T.C.
BANDIRMA ONYEDİ EYLÜL ÜNİVERSİTESİ
SOSYAL BİLİMLER ENSTİTÜSÜ
YÖNETİM BİLİŞİM SİSTEMLERİ ANABİLİM DALI

Yüksek Lisans Tezi

E-TİCARET VERİSİ KULLANILARAK MAKİNE ÖĞRENMESİ ALGORİTMALARININ KARŞILAŞTIRILMASI

Melisa GEYİK
205052009

Tez Danışmanı:
Dr. Öğr. Üyesi Zeynep ÖZER

Bandırma 2024

T.C
BANDIRMA ONYEDİ EYLÜL ÜNİVERSİTESİ
SOSYAL BİLİMLER ENSTİTÜSÜ
YÖNETİM BİLİŞİM SİSTEMLERİ ANABİLİM DALI

Yüksek Lisans Tezi

E-TİCARET VERİSİ KULLANILARAK MAKİNE ÖĞRENMESİ ALGORİTMALARININ KARŞILAŞTIRILMASI

Melisa GEYİK
205052009

Tez Danışmanı:
Dr. Öğr. Üyesi Zeynep ÖZER

Bandırma 2024

ETİK BEYAN
TÜRKİYE CUMHURİYETİ
BANDIRMA ONYEDİ EYLÜL ÜNİVERSİTESİ
SOSYAL BİLİMLER ENSTİTÜSÜ MÜDÜRLÜĞÜNE

Bu belge ile, bu tezdeki bütün bilgilerin akademik kurallara ve etik ilkelere uygun olarak toplanıp sunulduğunu beyan ederim. Bu kural ve ilkelerin gereği olarak, çalışmada bana ait olmayan tüm veri, düşünce ve sonuçları andığımı ve kaynağını gösterdiğimi ayrıca beyan ederim. (08/02/2024)

Melisa GEYİK
İmzası

ÖZET

E-TİCARET VERİSİ KULLANILARAK MAKİNE ÖĞRENMESİ ALGORİTMALARININ KARŞILAŞTIRILMASI

Melisa GEYİK

Her geçen gün gelişen teknoloji ticaret, pazarlama gibi birçok alana yansımış ve işletmeler rakiplerinden geri kalmamak için yenilikleri yakalamak zorunda kalmışlardır. Bu çalışmada Avrupa da ticaret faaliyetini sürdüren bir internet sitesinin verilerinden yararlanarak makine öğrenmesi algoritmaları karşılaştırılmıştır. Söz konusu algoritmalar; Naive Bayes, Destek Vektör Makineleri, K-MEANS, kNN, Regresyondur. Çalışmada kullanılan algoritmalar büyük veri seti ile incelendiğinde en yüksek başarı oranı Destek Vektör Makineleri olurken sırasıyla Naive Bayes, kNN ve Lojistik Regresyon onu takip etmektedir. Haldout yöntemi kullanılan bu çalışmada sonuçlar doğruluk, kesinlik, duyarlılık, kappa değeri ve auc ile değerlendirilmiştir. Sınıf dengeleme yöntemlerinden smote, ros ve rus kullanılmıştır. Çalışmada kullanılan büyük veri de en yavaş sonuç veren algoritmanın destek vektör makineleri olduğu gözlemlenmiştir.

Anahtar Sözcükler: Elektronik Ticaret, Makine Öğrenmesi Yöntemleri, Sınıflandırma, Veri Analizi, Veri Madenciliği

ABSTRACT

COMPARISON OF MACHINE LEARNING ALGORITHMS USING E-COMMERCE DATA

Melisa GEYİK

Ever-evolving technology has been reflected in many areas, such as commerce, marketing, and businesses have had to capture innovation to avoid falling behind their competitors. This study compared machine learning algorithms using data from a website that continues to operate in Europe. The algorithms in question are;Naive Bayes is Support Vector Machines, K-MEANS, kNN, Regression. When the algorithms used in the study are examined with the large dataset, the highest success rate is Support Vector Machines, while Naive Bayes, kNN and Logistics Regression follow, respectively. In this study, the results were evaluated with accuracy, precision, sensitivity, kappa value and auc in this study using the Haldout method. Of the class balancing methods, smote, ros and rus were used. In the big data used in the study, it was observed that the algorithm that gave the slowest results was support vector machines.

Keywords: Electronic Commerce, Machine Learning Methods, Classification, Data Analysis, Data Mining

ÖNSÖZ

Beni bu günlere getiren sevgili aileme, bana eğitim hayatımda katkı sağlayan değerli bütün eğitimcilere ve sevgili danışmanım Dr. Öğr. Üyesi Zeynep Özer'e teşekkür ediyorum.

Melisa GEYİK

Bandırma, 08.02.2024

İÇİNDEKİLER

ETİK BEYAN ii
ÖZET iii
ABSTRACT iv
ÖNSÖZ v
İÇİNDEKİLER vii
TABLOLAR ix
ŞEKİLLER x
GRAFİKLER xi
SEMBOLLER xii
KISALTMALAR xiii
GİRİŞ 1

BİRİNCİ BÖLÜM 3
GENEL KISIMLAR 3
1. LİTERATÜR TARAMASI 3
2. MAKİNE ÖĞRENMESİ 7
2.1. Makine Öğrenmesi Aşamaları 9
3. SINIFLANDIRMA 10
4. MODELLEME 12
4.1. Knn En Yakın Komşuluk Algoritması 13
4.2. Destek Vektör Makineleri (Dvm) 13
4.3. Naive Bayes Sınıflandırıcı 14
4.4. K- Means 15
4.5. Lojistik Regresyon 16
5. MODEL PERFORMANS DEĞERLENDİRME ÖLÇÜTLERİ 17
5.1. Doğruluk (Accuracy) 17
5.2. Kesinlik (Precision) 18
5.3. Kappa Değeri 18
5.4. Duyarlılık (Recall) 19
5.5. F Ölçütü 19
5.6. Auc 19

6. VERİ SETİ HAKKINDA 20

İKİNCİ BÖLÜM 21
YÖNTEM 21
1. KULLANILAN YAZILIM UYGULAMALARI 21
1.1. Kullanılan Kütüphaneler 22
2. VERİ SETİNDEKİ NİTELİKLER 22

ÜÇÜNCÜ BÖLÜM 28
BULGULAR 28
1. SINIF DENGELEME YÖNTEMLERİ 28
1.1. Sentetik Azınlık Yüksek Örnekleme Tekniği (Smote) 28
1.2. Rastgele Aşırı Örnekleme (Ros) 29
1.3. Rastgele Örneklem Azaltma (Rus) 30
2. ANALİZ SONUÇLARI 32
2.1. Knn Yakın Komşuluk Algoritması 32
2.2. Destek Vektör Makineleri 36
2.3. Naive Bayes 38
2.4. K-Means 40
2.5. Lojistik Regresyon 43
SONUÇ 47
KAYNAKÇA 49

TABLOLAR

Tablo 1 : Ülkeler ve Kodları 23
Tablo 2 : Veri Seti Nitelikler Hakkında Hesaplamalar 24
Tablo 3 : Nitelikler Hakkında Bilgiler 25
Tablo 4 : Smote 29
Tablo 5 : Ros 30
Tablo 6 : Rus 31
Tablo 7 : İki Sınıflı Örnek Karmaşıklık Matrisi 32
Tablo 8 : Knn Normal Hata Matrisi (n=5) 33
Tablo 9 : Knn Normal Hata Matrisi (n=3) 33
Tablo 10: Knn Performans Değerlendirme Ölçüt Sonuçları 34
Tablo 11: Destek Vektör Makineleri Normal Hata Matrisi 36
Tablo 12: Destek Vektör Makineleri Metrikleri 37
Tablo 13: Naive Bayes Algoritması Normal Hata Matrisi 38
Tablo 14: Naive Bayes Algoritması Haldout Metrikleri 39
Tablo 15: Regresyon Hata Matrisi 43
Tablo 16: Regresyon Haldout Metrikleri 45

ŞEKİLLER

Şekil 1: Makine Öğrenme Süreci Adımları 9
Şekil 2: Sınıflandırma Model Kurma Aşamaları 11
Şekil 3: Sınıflandırma Test Verileri ve Tahmin Örneği 12
Şekil 4: Destek Vektörleri ve Optimum Hiper-Düzlem 14
Şekil 5: Anaconda Navigator ile Kullanılan Bazı Uygulamalar 21
Şekil 6: Kullanılan Kütüphaneler 22
Şekil 7: DVM Normal Karışıklık Matrisi 37
Şekil 8: Naive Bayes Karışıklık Matrisi 39
Şekil 9: Regresyon Hata Matrisi Isı Haritası 44

GRAFİKLER

Grafik 1 : Ürün Renklerine Göre Tıklama Grafiği 26
Grafik 2 : Ürün Fiyatına Göre Tıklama Grafiği..... 26
Grafik 3 : Ürünlerin Sitedeki Konumuna Göre Tıklama Grafiği..... 27
Grafik 4 : RUS 2 Kde Grafiği..... 31
Grafik 5 : Knn Normal ROC Eğrisi..... 34
Grafik 6 : Naive Bayes Normal ROC Eğrisi..... 40
Grafik 7 : K-Means Küme Sayısını Belirlemek İçin Dirsek Yöntemi..... 41
Grafik 8 : Ülke ve Fiyatlara Göre Müşteri Kümeleme..... 41
Grafik 9 : Dirsek Yöntemi 2..... 42
Grafik 10: Ülke ve Ürün Konumuna Göre Kümeleme..... 42
Grafik 11: Regresyon..... 44
Grafik 12: Normal Regresyon Roc Eğrisi..... 46

SEMBOLLER

g(x) : Transformasyon Değişkeni

p : Başarı Durum Olasılığı

Pr(a) : Bütün Veriye Uyumu

Pr(e) : Uyumun Tesadüf Olma Olasılığı

x : Veri Seti

π : Pi Sayısı

β : Beta Sembolü

KISALTMALAR

DVM : Destek Vektör Makineleri

SMOTE : Sentetik Azınlık Yüksek Örnekleme Tekniği

ROS : Rastgele Aşırı Örnekleme

RUS : Rastgele Örneklem Azaltma

ROC : Alıcı İşletim Karakteristiği

RPL : Düşük Güç ve Kayıplı Ağ Yönlendirme Protokolü

TN : Doğru Negatif

TP : Doğru Pozitif

FN : Yanlış Negatif

FP : Yanlış Pozitif

GİRİŞ

Geçmişten günümüze ticaret faaliyetleri artış göstererek sürekli bir biçimde insan hayatında yer almaktadır. Kuruluşların varlıklarını sürdürebilmesi için ise müşterilerinin ne istediklerini bilmeleri ve gerekli tekniklerle yapılan araştırmalarla elde ettikleri verileri analiz etmeleri gerekmektedir. Veri analizi konusunda; makine öğrenimi algoritmaları, kuruluşların işlerine oldukça yarayacaklardır. Büyük veriyi analiz ederken, hangi algoritmaların hangi veri türleriyle ya da hangi durumlarda daha başarı göstereceği ve algoritmaların çeşitli amaçları olduğundan en uygun olanı seçilmesi yüksek önem arz etmektedir. Ticaret hızlı bir şekilde artış gösterirken, internetinde yaygın kullanımı ile birlikte ticaret ortamı internete entegre hale gelmiştir. Kuruluşların, rekabette geri kalmamak ve müşteri/tüketici zihninde doğru konumunu kaybetmemek için gelişen teknolojileri takip etmesi gerekir. Bir e-ticaret sitesinin özellikle tıklama verilerinin incelenmesi ve işletme amaçları doğrultusunda uygun bir şekilde yorumlanması müthiş bir olumlu katkı sağlayacaktır.

Makine öğrenimi algoritmalarını karşılaştırmak, hangi veri setinde hangi algoritmaların daha yüksek performans verdiğini bilmek çalışmalarda amaca ulaşma yolunda uygun algoritmayı seçmeye yarayacak ve avantaj haline gelecektir.

Bu çalışmada ele alınan veri seti Avrupa ülkelerinde hamile kadınlara yönelik satış yapan bir e-ticaret sitesinin tıklama verileridir. Söz konusu veri seti incelendiğinde işletme açısından ürünlerin bulundukları sayfadan, sayfa konumuna, ürün rengine ve cinsine bağlı olarak hangi ürünlerin daha çok tercih edildiği bilgisine erişebilmekte ve verileri kullanarak gelecek planlarına buna bağlı olarak yapma imkanı sağlamaktadır. İşletmenin Makine Öğrenmesinden yararlandığında kullanıcıların tercihleri doğrultusunda e-ticaret sitesinde ürün ve hizmet parametrelerinin kullanıcılar üzerinde detaylı incelenmesi mümkün hale gelmektedir.

Geçmişten günümüze çok hızlı bir şekilde ilerleyen teknolojiyle birlikte insanların gerek sosyal ağlar gerek diğer teknolojik cihaz ve uygulamalar ile etkileşimini arttırmış ve hızla genişleyen bir veri tabanı oluşturmaya başlamışlardır. İşletmeler için bilgi oldukça kıymetliyken bunca oluşan büyük veriyi kullanmakta önem arz etmektedir (Çelik ve Akdamar, 2018: 256).

Bu çalışmada Haldout yöntemi kullanılmıştır. kNN en yakın komşu algoritması, Destek vektör makineleri (DVM), Naive Bayes sınıflandırıcı, K- Means ve lojistik regresyon kullanılmıştır. Model performans ölçütleri olarak; doğruluk, kesinlik, kappa değeri, duyarlılık, F1 skor ve AUC kullanılmıştır. Sınıf dengeleme yöntemleri olarak ise; Sentetik azınlık yüksek örnekleme tekniği (SMOTE), Rastgele örneklem arttırma (ROS), Rastgele örneklem azaltma (RUS) kullanılmıştır.

Çalışmanın amacı büyük bir veri setinde Makine Öğrenmesi Algoritmalarının Karşılaştırılmasıdır. Çalışmada 165475 satırdan oluşan veri seti kullanılmıştır.

Çalışmanın önemi, literatür taraması yapıldığında Türkçe kaynaklarda e-ticaret tıklama verileri ile Makine öğrenmesi algoritmaları karşılaştırılmasında çok sayıda kaynak bulunamamış olduğundan Literatüre bu anlamda katkı sağlamak hedeflenmiştir.

BİRİNCİ BÖLÜM

GENEL KISIMLAR

1. LİTERATÜR TARAMASI

Çalışmanın bu alanında makine öğrenmesi alanında literatürde bulunan çalışmalar incelenmiştir. Yüksek Öğretim Kurulu Başkanlığı Ulusal Tez Merkezinde yapılan aramada makine öğrenmesi alanında ve makine öğrenmesi algoritmalarının karşılaştırılması filtrelenmesi yapılmıştır. Yüksek Öğretim Kurulu Başkanlığı Ulusal Tez Merkezinde yapılan filtrelemede, makine öğrenmesini içeren yüksek lisans tez sayısının izinli olarak seçildiğinde 820 olduğu görülürken, aynı içerikte doktora tezlerinde 196 adet sonuç olduğu görülmüştür. Makine öğrenmesi algoritmalarının karşılaştırılması filtrelemesi yapıldığında 5 adet sonuç olduğu bunlardan 5 adetinin de yüksek lisans tezi olduğu görüldü. Söz konusu 5 tezden 2 adedi Yönetim Bilişim Sistemden yer alırken, 2 adedi Bilgisayar Mühendisliği alanında ve son olarak Endüstri Mühendisliği Bilim Dallarındadır. Ayrıca Google Akademik kısmından yararlanılmıştır; yapılan filtrelemede "sağlık" kelimesini ve Makine Öğrenmesi kelimelerini içeren çalışmalar 796 adet iken "bilişim" kelimesi içeren makale sayısı 457 olduğu görülmüştür. Sağlık alanında Makine Öğrenmesinin pek çok alana göre daha popüler olduğu söylenebilir. Aynı şekilde "pazarlama" ve "tıklama" filtresi yapıldığında sırası ile 243 ve 35 makale sayısına ulaşılmıştır.

Sürekli gelişen teknoloji veri sayısını arttırmış bu durumda veri depolanmasının daha büyük hale getirmiştir. Büyük verinin kontrolü zorlaştığından gittikçe gürültülü hale gelebilme ihtimali doğurmuştur. Verinin değerli hale gelebilmesi için araştırmacılar tarafından analiz edilip gerçekleşmeyen verilerin tahmin edilmesi yararlı olacaktır. Tahmin edebilmek için makine öğrenmesi algoritmalarına ihtiyaç duyulduğu gibi bu alanlardan birisi de sağlık alanıdır. Şimdiki aşamada Pekel'in sağlık alanında diyabet verilerinin yer aldığı ve makine öğrenimi algoritmalarının karşılaştırdığı bir çalışma incelenmiştir (Pekel, 2018). Söz konusu çalışmada ilk başta sonuçlar DVM, Naive Bayes, Yapay Sinir Ağları ve Karar ağaçları algoritma sonuçları diyabet hakkında veri seti kullanıldığında en başarılı algoritmanın Naive Bayes olduğu yönünde bir sonuca ulaşılmıştır. Söz konusu çalışmada en kötü sonuç değerleri DVM algoritmasında aldığı

görülmüştür. Çalışmanın devamında Genetik Algoritma ile Melez Modeller kullanılması durumunda başarı oranlarının arttığına dair yer verilmiştir. Sonuç olarak ham veri yerine iyileştirme yapılarak verilerin kullanılmasının daha başarılı sonuçlar verebileceği yönünde öneride bulunulmuştur.

Makine Öğrenmesi pek çok farklı alanda kullanılmaktadır. Bu çalışmada e-ticaret verileri incelendiğinden literatürde tüketiciye yönelik makine öğrenmesi algoritma karşılaştırılması da araştırılmış olup şimdi inceleyeceğimiz Arı ya ait çalışma makine öğrenmesi algoritmalarının tüketici yorumlarına yönelik yapılmıştır. İnternet üzerinden yapılan alışverişler her geçen gün artmaktadır ve müşteriler satın aldıkları hizmet ve ürünlerin yorumlarını internet ortamında paylaşmaktadır. Bu paylaşım verileri giderek çoğalmakta ve büyük bir veri haline gelmektedir. Müşteriler yeni arayış içinde oldukları hizmet ve ürünlerin yorumlarından pozitif veya negatif olarak etkilenmektedirler. Çalışmada internet sitesinden, yorumların kelime sayıları ve fotoğraflı yorumlar gibi değişkenler alınmıştır. Söz konusu çalışmada makine öğrenmesi algoritmalarından en güçlü sonuca ulaşanın Karar Ağacı olmuştur. En başarısız sonucu veren algoritma ise Naive Bayes olarak bulunmuştur (Arı, 2022).

Makine Öğrenmesi algoritmalarının birçok alanda kullanıldığından Nesnelerin İnterneti Alanında da kullanılmıştır. Geçmişten günümüze internet kullanıcılarının sayıları artarken kullanıcılar Nesnelerin İnternetine de yönelmektedir. Cihazların birbirleri ile iletişim kurup belirlenen amaçlara işledikleri veriler ile ulaşabilmektedir. Çalışmanın bu bölümünde Kiraz'a ait Düşük güç ve kayıplı ağ yönlendirme protokolü (RPL) tabanı bulunan Nesnelerin İnterneti cihazlarının zayıflıklarını ele alıp bu konuda Makine Öğrenmesi algoritmalarının karşılaştırıldığı çalışma incelenmiştir (Kiraz, 2022). Söz konusu çalışmada RPL alanında farklı saldırılar incelenmiş olup Yapay Sinir Ağları ile birlikte K En Yakın Komşu Algoritmasının en güçlü başarı değerlerinin çıktığı ifade edilmektedir.

Çalışmanın bu alanında E-ticaret faaliyetlerinde Makine Öğrenmesi alanında duygu analizi yapılmış Kılıçer ve Samli'e ait bir makale incelenmiştir (Kılıçer ve Samli, 2023). Kullanılan veriler e-ticaret sitelerinde bulunan yorumlardır. Yorumlar olumlu, olumsuz ve nötr olarak sınıflandırılmıştır. Araştırmada 6 adet Makine Öğrenmesi algoritmasının kullanıldığı görülüyorken çalışmada Naive Bayes Multinominal

algoritmasının en güçlü başarı oranları vermiştir. Araştırmanın sonucunda araştırmacı nötr kelimelerle daha çok çalışılma yapılmasını önermektedir.

Dijitalleşen dünyada işletmeler için tıklama verileri ile kullanıcı analizi büyük önem arz etmektedir. Şimdi inceleyeceğimiz çalışma Yeşiladalı'ya ait olup kullanıcılardan gelen tıklama verileri ile dijital anlamda satın alma davranışlarının modellenmesine odaklanmaktadır. Ticaretin dijitale aktarılmasıyla işletme ve müşteri arasında etkileşim tarzı da değişmiştir. Artık müşteriler için bilgilere ulaşmak daha kolayken beraberinde yeni güçlükler getirmiştir. İşletmeler bu anlamda stratejilerini değiştirirken, yeni oluşan dijital pazar giderek hareketlenmeye başladı. Müşterilerin yürüttüğü mantığı analizlerle anlamaya çalışan söz konusu çalışmada ürünleri alınıp alınmaması noktasında sınıflandırma yöntemleri kullanılmıştır. Çalışmada Lojistik Regresyon 0.71 doğruluk sonucuna ulaşmıştır. Çalışmada en mühim satın alma ihtimalini etkileyen faktörlerin harcanmış miktar, kullanıcının en son görüntülediklerinin ortalaması ve üye dönüşüm oranı olarak ifade edilmektedir (Yeşiladalı, 2018).

Makine öğrenmesinin en yoğun kullanıldığı alanlardan birisinin sağlık alanı olduğunu belirtmiştik. Sağlık hizmeti sağlayıcıların daha doğru kararlar almasına olanak tanıyan kararlarına yönelik destek sistemleri bulunmaktadır. Şimdi inceleyeceğimiz çalışama Erdoğan'a ait olup acile gelen hastaların yatış yapıp yapmayacağını tahmin etmeye yönelik makine öğrenmesi algoritmalarının karşılaştırılacağı bir tezdir. Çalışmada 6 birbirinden farklı model kullanılmış olup, 3 adet makine öğrenimi algoritması karşılaştırılmıştır. Söz konusu çalışma da yapay sinir ağları en güçlü makine öğrenmesi algoritması olarak belirlenmiştir (Erdoğan, 2023).

İncelenen bir diğer çalışma Colley'e ait makine öğrenmesi algoritmalarının finansal değerlendirme durumları için kullanıldığı bir tez çalışmasıdır. Bir işletme açılışı durumunda yapılmış olan kredi skorları verilerini değerlendirmiştir. Risk durumunun çok olduğu ve sürekli olarak başvuruların arttığı dönemde bu başvuruların değerlendirilmesi giderek daha mühim hale gelmektedir. Ardından makine öğrenmesi algoritmaları kullanılmıştır. Sonuç olarak çok katmanlı algılayıcı ve destek vektör makinelerinin en yüksek başarı oranlarına ulaştığı ve verilerin dengeli olduğu ifade edilmiştir (Colley, 2019).

Bilişim dünyasında giderek güvenlik faktörü önem kazanmaktadır. Kullanıcı sayısındaki artışla beraber gelen güvenlik açıkları saldırı tespit sistemlerinin de gelişmesine olanak tanımıştır. Saldırı tespit sistemleri için gerçekleştirilen Özgür ve Erdem'e ait çalışma incelenmiştir (Özgür ve Erdem, 2012). Çalışmada algoritma sonuçları yüksek çıkmış ve algoritmalar sonuçları arasında mühim bir fark olmamakla birlikte birbirlerine çok yakın olduğu ifade edildiği gibi kullanılan veri setinin söz konusu alanda kullanımı tek başına uygun bulunmadığı da ifade edilmektedir. Sanal olmayan ortamda bu kadar yüksek sonuçlar verilmesinin beklenilmediğini de ayrıca ifade etmektedirler.

Sağlık hizmeti sağlayıcıların hastalarına erken teşhiste bulunması çok mühim olduğu gibi yine bu alanda da makine öğrenmesi kullanımı çokça artmıştır. Hastalık tespiti için risk elementlerini bulur ve tahminlemelerde bulunur. Bu alanda inceleyeceğimiz çalışma Kaba ve Kalkan'a ait olup 5 farklı makine öğrenmesi algoritması ile çalışılmış ve karşılaştırılmıştır (Kaba ve Kalkan, 2022). Kullanılan verilerde kalp rahatsızlıkları problemleri parametreleri bulunmaktadır. Söz konusu çalışmada Haldout yöntemi kullanılmıştır. En güçlü algoritma Rastgele Orman algoritması olduğu söylenirken, DVM'nin en başarısız sonuçlara ulaştığı ifade edilmektedir. Araştırmacı sağlık alanı söz konusu olduğunda en güçlü sonuçlara ulaştığı Rastgele Orman algoritmasını tavsiye etmektedirler.

Biyomedikal veriler ile makine öğrenmesi algoritmalarının karşılaştırıldığı Karakoyun ve Hacıbeyoğlu'na ait bir çalışmada 6 farklı algoritma kullanılmış ve yapay sinir ağları algoritmasının en güçlü sonuçlar verdiği ifade edilmektedir. (Karakoyun ve Hacıbeyoğlu, 2014). Söz konusu araştırmada küçük veriler için k en yakın komşu algoritmalarının hızlı sonuç verdiği de belirtilmiştir.

Tütüncü ve Gürsakal'a ait bir diğer çalışmada ekonomi alanında kredi temerrüt riskini önceden tahmin edebilmek için makine öğrenmesi algoritmalarını karşılaştırmışlardır (Tütüncü ve Gürsakal, 2023). Çalışmada en başarılı sonuçlara ulaşan algoritmanın Gradyan Arttırma olduğu ifade edilmektedir.

Makine Öğrenmesi algoritmalarının karşılaştırılmasını göğüs kanseri teşhisi için yapan Sevli'nin çalışmasında, çalışmada veriler rastgele bir şekilde eğitim ve test için

kullanılmıştır (Sevli, 2019). 5 farklı makine öğrenmesi algoritmasının kullanıldığı çalışmada en güçlü doğruluk oranları alarak Lojistik Regresyonun en iyi sonucu elde ettiği ifade edilmektedir.

Bir başka çalışmada otel odalarına ait fiyatların hangi faktörlere göre belirlendiğini makine öğrenmesi algoritmalarını karşılaştırarak hangi algoritma en doğru şekilde tahmin edebileceğine yönelik Ağca'ya ait çalışmadır (Ağca, 2021). Söz konusu çalışmada lojistik regresyonun verdiği sonuçların otel odalarının fiyatını etkileyen faktörlerin belirlenmesinde anlamlı sonuçlar elde edildiği ifade edilmektedir.

Makine öğrenmesi algoritmaları ile metin sınıflandırılması da gerçekleştirilebilmektedir. Şimdi inceleyeceğimiz çalışma duygu analizine odaklanmaktadır ve veriler Türkçe metinlerden oluşmaktadır (Toçoğlu, Çelikten, Aygün, ve Alpkoçak, 2019). Söz konusu çalışmada 4 farklı algoritma üzerinde çalışılmıştır. Veri setinde insani duygular kategorik olarak yer almaktadır. Algoritmaların doğruluk sonuçlarında en güçlü sonuçlar verenin Yapay Sinir Ağları olduğu ifade edilmiştir.

Bir başka çalışmada toplu taşınmaz değerleri için Makine Öğrenmesi algoritmalarının karşılaştırılması üzerine olmuştur (Aydınoğlu, Bovkır ve Çölkesen, 2023). Çalışmada 5 farklı algoritma kullanılmıştır. En başarılı sonuç verenin Rastgele Orman algoritması olduğu, en başarısız sonuçlar verenlerin ise Karar Ağaçları ve Genelleştirilmiş Doğrusal Model olduğu ifade edilmektedir.

Saldırı tespit sistemlerinin başarı oranlarına bakan makine öğrenmesi modellerinin karşılaştırılmasını sağlayan çalışmayı bu aşamada incelemekteyiz (Çebi, Bulut, Fırat, Karataş, & Şahingöz, 2019). Söz konusu çalışmada 7 makine öğrenmesi algoritması kullanılmıştır. Adaboost'un en başarılı sonuçlar verdiği görülürken, algoritmanın süresi hesaba katıldığında en başarılı olan algoritmanın Karar Ağacı olduğu ifade edilmektedir.

2. MAKİNE ÖĞRENMESİ

Teknolojinin her geçen gün ilerlemesi ve insanların teknolojik ürün kullanımlarının artmasıyla birlikte veri sayısı çoğalmaktadır. Oluşan büyük veriler makine öğrenimi algoritmaları ile incelenerek hedeflenen sonuca yönelik önemli avantajlar sağlamaktadır. Kullanıcıların bilgisayar kullanımı arttığı gibi veriler de

artmakta iken makine öğrenmesi kullanıcıların yaptıkları işleri kendi başına yapmakta ve kullanıcı verileri kullanarak kendini eğitmektedir. Tahminlerin yapılabilmesi için buna uygun veri setlerine ihtiyaç duymaktadır (Görmez, 2021).

Makine öğrenmesinde gerçekleşen verileri kullanarak gerçekleşmeyeni tahmin ederek analistlere yardımcı olmak amacıyla kullanılmaktadır (Çelik , 2009: 3).

Makine öğrenmesi, kullanıcıların karar verme tarzlarını taklit ederek, bilinen verilerden bilinmeyeni tahmin etmeye çalışır (Demirci, 2022: 1456).

Denetimli öğrenme yönteminde kullanılan veri seti, eğitim ve test için ayrılmaktadır. Doğru sonuca ulaşabilmek yolunda çalışmada yer edinen veri setinin gerekli miktarda büyük olması önerilmektedir (Onan ve Korukoğlu, 2016). Sınıflandırma algoritmalarını bu yönteme örnek olarak gösterebiliriz.

Denetimsiz öğrenmede ise bir talimat verilmeden ve kullanılan verilerde etikete yer verilmeden istenilen sonuca yönelik kümeler oluşmaktadır. Kümeleme algoritmaları denetimsiz öğrenme yöntemine örnek olarak gösterilebilir. Kümelenme yapılan grupların ögeleri arasında birbirine benzerlik bulunmaktadır. Yüz tanıma, görüntü işleme vb. alanlarda faaliyet gösterebilmektedir (Güldal, 2021).

Yarı denetimli öğrenmede sınıfların bilindiği ve bilinmediği veriler aynı anda kullanılır. Yani etiketli veriler bulunurken etiketsiz verilerde aynı anda çalışma göstermektedir. Etiketli veriler daha azdır ve etiketli olmayan verileri tahmin eder (Kara ve Şamlı, 2021: 416).

Makine öğrenmesi günlük yaşamda pek çok defa karşımıza çıkmaktadır. Örneğin sosyal mecralardan ürün ve hizmet tavsiye etmek gibi reklamcılık alanları, cihazların gelişmesi ile birlikte kullanıcı yüzü tanıma özelliği gibi özelliklerin günlük olarak karşımıza çıkabileceğini görmekteyiz (Günçe, 2021: 20).

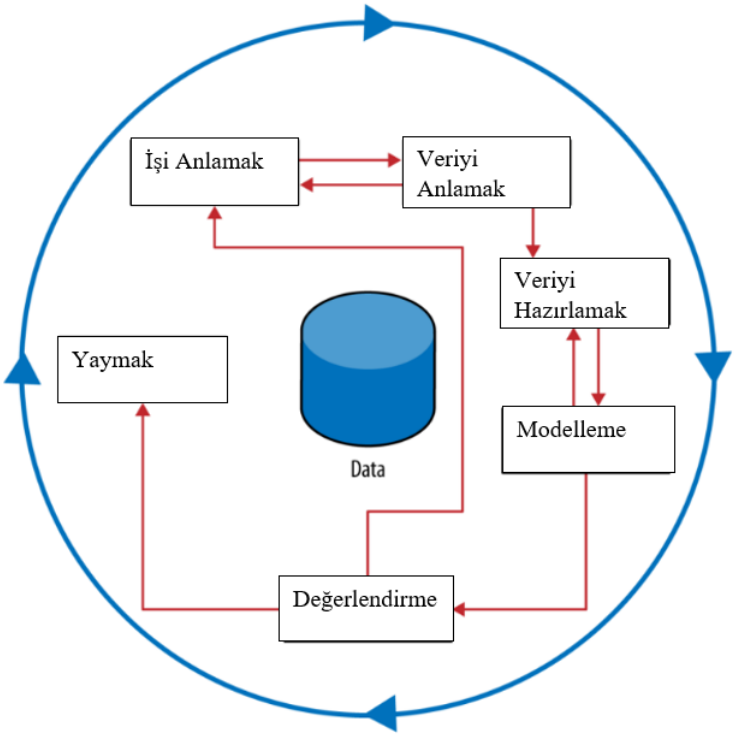
Makine öğrenmesinin en yaygın olarak tercih edildiği alanlar genellikler belirsizlik durumunun fazla olduğu durumlardır. Örnek olarak borsada sürekli değişim gerçekleşmektedir ve belirsizlik olduğundan bu alanda kullanımı yüksektir. Bir diğer örnek ise sağlık alanında tanı koyabilmek için sağlık hizmeti vericiler tarafından kullanılması olarak gösterilebilir (Keleş, Keleş ve Keleş, 2020: 73).

İhtiyaçlar doğrultusunda farklı alanlarda makine öğrenmesi algoritmaları geliştirilmiştir. Bunlar talebe göre sınıflandır ve kümeleme gibi işlemler yürütmektedir. Lojistik Regresyon, Naive Bayes, DVM gibi pek çok Makine Öğrenmesi algoritmaları geliştirilmiştir (Atalay ve Çelik, 2017: 161).

2.1. Makine Öğrenmesi Aşamaları

Veri analizi sürecinde, algoritmaların düzgün ve başarılı bir biçimde çalışmasını sağlayabilmek için temiz verilerle çalışmak önem arz etmektedir. Analiz safhasına gelmeden veri setinde düzenleme yapmak gerekebilir. Eksik, hatalı ya da gürültülü veriler önceden kontrol edilmeli ve başarılı analiz yapabilmek için gerekli veri ön işleme tekniklerinden yararlanılarak veri setini analize uygun hale getirilebilir.

Veriler arasında çok fazla uzak değerler olması durumunda normalizasyon veri ön işleme tekniği uygulanabilir. Farklı kaynaklardan veri kullanılacağı durumunda ise veri entegrasyonu ön işleme tekniği kullanılarak veri setleri arasındaki tutarlılık sağlanabilir (Özcan, 2020: 153).



Şekil 1. Makine Öğrenme Süreci Adımları
Kaynak: Provost, F.; Fawcett, T. (2013), "Data Science for Business: What You Need to Know about Data Mining and Data-Analytic Thinking", O'Reilly Medya, Inc. s.27.

Çalışılacak konunun ne olduğunu anlamak sonuca ulaşmak için çok fazla önem taşımaktadır. Analist öncelikle konunun ne olduğunu tam anlamıyla idrak etmelidir. Konuya uygun veri seti seçilmeli ve seçilen veri setinin analize uygun halde olup olmadığı kontrol edilmeli eğer değilse veri ön işleme teknikleri uygulanmalıdır. Modelleme aşamasında konuya uygun algoritma seçilir. Değerlendirmenin birkaç farklı yöntemi bulunmaktadır. Test edilen sonuçlar istatistiki olarak değerlendirilir, doğruluk payı uygun sonuçlar alındığında dağıtım aşamasına geçilmesi mümkündür (Provost ve Facwcett, 2013).

3. SINIFLANDIRMA

Birbirinden farklı olan verileri kategorize eden veri analizi yöntemlerinden birisidir. Boyutu yüksek olan verilerin daha başarılı analiz edilebilmesine olanak sağlamaktadır. Pek çok farklı alanda kullanılabilmektedir. Makine öğrenmesi algoritmalarından en çok kullanılan sınıflandırma yöntemlerinden Destek Vektör Makineleri ve Lojistik Regresyon örnek gösterilebilir (Pekel, 2018: 15).

Sınıflandırma yapılmadan önce verinin ön işlemeye tabi tutulması, parametrelerin seçimi gibi hususlara bağlı olarak başarı oranı uygulayıcıya bağlı olarak değişebilmektedir. Parametre seçimi her algoritmada aynı değildir (Coşkun ve Baykal, 2011: 2).

Veri madenciliğinde en sık kullanılmakta olan sınıflandırma ile mevcut verileri eğitim verileri olarak kullanır ve yeni kurallar oluşturur. En çok karar ağacında kullanılması ile birlikte regresyon gibi birçok araştırmada da kullanılır. Önce veri setine uygun model seçilir ve verinin rastgele seçilmiş bölümü ile eğitim gerçekleştirilir (Coşlu, 2013: 616).

Sınıflandırma da numerik verilerle işlem yapıldığı gibi kategorik verilerle de işlem yapılabilmektedir. Sınıflandırma denetimli öğrenmeler arasındadır (Kalaycı, 2018:871).

Sınıflandırılma yapılmadan önce verilere dikkat edilmelidir. Kategorize edilmek istenen veriler bölümlendirilmelidir (Uslu ve Akyol, 2021: 15).

Eğitim Verisi

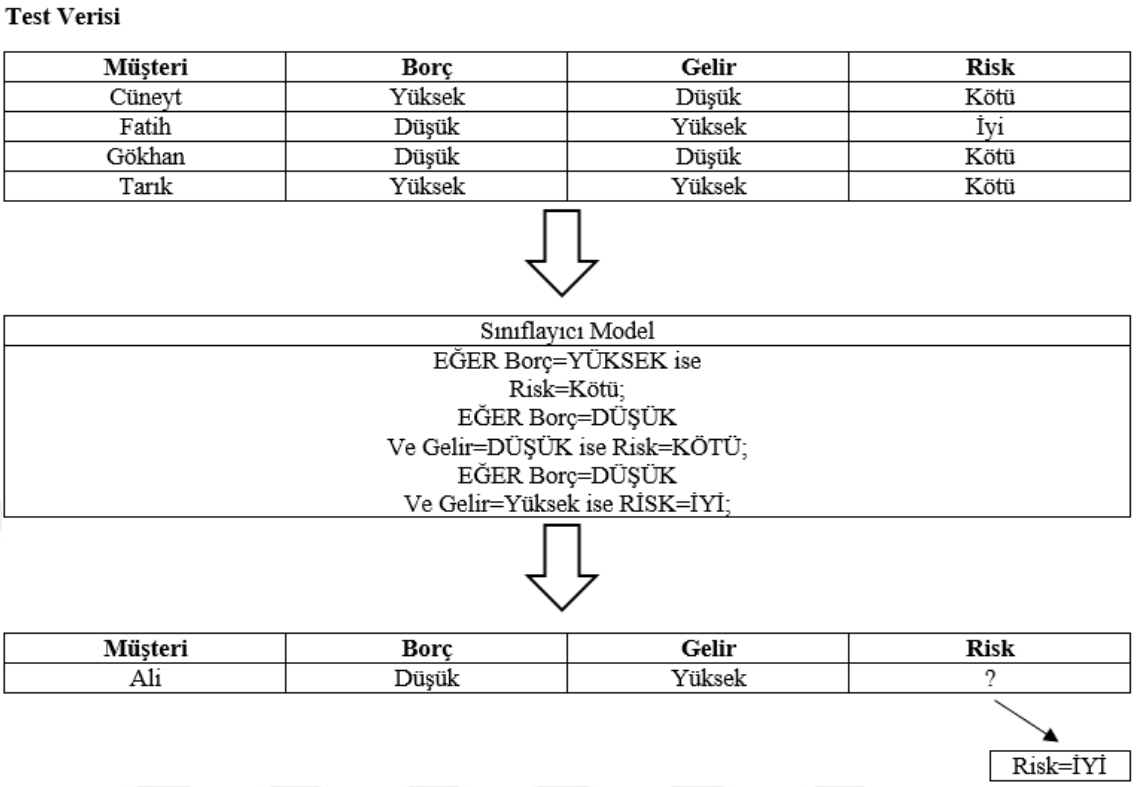
Müşteri	Borç	Gelir	Risk
Ali	Yüksek	Yüksek	Kötü
Ayşe	Yüksek	Yüksek	Kötü
Fatma	Yüksek	Düşük	Kötü
Fuat	Düşük	Yüksek	İyi
Ece	Düşük	Düşük	Kötü
Ayla	Düşük	Yüksek	İyi

Sınıflandırma Algoritması

Sınıflayıcı Model
EĞER Borç=YÜKSEK ise Risk=Kötü; EĞER Borç=DÜŞÜK Ve Gelir=YÜKSEK ise RİSK=İYİ

Şekil 2. Sınıflandırma Model Kurma Aşamaları
Kaynak: Coşlu, E. (2013), "Veri Madenciliği", Akademik Bilişim, s. 23-25.

Kullanılan bir diğer sınıflandırma yönteminde de önce model belirlenir ve daha sonra da test verilerinden yararlanarak tahmin ortaya çıkar (Coşlu, 2013: 617).



Şekil 3. Sınıflandırma Test Verileri ve Tahmin Örneği
Kaynak: Coşlu, E. (2013), "Veri Madenciliği", Akademik Bilişim, s. 23-25.

4. MODELLEME

Model bir çalışmanın temsilcisi ve anlaşılabilir halde olan görüntüsüdür. Çalışmadaki en önemli özellikleri içerir ve doğrulama teknikleri ile güvenirliği ölçülebilmektedir (Maria, 1997: 7).

Modeller oluşturulurken yeniden kullanıma uygun bir şekilde tasarlanmaktadır. Farklı araştırmalarda da kullanılabilecek genelleyebilir olması gerekmektedir (Olkun, Şahin, Akkurt, Dikkartın ve Gülbağcı, 2009: 67).

Test ve eğitim verilerinin de seçimine bağlı olarak model başarısı değişmektedir (Coşkun ve Baykal, 2011: 3).

Bu çalışmada Haldout yöntemi uygulanmıştır. Haldout yönteminde eğitim ve test veri setleri olarak veriler ayrılmaktadır. Çalışmaya uygun parametreler ve model performans değerlendirme ölçütleri kullanılmaktadır.

4.1. Knn En Yakın Komşuluk Algoritması

Denetimli makine öğrenmesi algoritmalarından olan kNN en yakın komşuluk algoritması, ilk başlarda sınıflandırma çalışmalarında kullanılmak üzere oluşturulmuştur ve daha sonra regresyon problemleri çözümünde ele alınmaya başlamıştır (Nacar ve Erdebilli, 2021).

Test verileri ile eğitim seti oluşturulmaktadır ve daha sonra çalışmaya uygun olarak seçilen k uzak noktası ile tüm verilerin uzaklıkları hesaplanmaktadır (Kaynar, Arslan, Görmez ve Işık, 2018: 178).

Kendi alanında en sık kullanılan algoritmalardan bir tanesidir ve kNN en yakın komşu algoritmasında en sık Öklid, Kosinüs ve Manhattan uzaklıkları kullanılmaktadır (Kılınç, Borandağ, Yücalar, Tunalı, Şimşek ve Özçift 2016: 90).

Çalışmada kullanılan veri setinde, belirlenen k uzaklığı baz alınmak üzere veriler arasındaki benzerlikler tespit edilerek sınıflandırma işlemi yapılması amaçlanmaktadır.

Bu algoritmanın sıkça tercih edilmesinde gürültülü veri analizinde güçlü olması ve uyarlanması kolay olması gibi avantajları yer almaktadır. Avantajları olduğu gibi beraberinde dezavantajlarını da getirebilmektedir. Dezavantajları için örnek verilmesi gerektiğinde; çok büyük veri ile çalışıldığı durumlarda yavaşlaması söz konusu olabilmektedir (Taşçı ve Onan, 2016).

Kısaca özetlemek gerekirse çalışmada ilk olarak k değeri belirlenir ve daha sonra en yakın uzaklıktaki sınıf etiketine ataması yapılır.

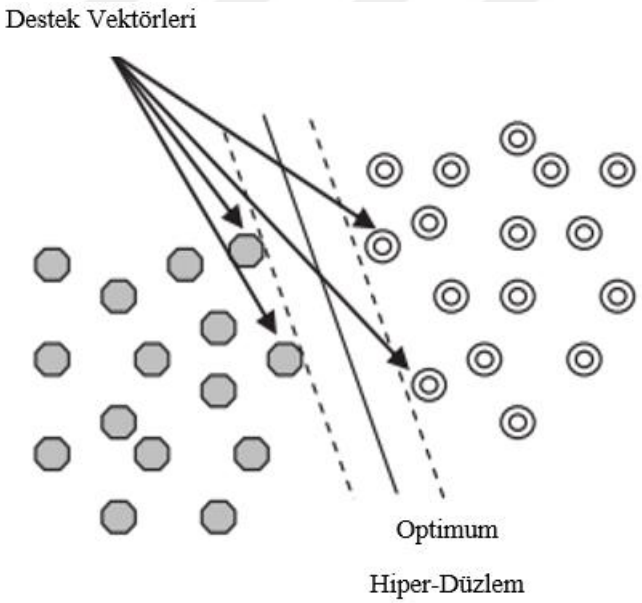
4.2. Destek Vektör Makineleri (Dvm)

Bu algoritma ilk olarak Vapnik tarafından geliştirilmiştir (Ayhan ve Erdoğmuş, 2014: 177).

Çok yüksek genelleme yapma özelliği bulunan Destek Vektör Makineleri aynı zamanda birçok farklı alanda kullanılmıştır. Çalışmada veriler doğrusal bir şekilde ayrıştırıla bilindiğinde işlemler daha da kolaylaşmaktadır. Verilerin doğrusal olmadığı durumda ise doğrusal sınıfa dönüştürülebilecekleri başka bir uzaya aktarılması gerekmektedir (Eray, 2008: 44).

DVM algoritması çalışmada kullanılan verileri iki boyutta veya daha fazla sınıfta ayırmaktadır (Ayhan ve Erdoğmuş, 2014:178). Veri kümesindeki değişkenler benzer ve bağımsız dağılım göstermektedir (Yakut, Elmas ve Yavuz, 2014: 143). Destek vektör makinası kullanıldığında hiper düzlem bulunması amaçlanır ve veriler en belirgin özelliklerine göre sınıflandırılır (Çalış, Gazdağı ve Yıldız, 2013: 3).

Destek Vektör Makinelerinde mevcut verilerden yararlanılarak, yapılması gereken işleri tahmin eden bir algoritmadır. Karar Destek Sistemleri arasında bulunur ve yapısal riski azaltması yaptığı işlemlerdendir. Sınıflandırma ve Regresyon alanlarında kullanıldığı çeşitleri bulunmaktadır. Görüntü işleme gibi pek çok alanda kullanıldığı bilinmektedir (Güner ve Çomak, 2011: 90).



Şekil 4. Destek Vektörleri ve Optimum Hiper-Düzlem
Kaynak: Kavzoğlu, T. ve Çölkesen , İ. (2010), "Destek Vektör Makineleri ile Uydu Görüntülerinin Sınıflandırılmasında Kernel Fonksiyonlarının Etkilerinin İncelenmesi", **Harita Dergisi**,Cilt 7, Sayı 144, s.76.

4.3. Naive Bayes Sınıflandırıcı

Naive Bayes, Bayes teoremine dayanmaktadır. Verilerin hangi sınıfta bulunduğunu ve bunun olasılığını hesaplamaktadır. Niteliklerin arasında önem farkı bulunmamaktadır. Naive Bayes sınıflandırıcıda kullanılan bütün niteliklerin birbiri arasında bağımsız kabul edilir (Solmaz, Günay ve Alkan, 2014: 4).

Naive Bayes sınıflandırıcı çalışmada değerlendirilen olayın gerçekleşme ihtimali belirlemede yardımcı olan bir algoritmadır (Kazan ve Karakoca, 2019: 19).

Denetimli öğrenme işlemini gerçekleştiren bir algoritmadır. Duygu analizi ve veri madenciliği gibi pek çok farklı alanlarda kullanılmaktadır (Şahinaslan, Dalyan ve Şahinaslan, 2022: 225).

Sınıflandırıcıda ilk başta iki niteliğe bakıldığında bağımlılık yüksek boyutta olabilirken, işin içine daha fazla nitelik girdiği zaman durum büsbütün değişebilmektedir (Zhang, 2004).

Naive Bayes, rastgele değişkenlere bakarak içerisinde bulunan olasılık dağılımı ve marjinal olan olasılıklar arasında bulunan ilişkiyi göstermektedir (Abdelsalam Algahani, 2021: 41).

4.4. K- Means

K-Means Algoritması, Makine Öğrenmesi alanında en çok kullanılan popüler algoritmalardan birisidir. Veri setinde bulunan özelliklere göre kümelemeler gerçekleşmektedir. Verileri k adet sınıflandırmak amaçlanırken, algoritma çalışması sonucunda veriler kümeler olarak sınıflandırılmaktadır. İlk başta belirlenen k adetine göre çalışacak algoritma olduğundan ilk adımlarda sonuç vermeyebilir ve uygun sonucu bulabilmek amacıyla farklı k adetleri kullanılması gerekebilir. Kullanılan veri seti sayısal olmalıdır ve gürültülü veri içermemelidir. Algoritma hızlı çalışır ve işlemler gerçekleştiğinde kümeleri grafik veya sayılar gibi çeşitli şekilde görebilmek mümkündür (Dinçer, 2006: 55).

Veri setindeki her eleman yalnızca bir kümeye ait olabilir ve algoritma gözetimsiz bir şekilde gerçekleşmektedir. Aynı kümede yer alan veriler birbirleri arasında birçok ortak özellik göstermekte ve çok benzemekte iken farklı kümelerdeki veriler birbirleriyle olabildiğince az düzeyde benzer özellikler göstermelidir. Verilerin aynı anda farklı kümelerde yer alamaması ve en başta belirtilen n sayıda verilen ve k adet belirtilen kümeye bölünmesi dolayısıyla keskin bir algoritmadır (Sarıman, 2011: 195).

4.5. Lojistik Regresyon

Lojistik regresyon analizi ilk olarak 1845'li yıllarda görülmektedir. İlk yıllarda nüfus artışını incelemek gibi alanlarda kullanılırken ilerleyen yıllarda bu analiz birçok farklı alanda kullanılırken en çok kullanılan alan ise tıp olmuştur (Çokluk, 2010: 1359).

Lojistik regresyon, doğrusal regresyon analizinin aksine birçok olasılık içermeyen bir regresyon analizidir. Diskriminant ve kümeleme analizleri gibi bazı analiz türlerine de benzemektedir (Şenel ve Alatlı, 2014: 36).

İki farklı değerli işlemlerin atanmasında da kullanımı uygundur. Analiz sonucunda iki değer, ilişkili kümeye atanmakta ve doğruluğu tespit edilmektedir. Başarılı ve başarısız durumların oranına lojistik dönüşüm oranı denilmektedir. Başarılı durumda aşağıda Denklem 1.1'deki formül kullanılır: (Oğuzlar, 2005: 22) (p: başarı durum olasılığı)

$$log\ it(\rho_i) = log\ (\frac{\rho_i}{1-\rho_i}) \quad \textbf{(1.1)}$$

Sınıflandırmanın amaçlandığı lojistik regresyon analizinde bağımlı ve bağımsız değişken arasındaki ilişki araştırılır ve atama yapılır. Süreklilik olması gerekmediği gibi risk olasılıklarını da dahil edilmektedir (Girginer ve Cankuş, 2008: 185).

Lojistik regresyonda, bağımlı değişkenin sürekli olması ve bağımsız değişkenlerinin çoklu normal dağılım göstermesi gerekmemektedir. Bu durum lojistik regresyonu diğer regresyon modellerinden ayıran önemli farklardan bir tanesidir. Bir diğer fark ise varyans- kovaryans matrislerinde bire birlik aramamasıdır (Ege ve Bayrakdaroğlu, 2009: 147).

Lojistik regresyon analizi üç farklı şekilde çalışmaktadır. Bunlardan ilki İkili (Binary) lojistik regresyon modelidir. Bu model bağımlı değişkenin ikili olması halinde kullanılır (örneğin: evet, hayır gibi). Bir diğer model ise Multinominal Lojistik Regresyon Modelidir. Bağımlı değişkenin çok seçenekli olması durumudur. Son model ise Sıralı (Ordinal) Lojistik Regresyon Modelidir. Bu durumda sıralı bir durum mevcuttur ve birçok farklı kategoride yer almaktadır (Şenel ve Alatlı, 2014: 36).

Lojistik regresyon formülü Denklem 1.2'de gösterilir: (Bircan, 2004: 189).

$$\pi(x)[1 + exp(-\beta_0 - \beta_1 X)]^{-1} \quad \textbf{(1.2)}$$

Logit transformasyonu formülü Denklem 1.3 ile gösterilir:(Bircan, 2004: 190).

$$g(x)\ In\ [\frac{\pi(x)}{l-\pi(x)}] = \beta_0 + \beta_1 x \quad \textbf{(1.3)}$$

Lojistik regresyonda test ve tahmin verilerinin karşılaştırılması log olabilirlik fonksiyonu ile mümkündür. (Bircan, 2004: 192).

$$D\ = -2\ In \frac{\text{Şu Anki Modelin Olabilirliği}}{\text{Doymuş Modelin Olabilirliği}} \quad \textbf{(1.4)}$$

Bağımsız değişkenin önemini belirleyebilmek için denklemin D olan ve olmayan halleri karşılaştırılır. Ortaya Denklem 1.5'teki gibi bir sonuç çıkmaktadır (Bircan, 2004: 192).

$$\text{G}\ =\ \text{D(Değişkensiz Model)} - \text{D(Değişkenli Model)} \quad \textbf{(1.5)}$$

5. MODEL PERFORMANS DEĞERLENDİRME ÖLÇÜTLERİ

Bu çalışmada performans değerlendirme ölçütlerinden; doğruluk, kesinlik, kappa değeri, duyarlılık, F1 skor ve AUC kullanılmıştır.

Karmaşıklık matrisinde;

TP: Doğru Pozitif

FN: Yanlış Negatif

FP: Yanlış Pozitif

TN: Doğru Negatif metrikleri kullanılmıştır.

5.1. Doğruluk (Accuracy)

Doğruluk; kullanılan araştırma yöntemlerinde, test sonuçları ve gerçek verilerin birbirlerine yakınlık derecesidir (Ertaş ve Kayalı, 2005: 46).

Doğruluk dünyada en çok kullanılan performans değerlendirme ölçülerindendir. Bir modelin doğruluk oranını gösterdiği gibi kullanılması oldukça basittir. Doğru sınıflandırılmış örnek sayısının toplam örnek sayısına bölünmesi ile sonuç elde edilmektedir. Denklem 1.6 'da ki gibi hesaplanmaktadır (Nizam ve Akın, 2014: 4).

$$\text{Doğruluk} = \frac{(\text{TP}+\text{TN})}{(\text{TP}+\text{FP}+\text{FN}+\text{TN})} \quad \textbf{(1.6)}$$

Doğruluk ile birlikte gerçek durumda pozitif verilerin tahminde de pozitif bulunması ve gerçekte negatif verilerin de negatif olarak tahmin edilmesi oranına bakılmaktadır (Gürkahraman ve Karakış, 2021: 1006).

5.2. Kesinlik (Precision)

Denklemde ifade edilen TP (Doğru Pozitif) pozitif olarak ifade edilen verilerin kesinlik değerleri sonucunda doğru oranda atanması oranını ifade etmektedir. FP (Yanlış Pozitif) ise gerçekte negatif olan fakat pozitif olarak tahmin edilmiş değerleri ifade etmektedir. Denklemin alt kesrine bakacak olursak bütün pozitif olarak atanmış verileri içermektedir. Denklem 1.7: (Nizam ve Akın, 2014: 4).

$$\text{Kesinlik} = \frac{\text{TP}}{(\text{TP}+\text{FP})} \quad \textbf{(1.7)}$$

Sıfır ile bir arasında değerler alan kesinlik değerinin bire yaklaştıkça doğruluk oranı yükselmektedir.

5.3. Kappa Değeri

Kappa değerinde iki birbirinden bağımsız olasılık hesaplanırken, iki değer arasında analiz yapar ve benzerliğin olup olmadığını inceleyen daha sonra ise bu durumun güvenliğine bakan bir istatistik yöntemidir (Aydemir, Işık ve Tuncer, 2021: 521).

$$\text{K} = \frac{\text{Pr(a)}-\text{Pr(e)}}{1-\text{Pr(e)}} \quad \textbf{(1.8)}$$

Denklem 1.8'de (Aydemir, Işık ve Tuncer, 2021:521) Pr(a) analiz sonucunda ortaya çıkan uyumun bütün veriye ne kadar uyumlu olduğunu gösterirken, Pr(e) ise ortaya çıkan bir uyum varsa bunun tesadüfen oluşma olasılığını ifade etmektedir.

Kappa değeri 0 ve 0'dan küçükse zayıf uyumu gösterirken, bu sayı 1'e yaklaştıkça yüksek uyum olarak yorumlanmaktadır. 0.00 ve 0.20 değerleri arası önemsiz derecede uyumu gösterirken, 0.21 ve 0.40 arası düşük, 0.41 ve 0.60 değerleri arası orta seviye uyum, 0.61 ve 0.80 değerleri arası önemli uyum, 0.81 ve 1.00 değerleri arası ise çok yüksek derecede uyum şeklinde yorumlanmaktadır (Benligül, Bektaş ve Arslan, 2022: 90).

5.4. Duyarlılık (Recall)

Duyarlılık, analiz sonucunda gerçekten pozitif olan verilerin doğru bir şekilde pozitif olarak tahmin edilmesidir (Gürkahraman ve Karakış, 2021: 1006).

$$\text{Duyarlılık} = \frac{\text{TP}}{(\text{TP}+\text{FN})} \quad \textbf{(1.9)}$$

Sıfır ile bir arasında sonuçlanan duyarlılık değeri bire yakınlaştıkça performans değeri artmaktadır.

Denklem 1.9'da doğru bir şekilde tahmin edilmiş pozitif örnek sayısının toplamda var olan pozitif veri sayısına oranı hesaplanmaktadır (Nizam ve Akın, 2014: 4).

5.5. F Ölçütü

Kesinlik ve Duyarlılık ölçütlerinin harmonik ortalaması F ölçütünü oluşturmaktadır. Denklem 1.10'da F Ölçütü Formülü yer almaktadır (Coşkun ve Baykal, 2011: 4).

$$\text{F Ölçütü} = \frac{2 \text{ X Duyarlılık X Kesinlik}}{\text{Duyarlılık}+\text{Kesinlik}} \quad \textbf{(1.10)}$$

Tek başına Kesinlik ve Duyarlılık ölçütleri anlamlı bir sonuç çıkmasında yeterlik olmadığından, bu ölçütlerin harmonik ortalaması alınarak F ölçütü geliştirilmiştir (Nizam ve Akın, 2014: 4).

Sıfır ile bir arasında sonuçlanan F Ölçütü bire yaklaştıkça yüksek performans şeklinde yorumlanmaktadır.

5.6. Auc

Roc eğrisinin altında kalan alan (Auc) 1950'li yıllarda radar sorunları için kullanılmaya başlanmış olup temeli istatistiksel karar teorisine dayanan bir eğridir. Deneysel psikoloji gibi çeşitli alanlarda da kullanımı mevcuttur (Tomak ve Bek, 2011: 58-59).

İki farklı yaklaşım şekli ile en iyi nokta bulunabilir. Bunlardan birisi eğrinin (0,1) koordinatlarına en kısa mesafede olan noktayı kesim noktası olarak ele almaktır. Diğer yaklaşımda ise en iyi noktayı bulabilmek için duyarlılık ve seçicilik gibi özelliklerin en yüksek çıktığı nokta ele alınabilir (Kılıç, 2013: 136).

Auc oluşturulduğunda koordinat sisteminde doğru pozitifler ve yanlış pozitifler yer almaktadır. Roc eğrisi ile birlikte işlemlerin gerçek özelliklerini ele alan içbükey bir eğri ile karşılaşılmaktadır (Akçay ve Demirel, 2011: 155).

Yapılan testlerin sınıflama alanında ne kadar iyi olduğunu görebilmek için Roc eğrisi altında kalan kısma bakılmalıdır. Söz konusu alan ne kadar büyükse sınıflama da o kadar başarılı demektir (Taşdemir, 2022).

Roc eğrisi altında kalan alan [0,1] aralığında değerlerini almaktadır ve yükseldikçe başarı durumu artmaktadır (Onan, 2017: 8).

6. VERİ SETİ HAKKINDA

Çalışmada kullanılan veri seti; hamile kadınlar için giyim ürünleri sunan bir e-ticaret sitesinin verilerinden oluşmaktadır ve UCI Machine Learning Repository adlı veri tabanı sitesinden alınmıştır. Söz konusu veri setinin adı Clicstream Data For Online Shopping'dir. 2008 yılının Nisan ayından başlayarak Ağustos ayına kadar olan veriler olmak üzere on dört nitelikten oluşmaktadır (Łapczyński ve Białowąs, 2013). Bu nitelikler; Yıl, Ay, Gün, Sipariş, Ülke, Oturum Kimliği, Kategori, Ürün, Renk, Fotoğrafın Sayfadaki Konumu, Fotoğraf Modeli, Ortalama Fiyat, Sayfa Numarasıdır.

Ülkeler arasında kırk iki ülke bulunmaktadır. Dört farklı ürün kategorisi bulunur. Bu kategoriler ise pantolon, bluz, etek ve indirimli ürün olarak ayrılmaktadır. Renk olarak on dört farklı seçenek bulunmaktadır.

Fotoğrafın sitede bulunduğu konum bilgisi de veri setindeki niteler arasında yer almaktadır. E-ticaret sitesinin ekranı altı farklı alana bölünmüş olup numaralandırılarak veri setinde yer edinmiştir. Fotoğraf modeli niteliğinde iki seçenek bulunur bunlarda fotoğrafın profilden mi yoksa doğrudan mı olduğu hakkındadır.

Ürün fiyatları Amerikan doları üzerindendir.

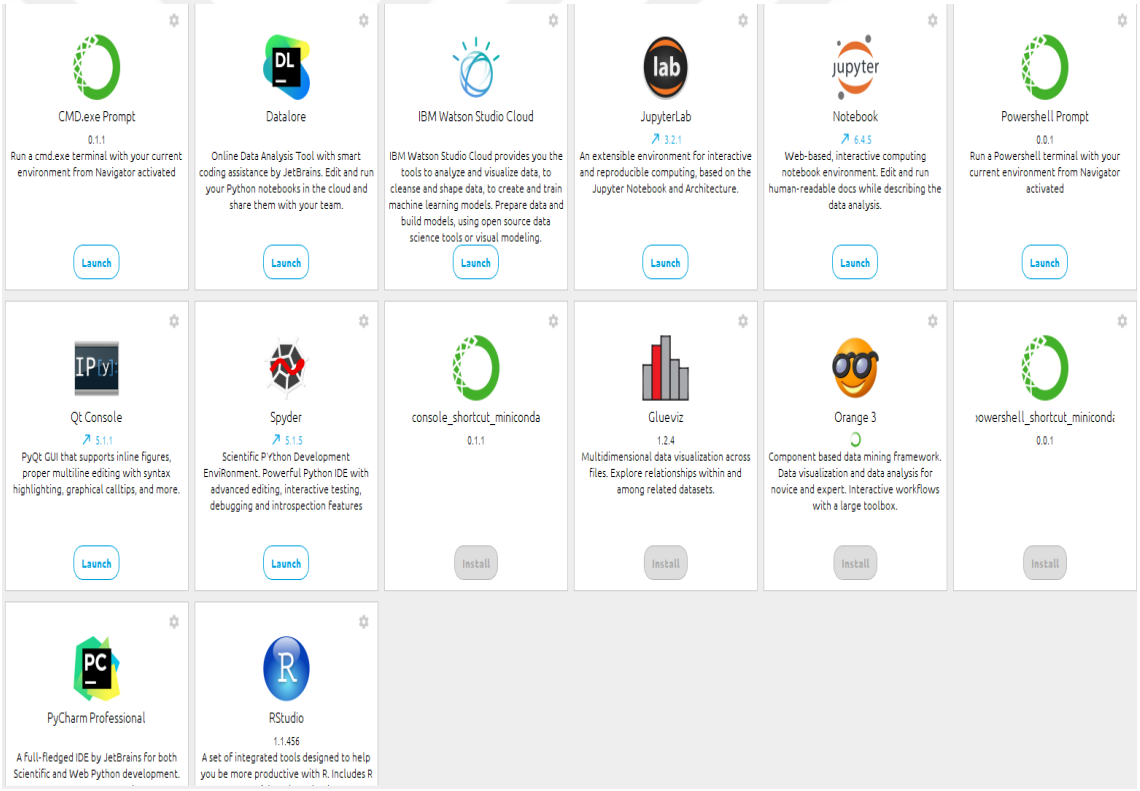
Veri setinde boş veri bulunmamaktadır. Toplam 165 bin 474 adet satır sayısı içermektedir.

İKİNCİ BÖLÜM

YÖNTEM

1. KULLANILAN YAZILIM UYGULAMALARI

Bu çalışmada Python 3.9.7 yazılım dili kullanılmıştır. Anaconda Navigator adlı açık kaynaklı dağıtım platformunda Spyder IDE 5.9.2 versiyonu kullanılmıştır. Anaconda dünyanın en popüler dağıtım platformlarından birisidir. Binlerce açık kaynaklı veri bilimi ve makine öğrenimi paketlerini bünyesinde bulundurmaktadır. Python dili için oluşturulmuş olsa da birçok farklı yazılım dil kullanılabilmektedir. Ayrıca bünyesinde RStudio, PyCharm Professional gibi uygulamaları bulundurur.



Şekil 5. Anaconda Navigator ile Kullanılan Bazı Uygulamalar

Ayrıca yukarıda bulunan resimde Anaconda ile elde edebileceğiniz bazı uygulamalar yer verilmiştir. Anaconda'ı ücretsiz olarak kendi web sitesinden indirebilme imkanı bulunur.

1.1. Kullanılan Kütüphaneler

Bu çalışmada aşağıdaki kütüphaneler kullanılmıştır:

```
#Kullanılan Kütüphaneler
import os
import missingno as msno
import seaborn as sns
import matplotlib.pyplot as plt
import numpy as np
import pandas as pd
import imblearn
import matplotlib.pyplot as plt
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import classification_report, confusion_matrix
from sklearn.metrics import accuracy_score, cohen_kappa_score, confusion_matrix
from sklearn.model_selection import train_test_split
from sklearn.metrics import r2_score, mean_absolute_error, mean_squared_error
from collections import Counter
from imblearn.over_sampling import SMOTE
from matplotlib import pyplot
from numpy import where
from imblearn import under_sampling, over_sampling
from sklearn.metrics import confusion_matrix
from sklearn.metrics import classification_report
from imblearn.over_sampling import RandomOverSampler, SMOTE
from imblearn.under_sampling import RandomUnderSampler, NearMiss
from sklearn.model_selection import train_test_split, GridSearchCV
from sklearn.metrics import mean_squared_error,r2_score
from sklearn.neighbors import KNeighborsRegressor
from sklearn.svm import SVC
from matplotlib.colors import ListedColormap
from sklearn import metrics
from sklearn.linear_model import LogisticRegression
from sklearn.preprocessing import StandardScaler
from sklearn.cluster import KMeans
from sklearn.naive_bayes import GaussianNB
from sklearn.metrics import accuracy_score
```

Şekil 6. Kullanılan Kütüphaneler

2. VERİ SETİNDEKİ NİTELİKLER

Çalışmada kullanılan veri setinde ilk sırada olan nitelik; verilerin hangi yılda alınmış olduğudur. Kullanılan veri setinde yalnızca 2008 yılının verileri bulunmaktadır. İkinci nitelik ise ayları göstermektedir. Bu aylar sırasıyla Nisan, Mayıs, Haziran, Temmuz ve Ağustos aylarından oluşmaktadır. Üçüncü nitelikte Ayın günlerini tek tek ele almaktadır. Dördüncü nitelik tıklama dizilerinden oluşmaktadır. Beşinci nitelikte işlemlerin gerçekleştiği ülkeler yer almaktadır.

Veri setinde kullanılan her ülkeye bir sayı atanmıştır.12 kodlu ülke veri seti açıklamasında tanımlanamayan olarak belirtilmiştir. Sırası ile bu ülkeler ve atanan sayıları:

Tablo 1. Ülkeler ve Kodları

1-Avustralya	2-Avusturya	3-Belçika
4-İngiliz Virgin Adaları	5-Cayman Adaları	6-Noel Adası
7-Hırvatistan	8-Kıbrıs	9-Çek Cumhuriyeti
10-Danimarka	11-Estonya	12-Tanımlanamayan
13-Faroe Adaları	14-Finlandiya	15-Fransa
16-Almanya	17-Yunanistan	18-Macaristan
19-İzlanda	20-Hindistan	21-İrlanda
22-İtalya	23-Letonya	24-Litvanya
25-Lüksemburg	26-Meksika	27-Hollanda
28-Norveç	29-Polonya	30-Portekiz
31-Romanya	32-Rusya	33-San Marino
34-Slovakya	35-Slovenya	36-İspanya
37-İsveç	38-İsviçre	39-Ukrayna
40-Birleşik Arap Emirlikleri	41-Birleşik Krallık	42-ABD

Altıncı nitelikte oturum kimliğini gösteren kısa kayıt bulunmaktadır. Yedinci nitelikte e-ticaret sitesindeki ana ürünler kategorisi bulunmaktadır. Ana ürünler sayılara atanmıştır. Ürünler ve atandıkları sayılar ise; ({1:pantolon, 2: etek, 3:bluz , 4: indirimli ürün}) şeklindedir.

Yedinci nitelikte her ürünün kodları bulunmaktadır ve bulunan ürün sayısı 217 adettir. Sekizinci nitelikte ürünlerin renkleri numaralandırılmıştır. Bahsedilen renkler ve kodları şöyledir: 1-bej 2-siyah 3-mavi 4-kahverengi 5-bordo 6-gri 7-yeşil 8-lacivert 9-birçok renkten 10-zeytin 11-pembe 12-kırmızı 13-menekşe 14-beyaz olarak kodlanmıştır.

Dokuzuncu nitelikte e-ticaret sayfasında fotoğrafların konumları bulunmaktadır ve e- ticaret sayfası ekranı 6 bölüme ayrılmıştır. Bu bölümlerin numaralandırılması şöyledir: 1-Sol üst 2-Üst ortada 3- Sağ üst 4- Sol alt 5-Alt ortada 6-Sağ alt olarak numaralandırılmıştır.

Onuncu nitelikte model fotoğrafı ile ilgilidir. Bu da iki numaraya ayrılmıştır; 1- yüz yüze ve 2- profil şeklindedir.

On ikinci nitelikte ürünlerin dolar üzerinden fiyatları bulunmaktadır.

On üçüncü nitelik belirli bir ürünün fiyatının tüm ürün kategorisinin ortalama fiyatından yüksek olup olmadığını bildirmektedir. Burada 1- evet 2- hayır şeklindedir.

Son nitelik olan on dördüncü nitelikte ise 1'den 5'e kadar web sitesine ait sayfa numaraları yer almaktadır. Ver setinde 14 adet değişken bulunmakta olup, 165474 satır veriden oluşmaktadır. Veri setinde toplamda 2316636 eleman bulunmaktadır. Veri setinde hiç boş eleman bulunmamaktadır. Veri setinde sayfa 2 (giyim modeli) değişkeninin tipi object olurken, geri kalan bütün değişkenler int64 tipindedir.

Tablo 2. Veri Seti Nitelikler Hakkında Hesaplamalar

Nitelik Adı	Min	Max	Varyans	Ortalama	Standart Sapma
yıl	2008	2008	0.000000	2008.000000	0.000000
ay	4	8	1.764009	5.585887	1.328160
gun	1	31	7.797550	14.524554	8.830374
dizi	1	195	1.816676	9.817476	13.478411
ülke	1	47	5.113238	26.952621	7.150691
kısa ID	1	24026	4.911794	12058.417056	7008.418903
sayfa 1 (ana kategori)	1	4	1.309697	2.400842	1.144420
sayfa 2 (giyim modeli)	A1	P9			
renk	1	14	1.794036	6.227655	4.235606
konum	1	6	2.935074	3.258198	1.713206
model fotoğrafı	1	2	1.924353	1.260071	0.438674
fiyat	18	82	1.574556	43.802507	12.548131
fiyat 2	1	2	2.498615	1.488167	0.499861
sayfa	1	5	9.651336	1.710166	0.982412

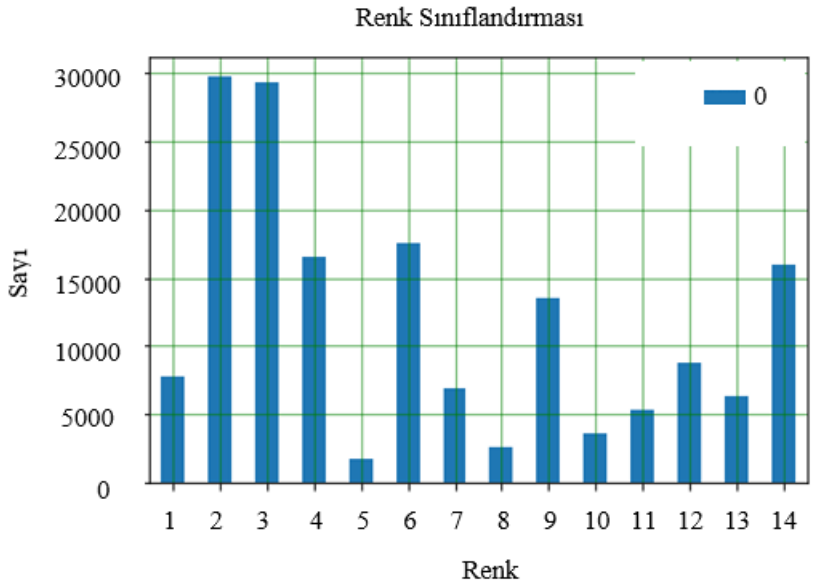
Tablo 2'de veri setinde bulunan nitelikler ile yapılan en düşük değer, en yüksek değer, varyans, ortalama ve standart sapma değerleri yer almaktadır. Sayfa 2(giyim modeli) kısmında veri setinin alındığı e-ticaret sitesindeki 217 ürünün kodları bulunmaktadır.

Tablo 3. Nitelikler Hakkında Bilgiler

Nitelik Adı	Çarpıklık	Basıklık	Medyan
yıl	0.000000	0.000000	2008.0
ay	0.261954	-1.191026	5.0
gun	0.175620	-1.183131	14.0
dizi	4.468848	32.481728	6.0
ülke	-1.382711	2.841221	29.0
kısa ID	0.005240	-1.225864	11967.5
sayfa 1 (ana kategori)	0.111052	-1.411372	2.0
renk	0.610796	-1.039410	4.0
konum	0.145440	-1.303740	3.0
model fotoğrafı	1.093895	-0.803404	1.0
fiyat	0.522185	-0.150328	43.0
fiyat 2	0.047344	-1.997783	1.0
sayfa	1.375776	1.228671	1.0

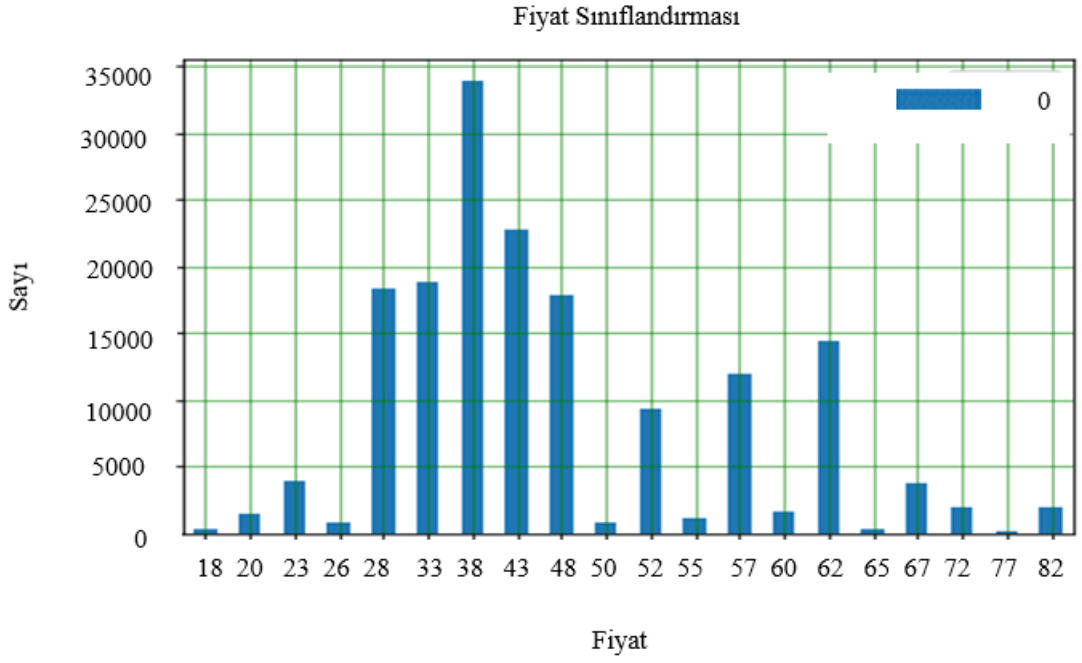
Tablo 3'te veri setinde yer alan nitelikler hakkında çarpıklık, basıklık ve medyan değerleri verilmiştir.

En çok tıklama 133963 adet olmak üzere Polonya'dan yapılmıştır. Polonya'yı Çek Cumhuriyeti takip etmektedir. Polonya ülkesinin kodu 29 olduğundan ve tıklamanın çok büyük kısmı Polonya'da gerçekleştiğinden ülke sınıfının ortalaması 29 olarak çıkmıştır. Ortalama (medyan) olarak ayın 14. Günü alışveriş yapılmasının yer aldığını görebilmekteyiz, yani ayın günleri verilerini ortadan bölen değer 14'tür.



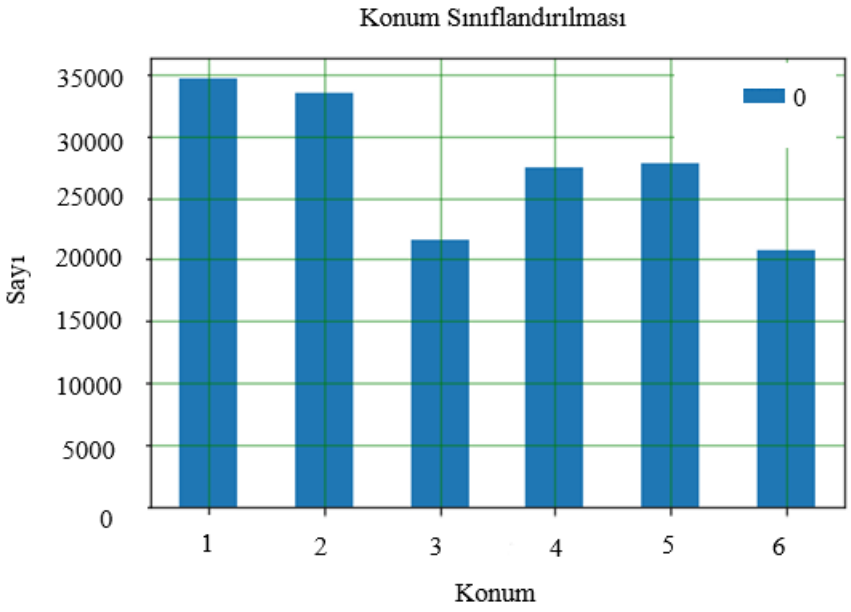
Grafik 1. Ürün Renklerine Göre Tıklama Grafiği

Grafik 1' de görüldüğü gibi en çok tıklanan renkler başta siyah olmak üzere mavi ve beyazdan oluşmaktadır. En az tıklanan renk ise bordo rengidir. E-ticaret sitesinde yer alan ürün renklerinin tıklanma grafiğine baktığımızda, tercih edilen renkler arasında dengesiz bir dağılım olduğu görülmektedir.



Grafik 2. Ürün Fiyatına Göre Tıklama Grafiği

Grafik 2'de, kullanıcıların en çok 38 birimden tıklama yapıldığı görülmektedir. 28 ve 48 birim arası ürünlerin en çok tıklandığı tespit edilirken daha düşük fiyatlı ürünlere tıklamanın daha az olduğu görülmektedir.



Grafik 3. Ürünlerin Sitedeki Konumuna Göre Tıklama Grafiği

Veriler incelendiğinde Grafik 3'te görüldüğü gibi en çok tıklama yapılmayı tercih edilen e-ticaret sitesinin ekran konumları sırası ile sol üst, üst orta, alt orta, alt sol şeklinde ilerlemektedir. Öncelikle sol tarafta bulunan ürünler ve sonrasında orta kısımda yer alan ürünler, en son olarak ise sağ bölümde yer alan ürünler tercih edilmiştir. Sol kısımda bulunan ürünlerin tıklama oranları sağ kısma göre fazladır. Bu durum ise kullanıcıların gördükleri ilk sırada yer alan ürünlere karşı tıklama eğiliminin daha fazla olduğu şeklinde yorumlanabilir.

E-ticaret sitesinde ürün tıklama oranları ise; 49742 tıklama ile birinci sırada pantolon, 38747 tıklama ile ikinci sırada etek, 38577 tıklama ile üçüncü sırada bluz, 38408 tıklama ile dördüncü ve son sırada indirimli ürünler bulunmaktadır.

ÜÇÜNCÜ BÖLÜM
BULGULAR

Çalışmanın bu bölümünde sınıf dengeleme yöntemleri olan Sentetik Azınlık Yüksek Örnekleme Tekniği (Smote), Rastgele Örneklem Azaltma (Rus), Rastgele Aşırı Örnekleme(Ros), Naive Bayes, Destek Vektör Makineleri, kNN En Yakın Komşu, K-MEANS, Lojistik Regresyon analiz sonuçlarına yer verilmiştir.

1. SINIF DENGELEME YÖNTEMLERİ

Çalışmada yer alan veri setindeki nitelikler kısmındaki grafiklerden de veri seti denge oranları yansımaktadır. Sınıf dengeleme yöntemleri olarak; Sentetik azınlık yüksek örnekleme tekniği (Smote), Rastgele aşırı örnekleme (Ros), Rastgele örneklem azaltma (Rus) kullanılmıştır.

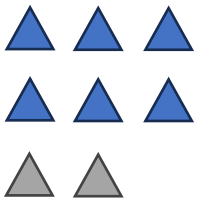
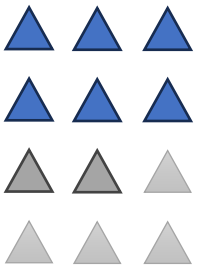
1.1. Sentetik Azınlık Yüksek Örnekleme Tekniği (Smote)

SMOTE sıklıkla kullanılmakla birlikte genellikle yüksek performans gösteren bir sınıf dengeleme yöntemidir. İlk olarak 2002 yılında ortaya çıkmıştır. Bu algoritmada azınlıktaki örneklem arttırılmaktadır. Azınlık örneklemleri, k en yakın komşu olmak üzere dengeli bir şekilde sentetik veriler ile arttırılır (Yavaş, Güran ve Uysal, 2020: 261).

Dengesiz veri setlerinde kullanılan bir yöntemdir. Dengesiz veriler doğru sonuç vermekte yeterli olmadığından sınıf dengeleme yöntemleri ortaya çıkmıştır (Akın ve Terzi, 2020: 22).

Smote veri dengeleme yöntemi kullanıldıktan sonra veri seti [(0, 63484), (1, 63484)] şeklinde dengelenmiştir.

Tablo 4. Smote

Orijinal Veri	Smote
	

Tablo 4'te mavi renkte olan üçgenler çoğunluk verisini temsil ediyorken, gri renkte bulunan üçgenler ise azınlık verilerini, açık gri üçgenler ise yapay verileri temsil etmektedir. Smote veri dengeleme yöntemi kullanıldıktan sonra azınlık verileri yapay veriler ile çoğunluk verilerine eşitlenmiştir.

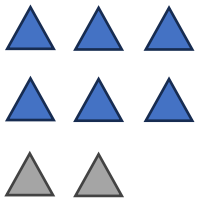
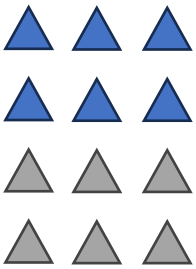
1.2. Rastgele Aşırı Örnekleme (Ros)

Rastgele Aşırı Örnekleme ya da İngilizcesi ile Random Over Sampling yönteminde; veri setini dengeleyebilmek için rastgele olarak, dengelenene kadar azınlık sınıfını çoğunluk sınıfına yaklaştırmak için veri üretir (Dal, Gümüş, Güldal ve Yavaş, 2021: 346).

Ros kullanıldığında mevcut veri setindeki azınlıkta bulunan veriler çoğunluk verilerine eşitlenirken, azınlık verilerine eklenen veriler, azınlık verilerinin kendisine ait verileri kopyalar ve çoğaltır. Kendi verileriyle arttırılan azınlık sınıfının uyum oranı da artabilir (Öztürk, 2022: 7).

Ros kullanıldıktan sonra veri seti [(0, 63484), (1, 63484)] şeklinde dengelenmiştir. Ros kullanılırken nicel verilerden yararlanılmıştır.

Tablo 5. Ros

Orijinal Veri	Ros
	

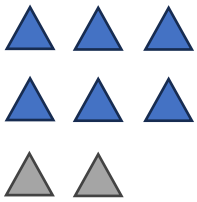
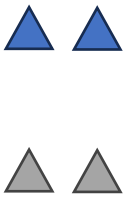
Tablo 5'te mavi renkte olan üçgenler çoğunluk verisini temsil ediyorken, gri renkte bulunan üçgenler ise azınlık verilerini temsil etmektedir. Ros veri dengeleme yöntemi kullanıldıktan sonra azınlık verisinin kendine ait verileri kopyalayarak çoğaldığını ve çoğunluk verileri ile eşitlendiği görülmektedir.

1.3. Rastgele Örneklem Azaltma (Rus)

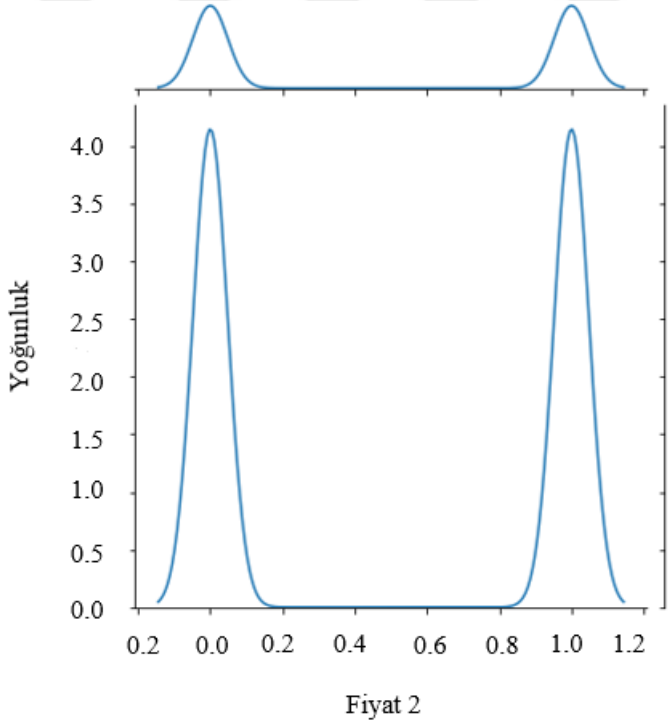
Sınıf dengeleme yöntemlerinden biri olan Rastgele Örneklem Azaltma ya da Random Under Sampling, dengeleme işlemi yapabilmek için çoğunluk sınıfında yer alan verileri rastgele olarak azaltma işlemi gerçekleştirir. Azalan veri sayısı ile birlikte sınıflandırma vakti de azalmaktadır. Çoğunluk sınıfını rastgele olarak elemesi önemli verilerinde ortadan kaybolacağı anlamına geldiğinden bu anlamda dezavantaj sağlamaktadır (Öztürk,2022: 6).

Rastgele örneklem azaltmada [(0, 60621), (1, 60621)] oranlarında çoğunluk sınıfı azınlık sınıfı ile dengelenmiştir.

Tablo 6. Rus

Orijinal Veri	Rus
	

Tablo 6'da mavi renkte olan üçgenler çoğunluk verisini temsil ediyorken, gri renkte bulunan üçgenler ise azınlık verilerini temsil etmektedir. Orijinal veri setine kıyasla Rus veri dengeleme yöntemi kullanıldıktan sonra çoğunluk verisinin azalarak azınlık verisine eşitlendiği görülmektedir.



Grafik 4. RUS 2 Kde Grafiği

Rus 2 kde grafiği Grafik 4'teki gibidir.

2. ANALİZ SONUÇLARI

Tezin bu kısmında kNN En Yakın Komşu Algoritması, Naive Bayes Sınıflandırıcı, Destek Vektör Makineleri, K- Means algoritmaları, Lojistik Regresyon ile Avrupa e-ticaret sitesi verileri kullanılarak yapılmış işlemlerin analiz sonuçlarına yer verilmektedir. Tezde Python 3.9.7 yazılımı kullanılmıştır.

Tablo 7. İki Sınıflı Örnek Karmaşıklık Matrisi

Karmaşıklık Matrisi		**Verinin Tahmin edilen Sınıfı**	
		E-Ticaret Verileri	Gözlem
Gerçek Veri Sınıfı	E-Ticaret Verileri	Doğru Negatif (TN)	Yanlış Pozitif (FP)
	Gözlem	Yanlış Negatif (FN)	Doğru Pozitif (TP)

İki sınıflı örnek karmaşıklık matrisinde yer alan;

TP: Doğru Pozitif (Doğruya doğru demek)

FN: Yanlış Negatif (Yanlışa doğru demek)

FP: Yanlış Pozitif (Doğruya yanlış demek)

TN: Doğru Negatif (Yanlışa yanlış demek) anlamlarına gelmektedir.

2.1. Knn Yakın Komşuluk Algoritması

kNN algoritmasında veri setinin doğal hali ile daha sonra SMOTE, ROS ve RUS sınıf dengeleme yöntemleri kullanılmıştır.

Tablo 8. Knn Normal Hata Matrisi (n=5)

Hata Matrisi		Verinin Tahmin edilen Sınıfı	
		Negatif	Pozitif
Gerçek Veri Sınıfı	Negatif	TN-17297	FP-2861
	Pozitif	FN-2762	TP-18449

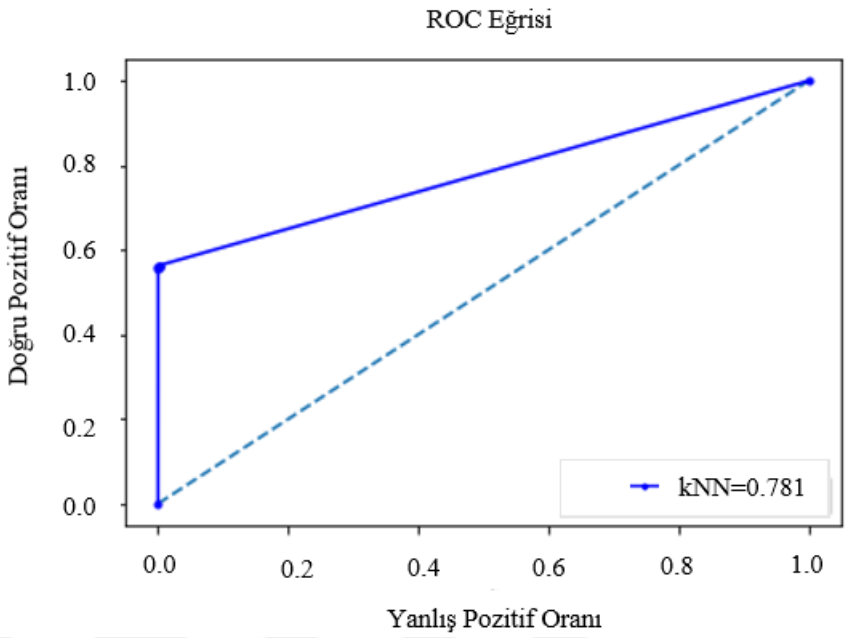
Tablo 8'de yer alan matriste henüz dengeleme yöntemleri kullanılmamış kNN algoritmasının n=5 normal hali ile hata matrisini görmekteyiz.

Tablo 8'de TN-17297, TP-18449, FN-2762 ve FP-2861 verileri bulunmuştur. Algoritmada 35.746 veri doğru tahmin edilmiş ve 5623 veri yanlış tahmin edilmiştir.

Tablo 9. Knn Normal Hata Matrisi (n=3)

Hata Matrisi		Verinin Tahmin edilen Sınıfı	
		Negatif	Pozitif
Gerçek Veri Sınıfı	Negatif	TN- 17220	FP- 2938
	Pozitif	FN- 2730	TP- 18481

Tablo 9'da henüz dengeleme yöntemleri kullanılmamış kNN algoritmasının n=3 normal hali ile hata matrisini görmekteyiz. Tabloda TN-17220, TP-18481, FN-2730 ve FP-2938 verileri bulunmuştur. Algoritmada 35.701 veri doğru tahmin edilmiş ve 5688 veri yanlış tahmin edilmiştir. Karşılaştırıldığında n=5 alındığında n=3'e göre 45 adet daha fazla doğru bulunmaktadır.



Grafik 5. Knn Normal ROC Eğrisi

Roc eğrisi geniş denilebilir ölçüdedir. Knn, 0.781 oran ile iyi bir sonuç elde etmiştir.

Tablo 10. Knn Performans Değerlendirme Ölçüt Sonuçları

		Doğruluk	Kesinlik	Duyarlılık	F Değeri
Normal	**n = 3**	0.86	0.86	0.86	0.86
	n = 5	0.86	0.86	0.86	0.86
Smote	**n = 3**	0.86	0.86	0.86	0.86
	n = 5	0.86	0.86	0.86	0.86
Ros	**n = 3**	0.87	0.87	0.87	0.87
	n = 5	0.87	0.87	0.87	0.87
Rus	**n = 3**	0.85	0.77	0.76	0.77
	n = 5	0.85	0.77	0.76	0.77

Doğruluk ölçütünde kNN çalışmasında n=3 kullanıldığında sırasıyla Normal veri seti, Smote, Ros, Rus ile dengelenmiş veri setlerinde sonuçlar; 0.86, 0.86, 0.87, 0.85 olarak çıkmıştır.

Kesinlik ölçütünde kNN çalışmasında n=3 kullanıldığında sırasıyla Normal veri seti, Smote, Ros, Rus ile dengelenmiş veri setlerinde sonuçlar; 0.86, 0.86, 0.87, 0.77 olarak çıkmıştır.

Duyarlılık ölçütünde kNN çalışmasında n=3 kullanıldığında sırasıyla Normal veri seti, Smote, Ros, Rus ile dengelenmiş veri setlerinde sonuçlar; 0.86, 0.86, 0.87, 0.76 olarak çıkmıştır.

F Değerinde kNN çalışmasında n=3 kullanıldığında sırasıyla Normal veri seti, Smote, Ros, Rus ile dengelenmiş veri setlerinde sonuçlar; 0.86,0.86, 0.87, 0.77 olarak çıkmıştır.

Doğruluk ölçütünde kNN çalışmasında n=5 kullanıldığında sırasıyla Normal veri seti, Smote, Ros, Rus ile dengelenmiş veri setlerinde sonuçlar; 0.86, 0.86, 0.87, 0.85 olarak çıkmıştır.

Kesinlik ölçütünde kNN çalışmasında n=5 kullanıldığında sırasıyla Normal veri seti, Smote, Ros, Rus ile dengelenmiş veri setlerinde sonuçlar; 0.86, 0.86, 0.87, 0.77 olarak çıkmıştır.

Duyarlılık ölçütünde kNN çalışmasında n=5 kullanıldığında sırasıyla Normal veri seti, Smote, Ros, Rus ile dengelenmiş veri setlerinde sonuçlar; 0.86, 0.86, 0.87, 0.76 olarak çıkmıştır.

F Değerinde kNN çalışmasında n=5 kullanıldığında sırasıyla Normal veri seti, Smote, Ros, Rus ile dengelenmiş veri setlerinde sonuçlar; 0.86,0.86, 0.87, 0.77 olarak çıkmıştır.

N değeri 3 alındığında da 5 alındığında da sonuçlar değişmemektedir. Sınıf dengeleme yöntemlerin kullanımı da sonuçları değiştirmeye etki etmemiştir.

Veri setindeki dağılım: ({0: 60621, 1: 63484})

Smote veri seti dağılımı: ({0: 63484, 1: 63484})

Ros veri seti dağılımı: ({0: 63484, 1: 63484})

Rus veri seti dağılımı: ({0: 60621 1: 60621})

Smote veri dengeleme yöntemi kullanıldığında azınlık verisinin, çoğunluk verisine eşitlendiğini görmekteyiz.

Ros veri dengeleme yöntemi kullanıldığında azınlık verisinin, çoğunluk verisine eşitlendiğini görmekteyiz.

Rus veri dengeleme yöntemi kullanıldığında çoğunluk verisinin, azınlık verisine eşitlendiğini görmekteyiz.

2.2. Destek Vektör Makineleri

Destek vektör makinesi algoritmasını kullanırken Radyal temelli ve doğrusal çekirdek kullanılmıştır.

svm_parametre={"C":np.arange(1,5),"kernel":["Linear","rbf"]}

Destek vektör makinesinde hem normal hali ile işlem yapılmış hem de Smote, Rastgele örneklem azaltma ve Rastgele aşırı örnekleme dengeleme yöntemleri ile başarı ölçümü gerçekleştirilmiştir.

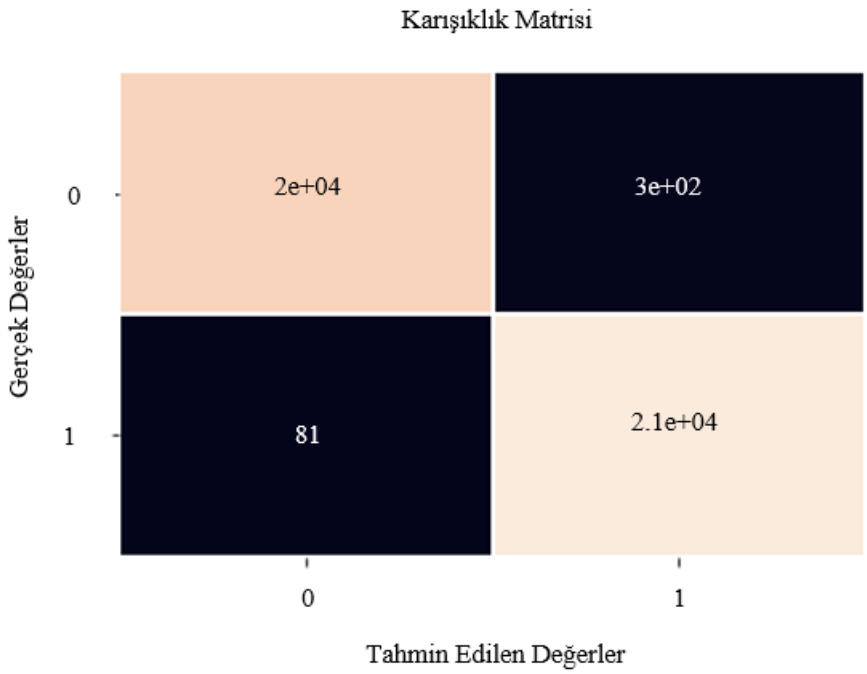
0.25 test boyutu tercih edilmiştir.

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.25, random_state = 0)

Tablo 11. Destek Vektör Makineleri Normal Hata Matrisi

Hata Matrisi		Verinin Tahmin edilen Sınıfı	
		Negatif	Pozitif
Gerçek Veri Sınıfı	Negatif	TN-19857	FP-301
	Pozitif	FN-81	TP-21130

Tablo 11'de yer alan matriste henüz dengeleme yöntemleri kullanılmamış DVM algoritmasının normal hali ile hata matrisini görmekteyiz. Tabloda TN-19857, TP-21130, FN-81 ve FP-301 verileri bulunmuştur. Algoritmada 40987 veri doğru tahmin edilmiş ve 382 veri yanlış tahmin edilmiştir. Çok iyi derecede iyi tahmin yaptığı görülmektedir.



Şekil 7. DVM Normal Karışıklık Matrisi

Tablo 12. Destek Vektör Makineleri Metrikleri

	Normal	Smote	Ros	Rus
Doğruluk	0.99	0.99	0.99	0.99
Kesinlik	0.99	0.99	0.99	0.99
Duyarlılık	1.00	1.00	1.00	1.00
F-Ölçütü	0.99	0.99	0.99	0.99

Tablo 12'de Doğruluk değerleri Normal veri seti ve Smote, Ros, Rus dengeleme yöntemleri için sırası ile 0.99, 0.99, 0.99, 0.99'dur. Hepsinin eşit olduğu görülmektedir. Ayrıca normal veri için kappa değeri 0.98 olarak bulunmuştur.

Kesinlik değeri Normal veri seti ve Smote, Ros, Rus dengeleme yöntemleri için sırası ile 0.99, 0.99, 0.99, 0.99'dur. Hepsinin eşit olduğu görülmektedir.

Duyarlılık değeri Normal veri seti ve Smote, Ros, Rus dengeleme yöntemleri için sırası ile 1.00, 1.00, 1.00, 1.00'dir. Hepsinin eşit olduğu görülmektedir.

F ölçütü Normal veri seti ve Smote, Ros, Rus dengeleme yöntemleri için sırası ile 0.99, 0.99, 0.99, 0.99'dur. Hepsinin eşit olduğu görülmektedir.

Değerlendirme ölçütleri hem normal veri setinde hem de dengelenmiş veri setlerinde çok yüksek başarı oranı ile beraber her biri aynı sonuçları elde etmiştir.

2.3. Naive Bayes

Naive Bayes algoritması Python programlama dilinde ‘GaussianNB’ fonksiyonu kullanılarak çalıştırılmıştır. ‘GaussianNB’ fonksiyonu Gauss Algoritmasından gelmektedir.

nb = GaussianNB()

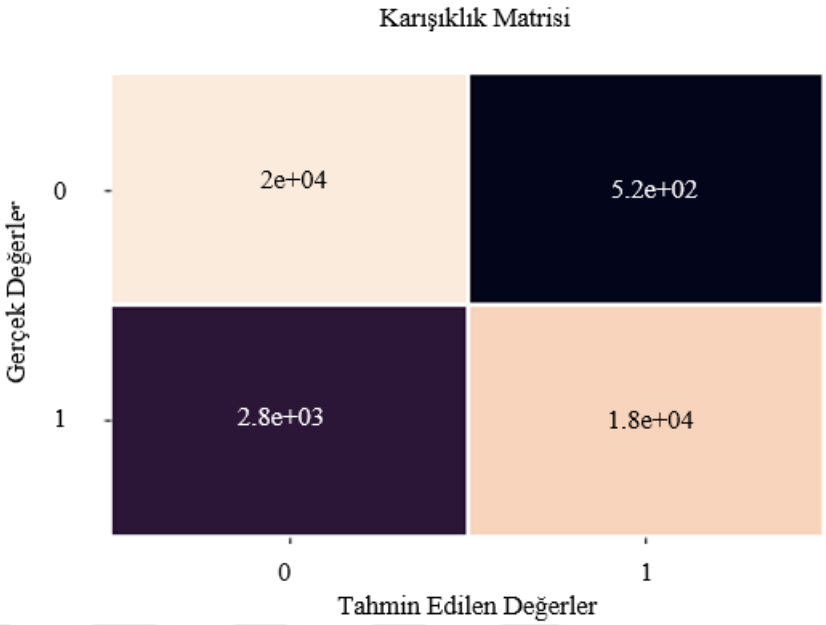
Veri setinde eksik veri yoktur. Naive Bayes Algoritması kullanılmadan önce veri setinde fiyat 2 başlığı kısmında 1 ve 2 rakamlarından oluşan verilerden 2 rakamı 0’a dönüştürülmüştür. Veri setinde başka değişiklik bulunmamaktadır.

Naive Bayes algoritmasında sentetik azınlık yüksek örnekleme tekniği (Smote), rastgele örneklem arttırma (Ros), rastgele örneklem azaltma (Rus) sınıf dengeleme yöntemleri kullanılmıştır.

Tablo 13. Naive Bayes Algoritması Normal Hata Matrisi

Hata Matrisi		Verinin Tahmin edilen Sınıfı	
		Negatif	Pozitif
Gerçek Veri Sınıfı	Negatif	TN-19641	FP-517
	Pozitif	FN-2750	TP-18461

Tablo 13’te henüz dengeleme yöntemleri kullanılmamış Naive Bayes algoritmasının normal hali ile hata matrisini görmekteyiz. Tabloda TN-19641, TP-18461, FN-82750 ve FP-517 verileri bulunmuştur. Doğru sınıflanan sayı 38102 olarak tespit edilmiştir. Yanlış sınıflanan veriler ise 3267 olarak tespit edilmiştir. Doğru ve yanlış sınıflanan veriler arasında 34835 adet fark bulunmakla birlikte doğru olarak sınıflanan verilerin yanlış olarak sınıflanan verilere kıyasla iyi bir sonuç elde ettiği görülmektedir.



Şekil 8. Naive Bayes Karışıklık Matrisi

Bağımsız değişken olarak aldığımız ‘fiyat 2’ verilerinde bir ürünün ortalama fiyattan yüksek olup olmadığı verisini Şekil 8’de sunmaktadır. 1’ler evet 0’lar hayır demektir.

Doğru sınıflanan sayı 38102 olarak tespit edilmiştir. Yanlış sınıflanan veriler ise 3267 olarak tespit edilmiştir.

Tablo 14. Naive Bayes Algoritması Haldout Metrikleri

	Normal	**Smote**	**Ros**	**Rus**
Doğruluk	0.91	0.92	0.92	0.92
Kesinlik	0.96	0.97	0.97	0.97
Duyarlılık	0.87	0.87	0.87	0.87
F-Ölçütü	0.92	0.92	0.91	0.91

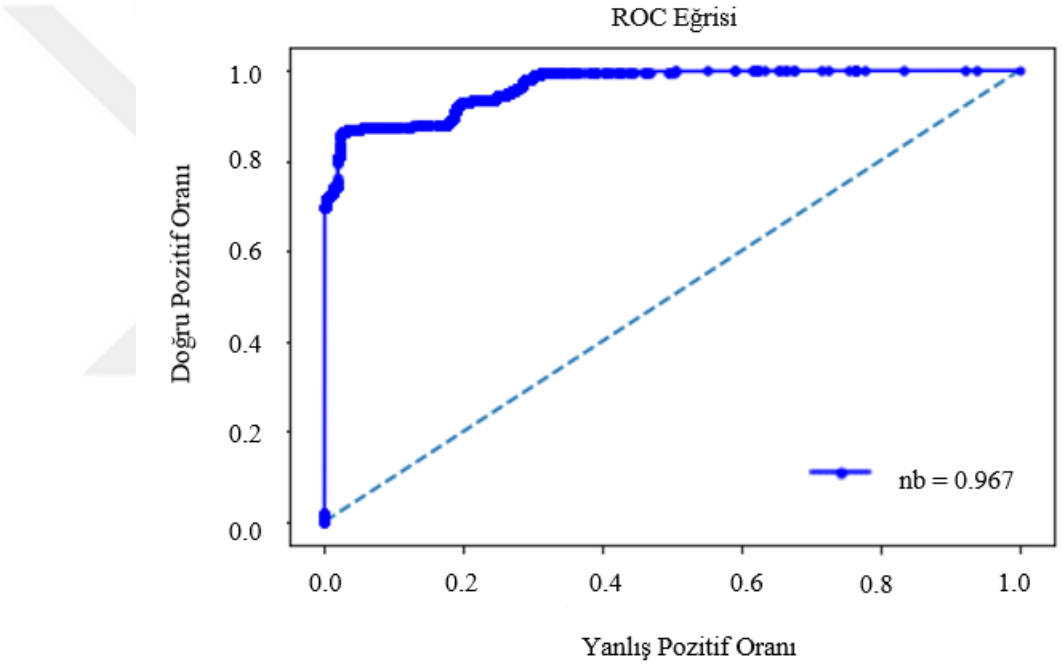
Doğruluk değerleri Normal veri seti ve Smote, Ros, Rus dengeleme yöntemleri için Tablo 14’te sırası ile 0.91, 0.92, 0.92, 0.92’dir. Hepsinin eşit olduğu görülmektedir. Ayrıca normal veri seti için kappa değeri 0.84 olarak bulunmuştur.

Kesinlik değeri Normal veri seti ve Smote, Ros, Rus dengeleme yöntemleri için sırası ile 0.96, 0.97, 0.97, 0.97’dir. Normal veri setinden en düşük kesinlik değerine ulaşılırken SMOTE, ROS ve RUS dengeleme yöntemlerinde eşit derecede en yüksek değer görülmektedir.

Duyarlılık değeri Normal veri seti ve Smote, Ros, Rus dengeleme yöntemleri için sırası ile 0.87, 0.87, 0.87, 0.87. Hepsinin eşit olduğu görülmektedir.

F ölçütü Normal veri seti ve Smote, Ros, Rus dengeleme yöntemleri için sırası ile 0.92, 0.92, 0.91, 0.91'dir. Normal veri ve Smote ile en yüksek değerin eşit olarak bulunmakta olduğu görülürken, Ros ve Rus dengeleme yöntemlerinde daha düşük değere ulaşılmıştır fakat aradaki fark yüksek değildir.

Görüldüğü gibi veri setinin normal hali ile dengelenmiş halleri arasında bu algoritmada pek fazla fark bulunmamaktadır, hatta fark yok denecek kadar azdır.

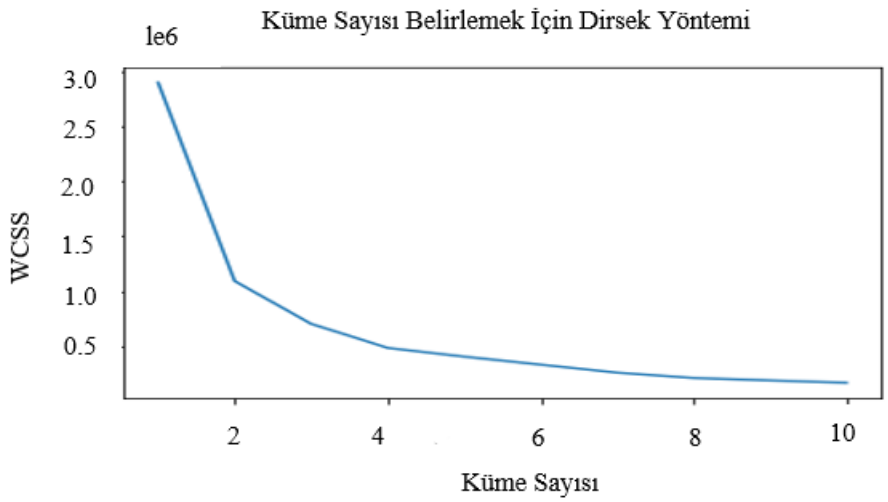


Grafik 6. Naive Bayes Normal ROC Eğrisi

Orijinal veri seti ile yapılmış Grafik 6'da bulunan, naive bayes ile roc eğrisine bakıldığında çok iyi bir sonuç çıktığı görülmektedir. roc eğrisi ne kadar genişse sonuç o kadar iyi demektir. Naive bayes algoritmasında roc eğrisinde 0.967 sonuç çıkması mükemmel derecede olarak söylenebilir.

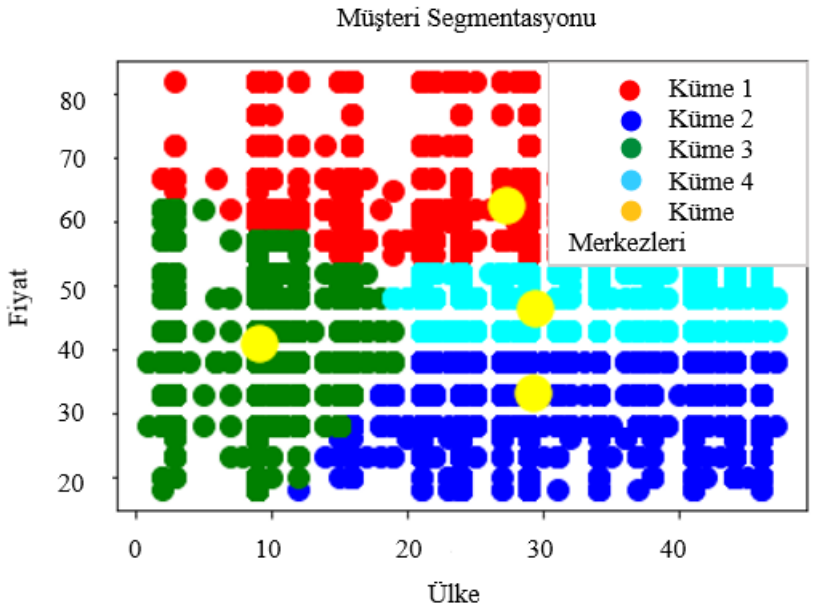
2.4. K-Means

Bu çalışmada ilk olarak ülke ve fiyat verilerini bağımsız değişken olarak alındı. Küme sayısını belirlemek için dirsek yöntemi kullanıldı.



Grafik 7. K-Means Küme Sayısını Belirlemek İçin Dirsek Yöntemi

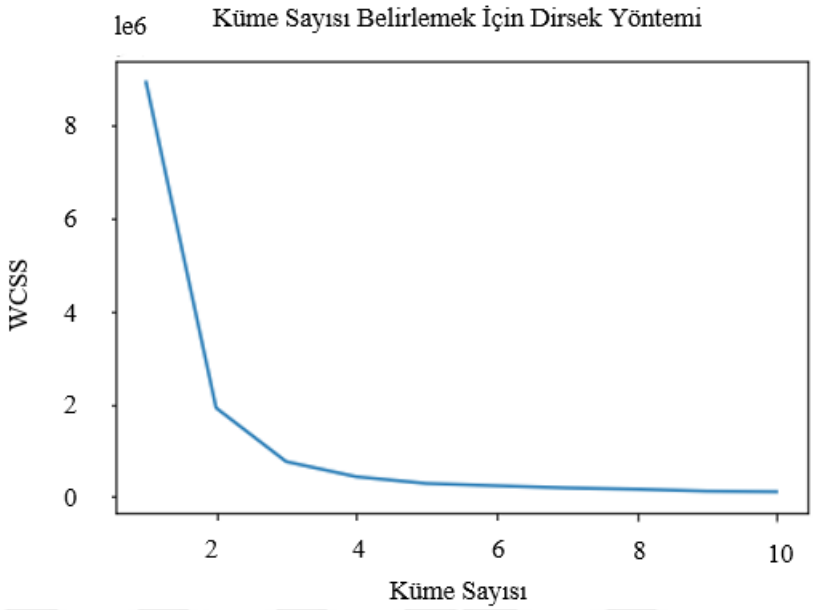
Dengeleme yöntemi kullanıldıktan sonra oluşan Grafik 7'de 4 adet küme oluşturulmasına karar verilmiştir.



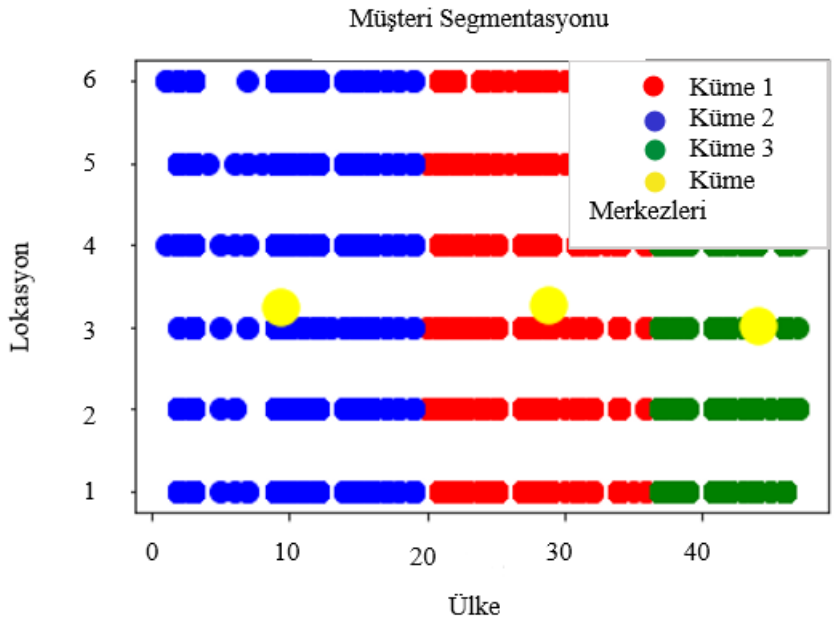
Grafik 8. Ülke ve Fiyatlara Göre Müşteri Kümeleme

Veri setinde 42 ülke bulunmaktadır. Ülkelerdeki hamile kadınların e-ticaret sitesinde tıklamalarını incelendiğinde Küme 1 merkezinin 60 dolaylarında, Küme 2 merkezinin 30 dolaylarında, Küme 3 merkezinin 40 dolaylarında ve son olarak Küme 4 merkezinin 50 birim dolaylarında olduğu görülmektedir.

Bağımsız değişken olarak ülke ve ürünün ekran konum verileri alındığında, dirsek yöntemi ile 3 küme oluşturulmasına karar verilmiştir.



Grafik 9. Dirsek Yöntemi 2



Grafik 10. Ülke ve Ürün Konumuna Göre Kümeleme

Lokasyon açılımları sırası ile 1-Sol üst, 2-Üst ortada, 3- Sağ üst, 4- Sol alt, 5-Alt ortada, 6-Sağ alt olarak numaralandırılmıştır.

Küme merkezleri grafik 10'da incelendiğinde 3 numaralı ekran lokasyonu olan sağ üst köşede bulunan ürünlerin daha fazla tıklandığı görülmektedir.

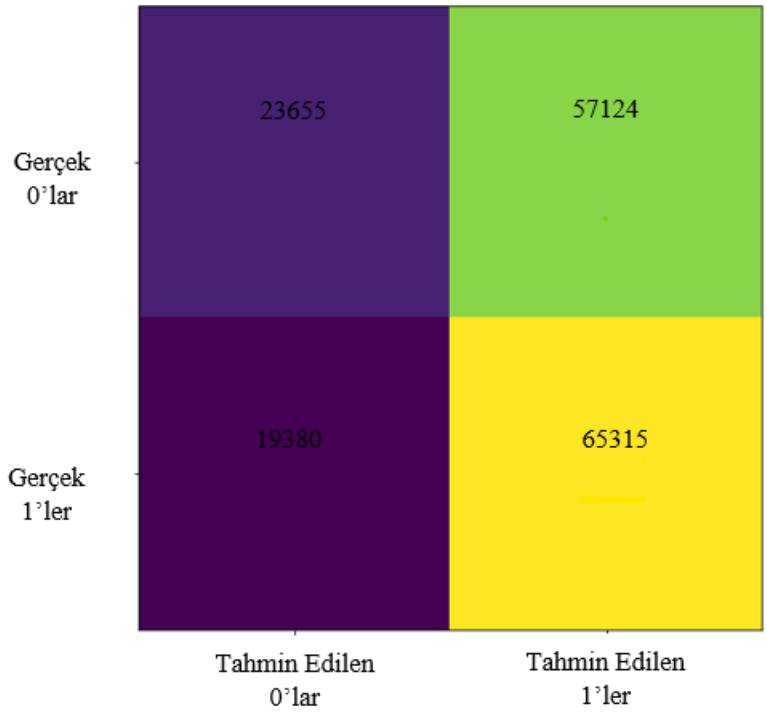
2.5. Lojistik Regresyon

Lojistik Regresyonda kullanılan parametreler fiyatı ortalama fiyattan yüksek mi? ve model profili üzerinde yapılmıştır. Modeli kesim noktası 0.46 olarak bulunmuştur.

Tablo 15. Regresyon Hata Matrisi

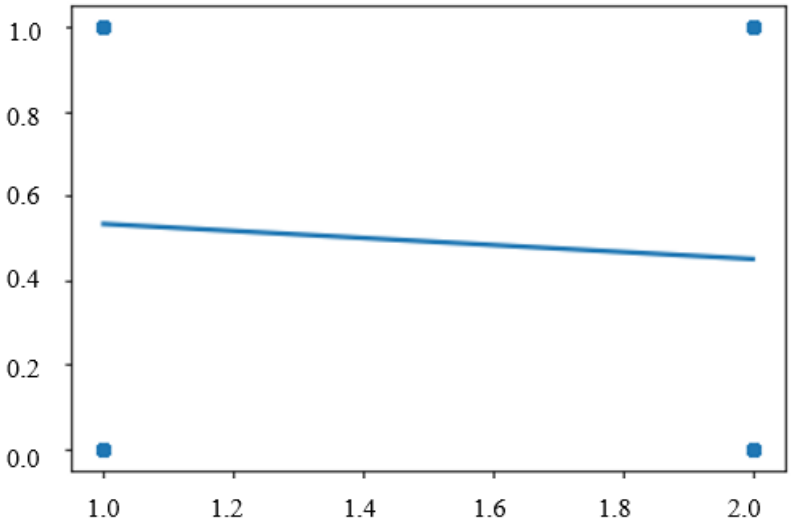
Hata Matrisi		Verinin Tahmin edilen Sınıfı	
		Negatif	Pozitif
Gerçek Veri Sınıfı	Negatif	TN-23655	FP-57124
	Pozitif	FN-19380	TP-65315

Tablo 15'te yer alan verilerde henüz dengeleme yöntemleri kullanılmamış Lojistik Regresyon algoritmasının normal hali ile hata matrisini görmekteyiz. Tabloda TN-23655, TP-65315, FN-19380 ve FP-57124 verileri bulunmuştur. Doğru sınıflanan sayı 88970 olarak tespit edilmiştir. Yanlış sınıflanan veriler ise 76505 olarak tespit edilmiştir. Doğru sınıflanan veriler, yanlış sınıflanan verilere göre 12465 adet daha fazladır. Arada bu kadar fark olmasına rağmen yine de yanlış sınıflanan verilerin 76505 adet olması da epey fazladır ve bu durumda çok iyi bir sonuca ulaştığı söylenemez.



Şekil 9. Regresyon Hata Matrisi Isı Haritası

Şekil 9'da renklerin birbirine benziyor olması sayıların da birbirine yakın olduğunu ifade eder. Görüldüğü üzere mor tonlarında yer alan alanlardaki rakamlarda veriler arasında çok büyük fark bulunmamaktadır. Sayılar arttıkça renk tonlarının açıldıkları görülmektedir.



Grafik 11. Regresyon

Accuracy Score 0.3947 oranı ile düşük çıkmıştır. Model skoru 0.53 çıkmıştır.

Tablo 16. Regresyon Haldout Metrikleri

	Normal	Smote	Ros	Rus
Doğruluk	0.39	0.54	0.54	0.54
Kesinlik	0.57	0.54	0.54	0.54
Duyarlılık	0.39	0.53	0.54	0.53
F-Ölçütü	0.47	0.51	0.54	0.51

Doğruluk değerleri Normal veri seti ve Smote, Ros, Rus dengeleme yöntemleri için Tablo 16'da sırası ile 0.39, 0.54, 0.54, 0.54'tür. Dengeleme yöntemlerinin hepsinin eşit olduğu görülürken Normal hali ile doğruluk değerinin daha düşük olduğu görülmektedir. Ayrıca normal veriler ile yapılan Regresyonda kappa değeri 0.0025 olarak bulunmuştur.

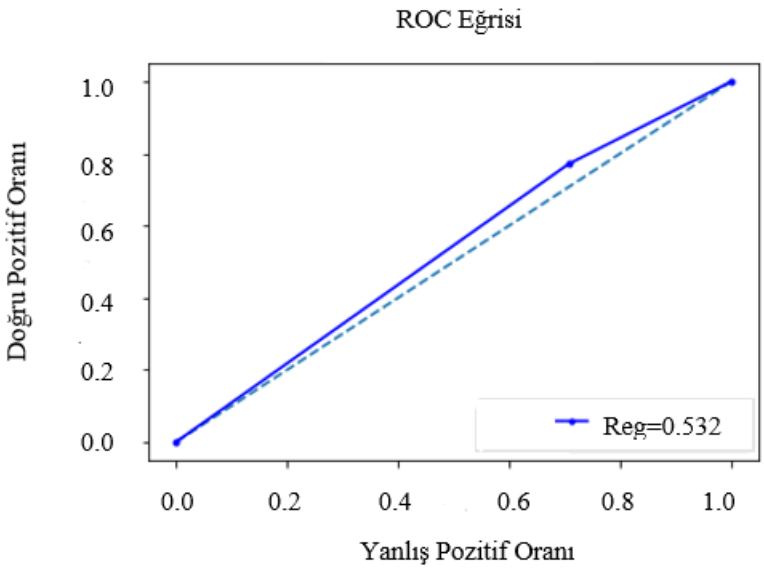
Kesinlik değeri Normal veri seti ve Smote, Ros, Rus dengeleme yöntemleri için sırası ile 0.57, 0.54, 0.54, 0.54'tür. Normal veri setinden en yüksek kesinlik değerine ulaşılırken Smote, Ros ve Rus dengeleme yöntemlerinde eşit derecede en düşük değer görülmektedir.

Duyarlılık değeri Normal veri seti ve Smote, Ros, Rus dengeleme yöntemleri için sırası ile 0.39, 0.53, 0.54, 0.53.Dengeleme yöntemlerinin hepsinin eşit olduğu görülmektedir. Normal veri hali ise en düşüktür ve fark yüksektir.

F ölçütü Normal veri seti ve Smote, Ros, Rus dengeleme yöntemleri için sırası ile 0.47, 0.51, 0.54, 0.51'dir. Normal veri seti en düşük değeri alırken, en yüksek değeri ROS veri dengeleme yöntemi almıştır.

Haldout metriklerine baktığımızda sınıf dengeleme yöntemleri kullanıldığında daha yüksek ve dengeli sonuçlar elde edildiği görülmektedir.

En iyi sonuçlar veri dengeleme yöntemi olan Ros kullanıldığında ortaya çıktığı görülmektedir, fakat; kullanılan diğer dengeleme yöntemleri olan Smote ve Rus ile arasında büyük bir fark bulunmamaktadır. Normal verilerle yapılan Regresyona göre Ros dengeleme yönteminde doğruluk ve duyarlılık verileri daha iyi olduğu görülmektedir.



Grafik 12. Normal Regresyon Roc Eğrisi

Grafik 12'de yer alan Normal verilerle yapılmış Lojistik Regresyon Roc Eğrisinde, eğrinin çok hafif eğimli olduğu görülmektedir. Bu durum başarı oranının düşük olduğunu ifade etmektedir. Eğri 0.0 noktasına çok yakın bulunmaktadır ve geniş değildir. Lojistik Regresyon Roc eğrisinin başarı durumu düşük olduğu söylenebilir.

SONUÇ

Her geçen gün internet üzerinden alışveriş artmaktadır ve bu durum da büyük veriyi beraberinde getirmektedir. Kullanıcılar için kısa sürede ulaşabilecekleri çok fazla seçenek bulunmaktadır ve başka kullanıcıların ürün ve hizmetler hakkında fikirlerine ulaşmaları ve bu durumdan etkilenmeleri ise kaçınılmazdır. İşletmelerin rakiplerinden geri kalmamak için stratejik olarak hareket etmeleri gerekmektedir. İşletmeler e-ticaret sitelerinde ürün ve hizmetlerin türü, ürün ve hizmetlerin ekranda bulundukları konum gibi faktörlerin kullanıcı üzerinde etkilerini bildikleri takdirde buna yönelik planlar gerçekleştirip daha başarılı sonuçlar elde edebilirler.

Bu çalışmada makine öğrenmesi algoritmaları, Avrupa'da hamilelere yönelik satış yapan e-ticaret sitesi verileri üzerinde karşılaştırılmıştır. Kullanılan algoritmalar; Naive Bayes, kNN en yakın komşuluk, Destek Vektör Makineleri (DVM), K- Means ve Lojistik Regresyon kullanılmıştır. Çalışma Python programlama dili üzerinden yürütülmüştür. Kullanılan veri setinde eksik veri olmaması avantaj yaratmıştır. Kullanılan veri setinde Türkiye'de bir çalışma olmaması ve sınıf dengeleme yöntemleri kullanılmaması çalışmada alt yapı yetersizliğine sebep olup çalışmada kısıt olarak karşımıza çıkmaktadır.

Analiz sonuçlarını değerlendirecek olursak kNN En Yakın Komşu Algoritması için n=3 ve n=5 için değerlendirilmiş olup 35746 doğru tahmin ile n=5 daha başarılıdır. Veri dengeleme yöntemleri kullanıldıktan sonra ROS 0.87 değerleriyle en başarılı sonuçları bulan dengeleme yöntemi olmuştur. ROC eğrisinin altında kalan alan 0.781 ile geniş bir alandır ve bu alan ne kadar genişse o kadar iyi sonuç elde edildiğinden çıkan sonuç pozitif anlamda yüksektir.

Destek vektör makineleri için Normal veriler ve Smote, Ros, Rus halleri arasında eşitlik görülmektedir. Alınan değerler 0.99 ile 1.00 arasında olup sonuçlar çok yüksek çıkmıştır.

Naive Bayes algoritmasında en iyi sonuçlar veriler Smote ile dengelendikten sonra ortaya çıkmıştır. Fark burada kesinlik ve f ölçütü ile ortaya çıkmıştır. ROC eğrisi altında kalan alan 0.967 ile mükemmel olarak adlandırılan derecede pozitif yüksek olarak çıkmıştır.

Lojistik regresyonda en yüksek oranlar ROS ile veri dengelendikten sonra ortaya çıkmıştır ve ROC eğrisi altında kalan alan ise 0.532 ile çok düşük çıkmıştır.

Bu algoritmalar incelendiğinde hacimli bir veri ile en çok Destek Vektör Makinelerinde başarı sağlanmıştır. Sırasıyla Naive Bayes, kNN ve Lojistik Regresyon devam etmektedir. En önemli derecede iyi sonuçlar verenler Destek Vektör Makineleri ve Naive Bayes olmuştur. Dezavantaj olarak Destek Vektör Makinelerinin kullanılan büyük veride çok yavaş sonuç elde ettiği görülmüştür. Gelecekte yapılacak olan çalışmalar farklı veri setlerinde, farklı alanlarda işlem gerçekleştirebilir.

KAYNAKÇA

ABDELSALAM ALGAHANI, M. (2021) "Bayes Teoremi ve Naive Bayes Sınıflandırıcısı Kullanılarak İyonkürenin İstatistiksel Analizi", Yüksek Lisans Tezi, **Kastamonu Üniversiteri F.B.E,** Kastamonu.

AĞCA, Y. (2021) "Otel Oda Fiyatlarını Açıklamada Makine Öğrenmesi Algoritmalarının Kıyaslanması", **İşletme Araştırmaları Dergisi**, Cilt 13, Sayı 1, ss. 450-463.

AKÇAY, E.; A. DEMİREL (2011), "Pyometralı Köpeklerde Bazı Kan Parametrelerinin Optimal Pozitiflik Eşiğinin Özgün Oranlar ve ROC (Receiver Operating Characteristic) Eğrisi Yöntemi ile Belirlenmesi", **Erciyes Üniversitesi Veteriner Fakültesi Dergisi**, Cilt 8, Sayı 3, ss. 153 - 163.

AKIN, P. ; Y. TERZİ (2020), "Dengesiz Veri Setli Sağkalım Verilerinde Cox Regresyon ve Rastgele Orman Yöntemlerin Karşılaştırılması", **Veri Bilimi Dergisi**, Cilt 3, Sayı 1, ss. 21-25.

ARI, O. (2022) "Tüketici Yorumlarının Fayda Düzeyinin Tahminlenmesine Yönelik Bir Araştırma: Makine Öğrenmesi Algoritmalarının Karşılaştırılması", Yüksek Lisans Tezi, **Sakarya Üniversitesi İşletme Enstitüsü,** Sakarya.

ATALAY, M.; E. ÇELİK (2017), "Büyük Veri Analizinde Yapay Zekâ ve Makine Öğrenmesi Uygulamaları", **Mehmet Akif Ersoy Üniversitesi Sosyal Bilimler Enstitüsü Dergisi**, Cilt 9, Sayı 22, ss. 155-172.

AYDEMİR, E., M. IŞIK; T. TUNCER (2021), "Türkçe Haber Metinlerinin Çok Terimli Naive Bayes Algoritması Kullanılarak Sınıflandırılması", **Fırat Üniversitesi Mühendislik Bilimleri Dergisi**, Cilt 33, Sayı 2, ss. 519-526.

AYDINOĞLU, A. Ç., R. BOVKIR; İ. ÇÖLKESEN (2023),"Toplu Taşınmaz Değerlemede Makine Öğrenme Algoritmalarının Kullanımı ve Konumsal/Konumsal Olmayan Özniteliklerin Tahmin Doğruluğuna Etkilerinin Karşılaştırılması", **Jeodezi ve Jeoinformasyon Dergisi**, Cilt 10,Sayı 1, ss. 63-83.

AYHAN, S.; Ş. ERDOĞMUŞ (2014), "Destek Vektör Makineleriyle Sınıflandırma Problemlerinin Çözümü İçin Çekirdek Fonksiyonu Seçimi", **Eskişehir Osmangazi Üniversitesi İibf Dergisi,** Cilt 9, Sayı 1, ss. 175-201.

BENLİGÜL, E. M., M. BEKTAŞ; G. ARSLAN (2022), "Meta-Analizi Anlamak ve Yorumlamak: Hemşireler İçin Öneriler", **Dokuz Eylül Üniversitesi Hemşirelik Fakültesi Elektronik Dergisi**,Cilt 15,Sayı 1, ss. 86-98.

BİRCAN, H. (2004), "Lojistik Regresyon Analizi: Tıp Verileri Üzerine Bir Uygulama", **Kocaeli Üniversitesi Sosyal Bilimler Enstitüsü Dergisi**, Sayı 8, ss. 185-208.

COLLEY, W. (2019), "Comparison Of Machine Learning Algorithms For Financial Evaluations", Yüksek Lisans Tezi. **Gebze Teknik Üniversitesi, F.B.E**.

COŞKUN, C.; A. BAYKAL (2011), "Veri Madenciliğinde Sınıflandırma Algoritmalarının Bir Örnek Üzerinde Karşılaştırılması", **Akademik Bilişim,** ss. 1-8.

COŞLU, E. (2013), "Veri Madenciliği", **Akademik Bilişim**, s. 23-25.

ÇALIŞ, K., O. GAZDAĞI ; O. YILDIZ (2013), "Reklam İçerikli Epostaların Metin Madenciliği Yöntemleri ile Otomatik Tespiti", **Bilişim Teknolojileri Dergisi**, Cilt 6, Sayı 1, ss. 1-7.

ÇEBİ, C. B., F. S. BULUT, H. FIRAT, G. KARATAŞ; Ö. K. ŞAHİNGÖZ (2019), "Saldırı Tespit Sistemlerinde Makine Öğrenmesi Modellerinin Karşılaştırılması", **Erzincan Üniversitesi Fen Bilimleri Enstitüsü Dergisi**, Cilt 12, Sayı 3, ss. 1513-1525.

ÇELİK, M. (2009), "Veri Madenciliğinde Kullanılan Sınıflandırma Yöntemleri ve Bir Uygulama", Yayınlanmamış Yüksek Lisans Tezi. **İstanbul Üniversitesi, S.B.E.**

ÇELİK, S.; E. AKDAMAR (2018), "Büyük Veri ve Veri Görselleştirme",**Akademik Bakış Uluslararası Hakemli Sosyal Bilimler Dergisi**, Sayı 65, ss. 253 - 264.

ÇOKLUK, Ö. (2010), "Lojistik Regresyon Analizi: Kavram ve Uygulama", **Kuram ve Uygulamada Eğitim Bilimleri**, Cilt 10, Sayı 3, ss. 1357-1407.

DAL, A., İ. GÜMÜŞ , S. GÜLDAL; M. YAVAŞ (2021), "A New Resampling Approach Based On Weighted Geometric Mean For Unbalanced Data", **Adıyaman Üniversitesi Mühendislik Bilimleri Dergisi**, Cilt 8, Sayı 15, ss.343-352.

DEMİRCİ, V. G. (2022), "Dijital Reklamcılıkta Makine Öğrenmesi ve Veri Gizliliği", **Kent Akademisi Dergisi**, Cilt 15, Sayı 3, ss. 1455-1474.

DİNÇER, E. (2006), "Veri Madenciliğinde K-Means Algoritması Ve Tıp Alanında Uygulanması",**Kocaeli Üniversitesi F.B.E**,Kocaeli.

EGE, İ.; A. BAYRAKDAROĞLU (2009), "İmkb Şirketlerinin Hisse Senedi Getiri Başarılarının Lojistik Regresyon Tekniği ile Analizi", **Uluslararası Yönetim İktisat ve İşletme Dergisi**, Cilt 5, Sayı 10, ss. 139-158.

ERAY, O. (2008), "Destek Vektör Makineleri ile Ses Tanıma Uygulaması", Yüksek Lisans Tezi. **Pamukkale Üniversitesi F.B.E ,**Denizli.

ERDOĞAN, A. M. (2023), "Comparison Of Machine Learning Algorithms For İmproved Admission Prediction Of The Emergency Department Patients" Yüksek Lisans Tezi, **İzmir Bakırçay Üniversitesi, L.E.E,** İzmir.

ERTAŞ, Ö. S.; A. KAYALI (2005), "Analitik Yöntem Geçerliliğine Genel Bir Bakış", **Ankara Eczacılık Fakültesi Dergisi**, Cilt 3, Sayı 1, ss. 41 - 57.

GİRGİNER, N.; B. CANKUŞ (2008), "Tramvay Yolcu Memnuniyetinin Lojistik Regresyon Analiziyle Ölçülmesi: Estram Örneği", **Yönetim ve Ekonomi Dergisi**, Cilt 15,Sayı 1, ss. 181-193.

GÖRMEZ, B. (2021), "Adli Bilişimde Makine Öğrenmesi: Makine Öğrenmesi Algoritmaları İle Terör Olaylarının Tahmin Edilmesi Çalışması", Yüksek Lisans Tezi. **Ankara Üniversitesi, S.B.E**, Ankara.

GÜLDAL, S. (2021), "Unsupervised Machine Learnin Algorithms To Find 3-Kcp Solution: Modularity, Clique Percolation, Spectral, Centrality, And Hierarchical Clustering", **Adıyaman Üniversitesi Mühendislik Bilimleri Dergisi**,Cilt 8, Sayı 14, ss. 169-178.

GÜNÇE, E. (2021), "Twitter Platformu Üzerinden Makine Öğrenmesi Algoritmaları ile Cinsiyet ve İlgi Alanı Analizi", Yüksek Lisans Tezi, **Trakya Üniversitesi, F.B.E**, Edirne.

GÜNER, N.; E. ÇOMAK (2011), "Mühendislik Öğrencilerinin Matematik I Derslerindeki Başarısının Destek Vektör Makineleri Kullanılarak Tahmin Edilmesi", **Pamukkale Üniversitesi Mühendislik Bilimleri Dergisi**, Cilt 17, Sayı 2, ss. 87-96.

GÜRKAHRAMAN, K.; R. KARAKIŞ (2021), "Brain Tumors Classification With Deep Learning Using Data Augmentation", **Journal of the Faculty of Engineering and Architecture of Gazi University**, Cilt 36, Sayı 2, ss. 997-1011.

KABA, G.; S. B. KALKAN (2022), "Kardiyovasküler Hastalık Tahmininde Makine Öğrenmesi Sınıflandırma Algoritmalarının Karşılaştırılması", **İstanbul Ticaret Üniversitesi Fen Bilimleri Dergisi**,Cilt 21, Sayı 42, ss. 183 - 193.

KALAYCI, T. E. (2018), "Kimlik Hırsızı Web Sitelerinin Sınıflandırılması için Makine Öğrenmesi Yöntemlerinin Karşılaştırılması", **Pamukkale Üniversitesi Mühendislik Bilimleri Dergisi**, Cilt 24, Sayı 5, ss. 870-878.

KARA, Ş. E.; R. ŞAMLI (2021), "Yazılım Projelerinin Maliyet Tahmini için WEKA'da Makine Öğrenmesi Algoritmalarının Karşılaştırmalı Analizi", **Avrupa Bilim ve Teknoloji Dergisi** , Sayı 23, ss.415-426.

KARAKOYUN, M.; M. HACIBEYOĞLU (2014), "Biyomedikal Verİ Kümeleri ile Makine Öğrenmesi Sınıflandırma Algoritmalarının İstatistiksel Olarak Karşılaştırılması", **Deü Mühendislik Fakültesi Mühendislik Bilimleri Dergisi**, Cilt 16, Sayı 48, ss. 20-42.

KAVZOĞLU, T.; İ. ÇÖLKESEN (2010), "Destek Vektör Makineleri ile Uydu Görüntülerinin Sınıflandırılmasında Kernel Fonksiyonlarının Etkilerinin İncelenmesi", **Harita Dergisi**, Cilt 7, Sayı 144, ss. 73-82.

KAYNAR, O., H. ARSLAN , Y. GÖRMEZ; Y. E. IŞIK (2018), "Makine Öğrenmesi ve Öznitelik Seçim Yöntemleriyle Saldırı Tespiti", **Bilişim Teknolojileri Dergisi**, Cilt 11, Sayı 2, ss.75-185.

KAZAN, S.; H. KARAKOCA (2019), "Makine Öğrenmesi ile Ürün Kategorisi Sınıflandırma", **Sakarya University Journal of Computer and Information Sciences**, Cilt 2, Sayı 1, ss.18-27.

KELEŞ, M. B., A. KELEŞ; A. KELEŞ (2020), "Makine Öğrenmesi Yöntemleri ile Uçuş Fiyatlarının Tahmini", **Euroasia Journal of Mathematics, Engineering, Natural and Medical Sciences**, Cilt 7, Sayı 11, ss. 72-78.

KILIÇ, S. (2013), "Klinik Karar Vermede ROC Analizi", **Journal of Mood Disorders**, Cilt 3, Sayı 3, ss. 135-40.

KILIÇER, S.; R. SAMLİ (2023), "E-Ticaret Sitelerindeki Türkçe Ürün Yorumları Üzerine Makine Öğrenmesi Algoritmaları ile Duygu Analizi", **Veri Bilimi Dergisi**, Cilt 6, Sayı 2,ss. 15-23.

KILINÇ, D., E. BORANDAĞ, F. YÜCALAR, V. TUNALI , M. ŞİMŞEK; A. ÖZÇİFT (2016), "KNN Algoritması ve R Dili ile Metin Madenciliği Kullanılarak Bilimsel Makale Tasnifi", **Marmara Fen Bilimleri Dergisi**, Cilt 28, Sayı 3, ss. 89-94.

KİRAZ, M. U. (2022), "Rpl Tabanlı Iot Cihazların Zafiyetinin Tespiti İçin Makine Öğrenmesi Algoritmalarının Karşılaştırılması", Yüksek Lisans Tezi, **Milli Savunma Üniversitesi, Hezârfen Havacılık ve Uzay Teknolojileri Enstitüsü**. İstanbul.

ŁAPCZYŃSKI, M.; S. BIAŁOWĄS (2013), "Discovering Patterns of Users' Behaviour in an E-shop-Comparison of Consumer Buying Behaviours in Poland and Other European Countries", **Studia Ekonomiczne**, Issue 151, ss. 144-153. doi:https://doi.org/10.24432/C5QK7X

MARİA, A. (1997), "Introduction To Modeling And Simulation", **Proceedings of the 1997 Winter Simulation Conference**, ss. 7 -13.

NACAR, E. N.; B. ERDEBİLLİ (2021), "Makine Öğrenmesi Algoritmaları İle Satış Tahmini", **Journal of Industrial Engineering**, Cilt 32, Sayı 2, ss. 307-320.

NİZAM, H.; S. S. AKIN, (2014), "Sosyal Medyada Makine Öğrenmesi ile Duygu Analizinde Dengeli ve Dengesiz Veri Setlerinin Performanslarının Karşılaştırılması", **XIX. Türkiye'de İnternet Konferansı**, Cilt 1, ss. 6.

OĞUZLAR, A. (2005), "Lojistik Regresyon Analizi Yardımıyla Suçlu Profilinin Belirlenmesi",**Atatürk Üniversitesi İktisadi ve İdari Bilimler Dergisi**, Cilt 19, Sayı 1, ss. 21-35.

OLKUN, S., Ö. ŞAHİN, Z. AKKURT , F. T. DİKKARTIN; H. GÜLBAĞCI (2009), "Modelleme Yoluyla Problem Çözme ve Genelleme: İlköğretim Öğrencileriyle Bir Çalışma", **Eğitim ve Bilim**, Cilt 34, Sayı 151, ss.65-73.

ONAN, A. (2017), "Twitter Mesajları Üzerinde Makine Öğrenmesi Yöntemlerine Dayalı Duygu Analizi", **Yönetim Bilişim Sistemleri Dergisi**,Cilt 3, Sayı 2, ss. 1-14.

ONAN, A.; S. KORUKOĞLU (2016), "Makine Öğrenmesi Yöntemlerinin Görüş Madenciliğinde Kullanılması Üzerine Bir Literatür Araştırması", **Pamukkale Üniversitesi Mühendislik Bilimleri Dergisi**,Cilt 22, Sayı 2, ss.111-122.

ÖZCAN, T. (2020), "Yığınlanmış Özdevinimli Kodlayıcılar İle Göğüs Kanserinin Sınıflandırılması Ve Klasik Makine Öğrenme Metotları İle Performans Karşılaştırması", **Erciyes Üniversitesi Fen Bilimleri Enstitüsü Dergisi**, Cilt 36, Sayı 2,ss. 151-160.

ÖZGÜR, A.; H. ERDEM, (2012), "Saldırı Tespit Sistemlerinde Kullanılan Kolay Erişilen Makine Öğrenme Algoritmalarının Karşılaştırılması", **Bilişim Teknolojileri Dergisi**, Cilt 5,Sayı 2, ss. 41 - 48.

ÖZTÜRK, H. (2022), "Dengesiz Veri Setlerinde Farklı Dengeleme Algoritmalarının Optimum Denge Oranlarının Sınıflandırma ve Regresyon Ağaçları Yöntemi ile İncelenmesi: Simülasyon Çalışması ", Doktora Tezi, **Aydın Adnan Menderes Üniversitesi Sağlık Bilimleri Enstitüsü**, Aydın.

PEKEL, E. (2018),"Farklı Makine Öğrenmesi Algoritmalarının Karşılaştırılması", Yüksek Lisans Tezi, **Ondokuz Mayıs Üniversitesi F.B.E**, Samsun.

PROVOST, F.; T. FAWCETT (2013), "Data Science for Business: What You Need to Know about Data Mining and Data-Analytic Thinking", **O'Reilly Medya, Inc.**

SARIMAN, G. (2011),"Veri Madenciliğinde Kümeleme Teknikleri Üzerine Bir Çalışma: K-Means ve K-Medoids Kümeleme Algoritmalarının Karşılaştırılması", **Süleyman Demirel Üniversitesi Fen Bilimleri Enstitüsü Dergisi**, Cilt 15, Sayı 3, ss.192-202.

SEVLİ, O. (2019), "Göğüs Kanseri Teşhisinde Farklı Makine Öğrenmesi Tekniklerinin Performans Karşılaştırması", **Avrupa Bilim ve Teknoloji Dergisi**, Sayı 16, ss. 176-185.

SOLMAZ, R., M. GÜNAY; A. ALKAN (2014), "Fonksiyonel Tiroit Hastalığı Tanısında Naive Bayes Sınıflandırıcının Kullanılması", **Akademik Bilişim Konferansı**, ss. 891-897. Mersin.

ŞAHİNASLAN, Ö., H. DALYAN; E. ŞAHİNASLAN (2022), "Naive Bayes Sınıflandırıcısı Kullanılarak YouTube Verileri Üzerinden Çok Dilli Duygu Analizi", **Bilişim Teknolojileri Dergisi**, Cilt 15, Sayı 12,ss. 221 - 229.

ŞENEL, S.; B. ALATLI (2014), "Lojistik Regresyon Analizinin Kullanıldığı Makaleler Üzerine Bir İnceleme", **Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi**, Cilt 5, Sayı 1,ss. 35-52.

TAŞCI, E.; A. ONAN (2016), "K-En Yakın Komşu Algoritması Parametrelerinin Sınıflandırma Performansı Üzerine Etkisinin İncelenmesi", **Akademik Bilişim**, Cilt 1, Sayı 1, ss. 4-18.

TAŞDEMİR, F. (2022), "ROC Analizi ve R Yazılımı ile Verilerin Sınıflama Doğruluklarının Karşılaştırılması", **National Journal of Education Academy**, Cilt 6, Sayı 2, ss. 230-241.

TOÇOĞLU, M. A., A. ÇELİKTEN, İ. AYGÜN; A. ALPKOÇAK (2019), "Türkçe Metinlerde Duygu Analizi Için Farklı Makine Öğrenmesi Yöntemlerinin Karşılaştırılması", **Dokuz Eylül Üniversitesi Mühendislik Fakültesi Fen ve Mühendislik Dergisi**,Cilt 21, Sayı 63, ss. 719-725.

TOMAK, L.; Y. BEK (2011), "İşlem Karakteristik Eğrisi Analizi ve Eğri Altında Kalan Alanların Karşılaştırılması", **Journal of Experimental and Clinical Medicine**, Cilt 27, Sayı 2, ss. 58-65.

TÜTÜNCÜ, T. E.; S. GÜRSAKAL (2023), "Kredi Temerrüt Riskini Tahmin Etmede Makine Öğrenme Algoritmalarının Karşılaştırılması", **Avrupa Bilim ve Teknoloji Dergisi**, Sayı 50, ss. 14-22.

USLU, O.; S. AKYOL (2021), "Türkçe Haber Metinlerinin Makine Öğrenmesi Yöntemleri Kullanılarak Sınıflandırılması", **Eskişehir Türk Dünyası Uygulama ve Araştırma Merkezi Bilişim Dergisi**, Cilt 2, Sayı 1, ss. 15-20.

YAKUT, E., B. ELMAS; S. YAVUZ (2014), "Yapay Sinir Ağları ve Destek Vektör Makineleri Yöntemleriyle Borsa Endeksi Tahmini", **Süleyman Demirel Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi**, Cilt 19, Sayı 1, ss. 139-157

YAVAŞ, M., A. GÜRAN; M. UYSAL (2020), "Covid-19 Veri Kümesinin SMOTE Tabanlı Örnekleme Yöntemi Uygulanarak Sınıflandırılması", **Avrupa Bilim ve Teknoloji Dergisi,**Özel Sayı, ss. 258-264.

YEŞİLADALI, B. (2018), "Modeling Customers' Online Purchasing Behavior Using Clickstream Data", Yüksek Lisans Tezi, **Boğaziçi Üniversitesi, F.B.E**, İstanbul.

ZHANG, H. (2004), "The Optimality of Naive Bayes", **American Association for Artificial Intelligence**,Cilt 1,Sayı 2,ss. 3.