiz olarak adlandırılmasının sebebi analiz sonucunda denetlenecek bir yanıt (y_i) değişkeninin olmamasından kaynaklanmaktadır. Burada herhangi bir eğitim verisi bulunmamaktadır, makine daha önceden görmediği veriler ile karşılaşarak, anlamlı bir bağ kurmak amaçlı veriler arasındaki yakınlık, uzaklık, ilişki (korelasyon) gibi kıstaslara bakarak bir model kurar ve bu model sonucunda x_i girdilerine karşılık bir y_i çıktı değerini kendisi tahmin etmeye çalışır.

Veriler makine tarafından bilinmediği için anlamlı bir şekilde veriyi gruplayabilmek amacıyla kümeleme yöntemleri kullanılabilir. Örneğin; bir web sitesindeki, bir kullanıcının web tıklamalarını ve demografik verilerini kullanarak, farklı türdeki kullanıcıları bir araya getirebilmektedir. Web sitesi daha sonra bu farklı türlere göre kişiselleştirilebilir. Diğer durumlarda, amaç verilerin boyutunu daha yönetilebilir bir değişken kümesine düşürmek olabilir. Denetimsiz öğrenme, çok sayıda değişken ve kayıtlarla karşı karşıya kalınan durumlara keşifsel veri analizinin bir uzantısı olarak bakılabilir. Amaç, bir dizi veri ve farklı değişkenlerin birbiriyle nasıl ilişkili olduğu hakkında fikir sahibi olmaktır. Denetimsiz makine öğrenmesi, bu değişkenleri incelemek, analiz etmek ve ilişkileri keşfetmek için kullanılan bir yöntemdir (Bruce, 2017). Boyut küçültme ve kümeleme denetimsiz öğrenme yöntemi ile en çok yapılan işlemlerdendir.



Şekil 4.4. Denetimsiz Öğrenme

Şekil 4.4’te iki grubu içeren bir kümeleme işlemi görülmektedir. Buradaki veriler iki gruba ayrılacak şekilde dağılım göstermiştir. Bu durumda verilerin birbirlerine olan yakınlıklarına bakılarak kümeleme yapılabilir, iki grup başarılı bir şekilde ayrılabilir.

4.3.3. Yarı denetimli öğrenme

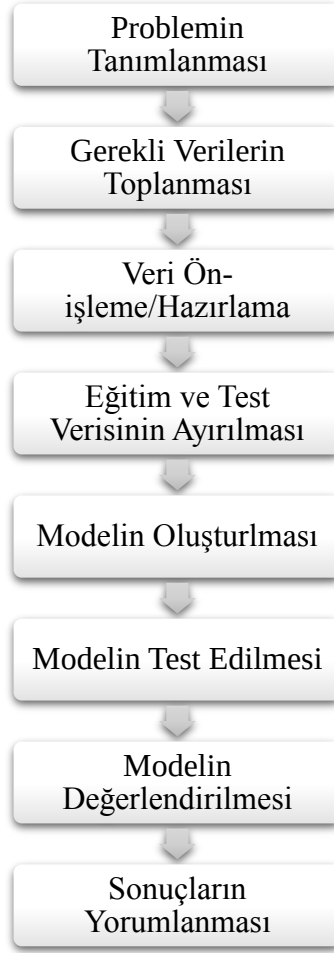
Bazı durumlarda kullanılacak yöntemin denetimli mi yoksa denetimsiz mi olduğu bilinmemektedir. Örneğin, bir dizi n gözlemin olduğu veri setinde, gözlemlerin m miktarı için, $m < n$ olduğunda, hem girdi değerleri hem de bu değerlere karşılık gelen çıktı değerleri bulunmaktadır. Kalan $n - m$ adet gözlem için girdi değerleri bulunmakta ancak çıktı değerleri bulunmamaktadır. Sahip olunan veri seti içerisinde, her bir x_i değerine karşılık gelen bazı y_i değerlerinin olması ve bazı y_i değerlerinin olmaması gibi karmaşık bir durum ile karşılaşmak mümkündür. Bu durum yarı denetimli makine öğrenme yöntemi olarak adlandırılmaktadır. Böyle bir durumda, çıktı değerlerinin mevcut olduğu m gözlemlerini ve bunların olmadığı $n - m$ gözlemlerini birleştirebilecek istatistiksel bir öğrenme yöntemi kullanmak tercih edilmektedir (James vd., 2013).

4.3.4. Takviyeli öğrenme

Takviyeli öğrenme, kararları kendi kendine alan bir öğrenme yöntemidir. Doğru hedefe ulaşmak için nasıl karar vereceğini nasıl öğrenebileceğini kendisi geliştirir. Takviyeli öğrenme yöntemi çoğunlukla robotik, oyun programlama, hastalık teşhisi ve fabrika otomasyonları gibi alanlarda kullanılmaktadır. Bu öğrenme yönteminde bir eğitmen mevcut olsa da sisteme ayrıntı verilmemektedir. Bu sebepten sistem karar verdiğinde, karar doğruysa ödül, yanlışsa ceza vermektedir. Amaç, denenen tüm durumları hatırlayarak kontrol edilmeye çalışılan durumların hedeflenip hedeflenmediğini kontrol etmektir (Çelik E. , 2019).

4.4. Makine Öğrenmesi Uygulama Süreci

Makine öğrenmesi uygulama süreci birçok aşamadan oluşmaktadır. Her aşama doğru ve başarılı bir sonuca ulaşmak için çok önemli detayları içermektedir. Her aşamanın belirli gereklilikleri mevcuttur ve bu aşamalar tamamlandığı takdirde bir sonraki adımı uygulamak gerekmektedir. Makine öğrenmesindeki her adım veya aşama bir sonraki sürecin daha kolay bir şekilde yapılmasına olanak sağlamaktadır. Genel çerçevesi itibari ile bir makine öğrenmesi süreci Şekil 4.5'te gösterildiği gibidir.



Şekil 4.5. Makine öğrenmesi uygulama süreci aşamaları

4.4.1. Problemin tanımlanması

Problemin tanımlanması yapılacak olan çalışmanın birinci ve en önemli aşamasıdır. Yapılacak olan çalışmanın amacı tam olarak belirtilmeli ve bu doğrultuda çalışmaya yönelik plan haritası çizilmelidir. Çözülmesi veya cevaplanması istenilen probleme yönelik ön çalışmalar, uygulanacak adımlar, gerekli analiz ve modellerin seçilmesi ve bu doğrultuda çalışmanın devam ettirilmesi gerekmektedir. Problemin amacına yönelik yapılmayan çalışmalar yanlış sonuçlar verebilmekte ve karar verme süreçlerini olumsuz etkileyebilmektedir.

4.4.2. Verilerin toplanması süreci

Çalışmanın amacı belirlendikten sonra, çalışmanın amacına yönelik verilerin toplanması sürecine geçilmektedir. Çalışma kapsamı ve bu çalışma kapsamına bağlı olarak problemin sonucuna ulaşmayı sağlayan örnek kitlenin en doğru şekilde seçilmesi gerekmektedir.

Verinin toplanacağı kitlenin taşınması gereken en önemli özelliği, araştırmanın konusuna yönelik özellikleri bulundurması gerektiğidir. Araştırma sonuçlarının genel yargılar içerebilmesi için yeterli büyüklükte veri toplanması gerekmektedir. Verilerin seçilmesi ve hacminin belirlenmesi istatistiksel yöntemler ile de belirlenebilmektedir.

Verilerin toplanacağı kaynakların güvenilir ve doğru kaynaklar olması gerekmektedir. Farklı kaynaklardan alınacak verilerin, oluşturulacak veri seti için aynı standartlarda veya ölçülerde olması gerekmektedir. Örneğin; seçilen değişkenlerin farklı gözlem sayıları içermesi eksik veri ve verilerin uyuşmamasına neden olabilmektedir. Yanlış toplanan veriler yapılan çalışmanın ilerleyen adımlarında modelin kurulması ve analizlerin uygulanması aşamasında, yanlış sonuçların çıkmasına neden olacak ve zaman açısından da büyük sorunlar ortaya çıkacaktır. Bu sebepten veri toplama süreci çalışma kapsamına uygun bir şekilde yapıldığında diğer adımların kolaylıkla gerçekleşmesi sağlanmaktadır.

4.4.3. Verilerin hazırlanması süreci

Veri hazırlık sürecinde, elde edilen tüm verilerin analize uygun hale getirilmesi gerekmektedir. Farklı veri tiplerinin bulunması farklı veri ölçümleri, eksik verilerin olması, aykırı değerlerin bulunması, veri dağılımlarının farklılık göstermesi veya mevcut veri tiplerinin kesikli, sürekli olması gibi durumlar hangi analiz yönteminin kullanılması gerektiğine karar verme sürecinde etkilidir.

Öncelikle toplanan verilerin farklı kaynaklardan toplanması durumu vs. göz önünde bulundurularak aynı format ve boyuta getirilmesi gerekmektedir. Daha sonra farklı

ölçüm değerinde olan verilerin birbirleri ile uyum içerisinde olmaması durumunda normalleştirme gibi işlemler yapılabilir. Mevcut veri durumuna göre eksik gözlem var ise, çalışma kapsamında eksik verilerin önem derecesine göre eksik veri analizleri uygulanabilir veya tekrardan önem derecesine göre eksik gözlemler veri setinden çıkartılabilir. Diğer taraftan verilerin karakteristik özellikleri incelenmeli ve ilerleyen süreçte analizin uygulanması aşamasında sorun yaratabilecek veri seti ile uyum sağlamayan, çok büyük veya çok küçük değerler içeren aykırı değerlerin kontrolü ve bu doğrultuda müdahale edilmesi gerekebilir. Çok büyük veya çok küçük değer içeren gözlemler model parametrelerini ve model sonucu yanlış etkileyebilirler. Verilerin doğru işlenmemesi tüm süreci olumsuz etkileyebilecek en önemli adımdır.

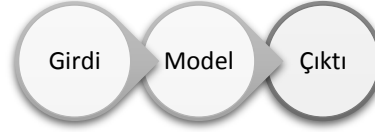
Verilerin doğru bir şekilde tanımlanması, düzenlenmesi ve veri seti üzerinde yapılan işlemlerin her biri analizin yapılabilmesi, oluşturulması, açıklanabilmesi ve yorumlanabilmesi açısından büyük önem taşımaktadır.

4.4.4. Eğitim ve test verisinin belirlenmesi

Denetimli makine öğrenmesi sınıflandırma yöntemlerinde uygulanan, mevcuttaki tüm veri setinin eğitim ve test olarak ayrılması süreci, yapılmak istenilen işlemin makineye hem öğretilmesi hem test edilmesi hem de doğru olup olmadığını sınamak için temel adımlardandır. Veri setinin büyük bir oranı eğitim küçük bir oranı test olmak üzere ikiye ayrılır. Eğitim verisi makineye tanıtılarak, eğitim verisi ile bir model oluşturulur ve daha sonra makineye tanıtılmayan test verisi ile model çalıştırılır. Burada amaç eğitim verisi ile en doğru sonuçları veren modeli kurmak ve yeni gelecek olan veriler üzerinde bu modeli uygulayarak doğru sonuçlar elde edebilmektir. Aynı zamanda uygulanan modelin ne kadar doğru sonuçlara ulaştığı da kontrol edilebilmektedir. Yanlış ayrılan veri setinin hem doğru modeli oluşturması hem de doğru sonuçlar vermesi beklenemez. Mevcut duruma göre eğitim ve test verisi farklı oranlar ile ayrılıp denenebilir.

4.4.5. Makine öğrenmesi yönteminin uygulanması/model oluşturma

Bu aşama bir önceki süreçlerde yapılmak istenilen çalışmanın amacına yönelik, hangi analizin kullanılacağına karar verildikten sonra uygulanması aşamasıdır. Bir önceki süreçlerde ön hazırlığın yapıldığı ve modelin oluşturulması için ayrılan eğitim verisi üzerinde makine öğrenmesi modeli oluşturulur. Eğitim verisi içerisindeki veriler modele ait parametrelerin belirlenmesini sağlamaktadır. Şekil 4.6’da gösterildiği gibi, uygulanan modelin amaca yönelik en iyi sonuçlara götürmesi beklenmektedir. Modelin doğruluğu ve başarısı ayrılan test verisi üzerinde uygulanarak model başarı ölçütlerine göre değerlendirilebilmektedir. Model başarı ölçütleri tez çalışmasının ilerleyen bölümlerinde açıklanmaktadır.



Şekil 4.6. Girdi, model, çıktı uzayı gösterimi

4.4.6. Modelin yorumlanması ve doğruluğunun ölçülmesi

Uygulanan denetimli makine öğrenmesi yöntemi sonucunda elde edilen model üzerinde test verisi uygulandıktan sonra, elde edilen sonuçlar gerçek değerler ile karşılaştırılmaktadır. Bu karşılaştırmalar sonucunda modelin doğruluk oranı gibi başarı oranını gösteren sonuçlar elde edilebilmektedir. Aynı veri üzerinde farklı sınıflandırma yöntemleri uygulanacağı için bu ölçüm değerleri hem modeli tek başına hem de yöntemlerin birbirleri ile karşılaştırılmasını sağlamaktadır. Karşılaştırma sonrasında çıkan ölçümler yorumlanarak model ve problemin amacına yönelik sonuçların yorumlanması yapılabilmektedir.

5. MAKİNE ÖĞRENMESİ SINIFLANDIRMA YÖNTEMLERİ

Makine öğrenmesi yöntemleri geleceği yönelik tahminler yapma ve doğru sınıflandırma algoritmaları kullanarak veri bilim sürecine katkıda bulunmaktadır. Makineye öğretilecek olan algoritmaların seçilirken ve tasarlanırken en iyi performansı göstermeleri beklenmektedir. Günümüzde bulunduğumuz konum itibari ile yapılan çalışmalarda neredeyse 5 yıllık süre içinde yeni teknolojik dönemlere girilmektedir. Veriye ulaşmak kolay olduğu gibi doğru analiz etme ve sınıflandırma yapmak da oldukça önem taşımaktadır (Baran, 2020). Sınıflandırma yöntemleri denetimli makine öğrenmesi yöntemleri olarak geçmektedir. Tez çalışmasının bu bölümünde kullanılan makine öğrenmesi yöntemleri, bu yöntemlere ait kısa tarihçeler, yöntemlerin algoritmaları ve yöntemlerin karşılaştırılması için önemli olan performans ölçütleri anlatılmaya çalışılmıştır.

5.1. Lojistik Regresyon

Bir olayın değerlerine ilişkin tahmin yapmak, bu tahmini yapabilmek için hangi değişkenlerin daha iyi sonuçlar vereceğini ve kullanılması gerektiğini bilmek yapılan çalışma açısından önem taşımaktadır. Tahmin edilmek istenilen, bağımlı değişkenin kategorik yapıda olduğu durumlarda, örneğin; iki kategorili olduğu varsayılın, bu iki kategori bir olayın meydana gelip gelmeyeceğini belirtmektedir, lojistik regresyon kullanılmaktadır (Hosmer ve Lemeshow , 2013).

Regresyon modelleri değişkenler arasında ilişki olup olmadığını analiz etmek için kullanılan modellerdir. Klasik doğrusal regresyon modelinde bağımlı değişken sürekli yapıda olması, bağımlı değişken ile bağımsız değişkenler arasında doğrusallık olması, sabit varyans olması, normal dağılım olması varsayımları aranmaktadır. Aynı zamanda birçok alanda bağımlı değişkenin kategorik olduğu durumlar ile de karşılaşılmaktadır, bu durumlarda doğrusal regresyon modeli varsayımları sağlanamayacağından lojistik regresyon kullanılmaktadır (Cöhce, 2020).

Lojistik Regresyon birçok sorunla karşılaşılması durumunda, kolaylıkla kullanılabilir ve tercih edilebilecek bir sınıflama yöntemidir. Doğrusal regresyon modeli ile aynı amacı taşımaktadır fakat bahsedildiği gibi bağımlı değişkenin kategorik olması durumunda en çok kullanılan yöntemlerdendir. Bağımlı değişkenin kategori sayısı ve yapısına göre farklı LR yöntemleri de bulunmaktadır. Bağımlı değişkenin kategorik olmasına karşın bağımsız değişkenler için herhangi bir kistas, zorunluluk veya varsayım bulunmamaktadır. Bağımsız değişkenler hem sürekli hem de kesikli verilerden oluşabilmektedir. Bağımlı değişkenin kategorik olmasından kaynaklı olarak bağımlı ve bağımsız değişkenler arasında doğrusal bir ilişkinin olması zorunlu değildir.

5.1.1. Lojistik Regresyon'un kısa tarihçesi

Lojistik modelin tarihi 1840'lı yıllara kadar uzanmaktadır. Lojistik Regresyon analizi nüfus artışlarının takibi için matematiksel olarak açıklanabilir olması sebebiyle kullanılmıştır. 1944 yılında Berkson biyolojik deneyler için Lojistik Regresyonun kullanılmasını gerektiğini belirtmiştir. 1967 yılında, Truett ve arkadaşları diskriminant analizinde normal dağılım varsayımı sağlanmadığında Lojistik Regresyonun kullanılabilirliğini söylemiştir. Lojistik Regresyon modellerinin yaygın kullanımı sayesinde katsayıların tahmin etme yöntemlerinin geliştirilmesine ve daha açıklayıcı bir şekilde incelenmesine yönelik çalışmalar yapılmıştır. Cornfield, 1962 yılında diskriminant analizindeki yaklaşımları Lojistik Regresyondaki katsayıları tahmin etme işlemlerinde ilk kez kullanmıştır. Cox, 1970 yılında, biyolojik deneylerde analizi kullanmak için düzenlemeler yapmış ve uygulamaya başlamıştır. 1971 yılında Halperin ve arkadaşları, tekrardan diskriminant analizinde normal dağılım varsayımı sağlanmadığında Lojistik Regresyonun kullanılabilirliğini belirtmiştir. Lee, 1984 yılında basit dönüşümler yapmak amacıyla doğrusal lojistik modeller üzerinde çalışmalar yapmıştır. Bonney, 1987 yılında Lojistik Regresyon modelinin geliştirilmesi üzerinde çalışmalara devam etmiştir (Yüksel, 2019).

Cox ve Snell, 1989 yılında, iki kategorili bağımlı değişkene ait özellikler üzerinde durarak, bu tip bir değişkene ait kurulacak modelin lojistik analizini açıklamış, Çapraz

Çizelge ve birden fazla bağımsız değişkenli modellerde Lojistik Regresyon analizinin kullanılmasına yönelik çalışmalar yapmışlardır. Hosmer ve arkadaşları, 1989 yılında, çok sayıda açıklayıcı değişken içeren Lojistik Regresyon modelleri için en uygun değişken eşleşmeleri üzerinde durmuştur. 1989 yılında Hosmer ve Lemeshow, bugün bilinen basit LR modelinin tanıtımını yapmış, katsayılardan, tahmin yeteneğinden, çok değişkenli model olabileceğini, model aşamalarını, yorumlanmasını, çok kategorili bağımlı değişken olması durumunda modelin nasıl kurulması gerektiğine dair tüm teorik açıklamaları yapmıştır (Baydemir, 2014).

5.1.2. Lojit dönüşüm ve Lojistik Regresyon modeli

Doğrusal olasılık modelinde karşılaşılan sorunların çözümleri bulunurken en büyük sorun $0 < E(Y_i|X_i) < 1$ kısıtının sağlanamamasından kaynaklanmaktadır. Bu sebepten alternatif modellere yönelmek gerekmektedir, yapılan dönüşümlerde bağımsız değişken x_i değerleri $-\infty$ ile $+\infty$ arasında değişmekte ve bu sayede modelin koşullu olasılık değeri 0'a doğru azalmakta veya 1'e doğru artmaktadır ama hiçbir zaman 0 ve 1 değerinin dışına çıkmamaktadır. Bu istenilen düzen lojistik dağılım fonksiyonu ile sağlanabilmektedir (Pampel, 2000).

0 ve 1 arasında olasılık değerleri alabilen Eşitlik 5.1'deki denkleme 'doğrusal olasılık modeli' denilmektedir. Bağımsız değişkenlerin sınırsız değerler alması nedeni ile doğrusal olasılık modeli her zaman sağlanamamaktadır (Tatlıdil, 2002).

$$E(y_i) = P(y_i = 1) = \sum_{k=0}^p \beta_k x_{ik} \quad (5.1)$$

Doğrusal olasılık modelinin olasılık değerleri üzerinde lojit dönüşümü yapmak için $P/(1-P)$ olasılık dönüşümü yapılmalı, sonra bu oran değerinin doğal logaritması alınarak sonuç değişkeninin sınırları $-\infty$, $+\infty$ olarak belirlenmelidir. Eşitlik 5.1'de belirtilen doğrusal olasılık modeline lojit dönüşüm uygulandıktan sonra lojistik model ya da lojit olarak ifade edilen, aşağıdaki Eşitlik 5.2'de görülen model elde edilmektedir (Ekmekçi, 2019).

$$E(y_i) = L_i = \log \left(\frac{P_i}{1-P_i} \right) \sum_{k=0}^p \beta_k x_{ik} \quad (5.2)$$

P_i olasılık değeri;

$$P_i = \frac{\exp(\sum_{k=0}^p \beta_k x_{ik})}{(1 + \exp(\sum_{k=0}^p \beta_k x_{ik}))} \quad (5.3)$$

şeklinde Lojistik Regresyon fonksiyonu elde edilmektedir. Bu fonksiyona lojit dönüşüm uygulanarak doğrusallaştırma yapılabilir.

$$\eta = \ln \left(\frac{P_i}{1-P_i} \right) \quad (5.4)$$

$P_i / (1 - P_i)$: odds oranını vermektedir.

Dönüşümler uygulandıktan sonra;

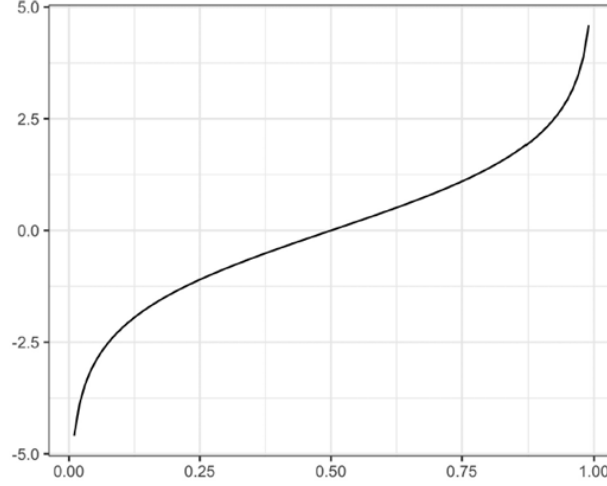
$$E(y_i) = \eta = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_k x_{ki} \quad (5.5)$$

$E(y_i)$, $-\infty$, $+\infty$ arasında yer almaktadır ve n birimlik bir örneklem için Lojistik Regresyon tahmin denklemi aşağıda Eşitlik 5.6'da görüldüğü gibidir:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + \dots + \hat{\beta}_k x_{ki} \quad (5.6)$$

Lojistik Regresyon denkleminde (Eşitlik 5.7) P_i değerinin 0,5’den büyük olup olmaması durumuna göre ikili sınıflandırmada; büyükse 1 küçükse 0 olarak şeklinde atama yapılmaktadır.

$$P_i = \frac{e^{y_i}}{1+e^{y_i}} \quad (5.7)$$



Şekil 5.1. Lojistik Regresyon fonksiyon grafiği

Lojistik Regresyon fonksiyon grafiği Şekil 5.1’de görüldüğü gibidir. Lojistik Regresyonun bağımlı değişkenin kategori sayısı veya sıralı olması durumuna göre bağımsız değişken veya değişkenler ile bağımlı değişkenin arasındaki ilişkiyi açıklamak için farklı türleri bulunmaktadır. Bağımlı değişkenin iki kategoriden oluşması durumunda İkili (Binary) LR Analizi, ikiden fazla kategori olması ve bu kategorilerin birbirinden bağımsız aralarında ilişki olmadığı durumlarda İsimsel (Nominal ve Multinomial) LR Analizi veya Çoklu Kategorik LR Analizi olarak da geçmektedir, bağımlı değişkenin kategorilerinin sıralı biçimde olması durumunda ise Çoklu Sıralayıcı (Multiordinal) LR Analizi kullanılmaktadır.

5.1.3. İkili (Binary) Lojistik Regresyon Analizi

Bağımlı değişkenin iki kategoriden oluşması durumunda İkili Lojistik Regresyon Analizi uygulanmaktadır. Bu iki kategori 0 ve 1’den oluşmakta ve genellikle olumsuzluk içeren değer 0 ve olması istenen değer 1 olarak atanmaktadır. Bu tez çalışması kapsamında da İkili Lojistik Regresyon Analizi kullanılmıştır ve satın alma yapan müşteriler için 1, satın alma yapmayan müşteriler için 0 değeri atanmaktadır. Bağımsız değişken için herhangi bir kistas yoktur sürekli veya kesikli kategorik değişkenler içerebilmektedir.

Lojistik Regresyonda bağımlı değişkenin 1 değerini alması durumunun olasılığı tahmin edilmeye çalışılır. Çıkacak olan sonuç olasılık değeri olduğu için bu olasılık değeri 0 ile 1 arasında değer alabilir (Alpar, 2017). x ’in değeri bilindiği durumda olayın olma olasılığı $Y=1$, η ile gösterilsin. η ’yi tanımlayabilmek için doğrusal model kullanılarak Eşitlik 5.8’deki model oluşturulabilir (Alpar, 2013):

$$\eta(x) = P\left(Y = \frac{1}{x} = x\right) = \beta_0 + \beta_1 x + e \quad (5.8)$$

Burada fonksiyon doğrusal olduğu için ve aynı zamanda sonuç 0 ile 1 arasında olmadığı için olasılığı modellemek için kullanılamaz. Bu sorunun ortadan kalkması için ağırlıklandırılmış en küçük kareler yöntemi kullanılabilir fakat burada da η bilinmediği için sorun yaratmaktadır. İki kategorili Lojistik Regresyon analizinde katsayı tahminlerinin yapılması için en çok olabilirlik, ağırlıklandırılmış en küçük kareler, minimum lojit ki kare yöntemleri kullanılmaktadır (Alpar, 2017).

5.1.4. Çoklu Sınıflayıcı (Multinomial) Lojistik Regresyon Analizi

Çoklu Sınıflayıcı ya da İsimli Lojistik Regresyon Analizi, bağımlı değişkenin iki den fazla kategorili yapıda olması durumunda kullanılan analizdir. Bu kategoriler arasında her hangi bir sıralama veya üstünlük bulunmaz. Bağımlı değişkenin kategorilerinin olasılıklarının toplamı yine 1’dir.

5.1.5. Çoklu Sıralayıcı (Multiordinal) Lojistik Regresyon Analizi

Bağımlı değişkenin yine iki den fazla kategoriye sahip olması durumunda kullanılabilir analiz türüdür. Buradaki Çoklu Sınıflayıcı Lojistik Regresyondan ayıran özellik kategoriler arasında bir sıralamanın bulunuyor olmasıdır.

Çoklu Sıralayıcı Lojistik Regresyon Analizinde, kategorilerin kendi aralarında kendilerine özgü bir sıralamaları olması ile birlikte paralel varsayımı da bulunmaktadır. B katsayıları farklı kategoriler için farklı kesme noktalarında değişiklik göstermez. Bu yüzden modeller arasında paralellik varsayımı bulunmaktadır. Bu varsayım yöntemin en belirgin özelliğidir (Yazar, 2018).

5.2. Naive Bayes

Naive Bayes sınıflandırması, bayes teoremine göre olasılıksal hesaplamalar yaparak sınıflandırma işlemi yapan bir yöntemdir. Naive Bayes sınıflandırmasında veriler sistemde belli bir dereceye kadar öğretilir. Öğretmek için mevcut verilerin bir sınıfı ya da kategorisi olmalıdır. Eğitim verisi üzerinde mevcuttaki her durumun olasılığına bakılır ve test verisi daha önceden bulunan olasılık değerlerine göre çalıştırılarak, bu şekilde bir sınıflandırma yapılmaya çalışılır. Öğretilen verilerin sayısı ne kadar çoksa, gerçek test verilerinin kategorisini belirlemek o kadar doğru olmaktadır. Naive Bayes sınıflandırma yönteminin birçok kullanımı mevcuttur, fakat önemli olan neyin sınıflandırıldığından daha çok nasıl sınıflandırıldığıdır. Yani öğrenilmesi istenen veriler ikili kategorili veya metin verileri olabilir, burada bu veriler ile hangi tür veriler arasında orantılı şekilde bir ilişki kurulduğu önemlidir (Çelik E. , 2019). Naive Bayes sınıflama yöntemi, bayes teoremine dayanan, makine öğrenmesi sınıflandırma yöntemleri içerisinde basit ve kullanışlı olarak bilinen bir yöntemdir. Bayes teoreminin bulunmasından sonra istatistiksel yaklaşımlarda klasik ve bayesyen olarak ayrılmıştır.

Bayesyen yaklaşımda, klasik istatistiksel yaklaşımdan farklı olarak tüm durumlar için sadece bayes teoremi kullanılır. İngiliz matematikçi Thomas Bayes (1702-1761) tarafından ortaya atılan bayes teoremi, tesadüfî bir değişkeni, olasılık kuralları içerisinde şartlı olasılıkları ile marjinal olasılıkları arasındaki ilişkiyi gösteren, deneme sonuçlarına göre elde edilen ek bilgiyi kullanarak son olasılıkları hesaplamaya dayalı teoremdir. Bayesyen yaklaşım, hem sahip olunan ilk bilgiyi hem de veriden elde edilen ek bilgiyi bayes teoremini kullanarak birleştirir (Zeytinoğlu, 2019).

Thomas Bayes vefat ettikten sonra arkadaşı Richard Price teorem hakkında çalışmalar yapmaya devam etmiştir ve 1763 yılında “Essay Towards Solving a Problem in the Doctrine of Chances” adlı bir çalışma yapmıştır. Daha sonrasında binom dağılımının olasılık parametresi için bir dağılımın nasıl hesaplanacağına dair çalışmalar yapılmaya başlanmıştır. Pierre-Simon Laplace 1812 tarihli “Théorie analytique des probabilités” başlıklı makalesi ile birlikte bayes teoremini geliştirmiştir. Bu gelişmeler sonrasında 1980’lerde finans piyasaları, tıp, sosyoloji, teknoloji, ekonomi, imalat ve diğer alanlarda uygulamalı istatistiklere artan ilgi ve uygulamalar yapılmaya başlanmıştır (HolyPython, 2020).

A olayının gerçekleşme olasılığı biliniyor ve B olayının gerçekleşme olasılığı A olayına bağlı ise Eşitlik 5.9’daki gibi gösterilmektedir (Turanlı, 1984):

$$P(B|A) = \frac{P(A \cap B)}{P(A)} \quad (5.9)$$

buradaki eşitlikten yola çıkarak A ve B durumlarının kesişimlerinin olasılığı Eşitlik 5.10’daki gibidir:

$$P(A \cap B) = P(B|A).P(A) \quad (5.10)$$

aynı şekilde A olayına ilişkin olasılık ve sonucunda kesişim olasılığı Eşitlik 5.11’deki gibi elde edilir:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \quad (5.11)$$

$$P(A \cap B) = P(A|B) \cdot P(B) \quad (5.12)$$

böylece; B olayının koşullu olasılık denklemi Eşitlik 5.13'deki gibi elde edilmektedir.

$$P(B|A) = \frac{P(A|B) \cdot P(B)}{P(A)} \quad (5.13)$$

P(A): A olayının gerçekleşme olasılığı

P(B): B olayının gerçekleşme olasılığı

P(A|B): B olayı gerçekleştiğinde A olayının gerçekleşme olasılığı

P(B|A): A olayı gerçekleştiğinde B olayının gerçekleşme olasılığı (Turanlı, 1984)

Bayes teoremi, eşikliklerde görüldüğü gibi istatistikte kullanılan koşullu olasılıkları kullanan bir teoremdir. Bu teoremden faydalanılarak sınıflandırma işlemi yapılabilmektedir (Erken, 2017).

Açıklama kolaylığı sağlamak için, $i = 0, 1$ olarak iki sınıf olduğu varsayalım. Amaç, daha büyük puanların sınıf 1 nesneleriyle ve daha küçük puanlar; sınıf 0 nesneleriyle ilişkilendirilecek şekilde bir puan oluşturmak için bilinen sınıf üyeliklerine sahip ilk nesne kümesini (eğitim kümesi) kullanmaktır. Sınıflandırma daha sonra bu puanın bir eşik t değeri (0,5) ile karşılaştırılmasıyla elde edilir. $P(i|x_j)$ fonksiyonu $x_j = (x_1, x_2, \dots, x_p)$ vektöründen oluşmaktadır ve i 'ye ait olma olasılığını göstermektedir.

Bu durumda $P(i|x)$ bir skor oluşturacaktır. Bu skor özellikle $\frac{P(1|x)}{P(0|x)}$ oranı şeklinde olacaktır. Temel olasılık $P(i|x)$ fonksiyonunu $f(x|i)P(i)$ ile orantılı olarak ayrıştırılabileceğini söylemektedir, burada $f(x|i)$, x 'in i sınıfı için koşullu dağılımını vermektedir ve $P(i)$ hakkında bir şeyler bilinmiyorsa bu dağılım x 'in i sınıfına ait olma olasılığını göstermektedir. Böylelikle oran Eşitlik 5.14'deki gibi yazılabilir (Wu vd., 2008):

$$\frac{P(1|x)}{P(0|x)} = \frac{f(x|1)P(1)}{f(x|0)P(0)} \quad (5.14)$$

Burada sınıflandırma yapabilmek için $f(x|i)$ ve $P(i)$ 'nin tahmin edilmesi gerekmektedir. $P(i)$ eğitim verisi içerisindeki i nesnelerin oranından tahmin edilebilir. $f(x|i)$ 'yi tahmin etmek için bayes yöntemi ile ve x 'in bileşenlerinin birbirinden bağımsız olduğu varsayılarak, $f(x|i) = \prod_{j=1}^p f(x_j|i)$ ve ardından tek değişkenli dağılımların her birini $f(x_j|i)$ ayrı ayrı tahmin edilir, burada $j = j_1, j_2, \dots, j_p$ ve $i = 0,1$ 'dir. Böylece p boyutlu çok değişkenli problem, p tek değişkenliye indirgenmiştir (Wu vd., 2008).

Marjinal dağılımlar $f(x_j|i)$ bağımsızsa ve her x_j sadece birkaç değer alıyorsa, o zaman $\hat{f}(x_j|i)$ tahmini, çok terimli bir histogram tipi tahmin edicisidir ve bu tahmin edici her x_j 'nin her hücreye düşen oranını gösterir (Wu vd., 2008).

Bağımsızlık varsayımı verildiğinde, oran Eşitlik 5.15'deki gibi olmaktadır:

$$\frac{P(1|x)}{P(0|x)} = \frac{\prod_{j=1}^p \frac{f(x_j|1)P(1)}{f(x_j|0)P(0)}}{\prod_{j=1}^p \frac{f(x_j|1)P(1)}{f(x_j|0)P(0)}} = \frac{P(1)}{P(0)} \prod_{j=1}^p \frac{f(x_j|1)}{f(x_j|0)} \quad (5.15)$$

Amaç $P(i|x)$ ile monoton olarak ilişkili bir skor üretmek olduğu için, Eşitlik 5.15'in loglarını alınmaktadır (log monoton artan bir fonksiyondur). Bu işlem Eşitlik 5.16'daki gibi gösterilmektedir:

$$\ln \frac{P(1|x)}{P(0|x)} = \ln \frac{P(1)}{P(0)} + \sum_{j=1}^p \ln \frac{f(x_j|1)}{f(x_j|0)} \quad (5.16)$$

Burada ise $w_j = \ln \left(\frac{f(x_j|1)}{f(x_j|0)} \right)$ ve sabit $k = \ln \left(\frac{P(1)}{P(0)} \right)$ olarak tanımlanırsa:

$$\ln = \frac{P(1|x)}{P(0|x)} = k + \sum_{j=1}^p w_j \quad (5.17)$$

denklemini elde edilir ve böylece yapılan sınıflandırıcı işlem sayesinde model basit bir yapıya dönüştürülür (Wu vd., 2008).

Naive Bayes sınıflandırıcısının çalışma şekli, her bir x_j elemanı için koşullu dağılımın bağımsız tek boyut olarak, yani her durumun olasılığını hesaplayarak ve bu olasılık değerleri içerisinde en yüksek olana göre sınıflandırma yaparak, gerçekleştirdiği için uygulaması oldukça hızlı ve kolay bir yöntemdir. Bu yöntem her bir dağılımı bağımsız olarak ele aldığı için boyutsallıktan kaynaklanan sorunlara daha az rastlanılmakta ve çok fazla eğitim verisine ihtiyaç duymadan, az sayıda oluşan veri setlerinde de başarılı sonuçlar elde edebilmektedir. Çok büyük veri setleri için de kullanışlı bir yöntem olan NB aynı zamanda çok sınıflı tahminlerde başarılı sonuçlar vermektedir. Dağılıma yönelik varsayımlarına ve mevcuttaki verilere göre Gaussian NB, Multinomial NB, Bernoulli NB olarak farklı türleri bulunmaktadır.

5.3. Destek Vektör Makineleri

Destek vektör makineleri, regresyon ve sınıflandırma amacıyla kullanılan ilişkili öğrenme algoritmalarıyla çalışan bir yöntemdir. Bu yöntem eğitim verisi üzerinde gözlemleri kategorilere atayarak model oluşturmakta ve bu modeli olasılıksız ikili doğrusal sınıflandırıcı ile yapmaktadır.

SVM modeli haritalama yönteminden yararlanır, örnekleri uzaydaki noktalar olarak gösterir ve bu sebepten gözlemler geniş ve aralıklarla bölünmektedir. Test verisindeki gözlemler bu aralığın hangi tarafında bulunduğu göre sınıflandırma işlemi yapılmaktadır. (Jordan vd., 2008).

SVM yöntemi eğitim verisi üzerinde belirtilen iki sınıfı ayırmak için bu iki sınıf arasından geçen hiper düzlemler belirler ve bu hiper düzlemler içerisinde en iyi ayırıcı

hiper düzlemi bulmaya çalışır. Hiper düzlemi bulmaya çalışırken gözlem dağılımlarının karmaşıklık ve saçılımlarına göre hiper düzlemler ile gözlemler arasında marj uzaklıkları hesaplar. Eğitim verisi üzerinde en iyi ayırıcı karar fonksiyonu yardımıyla doğrusal olarak verileri ayırarak doğru sınıfı bulmayı hedefler ve verileri gruplamak için en uygun ayırma şeklini sağladıktan sonra test verisi üzerinde bu kriteri uygulamakta ve böylece sınıflandırma işlemi gerçekleşmiş olmaktadır.

5.3.1. Destek Vektör Makineleri'nin kısa tarihçesi

Destek Vektör Makineleri hakkında ilk çalışmaların yapılması, geliştirilmesi Vapnik tarafından sağlanmıştır. Bu dönemde Lojistik Regresyon, sinir ağları gibi kavramlar ortaya çıkmış ve bu alanlarda çalışmalar yapılmaktayken, verilerin sınıflandırılması üzerine de çalışmalar devam etmiştir.

1963 yılında Vladimir Vapnik destek vektör makinesi olan başka bir sınıflandırma yöntemi geliştirmiştir. Vapnik ve Lerner, SVM ile uygulanan bir algoritma ile Genelleştirilmiş Portre algoritmasını geliştirmiştir. 1964 yılında Vapnik, Chervonenkis ile birlikte bulduğu Genelleştirilmiş Portre algoritması üzerinde çalışmalar yaparak bu algoritmayı geliştirmeye devam etmiştir. Aynı yıl içerisinde yöntem içerisindeki çekirdek kavramı Aizerman, Braverman ve Rozonoer tarafından uzay içerisinde iç ürün olarak çekirdeklerin geometrik tanımlarını yapılmaya başlanmıştır (Chervonenkis, 2013).

1968 yılında Smith tarafından, ayrılmayan değişkenler ve gürültülü değişken sorunlarına yönelik gevşek değişken kavramı ortaya konulmuştur. Vapnik 'İstatistiksel Öğrenme Teorisi' üzerinde çalışmalar yapmaya sürekli olarak devam etmiş ve bu teorileri sürekli olarak geliştirmiş ve yaptığı çalışmalar zamanla Rusçadan, İngilizce ve Almancaya çevrilmiştir.

Bu süreç içerisinde marj kavramı ve hiper düzlem kavramları bir çok kişi tarafından tartışılmaya başlanmış ve bir kaç tane istatistik makalesinde, girdi verileri üzerinde

büyük marjlı hiper düzlemlerin kullanılması gerektiği üzerinde durulmuştur. Daha sonra çekirdek kavramı ve gevşek değişken kavramları üzerinde çeşitli çalışmalar yapılmıştır.

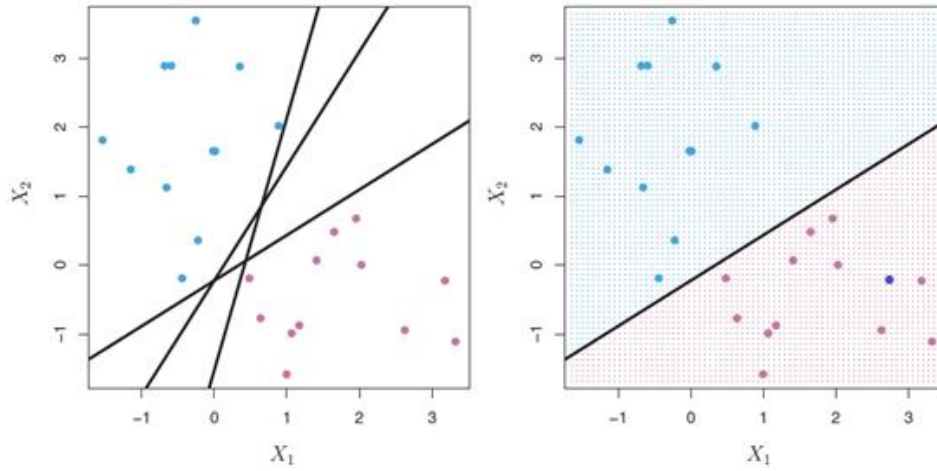
1995 yılında çalışmalarına devam eden Vapnik, Cortes ile birlikte yumuşak marj sınıflandırıcısından bahsetmiş ve bu kavramın regresyonun doğasında olduğu şeklinde de yorumlar yapmıştır. 1998 yılında, istatistik alanında çalışmalar yapan Bartlett, Shawe-Taylor kesin marjlı SVM'lerin istatistiksel sınırlarını belirlemiştir. Bartlett (1998) ve Shawe-Taylor, vd. (1998), kesin marjlı SVM'lerin genelleştirilmesine ilişkin istatistiksel sınırı belirtmiştir. 2000 yılında Shawe-Taylor ve Cristianini yumuşak kenar boşluğu ve regresyon için çalışmalar yapmış ve belirli sınırların olduğunu göstermiştir.

5.3.2. Doğrusal Destek Vektör Makineleri

Doğrusal destek vektör makinelerini incelemekten önce hiper düzlem kavramının açıklanması gerekmektedir. Bir p boyutlu uzay olduğu varsayıldığında, hiper düzlemin alt uzayı $p-1$ 'dir. Örneğin; 3 boyutlu bir uzayda 2 adet hiper düzlem bulunmaktadır. Bir p boyutlu uzay denklemini Eşitlik 5.18'de görüldüğü gibi ifade edilebilir:

$$\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p = 0 \quad (5.18)$$

Buradaki β katsayıları hiper düzlemi temsil etmektedir. $x = x_1, x_2, \dots, x_p$ p boyutlu uzayda hiper düzlemin geçtiği noktalar. Denklemden de görüldüğü gibi bu denklem bir doğru denklemdir. p boyutlu uzayı ikiye bölen hiper düzlem bu şekilde ifade ediliyorsa, amaç sınıflandırmayı yapabilmek için gözlemin hiper düzlemin sağında veya solunda kalması durumuna bakmak ve daha sonrasında hangi sınıfa dahil olduğunu belirtmektir. Diğer analizlerde olduğu gibi, burada da amaç mevcut verilerin öznitelik ölçümlerini kullanarak en iyi sınıflandırmayı yapmaktır. Diğer yöntemlerden farklı olarak SVM'de hiper düzlemler ile sınıfları ayırma işlemi yapılmaktadır. Şekil 5.2'de ise bu hiper düzlem ve iki sınıfa ayırması görsel olarak gösterilmektedir.



Şekil 5.2. Veri seti içerisindeki birden fazla hiper düzlemin gösterimi (James vd., 2013)

Şekil 5.2’de de görüldüğü gibi birçok hiper düzlem bulunmakta ve amaç en iyi ayırıcı hiper düzlem fonksiyonunu bulup, sonra hiper düzlemin sağında veya solunda kalma durumuna bakılarak en iyi sınıflandırmayı yapabilmektir. Bu durum matematiksel olarak aşağıdaki eşitliklerde gösterilmiştir:

$$i = 1, 2, \dots, n$$

$$y = 1 \text{ ise}$$

$$\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_p X_{ip} > 0 \quad (5.19)$$

$$y = -1 \text{ ise}$$

$$\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_p X_{ip} < 0 \quad (5.20)$$

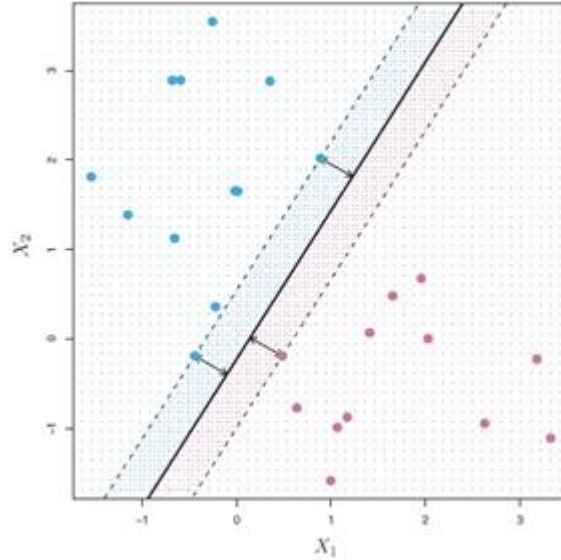
Ayırıcı hiper düzlem:

$$y_i(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_p X_{ip}) > 0 \quad (5.21)$$

Şekil 5.2’de mor alan ile belirtilen sınıf, $y_i = -1$ olarak tanımlanırsa Eşitlik 5.19 ve Eşitlik 5.20’de görüldüğü gibi 0’dan büyük olması durumunda mavi sınıfa ait olduğu, 0’dan küçük olması durumunda ise mor sınıfa ait olduğu görülmektedir.

Veriler içinde oluşan sonsuz sayıda hiper düzlem vardır, bu sonsuz hiper düzlemlerden bizim için en doğru ayırıcı olana karar verilmesi gerekmektedir. Bunun için, her bir eğitim gözleminden ayırıcı hiper düzleme olan (dik) mesafesi hesaplanır ve bu mesafe marj olarak adlandırılır. Eğitim gözleminin hiper düzleme olan minimum uzaklığı en çok mesafeye sahip olan hiper düzlem maksimal marj hiper düzlem (aynı zamanda optimal ayırıcı hiper düzlem) olarak adlandırılır. Sınıflandırma yapabilmenin temelinde de bu maksimal marj hiper düzlemi bulunmaktadır. Bu aynı zamanda maksimum marj sınıflandırıcı olarak da bilinmektedir. Maksimal ayırıcı hiper düzlem bulunursa bu durumda, doğru sınıflandırma yapılabilmesi beklenmektedir. Maksimal marj sınıflandırıcı genellikle başarılı sonuçlar vermektedir ve aynı zamanda aşırı uyuma neden olabilmektedir (James vd., 2013).

$\beta_1, \beta_2 \dots \beta_p$ maksimal marj sınıflandırıcının katsayıları ise test verisindeki x_i 'ye bağlı gözlemler için fonksiyonu $f(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots \beta_p x_p$ şeklinde olacaktır. Bu anlatılanlar Şekil 5.3'te görüldüğü gibidir:

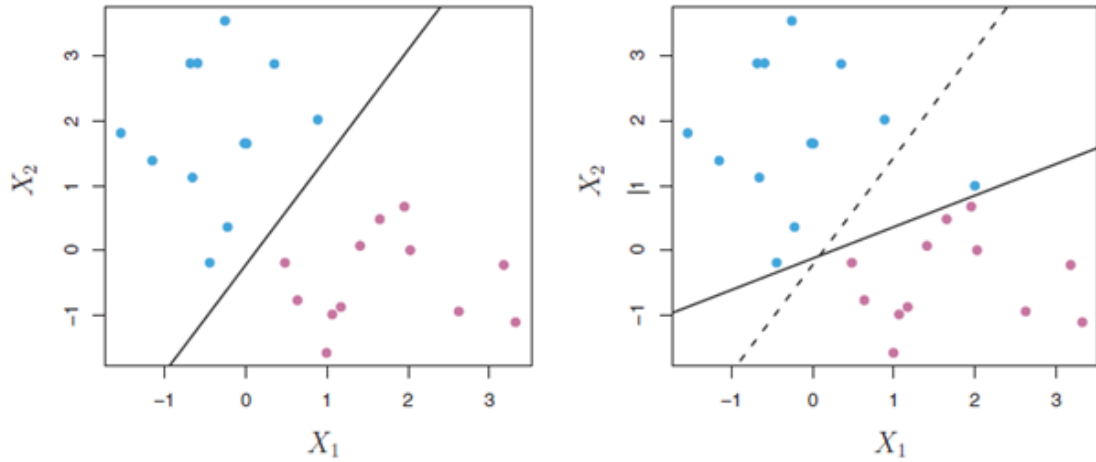


Şekil 5.3. Maksimal hiper düzlemi kenar boşluğu destek vektörleri gösterimi (James vd., 2013)

Şekil 5.3, veri setindeki maksimal hiper düzlemini göstermektedir. Burada karar verici hiper düzleme paralel olan her bir sınıfın en yakın veri noktalarına dokunan iki hiper düzlem arasındaki mesafe (kenar boşluğu olarak adlandırılır) belirler. En iyi karar hiper düzlemin, maksimum marjlı olanıdır. Hiper düzlemi bu şekilde tanımlayarak buradan, SVM yöntemi ile görünmeyen örnekler oldukça etkili bir şekilde genellenebilir (Kuramochi ve Karypis, 2005).

Maksimal hiper düzleme paralel olan iki düzleme (kesik çizgiler ile gösterilen) destek vektörleri denir. Maksimal hiper düzlem, gözlemlere yapılan herhangi bir müdahale sonucunda etkilenmez fakat destek vektörleri noktaları değişirse etkilenerek hareket eder.

Diğer taraftan her zaman veriler düzgün dağılmayabilir ve ayırıcı hiper düzlem olsa bile sınıflama yaparken yanlış sonuç çıkartılabilecek durumlar görülebilir. Şekil 5.4’de görüldüğü gibi, bir gözlemin eklenmesi maksimal hiper düzlemde değişikliğe yol açmaktadır. Yeni hiper düzlem tatmin edici olmayabilir. Bir hiper düzlemin, bir gözleme göre hassas olması (Şekil 5.4’de sağ tarafta gösterildiği gibi), eğitim verisine fazla uymuş olduğunu da gösterebilir. Bu durumda, iki sınıfı mükemmel bir şekilde ayırmayan bir hiper düzleme dayalı sınıflandırıcıyı dikkate almak gerekmektedir.



Şekil 5.4. Veri setine eklenen yeni gözlemin hiper düzlem üzerinde etkisi (James vd., 2013)

Böyle durumlarda sınıflandırmada daha iyi bir sonuç çıkarmak için birkaç eğitim gözlemini yanlış sınıflandırmak faydalı olabilir. Destek vektör sınıflandırıcısı, bazen yumuşak kenar boşluğu sınıflandırıcısı olarak da adlandırılır. Gözlemin, hem hiper düzlemin hem de marjın doğru tarafında olması her durum için sağlanmadığında bazı gözlemler için hiper düzlemin yanlış tarafında olması izin verilen bir durumdur. Hiper düzlem ve marja ilişkin verilen bu bilgiler ışığında aşağıda verilen eşitlikler yazılabilir:

$$\beta_1, \beta_2 \dots \beta_p, \epsilon_1, \dots \epsilon_n, M \quad (5.22)$$

$$\sum_{j=1}^p \beta_j^2 = 1 \quad (5.23)$$

$$y_i(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_p X_{ip}) \geq M(1 - \epsilon_i) \quad (5.24)$$

$$\epsilon_i \geq 0, \sum_{i=1}^n \epsilon_i \leq C \quad (5.25)$$

Bu eşitliklerde C negatif olmayan bir ayar parametresi, M ise marj genişliğini temsil etmektedir. Eşitlik 5.24’de bulunan $\epsilon_i = \epsilon_1, \epsilon_2, \dots \epsilon_n$ bireysel gözlemlerin açık olmasına izin veren gevşek değişkenlerdir. Eşitlik 5.21 ve 5.24 eşitliklerini çözüp gözlemlerin hiper düzlemin neresinde olduğuna bakılmaktadır. Test gözlemi $f(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots \beta_p x_p$ formülüne göre sınıflandırılmaktadır. C, ϵ_i ‘lerin toplamı ve yapılan ihlallerin şiddetini belirleyerek, marjın ne kadar tolere edilebileceğini göstermektedir. C arttıkça marjın ihlalleri tolere etmesi artar ve bu nedenle marjlar genişler, C azaldıkça, marjın ihlalleri tolere etmesi azalır ve bu nedenle marj daralır. Uygulamada, C genellikle çapraz doğrulama yoluyla seçilen bir ayar parametresi olarak kabul edilir. C istatistiksel öğrenme tekniğinin sapma-varyans dengelemesini kontrol eder (James vd., 2013).

5.3.3. Doğrusal Olmayan Destek Vektör Makineleri

Veri setini her zaman bir hiper düzlem ile doğrusal olarak ayırmak mümkün değildir. Bu sebepten sınıflandırma, ayırma eğrisinin tahmini ile yapılmaktadır (Ayhan ve Erdoğan, 2013).

2014). Sınıflar arası doğrusal olmayan sınırlar polinomial, üssel gibi dağılımlara sahip olabilirler. Örneğin, p özelliklerini kullanarak bir destek vektör sınıflandırıcısı $x_1, x_2, \dots, x_p$ ve $x_1, x_1^2, x_2, x_2^2, \dots, x_p, x_p^2$ şeklinde olabilir. Burada SVM kullanılarak, özellik alanını belirli bir şekilde genişletmesinden kaynaklanan, destek sınıflandırıcısı bulunabilir ve bu sınıflandırıcı çekirdek yaklaşımı tanımlanmaktadır. Yukarıda belirtilen p uzayında $x_{ij}x_{i'j}$ şeklindeki vektörel gözlemler için aşağıdaki formüller tanımlanmaktadır:

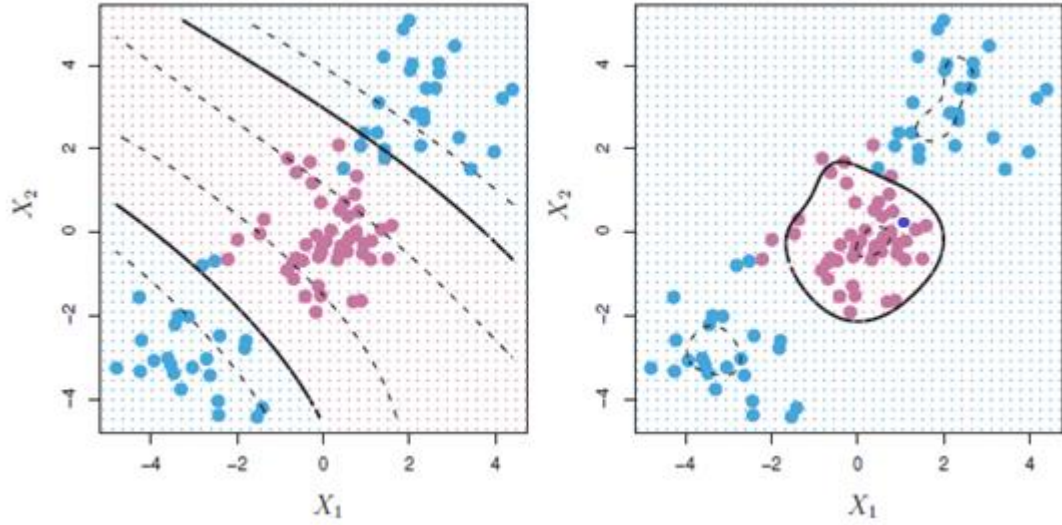
$$\langle x_{ij}x_{i'} \rangle = \sum_{j=1}^p x_{ij} x_{i'j} \quad (5.26)$$

$$f(x) = \beta_0 + \sum_{i \in S} a_i \langle x, x_i \rangle \quad (5.27)$$

Yukarıdaki Eşitlik 5.27'da görülen S değeri, bu destek noktalarının endekslerinin toplamını vermektedir ve burada doğrusal sınıflandırıcı $f(x)$ 'i temsil etmede ve hesaplamada en önemli olan tek şey eşitlikte görüldüğü gibi iç çarpımlardır. Diğer taraftan daha önceki açıklamalarda verilen çekirdek kavramında en çok kullanılan Kernel çekirdek'dir, K ile gösterilir ve Eşitlik 5.28'de görüldüğü gibidir:

$$K(x_{ij}x_{i'}) \quad (5.28)$$

Eşitlik 5.28'de görülen ifade kullanılarak çekirdek formülü ile farklı dağılımlara sahip ve farklı yapıda olan, doğrusal olmayan veri setleri üzerinde sınıflandırma işlemi yapmak mümkündür. Şekil 5.5'deki sağdaki grafiğe bakılırsa SVM'nin kernel çekirdek ile doğrusal olmayan veride de güzel bir şekilde ayırıp iki sınıf çıkardığı görülmektedir.



Şekil 5.5. Doğrusal olmayan SVM sınıflandırması (James vd., 2013)

5.4. Rastgele Ormanlar

Rastgele Ormanlar yöntemi, diğer sınıflandırma yöntemleri gibi denetimli makine öğrenmesi yöntemi olarak sınıflandırma ve regresyon uygulamak amacıyla kullanılmaktadır. Bu yöntem eğitim verisi içerisinde birbirinden bağımsız rastgele örneklemelere dayanan karar ağaçlarından oluşmaktadır. Karar ağacı gibi yöntemlerde en iyi özelliklere göre dallandırma yapılırken, rastgele ormanlar yönteminde bir düğümü parçalara ayırmada rastgele bir alt küme kullanılmaktadır. Rastgele Ormanlar yöntemi ve karar ağaçları, adaboost, boosting gibi yöntemleri Breiman temelde benzer özelliklerden yola çıkarak geliştirmiştir ve bu yöntemler topluluk (ensemble) algoritmaları olarak literatürde yer almaktadır.

Breimen bu yöntemi oylama yapma örneği ile literatüre kazandırmıştır. Bu sebepten oylamaya yönelik tanımlar kaynaklarda yer almaktadır. Rastgele Ormanlar yönteminde model oluşturmak için veri seti içerisinde alt uzaylar oluşturulmakta ve rastgele seçim yapılarak ağaçlar oluşturulmaktadır. Diğer bir ifade ile eğitim verisi üzerinde birçok karar ağacı oluşturulmakta ve bu ağaçlandırmalardan yararlanılarak çoğunluk oylara göre sınıflandırma sonuçlarına ulaşılmaya çalışılmaktadır. Rastgele Ormanlar

yönteminde aykırı değer, aşırı gürültü ve aşırı uyum sorunları yaşanmamakta ayrıca Karar ağaçları ve AdaBoost yönteminden daha doğru sonuçlar vermekte ve torbalama (bagging), boosting yöntemlerine göre daha hızlı çalışmaktadır (Breiman, 2001).

Karar ağaçları, verisetindeki değişkenler üzerinden bir ağaç oluşturarak tahminler elde eden bir yöntemdir. Veride bir değişken kök olarak seçilerek ağaçlandırmaya başlanır. Dallanma ise düğüm olarak adlandırılan noktalar ile sağlanmaktadır (Bell, 2014). **Torbalama (bagging)** yönteminde ise eğitim seti bootstrap örnekleme ile daha küçük parçalara bölünür ve sonrasında bu farklı eğitim setleri kullanılarak modeller oluşturulur. Model tahmin sonuçları oylanarak ya da ortalamaları alınarak birbirinden farklı veri ile oluşturulan eğitim modelinin tek bir eğitim veri setine göre daha iyi sonuç vereceği ve bu sayede varyansının azalacağı amaçlanmaktadır. **Adaboost** yöntemi, torbalama yöntemi gibi çalışmaktadır. Bu yöntem test sonucunda yanlış çıkan tahminleri tekrar test eder ve sadece doğru tahmin edilen sonuçları dikkate alır. **Boosting** yöntemi de torbalamaya benzer şekilde çalışır. Bu yöntemde ağaçlandırma yapılırken her aşamada elde edilen özellikler kullanılarak etkileşimli bir şekilde büyüme sağlanır (Breiman, 1996).

$Q_1, Q_2, \dots, Q_{j-1}$ aynı dağılıma sahip ve bağımsız rastgele vektörlerdir. j. ağaç için Q_k rastgele vektörü oluşturulur ve x 'in giriş vektörü olduğu $h(x, Q_j)$ sınıflandırıcı için, eğitim veri seti ve Q_j kullanılarak bir ağaç oluşturulmaktadır. Torbalama (bagging) yönteminde N eğitimli veri setinde kutulara atılan oyların N gözlemde elde edilen N miktarında sayımlar rastgele vektör Q ile oluşturulur. Rastgele Ormanlar yönteminde 1 ile j arasında rastgele bağımsız tam sayıdan Q vektörü oluşturulabilmektedir. Bu yöntemde hem torbalama hem de özellik seçilmesi birlikte kullanılmaktadır. Torbalamadan farklı olarak torba dışı hataları (Out-of-bag (OOB)) hesaplayabilmekte ve kullanılan rastgele değişken seçiminin doğruluk oranını artırdığı görülmektedir (Breiman, 2001).

Rastgele Ormanlar yöntemi, çoklu sınıflandırma ve regresyon yapabilme, büyük boyutlu veri setleri için kullanılabilme, eğitim ve test aşamalarında hızlı çalışabilme, farklı sınıflar için ağırlıklandırma yapabilme ve görselleştirme sunmak gibi özelliklerinden dolayı tercih edilen bir yöntemdir (Cutler vd., 2011).

5.4.1. Rastgele Ormanlar'ın kısa tarihçesi

Yukarıda açıklandığı gibi Rastgele Ormanlar yöntemi, 2001 yılında Breiman tarafından uygulanmaya başlanmıştır. Ancak 2001 yılından önce bu konuda birçok bilim adamı tarafından çeşitli çalışmalar yapılmıştır. Ho 1995 yılında, “Rastgele Orman” terimini kullanmış ve 1998 yılında her ağacı büyütmek için kullanılacak özelliklerin bir alt kümesinin rastgele bir seçimini yapan “rastgele alt uzay” yöntemi hakkında bir dizi makale yazmıştır. 1990 yılında Kleinberg ve daha sonra 1997 yılında Amit ve Geman aşırı uyumu önlemek için ortalamaların kullanılması fikrini öne sürmüştür. Daha sonra görüntüleri tanımlayabilmek için rastgele ağaçları sınıflandırma amaçlı kullanmışlardır. 1996 yılında Breiman, Rastgele Ormanların torbalama yöntemini tanıtmıştır. 1998 yılında Dietterich her düğümde en iyi k bölmeleri arasından rastgele bölme seçimi yapılmasına dair ek randomizasyon ile torbalama yöntemine provizyon yapmıştır. Friedman ve Hall, 2007 yılında, alt örnekleme (değiştirilmeden) torbalama için etkili bir alternatif olduğunu göstermiştir (James vd., 2013).

Günümüzde Rastgele Ormanlar yönteminin en belirgin şekilde uygulanmasını sağlayan Breiman herkes tarafından bilinmekte ve çalışmaları Rastgele Ormanlar yöntemi için ana kaynak olarak görülmektedir.

5.4.2. Rastgele Orman algoritması

Adından da anlaşılacağı gibi, rastgele bir orman, rastgele değişkenlerin bir yapıya bağlı olduğu ağaç tabanlı bir topluluktur. p boyutlu rastgele bir vektör için $X = (X_1, X_2, \dots, X_p)^T$ yanıtı temsil eden rasgele bir değişken y için ortak dağılım $P_{XY}(X, Y)$ şeklinde olduğu varsayılmaktadır. Amaç, y 'yi tahmin etmek için bir tahmin

fonksiyonu $f(X)$ bulmaktır. Tahmin fonksiyonu, bir kayıp fonksiyonu $L(Y, f(X))$ tarafından belirlenir ve kaybın beklenen değerini en aza indirmek için Eşitlik 5.29'daki gibi tanımlanır:

$$E_{XY}(L(Y, f(X))) \quad (5.29)$$

Eşitlik 5.29'da görülen $l(y, f(X))$, $f(X)$ 'in Y 'ye ne kadar yakın olduğunun bir ölçüsüdür; Y 'den uzak olan $f(X)$ değerlerini cezalandırır. L 'nin tipik seçimleri, regresyon için $L(Y, f(X)) = (Y - f(X))^2$ hata kaybının karesidir ve sınıflandırma için sıfır bir kayıptır.:

$$L(Y, f(X)) = I(Y \neq f(X)) = \begin{cases} 0, & Y = f(X) \\ 1, & \text{diğer durumlarda} \end{cases} \quad (5.30)$$

Kare hata kaybı için $E_{XY}(L(Y, f(X)))$ en aza indirmenin koşullu beklentiyi verdiği Eşitlik 5.31'deki gibi ortaya çıkmaktadır:

$$f(x) = E(Y|X = x) \quad (5.31)$$

aksi takdirde regresyon fonksiyonu olarak bilinir. Sınıflandırma durumunda, Y 'nin olası değerleri kümesi y ile gösterilir, sıfır bir için $E_{XY}(L(Y, f(X)))$ Eşitlik 5.32'deki gibi en aza indirilir:

$$f(x) = \operatorname{argmax}_{y \in Y} P(Y = y|X = x) \quad (5.32)$$

şeklinde. Aksi takdirde bayes kuralı olarak bilinir. Topluluklar, temel öğrenen $h_1(x), \dots, h_J(x)$ 'ler ile $f(x)$ oluşturulur. Sınıflandırma en çok tahmin eden sınıf $f(x)$ 'dir.

$$f(x) = \operatorname{argmax}_{y \in Y} \sum_{j=1}^J I(y = h_j(x)) \quad (5.33)$$

Rastgele Ormanlarda, j 'inci temel öğrenci, Q_j 'nin olduğu $h_j(X, Q_j)$ ile gösterilen bir ağaçtır rastgele değişkenlerin bir koleksiyonu ve Q_j 'ler $j = 1$ için bağımsızdır, j . rastgele bir ormanın tanımı çok genel olsa da, her zaman açıklanan şekilde uygulanırlar. Rastgele Orman algoritmasını anlamak için, temel öğrenenler olarak kullanılan ağaç türü hakkında temel bilgiye sahip olmak önemlidir (Cutler vd., 2011).

5.5. Sınıflandırma Model Performanslarının Değerlendirme Ölçütleri

Uygulanan makine öğrenmesi sınıflandırma yöntemlerinin hem tek başlarına hem de birbirleri arasında hangisinin daha başarılı sonuçlar verdiğini tespit edebilmek açısından bazı ölçüt değerlerine bakılması gerekmektedir. Bu değerler uygulanan sınıflandırma yönteminin başarısına dair değerlendirme yapılması ve sınıflandırma yöntemlerinin karşılaştırılması için kullanılmaktadır. Bu değerleri tek başına başarı ölçütü olarak değerlendirmek yanlış olacağı için bir den fazla değerlere bakılması gerekmektedir. Başarılı bir sınıflandırma yöntemi eğitim ve test verisi için en iyi doğru sonucu verecek şekilde ayrılmalıdır. Sınıflandırma yönteminin performansını ölçerken veri setinin dengeli olması model parametrelerini en iyi şekilde açıklaması gerekmektedir.

5.5.1. Holdout yöntemi

Denetimli makine öğrenmesi mevcut verileri eğitim ve test verisi olarak ikiye ayırmakta, ayrılmış olan eğitim verisi ile makine öğrenmesi algoritmalarını kullanarak bir model oluşturulmaktadır. Sonrasında makinenin daha önce görmemiş olduğu test verisi üzerinde bu model uygulanarak tahmin yapılmakta olduğuna dair önceki bölümlerde açıklamalar yapılmıştır. Veri setini eğitim ve test verisi olarak ayırmak için Holdout yönteminden farklı olarak Tekrarlı Holdout, Tabakalı Örnekleme, Üçlü Ayırma gibi farklı yöntemler de bulunmaktadır. Verilerin eğitim ve test olarak ayrılması veri setinin boyutuna ve dağılımına göre değişiklik göstermektedir. Veriyi doğru ayırmak analiz sonucunda elde edilebilecek doğruluk değerlerini de etkilemektedir. Doğru ayrılmayan bir veri seti, yanlış boyuttaki eğitim verisi ile oluşturulan model test verisi ile doğru sınıflandırma yapmayabilir ve modele ilişkin parametreler doğru belirlenmeyip

performans ölçütünü etkileyebilirler. Farklı veri setleri farklı oranlarda ayrılarak en doğru sonuçları verebilirler. Bu sebeplerden dolayı veri setini eğitim ve test olarak ayırmak oldukça önemli bir adımdır. Veri seti boyutuna göre veri %80 eğitim, %20 test, %70 eğitim %30 test verisi gibi oransal olarak ayrılabilir. Burada makinenin modeli oluşturduğu kısım eğitim verisi olduğu için büyük boyuttaki oran eğitim verisine ait olmaktadır.

5.5.2. Karışıklık matrisi

Sınıflandırma işlemlerinde eğitim veri seti ile oluşturulan model üzerinde test veri setindeki her bir gözlem modelde yerine konularak sınıflandırma tahmin sonuçları elde edilmektedir. Elde edilen tahmin sonuç değerleri gerçek değerler ile karşılaştırılarak oluşturulan bu modelin ne kadar iyi tahmin yaptığı, başarısı ve performansı test edilmektedir. Karışıklık matrisi modelin ne kadar doğru tahminler yaptığına dair sonuçları özetlemekte ve makine öğrenmesi sınıflama yönteminin ne kadar başarılı olduğu konusunda performans değerlendirme sonuçlarını da göstermektedir. Çizelge 5.1’de iki kategorili bir sınıflandırma yöntemi için karışıklık matrisi gösterilmektedir.

Çizelge 5.1. İki sınıflı karışıklık matrisi

		Tahmin Edilen Sınıf	
		0	1
Gerçek Sınıf	0	Doğru Negatif (DN)	Yanlış Negatif (YN)
	1	Yanlış Pozitif (YP)	Doğru Pozitif (DP)

Çizelge 5.1’de test veri setindeki her gözlem için tahmin yapılmakta ve bu tahmin değerleri gerçek değerler ile karşılaştırılmaktadır. Çizelgenin sol tarafındaki satırlar gerçek bir sınıfa karşılık gelen değerleri, ve diğer sütunlar tahmini değerleri temsil etmektedir. Doğru ve yanlış sınıflandırma ve tahmin sonuçları ile çizelgeye doldurulmaktadır. Karışıklık matrisi aynı zamanda hipotez testlerinde gösterilen hata matrisine de benzetilmektedir.

Çizelge 5.1’de görüldüğü gibi iki sınıflı karışıklık matrisini oluşturan değerler aşağıdaki gibi açıklanmaktadır:

- Doğru pozitif (DP); gerçek sınıf değeri 1 olan gözlemlerin doğru bir şekilde 1 olarak tahmin edilmesi durumunu gösterir.
- Doğru negatif (DN); gerçek sınıf değeri 0 olan gözlemlerin doğru bir şekilde 0 olarak tahmin edilmesi durumunu göstermektedir.
- Yanlış Negatif (YN); gerçek sınıf değeri 1 olan gözlemlerin tahmin sonucunda 0 olarak yanlış değerlendirildiğini göstermektedir.
- Yanlış pozitif (YP); gerçek sınıf değeri 0 olan gözlemlerin tahmin sonucunda 1 olarak yanlış değerlendirildiğini göstermektedir (Alan ve Karabatak, 2020).

Karışıklık matrisi sonuçlarını karşılaştırmak için, doğruluk oranı, hata oranı, kesinlik, duyarlılık, F1 ölçütü, ROC eğrisi ve AUC değerlerinin incelenmesi gerekmektedir. Çalışmanın bu bölümünde bu oranlar ve ölçütler açıklanmaya çalışılacaktır.

5.5.3. Doğruluk oranı (Accuracy rate)

Doğruluk oranı sınıflandırılan gözlem sayısının (DP+DN), toplam örnek sayısına (DP + DN + YP + YN) oranının alınması şeklinde hesaplanmaktadır. Bu bir sınıfın gerçek değeri 1 iken sınıflandırma modeli ile yapılan tahmin sonucu değerinin 1 olarak tahmin edilmesi ve sınıfın gerçek değeri 0 iken sınıflandırma sonrası tahmin edilen değerinin 0 olması durumunun birlikte değerlendirilmesini sağlamaktadır. Doğruluk oranı;

$$\text{Doğruluk Oranı (ACC)} = \frac{DP+DN}{DP+DN+YP+YN} \quad (5.34)$$

şeklinde gösterilmektedir.

5.5.4. Hata oranı (Error rate)

Hata oranı sınıflandırılan gözlem sayısının gerçek değeri 1 iken sınıflandırma modeli ile yapılan tahmin sonucu 0 olarak tahmin edilmesi (YN) ve gerçek değeri 0 iken sınıflandırma sonrası tahmin edilen değerinin 1 olması (YP) durumunun birlikte incelenmesidir. Doğruluk ve hata oranı değerleri uygulanan sınıflandırma yönteminin başarısını göstermektedir. Hata oranı;

$$Hata\ Oranı\ (ERR) = \frac{YN+YP}{DP+DN+YP+YN} \quad (5.35)$$

veya;

$$Hata\ Oranı\ (ERR) = 1 - ACC \quad (5.36)$$

şeklinde gösterilmektedir.

5.5.5. Kesinlik (Precision)

Bu ölçüt gözlem sayısının doğru olarak sınıflandırılan 1 değerlerinin (DP) veya pozitif girdilerin toplam pozitif değerlere (DP+YP) oranıdır ve bu oran pozitif öngörü oranı olarak da bilinmektedir. Bu ölçüt, veri seti içerisindeki doğru değerlerin ne kadar gerçeğe yakın olduğunu göstermektedir. Hassasiyet değerinin yüksek olduğu bir sınıflandırma modelinde kesinlik değerinin yüksek olması gerekir. Diğer taraftan Hassasiyet oranı düşük çıkan bir sınıflandırma modelinde kesinlik değeri doğru sonuç vermeyebilir. Kesinlik değeri yapılan doğru tahminin ne kadar doğru olduğunu gösteren bir ölçüt olduğu için yüksek çıkması beklenen ve istenilen bir durumdur.

Kesinlik formülü:

$$Kesinlik (Precision) = \frac{DP}{DP+YP} \quad (5.37)$$

şeklinde gösterilmektedir.

5.5.6. Duyarlılık/Hassasiyet (Sensitivity/Recall rate) oranı

Duyarlılık ölçütü doğru olarak sınıflandırılmış (DP) pozitif girdi değerlerinin toplam gerçek pozitif değerlere (DP + YN) oranıdır. Duyarlılık oranı;

$$Duyarlılık (Recall) = \frac{DP}{DP+YN} \quad (5.38)$$

şeklinde gösterilmektedir.

5.5.7. F1 ölçütü (F1-measure)

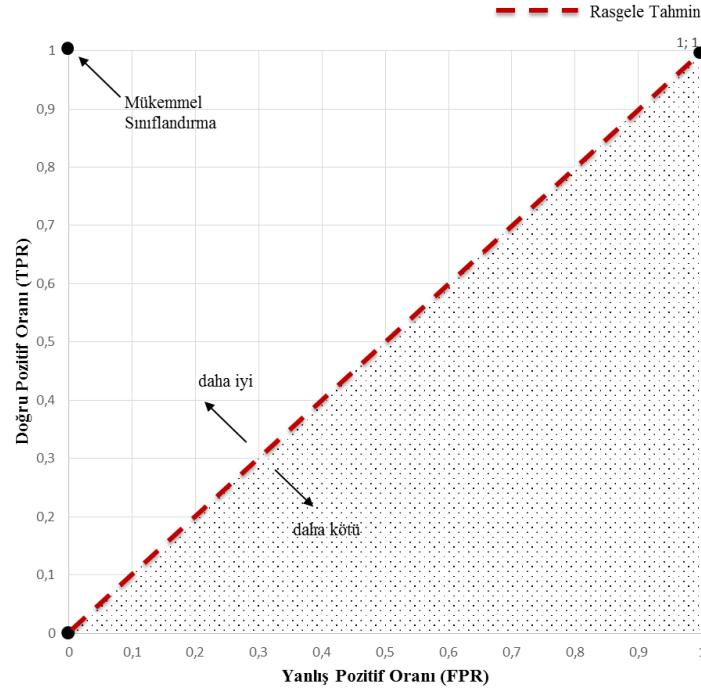
Duyarlılığın çok kesinliğin az olması veya tam tersi olması durumlarında iki modeli karşılaştırmak zorlaşmaktadır. Bu durumda F1 ölçütü duyarlılık ve kesinlik değerlerinin aşırı değerlerden kurtulmak için harmonik ortalamalarını alarak bu değerlerin aynı anda ölçülmelerini sağlamaktadır. F1 ölçütü;

$$F1 \text{ Ölçütü} = \frac{2 * Duyarlılık * Kesinlik}{Duyarlılık + Kesinlik} \quad (5.19)$$

şeklinde gösterilmektedir.

5.5.8. Alıcı işlem karakteristik (Receiver operating characteristic-ROC) eğrileri

ROC eğrileri iki kategorili sınıflandırma yöntemlerinde kullanılmaktadır. ROC eğrileri doğru pozitif ve yanlış pozitif oranların karşılaştırarak ve grafiği çizerek model başarısı hakkında bilgi vermektedir. Bu bilgiyi verirken, mükemmel test ve yararsız test ayırımı yapmaktadır. Yararsız test; satın alan ve almayan web site kullanıcılarını birbirinden ayıramayan testtir ve %50 olasılığa sahiptir. Mükemmel test; satın alma yapan ve yapmayan kullanıcıları tam anlamıyla ayırabilen testtir. DP=1, YP=0 olarak gösterilir. Testler yararsız test ile mükemmel test arasında bir başarı performansına sahiptir (Alpar, 2010). ROC eğrisi aşağıdaki Şekil 5.6'de gösterilmektedir.



Şekil 5.6. ROC uzayı (Fawcett, 2006)

İdeal durum için dikey olarak (0,0) noktasından (0,1) noktasına ve buradan grafiğin en sağ üstündeki (1,1) noktasına doğru çizilir. Bir çizgide (0,0) noktasından (1,1) noktasına doğru çizilir, bu çizgi rastgele yapılacak bir sınıflandırma tahmini ile sınıflandırma modeli ile yapılacak olan sınıflandırma tahmini arasında bir fark olup olmadığını ölçmek

için kullanılır. ROC eğrisinin altındaki alan hesaplanarak modelin tahmin başarısı değerlendirilir. ROC değerinin;

- 1'e yakın olması modelin sınıflandırma başarısının iyi olduğunu
- 0.5'e yakın olması modelin sınıflandırma başarısının rastgele alınan bir karardan daha iyi olmadığını
- 0'a yakın bir değer ise, modelin tüm örnekleri yanlış sınıflandırdığını ifade eder (Witten vd., 2011).

5.5.9. Eğri altındaki alan (AUC-area under curve)

AUC, ROC eğrisinin altındaki (0,0) ile (1,1) arasındaki tüm alanın ölçümünü göstermektedir. Bu ölçüm ROC eğrisi ile birlikte model performansını gösteren ölçümlerdendir. AUC değeri 0 ile 1 arasında değer almakta ve bu değer 1'e yakın olması durumu istenmektedir, bu durum modelin doğruluğunun fazla olduğunu göstermektedir. ROC eğrisi altındaki alanın büyük olması durumunda DN ve DP'ye ait dağılımlar kesişmemekte ve sınıflar birbirinden başarılı bir şekilde ayrılmaktadır.

6. UYGULAMA

Tez çalışmasının bu bölümünde makine öğrenmesi uygulama aşamaları takip edilerek e-ticaret verileri üzerinde makine öğrenmesindeki Lojistik Regresyon, Naïve Bayes, Destek Vektör Makineleri ve Rastgele Ormanlar sınıflandırma yöntemleri uygulanmış ve elde edilen sonuçlar değerlendirilmiştir. Çalışmanın uygulamasında kullanılan terimler ve hangi anlama geldikleri ilgili bölümlerde detaylı olarak açıklanmıştır. Çalışma kapsamında toplanan veriler istatistiksel modellemeler, veri görselleştirme, veri manipülasyonları gibi verinin etkili bir şekilde kullanılmasını ve işlem yapılabilmesini sağlayan Python 3.7 programlama dili kullanılmış ve Jupyter Notebook 6.1.4 ortamında test edilmiştir. Python makine öğrenimi ve veri bilimine ilişkin analizlerin yapılması amacı ile geliştirilmiş olup, bu alanlarda en çok tercih edilen programlama dilidir. Diğer taraftan Jupyter Notebook farklı program dillerinin bir arada kullanılmasını sağlayan ve akademik çalışmalar için oldukça kullanışlı bir programlama dilidir.

6.1. Problemin Tanımlanması

Günümüzde her bir bilgi veri kaynağı olmakta ve gelişen teknoloji, internet ile birlikte sürekli bir veri akışı sağlanmaktadır. E-ticaret de en çok veri kaynağı sağlayan ve makine öğrenmesi algoritmalarının en çok kullanıldığı alanlardandır. Elektronik ortamlarda ticaret yapan her kişi ve işletmeler için ürün satın alan müşterilerin tespit edilmesi çok önemli bir unsurdur. E-ticarete yapılan satışların takibi, satın alma yapan müşterilerin doğru tahmin edilmesi ile orantılı olarak satış teknikleri geliştirilebilmektedir. Bu tez çalışması kapsamında amaç, satın alma yapan müşterilerin tespiti için doğru tahmin yapabilmek için makine öğrenmesi sınıflandırma yöntemlerini kullanarak en iyi sonucu veren yöntemin belirlenmesidir.

6.2. Verinin Toplanması

Tez çalışması kapsamında Türkiye merkezli, yurtdışı satışlarına açık bir e-ticaret sitesi olan İncikcincik E-Ticaret Ltd. Şti. firma ismi ile varlığını sürdüren incikcincik.co verileri kullanılmıştır. Firma aksesuar ve yoğun olarak kişisel ürünler üzerine satış yapmaktadır. Veriler Google Analytics aracılığı ile 2019 Mart ve 2020 Mart dönemlerine ait web siteyi ziyaret eden 256.143 ziyaretçiye ait bilgileri içermektedir.

Google Analytics kullandığı javascript kod parçaları ile verileri toplamakta ve bu kodun web sitesi içerisine gömülü bir şekilde eklenmesi ile site üzerindeki ziyaretçilerin hangi sayfaları, hangi cihazdan giriş yaptıkları, web sitede geçirdikleri süreler ve yaptıkları işlemlere ait bilgileri toplamaktadır. Toplanan verileri Google içerisindeki sunucular ile işleyip anlamlı bir şekilde gruplama, kategorize etme, kümeleme gibi işlemler yapmaktadır. Bu veri ön işlemlerinden sonra site ile ilgili erişimler hakkında raporlar sunan bir platformdur.

Araştırma kapsamında toplanan veri seti içerisindeki, kullanılan değişkenler ve bunların hangi bilgileri kapsadıklarına dair açıklamalar Çizelge 6.1’de detaylı olarak gösterilmektedir.

Çizelge 6.1. Araştırma kapsamındaki değişkenler ve açıklamaları

Değişken	Açıklama
Müşteri Numarası	Web sitesini ziyaret eden her kullanıcı için eşsiz bir kullanıcı numarası atanmaktadır.
Oturum Sayısı	Web site üzerinde kullanıcının belirtilen tarih aralığında oluşturduğu ziyaret/oturum sayısıdır.
Ortalama Geçirilen Süre	Web site üzerinde kullanıcının belirtilen tarih aralığındaki ziyaretlerinde web sitesinde ortalama harcadığı süreyi göstermektedir.
Hemen Çıkma Oranı	Web siteyi ziyaret eden kullanıcının hiçbir aksiyon gerçekleştirmeden siteyi ilk 30 sn. için terk etmesinin tüm ziyaretine oranını göstermektedir.
İşlem sayısı	Kullanıcının belirtilen tarih aralığında vermiş olduğu sipariş sayısı (0 olanlar siparişi olmayanlar, 1 ve üzerindeki sipariş verenler)
Cinsiyet	Kullanıcıların cinsiyetlerini göstermektedir.
Yaş	Kullanıcıların yaşları belirli aralıklar ile aralıklı ölçek şeklinde toplanmaktadır. Yaş aralıkları 18-24, 25-34, 35-44, 45-54, 55-64, 65+ şeklindedir.
Edinme	Kullanıcıların belirtilen tarih aralığında siteye erişim sağladıkları kaynağı temsil etmektedir. Kaynaklar; direkt, sosyal medya, organik arama, referanslı ve diğer olarak kategorize edilmektedir.
Cihaz	Kullanıcıların belirtilen tarih aralığında siteye erişimlerini hangi cihaz yolu ile sağladıklarını göstermektedir. Cihazlar; Mobil, tablet ve bilgisayar olarak kategorize edilmektedir.
Satın alma	Kullanıcıların belirtilen tarih aralığında web sitesi üzerinden satın alma yapıp yapmadıklarına dair bilgiyi içermektedir. (0 olanlar siparişi olmayan kullanıcılar, 1 olanlar tutara bakılmaksızın satın alma işlemi yapan kullanıcılar)

Veri setinin genel görünümü Şekil 6.1’de gösterilmektedir.

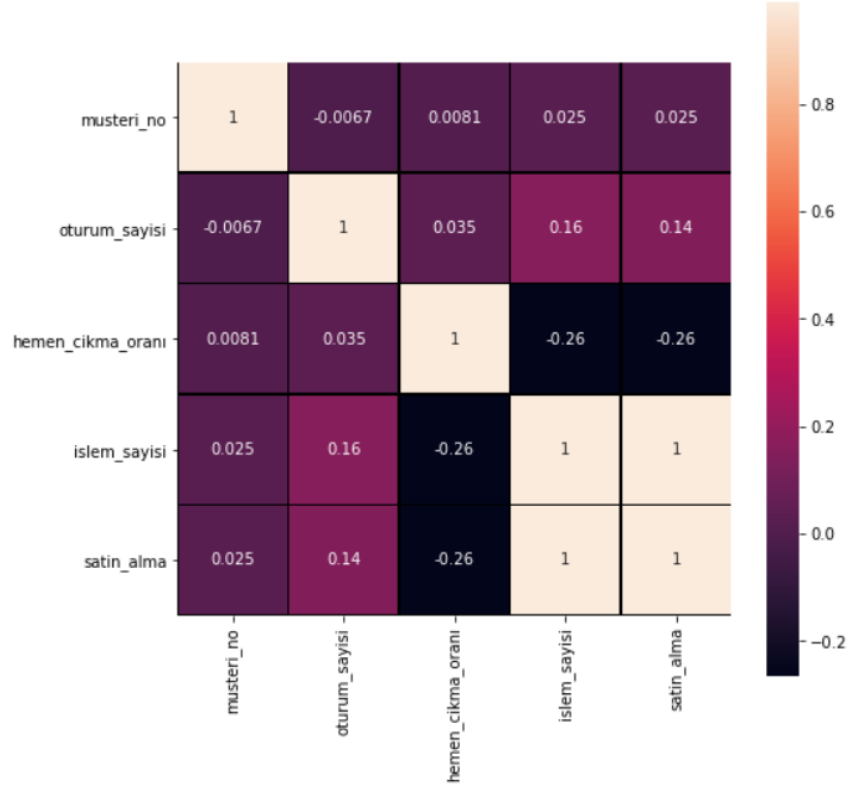
	oturum_sayisi	ort_oturum_suresi	hemen_cikma_orani	cinsiyet	yas	edinme	cihaz	satin_alma
0	1	00:04:37	0.0	Male	25-34	Direkt	Mobile	0
1	1	00:00:00	1.0	Female	65+	Social	Mobile	0
2	1	00:02:20	0.0	Male	25-34	Direkt	Mobile	0
3	1	00:04:37	0.0	Male	25-34	Direkt	Mobile	0
4	1	00:04:19	0.0	Male	18-24	Social	Mobile	0

Şekil 6.1. Veri setinin görünümü

6.3. Veri Ön-işleme/Hazırlama

Veri ön-işleme aşaması verinin sınıflandırma analizine uygun hale getirilmesi için yapılan işlemleri kapsamaktadır. Gözlem değerlerinin doğru bir şekilde ön işleme yapılmadığı durumlarda makine öğrenmesinin ilerleyen süreçlerinde yanlış sonuçlar elde edilebilir. Bu sebepten veri üzerinden bir takım işlemlerin yapılması ve makine öğrenmesi analizine uygun formatlara getirilmesi zorunlu bir aşamadır.

Veri seti içerisinde eksik veri gözlem bulundurmamaktadır ve veri seti içerisinde kategorik ve sürekli yapıda olan sayısal veri tipleri mevcuttur. Öncelikli olarak sürekli yapıda olan sayısal verilerin birbirleri ile aralarındaki ilişkiye bakmak amaçlı korelasyon matrisi oluşturulmuştur. Korelasyon matrisi aşağıda Şekil 6.2’de gösterilmektedir.



Şekil 6.2. Korelasyon matrisi

Şekil 6.2'deki korelasyon matrisinde iki değişken arasındaki ilişki siyahtan krem renge doğru ısı haritası gibi renk değişimleri ile gösterilmektedir. Renkler ve karşılık geldikleri sayısal değerler Şekil 6.2'nin sağ tarafında görüldüğü gibidir. Renkler ve sayısal değerlere göre, yapılan işlem sayısı ile satın alma durumu arasında tam ve doğru bir ilişki olduğu sonucuna varılmaktadır.

Çıkan sonuca göre kullanıcıların tek işlemde satın alma yaptıkları söylenebilmektedir. İşlem sayısını içeren değişkeni, çalışmanın ilerleyen adımlarında analizlerde yanlış sonuçlar vermemesi için analizin dışında bırakılmıştır. Diğer taraftan müşteri numarasını içeren değişken kullanıcılara yönelik sayısal bir ölçümü ifade etmediği için analize dail edilmemiştir.

Veri seti içerisinde kategorik değişkenler bulunmaktadır. Kategorik değişkenler öncelikle herhangi bir analiz içerisinde bulunabilmeleri için sayısallaştırılmaları

gerekmektedir. Burada önemli olan nokta kategorik değişkenlerin verimli bir şekilde kullanılabilmesi için doğru veri dönüşümleri yapılması gerektiğidir. Kategorik değerlere karşılık gelecek sayısal değerlerin sürekli değişkenler gibi bir ölçüm değeri olmadıkları için Python programında kukla değişken olarak adlandırılan şekle getirilmesi gerekmektedir. Veri setimizde bulunan kategorik veriler isimsel veri tipinde oldukları için kategoriler arasında herhangi bir üstünlük durumu bulunmamaktadır. Kategorik değişkenler için kukla değişkenler gerekli kodlama ile yapılmıştır.

Veri seti içerisindeki sürekli ve kesikli sayısal değişkenlerin veri tiplerine yönelik ön-işleme aşamaları gerçekleştirilmiş ve bir sonraki adımın uygulanabilmesi için veri seti hazırlanmıştır.

6.4. Eğitim ve Test Verisinin Ayrılması

Veri seti analize uygun bir şekilde hazırlandıktan sonra sınıflandırma yöntemlerinin uygulanabilmesi için eğitim ve test olarak ikiye ayrılması gerekmektedir. Bu çalışma kapsamına veri seti holdout yöntemi ile %70 eğitim ve %30 test veri seti olarak ayrılmıştır. Diğer eğitim ve test olarak ayırma oranları veri seti üzerinde denenmiş ve en iyi tahmin yapılmasını sağlayan ayırma oranlarının %70 - %30 olduğuna karar verilmiştir.

6.5. Modelin Oluşturması, Test Edilmesi ve Değerlendirilmesi

Çalışma kapsamında belirtilen dört makine öğrenmesi yöntemi eğitim verisi üzerinde uygulanmış ve bu yöntemlere yönelik modeller kullanılan programlama dili sayesinde oluşturulmuştur. Oluşturulan modeller test verisi üzerinde test edilmiş ve daha sonrasında yapılan tahmin sonucunda modellerin doğruluk oranları ve performans ölçütlerine dair Çizelge 6.2'deki sonuçlar elde edilmiştir.

Çizelge 6.2. Sınıflama yöntemlerinin karşılaştırılması

		Kesinlik (Precision)	Duyarlılık (Recall)	F1 Ölçütü	Doğruluk (Accuracy)	Hata Oranı (Error Rate)
Lojistik Regresyon	Satın almadı	0.90	0.93	0.91	0.89	0.11
	Satın aldı	0.89	0.83	0.86		
Naïve Bayes	Satın almadı	0.71	0.99	0.83	0.75	0.25
	Satın aldı	0.98	0.38	0.55		
Destek Vektör Makineleri	Satın almadı	0.90	0.93	0.92	0.90	0.10
	Satın aldı	0.89	0.85	0.87		
Rastgele Ormanlar	Satın almadı	0.94	0.96	0.95	0.94	0.06
	Satın aldı	0.94	0.91	0.92		

Çizelge 6.2’de dört sınıflandırma yöntemine göre yapılan, müşterilerin satın alma durumlarının tahmin sonuçları görülmektedir. Lojistik Regresyon analizi sonucunda satın alma yapmayan kullanıcıların %90’ı ve satın alma yapan müşterilerin %89’u doğru tahmin edilmiştir. Doğru tahmin edilen satın alma yapmayan müşterilerin toplam tahmine oranı %93, tahmin edilen satın alma yapan müşterilerin toplam tahminlerine oranı %83 olarak görülmektedir. Lojistik Regresyon modeli ile yapılan sınıflama algoritması genel olarak %89 doğruluk oranına ulaşmıştır ve dolayısı ile %11 hata oranı bulunmaktadır.

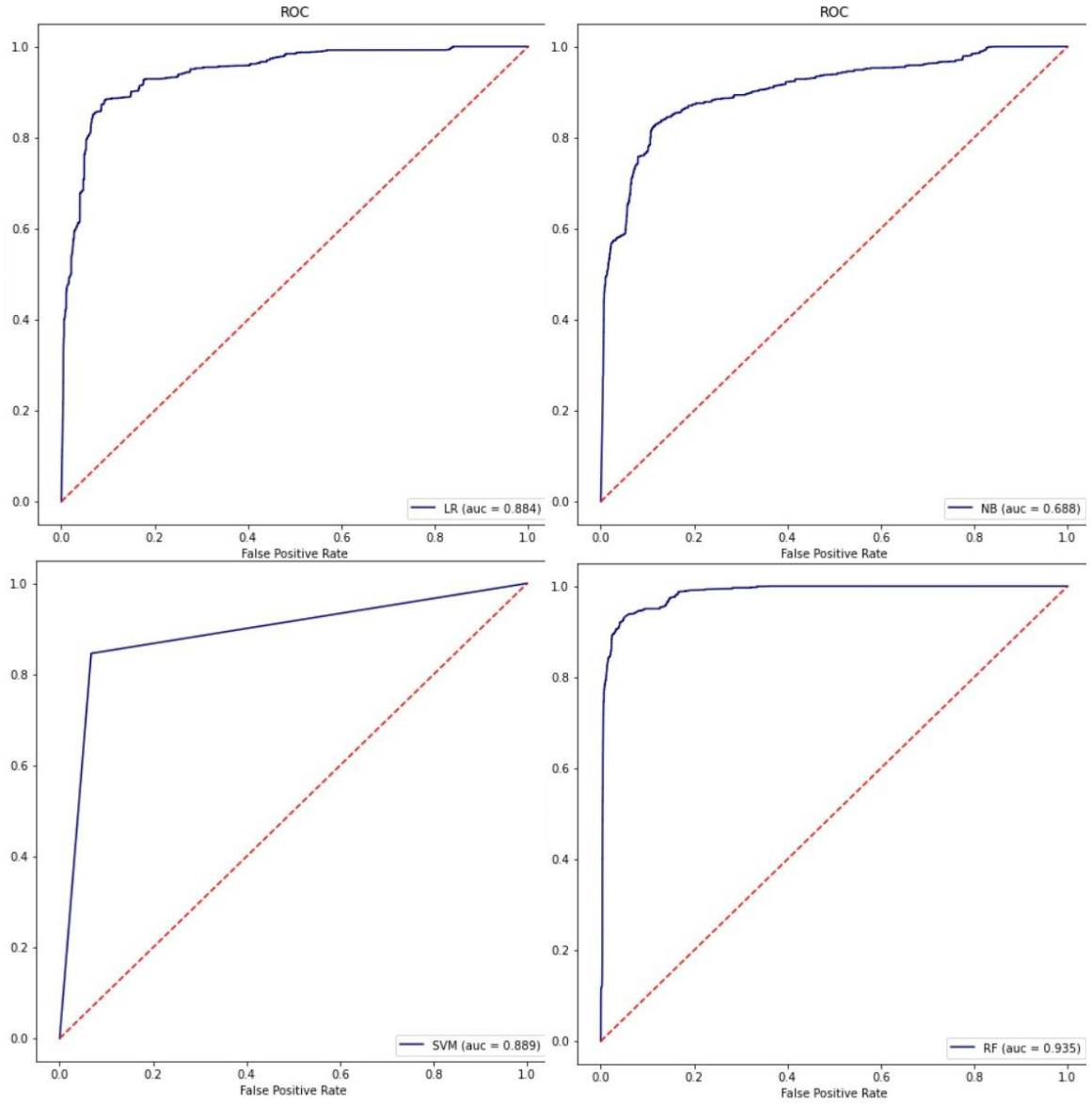
Naïve Bayes analizi sonucunda satın alma yapmayan kişilerin %71’i ve satın alma yapan müşterilerin %98’i doğru tahmin edilmiştir. Doğru tahmin edilen satın alma yapmayan müşterilerin toplam tahmine oranı %99, tahmin edilen satın alma yapan müşterilerin toplam tahminlerine oranı %38 olarak görülmektedir. Naive Bayes ile yapılan sınıflama

algoritması genel olarak %75 doğruluk oranına ulaşmıştır ve %25 hata oranı bulunmaktadır.

Destek Vektör Makinesi analizine bakıldığında ise; satın alma yapmayan kişilerin %90'ı ve satın alma yapan müşterilerin %89'u doğru tahmin edilmiştir. Doğru tahmin edilen satın alma yapmayan müşterilerin toplam tahmine oranı %93, tahmin edilen satın alma yapan müşterilerin toplam tahminlerine oranı %85 olarak görülmektedir. Destek Vektör Makinesi ile yapılan sınıflama algoritması genel olarak %90 doğruluk oranına ulaşmıştır ve %10 hata oranı bulunmaktadır.

Rastgele Ormanlar analizine bakıldığında ise; satın alma yapmayan kişilerin %94'ü ve satın alma yapan müşterilerin %94'ü doğru tahmin edilmiştir. Doğru tahmin edilen satın alma yapmayan müşterilerin toplam tahmine oranı %96, tahmin edilen satın alma yapan müşterilerin toplam tahminlerine oranı %91 olarak görülmektedir. Rastgele Ormanlar yöntemi ile yapılan sınıflama algoritması genel olarak %94 doğruluk oranına ulaşmıştır ve %4 hata oranı bulunmaktadır.

ROC eğrisi doğru pozitif ve yanlış pozitif değerlerini kullanarak, uygulanan makine öğrenmesi sınıflandırma yönteminin performansını değerlendiren ölçütlerden bir tanesidir. Farklı modellerin karşılaştırılmasında veya veri seti içerisinde dengesiz durumların bulunması halinde modelin performansının değerlendirilmesi açısından yaygın olarak bakılması gereken ve kullanılan bir ölçümdür. Aslında kullanılan modelin yaptığı tahminde ne kadar başarılı olduğunu gösteren bir olasılık eğrisidir. ROC analizi sonrası satın alma yapan kullanıcıların doğru tahmin edilmesi ve model başarısı değerlendirilen bir testin yararsız mı yoksa iyi bir test mi olduğu kararı verilebilir. Doğru tahmin edilen sınıfların, yanlış tahmin edilen sınıflara oranını göstermek için yöntemlerden elde edilen ROC eğrileri Şekil 6.3'de görüldüğü gibidir:



Şekil 6.3. Sınıflandırma yöntemlerinin ROC eğrileri. (Sol-üstte, LR, sağ-üstte, NB, sol altta, SVM, sağ-altta RF, ROC eğrileri bulunmaktadır.)

ROC eğrilerinin 1'e yakın olması durumu modelin sınıflandırma başarısının iyi olduğunu göstermektedir. Genel olarak dört sınıflandırma yöntemine bakıldığında da 1'e yakın olarak eğri oluşturdukları görülmektedir. ROC eğrileri ile birlikte AUC değerlerine bakılması gerektiği için AUC değerleri Çizelge 6.3'de gösterilmiştir.

Çizelge 6.3. AUC değerleri

	Lojistik Regresyon	Naïve Bayes	Destek Vektör Makineleri	Rastgele Ormanlar
AUC	0.884	0.688	0.889	0.935

AUC değeri 1 değerine doğru yaklaştıkça modelin doğru tahmin gücü yüksek olduğu söylenebilmektedir. Çizelge 6.3'te Lojistik Regresyon yönteminin ROC eğrisi altında kalan alanı; 0.884, Naïve Bayes yönteminin ROC eğrisi altında kalan alanı: 0.688, Destek Vektör Makinelerinin ROC eğrisi altında kalan alanı: 0.889, Rastgele Ormanlar yönteminin ROC eğrisi altında kalan alanı: 0.935 olarak çıkmaktadır. Bu sonuçlardan yola çıkarak Rastgele Ormanlar Makine öğrenmesi yöntemi ile oluşturulan sınıflandırma modelinin daha iyi tahminler yaptığı sonucuna ulaşılmaktadır.

7. SONUÇ ve ÖNERİLER

Çalışma kapsamında belirtilen teknolojik gelişmeler ve internetin yaygın kullanılması ile birlikte veri akışı ve boyutları her geçen zaman diliminde ciddi artışlar gerçekleştirmektedir. Veri akışını sağlayan etkenler incelendiğinde dönüşümlü olarak çevrimiçi alışveriş dikkat çekmektedir. Bu büyük boyuttaki verilerin ölçülebilir hale gelmesi, karmaşıklıkların çözülmesi, kullanılabilir hale gelebilmesi için matematik ve istatistiksel yöntemlerden faydalanan makine öğrenmesi yöntemlerine başvurulmaktadır. E-ticarete makine öğrenmesi yöntemleri ile veri kaynaklarının kullanılması sağlanmaktadır. E-ticaret farklı makine öğrenmesi analizlerinin kullanılması ve test edilmesine olanak sağlayan oldukça elverişli bir alandır. E-ticaret alanında, her bir çevrimiçi alışveriş adımı için farklı algoritmalar geliştirilebilir. E-ticarete satın alma yapan müşterilerin tespit edilmesi, geleceğe yönelik satış tahmin edilmesi, mevcut müşterilerin elde tutulması, müşterilere yönelik kampanyaların yapılması, mevcut müşteri profillerinin belirlenmesi gibi birçok açıdan büyük önem taşımaktadır.

Bu sebepten bu araştırma kapsamında Türkiye’de faaliyet veren ve yurtdışı satışlarına açık olan incikcincik.co web sitesinin 2019 Mart ve 2020 Mart tarihleri arasındaki web sitesini ziyaret eden kullanıcılara ait veriler kullanılarak, satın alma yapan müşterilerin tahmininin yapılabilmesi için doğru sonucu veren sınıflandırma yöntemi araştırılmıştır. Bu amaçlar doğrultusunda dört farklı makine öğrenmesi yöntemi karşılaştırılmaktadır. Veriler Google analytics aracılığı ile toplanmış ve Python 3.7 programlama dili ile Jupyter Notebook 6.1.4 ortamında test edilmiştir. 256.143 web site ziyaretçisine ait veriler holdout yöntemi ile %70 eğitim ve %30 test verisi olarak ayrılmıştır. Makine öğrenmesi sınıflandırma algoritmalarından Lojistik Regresyon, Naive Bayes, Destek Vektör Makineleri ve Rastgele Ormanlar yöntemleri Python programında yazılan kodlar yardımı ile farklı veri tiplerine yönelik ön işleme, veri hazırlık süreçleri ile veri seti daha anlaşılır ve analiz için hazır hale getirilmiştir.

Tez Çalışması kapsamında belirtilen kavramların ve ismi geçen makine öğrenmesi yöntemlerinin hepsi ilgili bölümlerde detaylı olarak anlatılmıştır. Çalışmada işletmeden

tüketiciye yapılan e-ticaret (B2C) türünde, müşterilerin satın alma yapıp yapmamasına dair doğru algoritmayı bulmak için Lojistik Regresyon, Naïve Bayes ve Destek Vektör Makinesi yöntemleri uygulanmıştır. Yapılan analizler sonucunda test edilen yöntemlerin genel başarıları doğruluk (accuracy) değeri ile karşılanmaktadır. Çalışma kapsamındaki en iyi doğruluk oranını veren model %94 doğruluk oranı ile Rastgele Ormanlar yöntemi olmuştur. SVM %90 ve LR %89 doğruluk değeri ile genel model başarısı açısından değerlendirildiğinde birbirlerine yakın sonuçlar vermiştir. Genel model başarı oranı en düşük olan %75 doğru tahmin yaparak NB sınıflandırıcısı olmuştur. Web site üzerinde satın alma işlemi yapan kullanıcıların sitede daha fazla zaman geçirdikleri gözlemlenmiştir.

Çalışmanın amacı satın alan müşterilerin doğru tahmin edilmesi olduğu için modellerin sonuçlarını, gözlem sayılarını doğru sınıflandıran değerlerin toplam pozitif değerlere oranını gösteren kesinlik (precision) değerlerine de bakılarak en iyi model hakkında bilgi sahibi olmak gerekmektedir. LR yöntemi satın alan müşterileri %90 oranında, NB yöntemi %71, SVM yöntemi %90 ve RF yöntemi ise %94 oranında doğru olarak tahmin etmiştir. ROC eğrisi ve AUC değerleri de incelendiğinde en yüksek tahmin değerlerinin Rastgele Ormanlar yöntemi olduğu tespit edilmiştir.

Müşteri profillerinin belirlenebilmesi, kişi veya işletmelerin geleceğe yönelik tahmin etme çalışmaları yapabilmesi, mevcut durumların değerlendirilebilmesi, kampanya ve fırsatlarını düzenleyebilmesi için mevcut müşterilerinin doğru tahmin modeli ile yeni müşterilerin satın alıp almayacaklarına dair tahmin yapılması gerekmektedir. Bu tez çalışması e-ticaret platformu üzerinde, satın alan müşterilerin en doğru hangi model ile tahmin yapılması gerektiğine dair çalışma yapmak isteyen kişilere fikirler sunulmuştur. Bu çalışmanın bir sonraki aşaması olarak bu müşterilere yönelik fiyat tahminlemesi yapılabilir. Müşterilerin satın alma yapıp yapmadıklarına yönelik doğru modelin bulunması için diğer sınıflandırma algoritmaları da denenebilir, buradaki çalışmada olduğu gibi doğru sınıflandırma modeli bulunduktan sonra farklı makine öğrenmesi algoritmaları, farklı boyuttaki veri setleri, farklı değişkenler kullanılarak amaca yönelik daha detaylı çalışmalar yapılabilir.

KAYNAKÇA

- Alan, A. (2020). Makine Öğrenmesi Sınıflandırma Yöntemlerinde Performans Metrikleri ile Test Tekniklerinin Farklı Veri Setleri Üzerinde Değerlendirilmesi, Fırat Üniversitesi, Yüksek Lisans Tezi, 82, Elazığ.
- Alan, A., Karabatak, M. (2020). Veri Seti-Sınıflandırma İlişkisinde Performansa Etki Eden Faktörlerin Değerlendirilmesi. Fırat Üniversitesi Mühendislik Bilimleri Dergisi, 32(2), 531-540.
- Alpar, R. (2010). Spor, Sağlık ve Eğitim Bilimlerinden Örneklerle Uygulamalı İstatistik ve Geçerlik Güvenirlik. Detay Yayıncılık, 338-42, Ankara.
- Alpar, R. (2017). Çok Değişkenli İstatistiksel Yöntemler. Detay Yayıncılık, 5. Baskı, Ankara.
- Alpaydın, E. (2010). Introduction to Machine Learning. Massachusetts Institute of Technology Press. United States of America.
- Alpaydın, E. (2016). Machine learning: the New AI. Massachusetts Institute of Technology Press. United States of America.
- Ayhan, S., Erdoğan, Ş. (2014). Destek Vektör Makineleriyle Sınıflandırma Problemlerinin Çözümü için Çekirdek Fonksiyonu Seçimi. Eskişehir Osmangazi Üniversitesi İİBF Dergisi, 9(1), 175- 198.
- Bagui, S., Fang, X., Kalaimannan, E., Bagui, S., & Sheehan, J. (2017). Comparison of machine-learning algorithms for Classification of VPN Network Traffic Flow Using Time-Related Features. Journal of Cyber Security Technology, <https://doi.org/10.1080/23742917.2017.1321891>, 1(2), 108-126.
- Ballouk, S. (2020), Erişim Tarihi: 01.15.2021). javapoint: <https://www.javatpoint.com/difference-between-artificial-intelligence-and-machine-learning> adresinden alındı.
- Baran, M. (2020). Makine Öğrenmesi Yöntemleriyle Çoklu Etiketli Verilerin Sınıflandırılması. Sivas Cumhuriyet Üniversitesi, Sosyal Bilimler Enstitüsü, Yüksek Lisans Tezi, 113, Sivas.
- Bartlett, J. (2015). The Dark Net: Inside The Digital Underworld. Melville House ISBN: 978-1-61219-489-9. United States of America.
- Baydemir, M. B. (2014). Lojistik Regresyon Üzerine Bir İnceleme. İnönü Üniversitesi, Sosyal Bilimler Enstitüsü, Yüksek Lisans Tezi, 83, Malatya.
- Bell, J. (2014). Machine Learning : Hands-On for Developers and Technical Professionals. John Wiley & Sons Inc, United States of America.

- Berkeley, E. C. (1949). *Giant Brains or Machines That Think* (First Edition). Yayınlayan: Chapman & Hall / John Wiley. New York, United States of America.
- Breiman, L. (1996). Bagging Predictors. . *Machine Learning*, 24(2), 123-140.
- Breiman, L. (2001). Random Forests. *Machine learning*, 45(1), 5-32.
- Bruce, P., & Bruce, A. (2017). Bruce P., Bruce A. (2017). *Practical Statistics for Data Scientists*. O'Reilly Media, ISBN: 9781491952962, 432-433, United States of America.
- Çelik, B. (2015). *An Exploratory Analysis of Online Shopping Behavior in Turkey*. Bahçeşehir Üniversitesi, Sosyal Bilimler Enstitüsü, Yüksek Lisans Tezi, 86, İstanbul.
- Çelik, E. (2019). *Customer Churn Analysis Based on Machine Learning by Using Data Mining Techniques in Telecommunication Sector*. Yeditepe Üniversitesi, Sosyal Bilimler Enstitüsü, Yüksek Lisans Tezi, 147, İstanbul.
- Chervonenkis, A. (2013). *Early History of Support Vector Machines*. Empirical Inference. içinde Berlin, Heidelberg, : Springer, ISBN: 978-3-64-241135-9, 978-3-64-241136-6, https://doi.org/10.1007/978-3-642-41136-6_3.
- Cöhce, M. O. (2020). *Lojistik Regresyon Analizi ile Üniversite Öğrencilerinin Kariyer Planlarını Etkileyen Faktörlerin Analizi*. Bursa Uludağ Üniversitesi, Sosyal Bilimler Enstitüsü, Yüksek Lisans Tezi, 110, Bursa.
- Coleman, M., Ganong, L. (2014). *The Social History of the American Family An Encyclopedia*. Sage Publications, ISBN:9781452286167, United States of America.
- Cornelius, P., Kirkman, G., Sachs, J., Scwab, K. (2002). *Country Profiles. The Global Information Technology Report 2001-2002: Readiness for the Networked World*. New York, NY: Oxford University Press, USA, UK: New York, Oxford: Oxford University Press. New York, Oxford.
- Cutler, A., Cutler, R., & Stevens, J. (2011). Random Forests. *Machine Learning Dergisi*, 45(1):157-176. DOI: 10.1007/978-1-4419-9326-7_5.
- Efendioğlu, A., & Sezgin, E. (2007). E-Devlet Uygulamalarında Bilgi ve Paylaşım Güvenliği. *Ç.Ü. Sosyal Bilimler Enstitüsü Dergisi*, 16(2), 219-236.
- Ekmekçi, H. İ. (2019). *Türkiye İşgücü Yapısının Özelliklerinin İncelenmesi: Lojistik Regresyon ve Karar Ağacı Uygulaması*. Hacettepe Üniversitesi. Fen Bilimleri Enstitüsü. Yüksek lisans Tezi, 192, Ankara.

- Eren, K. (2009). İnternet Tüketicisinin Satın Alma Davranışlarının İncelenmesi Üzerine Bir Araştırma. Çukurova Üniversitesi, Sosyal Bilimler Enstitüsü, Yüksek Lisans Tezi, 115, Adana.
- Erken, Ş. (2017). Veri Madenciliği Yöntemleri ve Optimizasyona Dayalı Modeller Üzerinde Bir Araştırma ve Bir Uygulaması. Dokuz Eylül Üniversitesi, Sosyal Bilimler Enstitüsü, Yüksek Lisans Tezi, 106, İzmir.
- Fawcett, T. (2006). An introduction to ROC Analysis. Pattern Recognition Letters, 27(8), 861–874.
- Gittigidiyor, (2021). Erişim Tarihi: 01.01.2021. <https://www.gittigidiyor.com/>
- Hepsiburada, (2021). Erişim Tarihi: 01.01.2021. <https://www.hepsiburada.com/>
- HolyPython, (2020). Naive Bayes Classifier History. Erişim Tarihi: 01.15.2021 <https://holypython.com/nbc/naive-bayes-classifier-history/>
- Hosmer, W., Lemeshow, S. (2013). Applied Logistic Regression. 3. Baskı, John Wiley & Sons, ISBN: 978-1-118-54835-6, 398, Canada.
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An Introduction to Statistical Learning with Applications in R. Springer Science, Business Media New York, ISBN 978-1-4614-7137-0. New York, United States of America.
- Kalaycı, C. (2008). Elektronik Ticaret ve KOBİ'lere Etkileri. Uluslararası İktisadi ve İdari İncelemeler Dergisi, 1(1), 140-150.
- Keskin, A. K. (2018). Makine Öğrenmesi Sınıflandırma Algoritmalarının İncelenmesi. Sinop Üniversitesi, Fen Bilimleri Enstitüsü, Yüksek Lisans Tezi, 42, Sinop.
- Kozak, R. (2007). Seyahat Acentalarının İş Süreçlerinde İnternetin Önemi (TURSAB-BYK Üyeleri Algılamaları Üzerine Bir Araştırma). Anadolu Üniversitesi Sosyal Bilimler Dergisi, 7(1).
- Kuramochi, M., Karypis, G. (2005). Gene Classification Using Expression Profiles: A Feasibility Study. International Journal on Artificial Intelligence Tools, 14(04), 641-660.
- Marangoz, M. (2011). Girişimciler İçin Sınırsız Ticaret. Girişimcilik ve Kalkınma Dergisi, 181-201.
- Markoff, J. (2005). What the Dormouse Said: How the Sixties Counterculture Shaped the Personal Computer Industry. Penguin Books, ISBN: 9780670033829. United States of America.

- Marr, B. (2016). A Short History of Machine Learning. Eriřim Tarihi: 01.15.2021 <https://www.forbes.com/sites/bernardmarr/2016/02/19/a-short-history-of-machine-learning-every-manager-should-read/?sh=1181438c15e7>
- Mık, C. (2018). Tüketicilerin Online Alıřveriř Davranıřlarını Etkileyen Faktörlerin Analiz Edilmesi. Bahçeřehir Üniversitesi, Yüksek Lisans Tezi, 76, İstanbul.
- Mohri, M., Rostamizadeh, A., Talwalkar, A. (2012). Foundations of machine learning. The MIT Press Cambridge, London.
- Nilsson, N. J. (2010). The Quest for Artificial Intelligence: A History of Ideas and Achievements. Cambridge, UK: Cambridge University Press, ISBN: 978-0-521-12293-1.
- Özbay, A., Devrim, J. (2000). E-Ticaret Rehberi. Hayat Yayıncılık, İstanbul.
- Pampel, F. C. (2000). Logistic Regression: A Primer. SAGE Publications Inc, United States of America.
- Statista. (2021). Statista. Number of digital buyers worldwide from 2014 to 2021: <https://www.statista.com/statistics/251666/number-of-digital-buyers-worldwide/>
- Statista. (2021). Statista. Retrieved from Statista, B2B e-commerce GMV worldwide 2013-2019: <https://www.statista.com/statistics/705606/global-b2b-e-commerce-gmv/>
- Tatlídil, H. (2002). Uygulamalı Çok Deęiřkenli İstatistiksel Analiz. Engin Yayınları, Ankara.
- Teknosa. (2021). Eriřim Tarihi: 01.01.2021. <https://www.teknosa.com/>
- Tian, Y., Stewart, C. (2008). History of E-Commerce. USA: Idea Group Inc., DOI: 10.4018/978-1-59904-943-4.ch001.
- Toprak, N. G. (2014). B2C E-Ticaret'in Tam Rekabet Piyasası Çerçevesinde Deęerlendirilmesi: Türkiye Örneęi. Kırklareli Üniversitesi İktisadi ve İdari Bilimler Fakóltesi Dergisi. (ISSN: 2146-3417), 3(1).
- Turan, A. (2008). İnternet Alıřveriři Tüketici Davranıřını Belirleyen Etmenler: Geliřtirilmiř Teknoloji Kabul Modeli (E-Tam) ile Bir Model Önerisi. Akademik Biliřim, 723-731.
- Turanlı, M. (1984). Pazarlama Yönetiminde Karar Alma. Beta Basım Yayım Dağıtım A.ř. 30-34, İstanbul.
- Turing, A. M. (1950). Computing Machinery and Intelligence. Mind, LIX(236), 433-460. doi:<https://doi.org/10.1093/mind/LIX.236.433>.

- Türker, S. (2019). Zararlı Android Yazılımlarının Makine Öğrenmesi ile Ailelerine Göre Sınıflandırılması. Hacettepe Üniversitesi, Fen Bilimleri Enstitüsü, Yüksek Lisans Tezi, 79, Ankara.
- Uygur, E. (2010). E-Ticaret ve Türkiye'deki Durumu. Atılım Üniversitesi, Sosyal Bilimler Enstitüsü, Yüksek Lisans Tezi, 86, Ankara.
- Witten, I., Frank, E., & Hall, M. (2011). Data Mining: Practical Machine Learning Tools and Techniques. Second Edition, Elsevier, San Francisco.
- World Stats, (2021). Erişim Tarihi: 01.01.2021.
<https://www.internetworldstats.com/stats.htm>
- Wu, X., Kumar, V., Quinlan, J., Ghosh, J., Yang, Q., Motoda, H., Steinberg, D. (2008). Top 10 Algorithms in Data Mining. Knowledge and Information Systems, DOI: 10.1007/s10115-007-0114-2., 14(1),1-37.
- Yazar, G. N. (2018). Aile İçi Kadına Yönelik Şiddeti Etkileyen Faktörlerin Lojistik Regresyon ile Analizi. Uludağ Üniversitesi, Sosyal Bilimler Enstitüsü, Yüksek Lisans Tezi, 158, Bursa.
- Yüksel, N. (2019). Üniversite Öğrencilerinin Yurtları Tercih Etme Nedenlerinin Lojistik Regresyon Analizi ile Belirlenmesi: Kırşehir Ahi evran Üniversitesi Örneği. Sivas Cumhuriyet Üniversitesi, Sosyal Bilimler Enstitüsü, Yüksek Lisans Tezi, 105, Sivas.
- Zeytinoğlu, F. (2019). Bayesyen İstatistik. Yalın Yayıncılık, ISBN:6054539892, 9786054539895, E-Book: <https://books.google.com.tr/books?id=mI-vDwAAQBAJ&lpg=PR9&dq=bayes%20teoremi&hl=tr&pg=PR9#v=onepage&q=bayes%20teoremi&f=false> syf:77-78. İstanbul.

ÖZGEÇMİŞ

Adı Soyadı : Rabia ÖZDEMİR
Doğum Yeri ve Yılı : Sivas, 21/12/1994
Medeni Hali : Bekar
Yabancı Dili : İngilizce
E-posta : rabiaozdemir1@gmail.com



Eğitim Durumu

Lise : Şükran Ülgezen Anadolu Meslek Lisesi, 2012
Lisans : İstanbul Ticaret Üniversitesi
İnsan ve Toplum Bilimleri Fakültesi, İstatistik Bölümü
Yüksek Lisans : İstanbul Ticaret Üniversitesi
Fen Bilimleri Enstitüsü, İstatistik Ana Bilim Dalı

Mesleki Deneyim

Çözüm Araştırma Eğitim ve Danışmalık Şirketi, Staj 2015
İstanbul Ticaret Odası, Staj 2016
İstanbul Ticaret üniversitesi Asistan Öğrenci 2016-2017
Affidea Sağlık Hizmetleri A.Ş. 2017-Devam ediyor

Yayınları

ÖZDEMİR, R. & TURANLI, M. (2021). Comparison of Machine Learning Classification Algorithms for Purchasing Forecast. Journal of Life Economics. 8(1), 59-68, DOI: 10.15637/jlecon.8.1.06