T.C.
SAKARYA ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

VERİ MADENCİLİĞİ YÖNTEMLERİYLE MÜŞTERİ KAYBI ANALİZİ: YAZILIM SEKTÖRÜ

YÜKSEK LİSANS TEZİ

Sena KASIM

Enstitü Anabilim Dalı	**:**	**BİLİŞİM SİSTEMLERİ MÜHENDİSLİĞİ**
Tez Danışmanı	**:**	**Dr. Öğr. Üyesi Levent ÇALLI**

Temmuz 2022

T.C.
SAKARYA ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

VERİ MADENCİLİĞİ YÖNTEMLERİYLE MÜŞTERİ KAYBI ANALİZİ: YAZILIM SEKTÖRÜ

YÜKSEK LİSANS TEZİ

Sena KASIM

Enstitü Anabilim Dalı : BİLİŞİM SİSTEMLERİ MÜHENDİSLİĞİ

Bu tez …/…/2022 tarihinde aşağıdaki jüri tarafından oybirliği / oyçokluğu ile kabul edilmiştir.

Doç. Dr.	**Doç. Dr.**	**Dr. Öğr. Üyesi**
Ad SOYAD	**Ad SOYAD**	**Ad SOYAD**
Jüri Başkanı	**Üye**	**Üye**

BEYAN

Tez içindeki tüm verilerin akademik kurallar çerçevesinde tarafımdan elde edildiğini, görsel ve yazılı tüm bilgi ve sonuçların akademik ve etik kurallara uygun şekilde sunulduğunu, kullanılan verilerde herhangi bir tahrifat yapılmadığını, başkalarının eserlerinden yararlanılması durumunda bilimsel normlara uygun olarak atıfta bulunulduğunu, tezde yer alan verilerin bu üniversite veya başka bir üniversitede herhangi bir tez çalışmasında kullanılmadığını beyan ederim.

Sena KASIM

20.05.2022

TEŞEKKÜR

Yüksek lisans eğitimim boyunca beni destekleyen, her zaman deneyimini ve bilgisini benimle paylaşan değerli danışman hocam Dr. Öğr. Üyesi Levent Çallı'ya teşekkürlerimi sunarım.

Tüm yüksek lisans eğitimim boyunca her zaman yanımda olan, bu süreci benimle birlikte yaşayan aileme ve arkadaşlarıma çok teşekkür ederim.

İÇİNDEKİLER

SİMGELER VE KISALTMALAR LİSTESİ

K-NN	: K-Nearest Neighbors
SaaS	: Software As A Service
ERP	: Enterprise Resource Planning
CRISP- DM	: Cross Industry Standart Process Model for Data Mining
STD	: Standart
MIN	: Minimum
MAX	: Maximum

ŞEKİLLER LİSTESİ

TABLOLAR LİSTESİ

ÖZET

Anahtar Kelimeler: Müşteri Kaybı Analizi, Karar Ağaçları, Rastgele Orman, Lojistik Regresyon, K-nn, Naive Bayes, Yapay Sinir Ağları

Rekabet gücünün arttığı ve hızlı büyüyen sektörlerde devamlılığın sağlanması için sadık müşterilerin artırılması önemli bir konudur. Yeni müşteri kazanmak için harcanan maliyet mevcut müşteriyi elde tutmak için harcanan maliyetten çok daha yüksektir ve bu açıdan mevcut müşterilerin davranışları incelenerek, firmayı bırakma ihtimali olan müşteriler belirlenip memnuniyet artırmaya yönelik çalışmaların yapılması gereklidir. Literatürde, müşteri işletmeyi terk etmeden önce bunu fark edebilecek önleyici yaklaşımlar önerilmektedir. Churn (kayıp) analizi yöntemi de bu yaklaşımlardan biridir. Telekominasyon, bankacılık ve daha çok üyelik sisteminin yoğun olduğu sektörlerde tercih edilen kayıp analizi, bu çalışmada farklı olarak yazılım sektöründe hizmet veren özel bir firma verileriyle yapılmıştır. Veri setinde bulunan 17 özellik analiz edilerek müşteri kayıp durumu ile aralarında anlamlı bir ilişki bulunan değişkenler çalışmada kullanılmıştır. Ürün sayısı, cari sayısı, sipariş sayısı, teklif sayısı, fatura sayısı, kasa kullanıcı sayısı, kargo kullanımı, mail kullanımı, özel rapor kullanımı ve kullanıcı sayısı ile müşteri kaybı arasında anlamlı ilişki bulunmuş ve çalışma bu özellikler kullanılarak yapılmıştır. Veri setimiz 1951 kayıttan oluşmaktadır, bu veriler 1463 adet eğitim verisi 488 adet test verisi olacak şekilde 2 parçaya ayrılmıştır. Veri seti incelenerek ve ön işlemler yapılarak analize uygun hale getirilmiştir. Karar Ağaçları, Rastgele Orman, Lojistik Regresyon, K-nn, Naive Bayes ve Yapay Sinir Ağları algoritmaları kullanılarak müşteri kayıp analizi yapılmış ve doğruluk oranları karşılaştırılmıştır. Veriler çapraz doğrulama(cross validation) ile optimize edilerek, tekrar analiz edilmiş ve iki analiz sonucunda da Rastgele Orman(Random Forest) algoritması en iyi sonucu veren algoritma olarak bulunmuştur.

CUSTOMER CHURN ANALYSIS WITH DATA MINING METHODS: SOFTWARE AS A SERVICE(SAAS) INDUSTRY

SUMMARY

Keywords: Customer Churn Analysis, Decision Tree, Random Forest, Logistic Regression, k-NN, Naive Bayes, Artificial Neural Networks

Ensuring customer continuity and increasing the number of loyal customers is an essential issue for industries with increasing competitiveness and rapidly growing. Since the cost of acquiring new customers is much higher than retaining existing customers, businesses must examine existing customer behaviors, identify customers likely to leave the business, and conduct marketing activities to increase satisfaction. In the literature, some approaches are suggested to detect customers who have the intention to leave the company. The churn analysis method is one of them and is mostly used in business-to-consumer business models such as telecommunications, banking, and retailing. This study considered a software company that provides services within the business-to-business model was considered. Seventeen features in the data set with 1951 records were analyzed, and the ten features (number of products, number of customers, number of offers, number of orders, number of invoices, cargo usage, number of users, custom report usage, number of cash register receipts, email connection) that found a significant relationship with customer churn variable were selected to analyze in the study. Customer churn analysis was performed using Decision Tree, Random Forest, Logistic Regression, k-nearest neighbors (k-NN), Naive Bayes, and Artificial Neural Networks algorithms. As a result, the Random Forest algorithm was found to give the best result.

BÖLÜM 1. GİRİŞ

Bir müşterinin, almış olduğu hizmet veya kullanmış olduğu ürünü bırakması durumunu analiz etmek için yapılan çalışmalar müşteri kayıp analizi (churn) olarak adlandırılır (Çelik, 2018). Genellikle çok kullanıcılı sistemler için yapılan bu çalışmalar sektörel olarak bankacılık, telekom veya sigortacılık sektörlerinde daha fazla tercih edilmektedir. Yapılan analizler ve tahminlerle müşteri kaybının önüne geçilmesi ve müşteri ilişkileri yönetiminde çözümler üretilebilir (Seker, 2016).

Artan rekabet ortamında, tüketici davranışları ve memnuniyeti son derece önemli noktaya ulaşmıştır. Yeni müşteri kazanma maliyeti, var olan müşterileri elimizde tutmaktan daha maliyetlidir. Müşteriyi elde tutma stratejisinin, yeni bir müşteri elde etme stratejisinden karlı olduğu defalarca gösterilmiştir (Keramati et al., 2014). Maliyetin fazla olacağı bu ortamda uzun vadeli müşteri memnuniyetinin sağlanması ancak etkili bir müşteri ilişkileri yönetimi ile sağlanabilir (Kaynar et al., 2017). Müşteriyi elde tutma şirkete rekabet ortamında avantaj sağlayan en önemli stratejilerden biridir (Kaptan, 2019).

Müşteri ilişkileri yönetimi sistemlerinde, müşterilere ait temel bilgiler, müşteri alışkanlıkları, müşteri memnuniyetleri, müşterinin kullanmış olduğu hizmet değerlendirmeleri gibi özellikler kayıt altına alınmaktadır. Toplanan bu veriler müşteri kayıp analizi için kullanılarak elde edilen sonuçlara göre müşteriler segmente edilebilir (Kaptan, 2019). Şirketler müşteriler için doğru stratejiler oluşturabilir.

Gelişen teknoloji ile her geçen gün elektronik ortama taşınan işlerin sayısı artmaktadır. Birçok işletme, süreçlerini yazılım programları kullanarak elektronik ortama taşımaktadır. Bu da yazılım sektöründeki popülerliğin artmasını sağlamış, bununla birlikte rekabet ortamı da artmıştır. Hizmet sektörü olarak yazılım (SaaS) en büyük

Pazar segmentlerindendir ve bu konumun korunması beklenmektedir (Amornvetchayakul & Phumchusri, 2020) . Yazılım sektörü bilgi çağında en popüler iş modellerinden biridir (Sukow & Grant, 2013). Müşterilerin varlığı bu sektör içinde önemli hale gelmiştir. İnternetin fazla kullanıldığı ve abonelik sisteminin uygulandığı yazılım hizmetleri de müşterileri elde tutmak için gerekli analizleri yaparak, gelecek için stratejiler oluşturmalıdır.

Müşteri kayıp analizi için veri madenciliği yöntemleri (Destek Vektör Makineleri, Naive Bayes, Yapay Sinir Ağları, Lojistik Regresyon, Karar Ağacı, Rastgele Orman) ve araçları(Orange, Rapidminer, R, Weka, KNime) kullanılabilir. Yapılan çalışmalarla müşterinin firmayı terk etmesine etki eden unsurlar belirlenebilir ve müşterinin kalıp gitme durumunu tahmin edilebilir.

Bu çalışmanın amacı; hizmet sektöründe bulunan özel bir yazılım firmasının müşteri verilerini kullanarak veri madenciliği yöntemleri ile müşteri analizini yapmak ve müşterilerin firmayı hangi sebeplerden ötürü terk ettiğini tahmin etmektir. Bu bağlamda konu itibari ile Türkiye'de yapılan ilk çalışmalardan biri olmasından verimli churn modeli tespitinden ve etkin özellik seçiminde akademik katkı sağlayacağı umulmaktadır. Pratik katkının ise şirket yöneticilerine müşteri devamlılığı açısından katkı sağlayacağı düşünülmektedir.

BÖLÜM 2. TEMEL KAVRAM VE YÖNTEMLER

Çalışmada özel bir yazılım firmasının Almanya ve Türkiye'deki müşterilerine ait veriler kullanılmıştır. 16 bağımsız özellik dikkate alınarak bağımlı değişken olarak müşteri kayıp özelliği kullanılmıştır. İlk aşamada 16 özelliğin seçimi için özellik seçimi yöntemlerinden ki- kare yöntemi, bilgi kazancı yöntemi, kazanç oranı yöntemi ve gini indeksi yöntemi veri setine uygulanarak 4 yöntem içinde ortak olan 10 özellik belirlenmiştir ve bu özellikler kullanılarak müşteri kayıp analizi yapılmıştır.

2.1. Yöntem

Özel bir yazılım firmasına ait müşteri verileri, makine öğrenmesi modelleri kullanılarak Müşteri kayıp(churn) analizi uygulaması Python programlama dili ile yapılacaktır. Veri setinde bulunan özelikler Orange (Orange3 for Windows, Version 3.25, 2020) programında özellik seçimin yapılarak, öncelikli özellikler belirlenecek. Veriler ön analiz süreçlerinden geçirilerek, eğitim ve test verisi olarak rastgele iki farklı gruba ayrılacaktır. Değişkenler arasındaki ilişki korelasyon matrisi ile gösterilecektir. Modelleme için uygun hale getirdiğimiz veri setine; Karar Ağacı, Lojistik Regresyon, Naive Bayes, K-NN, Random Forest ve Sinir Ağları sınıflandırma algoritmaları uygulanarak doğruluk oranları hesaplanarak ve en iyi sonucu veren algoritma ile müşteri kayıp analizi yapılacaktır.

2.1.1. Öznitelik seçimi

Büyük veri içerisinde gereksiz değişkenlerin veri kümesinden çıkartılması işlemi olarak ifade edebileceğimiz özellik seçimi, veri madenciliği çalışmalarının önemli aşamalarından biridir. Bağımlı değişkenimiz ile diğer değişkenler arasında ki ilişkiler incelenerek, gereksiz olan değişkenler özellik seçimi yöntemleri ile belirlenebilmekte

ve veri boyutunun azaltılması işlem ile analiz gücünde etkili olmaktadır. Budak (2018) özellik seçimi avantajlarını şu şekilde ifade etmektedir;

- Veri kümesinin boyutunu düşürür, çalışma hızını arttırır,
- Gereksiz veriyi çıkartır,
- Veriyi anlaşılır ve basit hale getirir,
- Depolama için gerekli olan alan azalır

Özellik seçiminde filtreleme (filter), sarmal (wrapper) ve gömülü (embedded) olmak üzere üç tür yöntem bulunmaktadır (Şener 2020). Filtreleme yöntemi; kullanılan en eski özellik seçimi yöntemlerindendir. Sarmalayıcı yöntemlere göre daha az maliyetli ve daha hızlıdır. Kullanımı basit olduğundan yüksek boyutlu verilerin analizinde tercih edilir. Sarmal yöntem; veri setinde farklı alt kümeler oluşturularak model oluşturulur. En iyi sonucu veren özellikleri seçilir. Her değişken için model çalıştırıldığından zaman ve maliyet açısından avantajlı değildir. Gömülü Yöntem ise veri seti içindeki alt kümelerde daha iyi performans gösteren kombinasyonu bulmak için tarama yaparak ilerler. Tüm kombinasyonlar deneneceği için maliyet ve zaman açısından avantajlı değildir (Şener 2020).

2.1.1.1. Ki-Kare(Chi- Square) testi

Ki-kare testi, iki veya daha fazla özelliğin birbirinden bağımsız olup olmadığını araştırmak için kullanılır. Ki-kare testi fazla ön koşulu bulunmadığından dolayı kolay uygulanmakta ve çalışmalarda tercih edilmektedir (Suner & Demirarslan, 2021). En önemli koşul verilerin kategorik olması ve gruplar birbirinden bağımsız olmasıdır.

Ki-kare testinde:

(H_0) Sıfır Hipotezi: İki kriterin bağımsız olduğunu
(H_A)Araştırma Hipotezi: İki kriterin bağımlı olduğunu ifade eder.
Ki-Kare testi gözlenen frekans değeri ile beklenen frekans değerlerinin karşılaştırılmasına dayanır. Formülü denklem 2.1 de gösterilmektedir.

$$X^2 = \sum \frac{(G-B)^2}{B} \quad (2.1)$$

G= Gözlenen değer

B= Beklenen değer

$$\text{Beklenen} = \frac{\text{Satır Top.X Sütun Top.}}{\text{Genel Top.}} \quad (2.2)$$

Ki-kare değerinin anlamlı bir fark olup olmadığını anlamak için serbestlik değerine ihtiyaç duyulur. Serbestlik değeri kategorik değişkenlerin kategori sayılarından 1 çıkartılarak bu sayıların birbiri ile çarpımından bulunur.

$$d = (\text{Satır Sayısı} - 1)x(\text{Sütun Sayısı} - 1) \quad (2.3)$$

2.1.1.2. Bilgi Kazancı(Information Gain) yöntemi

Bilgi kazancı yöntemi entropi teorisine dayanan bir yöntemdir. Entropi, düzensizlik veya belirsizliğin ölçüsüdür. A özelliğine bağlı olarak B özelliğinde ki entropi değerinin azalmasını gösterir (Emhan & Akın, 2019). Entropi 0 ve 1 arasında değer alır, 1'e yakın olması belirsizliğin artmış olduğunu gösterir. Karar ağaçlarının oluşmasında kullanılan önemli yöntemlerden biridir. Veri kümesinde ayırt edici özellikleri belirlemek için kullanılır (Yazıcı et al., n.d.). Denklem 2.4'te entropi, 2.5'te Bilgi kazancı gösterilmektedir.

$$H(Y) = -\sum_{k=1}^{m} p_i \log_i (p_i) \quad (2.4)$$

p_i : Y veri kümesindeki i sınıfının olasılığıdır.

$$BK(Y,X) = H(Y) - \sum_{k=1}^{n} p(Y_i)H(Y_i) \quad (2.5)$$

H(Y) : X'e bölünmeden önceki entropi değeri

H(Yi): i alt bölümünün X üzerinde bölünme olduktan sonraki entropisi

p(Yi): i alt bölümünün X üzerinde bölünme olduktan sonraki olasılığıdır (Yazıcı et al., n.d.).

Bilgi kazancı yöntemi ayırt edici özellikleri seçmek için özellik seçimlerinde kullanılır (Demir, 2021).

2.1.1.3. Kazanç Oranı (Gain Ratio) yöntemi

Kazanç oranı yöntemi, bilgi kazancı değerinin entropi değerine bölünmesi ile elde edilir. Denklem 2.6'da gösterilmiştir.

$$KO(Y,X) = \frac{BK(Y,X)}{H(Y,X)} \tag{2.6}$$

BK(Y,X): Bilgi kazancı oranı
H(Y,X): Entropi oranı

Kazanç oranı, doğruluk oranı açısından değerlendirildiğinde bilgi kazancı yöntemine göre daha performanslıdır (Demir, 2021).

2.1.1.4. Gini İndeksi (Gini Index) yöntemi

Gini indeksi, rastgele seçilen özelliğin ne sıklıkla yanlış tespit edildiğini ölçmek için kullanılan bir yöntemdir. Düşük gini oranı tercih edilmelidir. 0 ve 1 arasında değer alır. 0 en iyi eşitlik 1 en iyi eşitsizliktir. Denklemi 2.7'de gösterilmiştir.

$$Gini = 1 - \sum_j p_j^2 \tag{2.7}$$

Pj: X veri kümesindeki bir kaydın, L kümesine ait olma olasılığıdır.

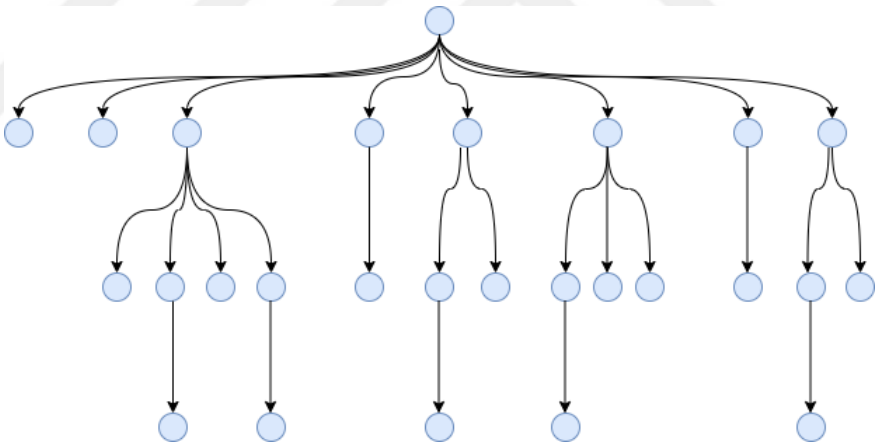
Gini indeksi, iki parçalı sonuç elde etmek istenilen kullanılan yöntemlerden biridir(M. Çelik, 2009).

2.1.2. Churn analizi için kullanılan tahminleme yöntemleri

Literatürde churn analizi ile ilgili yapılan çalışmalar incelendiğinde, kullanılan yöntemler Tablo 3.1.'de gösterilmiştir. Çalışmalarda ortak olarak kullanılan 6 yöntem seçilerek, çalışma gerçekleştirilmiştir. Kullanılan yöntemler aşağıda anlatılmıştır.

2.1.2.1. Karar Ağacı (Decision Tree)

Karar ağacı algoritması, sınıflandırma algoritmalarından biridir. Bir karar ağacı tanımlanmış hedef değişkenine sahiptir. Model yukarıdan aşağıya böl ve yönet metodu ile çalışır. Veri kümesi, karar verme kurallar uygulanarak daha küçük kümelere bölünür. Karar ağaçları kök, düğüm ve yapraklardan oluşur. Bölünme en iyi bölünmeyi gösteren kök hücre ile başlar. Bu işlem hedefe ulaşana kadar tekrarlı bir şekilde yinelenir.



Şekil 2.1. Örnek Karar Ağacı Modeli

En yaygın kullanılan karar ağacı modellerinden bazıları şunlardır: ID3, C4.5, CHAID, CART(Kaptan, 2019).

Karar ağacı algoritmalarının faydaları ise şunlardır:

- Anlama ve yorumlaması kolaydır. Kullanılan ağaç yapıları görselleştirilebilir.
- Kısıtlama olmadan hem kategorik hem sayısal veriler kullanılabilir.

- Diğer sınıflandırma tekniklerine göre daha az veri ile oluşturulabilir.(Keramati et al., 2014).

2.1.2.2. Lojistik Regresyon (Logistic Regression)

Lojistik regresyon makine öğrenmesinde en iyi olarak bilinen yöntemlerden biridir (Amornvetchayakul & Phumchusri, 2020). Sonucu belirleyen bir veya daha fazla bağımsız değişkenin bulunduğu veri setlerini analiz etmek için kullanılır. Denklem 3.4'de çoklu doğrusal regresyon tahmin fonksiyonu verilmiştir.

$$\ln\left(\frac{Pj}{Pk}\right) = b_0 + \sum_i b_i X_i \qquad (2.8)$$

Pj = hedef sınıf j'nin olasılığı
Pk = referans hedef sınıf k'nin olasılığı
b_0 = model sabiti
b_i= regresyon katsayıları
X_i = tahminleyiciler

Lojistik regresyon algoritmasının faydaları ise şunlardır:

- Uygulanması ve yorumlanması kolaydır.
- Veri seti doğrusal ise iyi performans gösterir.

2.1.2.3. Naive Bayes

Naive Bayes algoritması, tahminlerini yapabilmek için Bayes teoreminin matematiğini kullanır (Nath, 2014). Kolay uygulanabilen ve anlaşılabilen en basit makine öğrenme algoritmalarından biridir. Bir örnek için her durumun olasılığını hesaplar ve en yüksek değere göre sınıflandırır. Az veri seti kullanılarak başarılı sonuçlar elde edilebilir. Dengesiz veri setlerinde de kullanılabilir. Bayes teoremi 2.9 denkleminde verilmiştir.

$$P\left(\frac{H}{X}\right) = \frac{P\left(\frac{X}{H}\right)P(H)}{P(X)} \quad (2.9)$$

$P\left(\frac{H}{X}\right)$ = X olayı gerçekleştiği anda H olayının meydana gelme olasılığı

$P\left(\frac{X}{H}\right)$= H olayı gerçekleştiği anda X olayının meydana gelme olasılığı

P(H) = H olayının ön olasılığı

P(X) = X olayının ön olasılığı

Naive Bayes regresyon algoritmasının faydaları ise şunlardır:

- Basit ve kolay uygulanabilir.
- Yüksek boyutlu verilerde iyi çalışabilir.
- Hızlı olduğu içi gerçek zamanlı sistemlerde kullanılabilir.

2.1.2.4. K-NN (K-Nearest Neihbors)

K-NN algoritması kolay uygulanabilen sınıflandırma algoritmalarından bir tanesidir. Hem sınıflandırma hem de regresyon tarafında kullanılmaktadır. Algoritma sınıfları belli olan veri kümesinden yararlanılarak kullanılmaktadır. Veri setine katılacak olan yeni bir verinin, var olan verilere uzaklığı hesaplanır ve k sayıdaki yakın komşuluğuna bakılır. k seçimi uygulamada önemli bir kısımdır. K değeri küçük bir sayı olduğunda algoritmanın performansında bozulma gözükebilir(Akyiğit, 2021).

Uzaklık hesaplamaları için şu fonksiyonlar kullanılmaktadır: Eucliedan, Manhattan, Minkowski. Bu fonksiyonlar arasında yaygın olarak kullanılan Öklid(Eucliedan) fonksiyonudur (Akyiğit, 2021).

$$d(i, j) = \sqrt{\sum_{k=1}^{p} (x_{ik} - x_{jk})^2} \quad (2.10)$$

d(i, j) = i'nci ve j'inci nesneler arasındaki uzaklık

p= değişken sayısı

x_{ik}= i'nci nesnenin k'ıncı değişkendeki değeri

x_{jk}= j'nci nesnenin k'ıncı değişkendeki değeri

k= komşuluk değeri

K-NN algoritmasının faydaları ise şunlardır:

- Eğitim olmaması,
- Kolay gerçekleştirilebilir ve yerel bilgilere uyarlanabilir olması,
- Gürültülü eğitim verilerine karşı dirençli olması.

2.1.2.5. Rastgele Orman (Random Forest)

Karar ağacı modellerinin en önemli problemlerinden biri veriyi ezberleme ve aşırı öğrenmedir. Rastgele orman modeli bu problemi ortadan kaldırmak için veri setinden yüzlerce farklı alt ağaçlar seçerek bunları eğitmektedir. Bu yöntemle yüzlerce karar ağacı modeli oluşturulur ve oluşturulan karar ağaçları bireysel olarak tahminlemede kullanılır (Akyiğit, 2021).

Bir diğer önemli özellikte değişkenlerin ne kadar önemli olduğunu bize vermesidir. Bu algoritma ile belirlediğimiz sayıda ki en faydalı değişkenleri öğrenebiliriz.

$$\text{Gini} = 1 - \sum_{i=1}^{c} (p_i)^2 \tag{2.11}$$

c= seçilen veri

p_i= her bir verinin, kendisinden küçük ve kendisinden büyük eleman sayılarına bölüm karesi

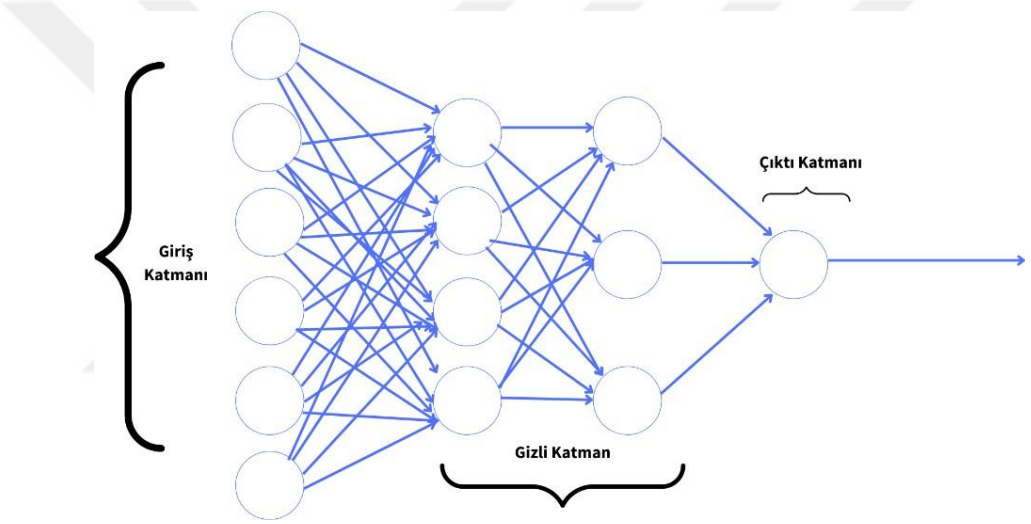
Rastgele Orman algoritmasının faydaları ise şunlardır:

- Hem sınıflandırma, hem regresyon problemlerinde kullanılabilir.
- Veri setinde ki özellikler içinden en önemli özelliği tanımlamamızı sağlar.
- Aşırı uyum problemi ihtimali azdır.

2.1.2.6. Yapay Sinir Ağları (Artificial Neural Network)

Yapay sinir ağları, insan sinir sistemi mantığı ile çalışan ve bunu taklit ederek öğrenmeyi hedefleyen bir algoritmadır. Tahminleme ve sınıflandırma algoritmalarında kullanılan en etkili makine öğrenme algoritmalarındandır (Kaptan, 2019).

Model farklı katmanlar halinde ki nöronlardan oluşmaktadır. Giriş katmanı, verilerin okunduğu katmandır. Her nöron farklı bir niteliği ifade eder ve her nitelik için bir nöron olmalıdır. Çıktı katmanı, hedefin belirlendiği katmandır (Kaynar et al., 2017).



Şekil 2.2. Örnek Yapay Sinir Ağı Modeli

Yapay Sinir Ağları algoritmasının faydaları ise şunlardır:

- Eş zamanlı olarak çalışarak karmaşık işler gerçekleştirir.
- Öğrenerek benzer olaylar karşısında karar verebilirler.
- Diğer yöntemlere göre daha hızlı ve başarılı sonuçlar vermektedir.
- Daha önce karşılaşmadığımız örnekler hakkında bilgi verebilirler.

BÖLÜM 3. KAYNAK ARAŞTIRMASI

Bu çalışmanın literatür taraması için müşteri kayıp analizi ile ilgili yapılan araştırmalar ve uygulamalar incelenmiştir. İncelenen bu çalışmalarda kullanılan yöntemlerin neler olduğu ve elde edilen doğruluk oranları bilgilerine ulaşılmıştır. Çalışmaların daha çok üyelik temelli gelir elde eden sektörler üzerine yapıldığı görülmüştür(Seker, 2016). Telekom, sigortacılık ve bankacılık gibi sektörlerin yanı sıra yazılım hizmetleri, perakendecilik, eğlence, oyun gibi farklı sektörlerde de uygulanmaktadır.

2009 yılında yapılan bir çalışma hava yolu sektöründe bulunan bir firma için yapılmış ve çalışmada kayıp müşteri tanımı ile kayıp müşteri öngörü modeli sunulmuştur. Veri madenciliği tekniklerinden karar ağacı, lojistik regresyon ve yapay sinir ağları kullanılarak modeller geliştirilmiş, geliştirilen modellerin doğruluk oranı karşılaştırılmıştır. Model sonucunda müşteri kaybını etkileyen faktörler açıklanmıştır (Kaptan, 2019).

Telekominasyon şirketleri; müşteri sayısı olarak bakıldığında birçok sektöre göre daha fazla üyeye sahiptir. Müşteri kayıp analizi ile ilgili bu sektörde oldukça fazla çalışma yapılmıştır (Ozcan et al., 2016). Bir telekominasyon şirketinde ki müşteri kaybını tahmin etmek için destek vektör makineler, yapay sinir ağları ve naive bayes yöntemleri kullanılarak bir analiz gerçekleştirilmiş. Toplamda 4667 müşteri ve 21 özellik bulunan veri seti ile çalışma yapılmış. Çalışma sonucunda yapay sinir ağları ile elde edilen sonuç diğer yöntemlere göre daha başarılı çıkmıştır (Kaynar et al., 2017). Tablo 3.1.‘de incelenen kaynakların genel bilgileri bulunmaktadır.

Tablo 3.1. Literatür Özet

Çalışma	Sektör	Model/Teknik	Kullanılan Özellik (Features)	Sonucu Etkileyen Özellikler	Veri	Sonuç	Yazarlar
2008	Telekominasyon	Karar Ağaçları	Cinsiyet, Yaşlı, Ortaklık, Muhtaçlık, Kullanım Hakkı Grup, Telefon hizmeti, Çoklu Hat, İnternet Hizmeti, Çevrimiçi güvenlik, Çevrimiçi yedekleme, Cihaz Koruma, Teknik Destek, Televizyon Akışı, Film Akışı, Sözleşme, Kağıtsız faturalama, Ödeme Yöntemi, Aylık Ücret, Toplam Ücret, Ayrılma	Sözleşme	7032		(Çelik, 2018)
2009		Lojistik Regresyon, Rastgele Orman	6 farklı veri seti kullanılmış (Bank1, Bank2, Mobile, Newspaper, Paytv, Supermarket)			Rastgele Orman	(Burez & Van den Poel, 2009)
2009	Banka	Karar Ağaçları, Rastgele Orman, Destek Vektör Makinesi, Yapay Sinir Ağları	27 değişken		20000	Rastgele Orman	(Xie et al., 2009)
2009	Telekominasyon ve yazılım destek karşılaştırma		tüm kullanıcılar, müşteri tarihi, uzak kullanıcılar, girişler, maksimum koltuklar, mkt segmenti, modül, geçerli kullanıcılar		600+		(Frank & Pittges, 2009)
2009	Yazılım	Karar Ağaçları, Rastgele Orman, Destek Vektör Makinesi, Lojistik Regresyon	23 değişken	currMonthTrans, prevMonthTrans, amountSpend, numAct , UserAct	1788	Rastgele Orman	(Amornvetchayakul & Phumchusri, 2020)
2009	Telekominasyon	Lojistik Regresyon, Karar Ağaçları	Müşteri Kimliği, Hedef Flag, Ön Ödemeli Flag, Tek Şapka Flag, Aol Günleri, Güncel Tarife Grubu, Ortalam Çağrı 5 dakika üstü, Ortalam Çağrı 60 300, Ortalam Çağrı Firması Dnum, Ortalam Çağrı Firması Dur, Ortalam Çağrı GSM Dnum, Ortalam Çağrı GSM Dur, Ortalam Çağrı Intcom Dur, Ortalam Çağrı MAX Dur, Ortalam Çağrı MIN Dur, Ortalam Çağrı N Intcom Dur, Ortalam Çağrı Pstn DurOrtalam Görüşme Toplam Süre, Ortalam Çağrı Toplam Fark Sayısı, Ortalama Görüşme Toplam Sayısı, Standart Çağrı 1 Dakika Aıtı, Standart Çağrı 5 Dak Üstü,Standart Çağrı 60 300,Standart Talep Firması Dnum	Sözleşmeli Olmak, Aynı servis sağlayıcıda bulunan abonelerden arama almak,	1000	Karar Ağacı Modeli	(Gürsoy, 2009)

Tablo 3.1. (Devamı)

2012	Telekominasyon	K-nn, Rastgele Orman	260 değişken	39 değişkene düşürülmüş	50000	Rastgele Orman	(Idris et al., 2012)
2013	Yazılım		Abonelik ücreti, Kayıp oranı, Edinme Oranı, Toplam Abone, Müşteri Ömrü, Müşteri Yaşam Boyu Değeri, Abonelik Ücreti(aylık), Aylık Yinelenen Gelir	Abonelik ücreti, Kayıp oranı, Edinme Oranı, Toplam Abone, Müşteri Ömrü, Müşteri Yaşam Boyu Değeri, Abonelik Ücreti(aylık), Aylık Yinelenen Gelir			(Sukow & Grant, 2013)
2014	Telekominasyon	Naive Bayes	Coğrafi ve nüfus bilgileri, Çağrı detayarı, Hizmet kalitesi, Paket özellikleri	Çift bant, Araba tipi, eğitim1, Etnik, Toplam Kabul, işgal1, Alan, Gelir, oturma boyutu, Prop tipi	50000	Naive Bayes	(Nath, 2014)
2014	Telekominasyon	Karar Ağaçları, Yapay Sinir Ağları, K-nn, Destek Vektör Makinesi	Çağrı Sayısı, Şikayet sayısı, Abonelik süresi, Ücret Tutarı, Kullanım saniyesi, Kullanım Sıklığı, Hizmet türü, yaş grubu, Duru, Churn	Kullanım sıklığı, toplam şikayet sayısı, kullanım saniyesi	3150		(Keramati et al., 2014)
2015	Telekominasyon	Lojistik Regresyon, Karar Ağaçları			50, 100 ve 608	Karar Ağacı Modeli	(Dahiya & Bhatia, 2015)
2017	Telekominasyon	Destek Vektör Makineleri, Yapay Sinir Ağları, Naive Bayes			4667	Yapay Sinir Ağları	(Kaynar et al., 2017)
2016	Telekominasyon	Literatür Araştırması					(Ozcan et al., 2016)
2017	Yazılım	Lojistik Regresyon, Random Forest, XGBoost	21 Değişken	A Tipi Kullanıcı Girişi Sayısı, Proje Sayısı, B Tipi Kullanıcı Girişi Sayısı, Dosya Sayısı, İç Yorum Sayısı	8256	XGBoost	(Ge et al., 2017)
2017	Telekominasyon	Lojistik Regresyon, Karar Ağaçları, Rastgele Orman, Naive Bayes					(Verbeke et al., 2014)

Tablo 3.1. (Devamı)

2019	Telekominasyon	Random Tree (RT), J48, Random For- est (RF), Decision Stump, AdaboostM1 + Decision Stump, Bagging + Random Tree, Naïve Bayes (NB), Multilayer Per- ceptron (MLP), Logistic Regression (LR), IBKandLWL	29 Değişken, 16 Değişken	Toplam aramalar, toplam dakikalar, toplam aramalar geri, çevrimiçi aramalar, ağ dakikaları, ağ devirleri, ağ dışı aramalar, kapalı dakikalar, ağ dışı devirler, gelen toplam aramalar, chrgd aramaları, chrgd mins, chgrd rev, ücretsiz aramalar, ücretsiz dakika geliri sms, yeniden toplam yük	64107, 3333	Rastgele Orman	(Ullah et al., 2019)
2019	Yazılım	Yapay Sinir Ağı, Destek vektör Makinesi, Random Forest	İş metrikleri, Özellik kullanımı, Platform kullanım metrikleri, Hizmet kalitesi metrikleri, Olay metrikleri	Platform kullanım metrikleri, Hizmet kalitesi metrikleri		Destek Vektör Makinesi	(Rautio, 2019)
2019	Yazılım	Lojistik Regresyon	12 Değişken		15000	Lojistik Regresyon	(Mutanen, 2006)
2019	Hava Yolu Endüstrisi	Karar Ağaçları, Lojistik Regresyon, Yapay Sinir Ağları	Id, Yaş, Cinsiyet, Üyelik sonrası Geçen Ay Sayısı, Üyelik Statüsü, Üyelik Durumu, Kazanılan Mil Miktarı, Harcanan Mil Miktarı, Pnr Adet, Son Aydaki Pnr Adeti, Son Yıldaki Ortalam Pnr Adeti, Ekonomi Segment Pnr Adet, Bilet Adet, Son Ay Bilet Adet, Uçulan BİIet Adet, Uçulmayan Bilet Adet, Biletleme Sıklığı, Rotarlı Uçus Adeti, Kayıp Baygaj Adet, Müşteri Kayıp Durumu	Son yıldaki Ortalama Pnr Adedi, Son Ay Bilet Adedi, RotarlıUçus Adedi, Bilet Adet, Uculan Bilet Adet, Üyelik Sonrası Geçen Ay	115021	Karar Ağacı Modeli	(Kaptan, 2019)
2021	Sigortacılık	K-nn, Rastgele Orman, Karar Ağaçları	Yaş, Cinsiyet Medeni Hal, Çalışma Durumu Meslek, Yaşadığı İl, Eğitim Seviyesi, Müşterinin Ödediği Toplam Prim, Plaka Sayısı, Marka, Marka Sayısı, Model, Model Sayısı, Kullanım Tarzı, Model Yılı, Plaka, İl Kodu, Ortalama Hasarsızlık Kademesi, Hasarsızlık Kademe, Araç Yaşı, Satış Kanalı, Unsur Tip, Yenileme	Yaş, Cinsiyet Medeni Hal, Çalışma Durumu Meslek, Yaşadığı İl, Eğitim Seviyesi, Müşterinin Ödediği Toplam Prim, Plaka Sayısı, Marka, Marka Sayısı, Model, Model Sayısı, Kullanım Tarzı, Model Yılı, İl Kodu, Ortalama Hasarsızlık Kademesi, Hasarsızlık Kademe, Araç Yaşı, Satış Kanalı, Unsur Tip	532723	Rastgele Orman	(Akyiğit, 2021)

Telekominasyon şirketi için Türkiye'de yapılmış bir diğer çalışmada, ayrılma eğilimi gösteren müşteriler belirlenmiş ve bu müşteriler için pazarlama stratejilerinin geliştirilmesi hedeflenmiştir. Müşteri kaybını belirlemek için karar ağaçları ve lojistik regresyon yöntemleri uygulanmış ve sonuçlar firma ile paylaşılmıştır (Gürsoy, 2009). 7032 müşteri kaydının ve 20 farklı özelliğin dikkate alınarak IBM şirketi tarafından yürütülen ve bir Telekom şirketine ait müşteri kayıp analizi araştırmasında ise karar ağaçları yöntemi kullanılarak çalışma gerçekleştirilmiş. Çalışma sonunda müşteri kaybını etkileyen önemli öz nitelik sözleşme olarak belirlenmiştir (Çelik, 2018).

Telekom sektöründe yapılan bir ankette yeni müşteri edinmenin var olan müşteriyi elde tutmaktan daha zor olduğu belirlenmiş ve bunun üzerine müşteri kayıp analizi yapılarak müşterilerin ayrılıp ayrılmayacağı yapılan çalışma ile belirlenmiştir. WEKA yazılımı kullanılarak lojistik regresyon ve karar ağaçları yöntemleri ile analiz yapılmış ve sonuçlar karşılaştırılmıştır. Karar ağaçları yöntemi daha etkili bir sonuç ortaya koymuştur (Dahiya & Bhatia, 2015).

Verbeke et al. (2014), sosyal ağ bilgilerinin müşteri kayıp tahmini üzerine etkisini araştırdıkları çalışmada telekominasyon verilerini kullanmışlardır. Çalışmada 2 farklı veri seti kullanılmıştır ve bu veri setlerini birbirinden ayıran en önemli nitelik ağ bağlantısıdır. Çalışma sonunda sosyal ağ etkilerinin müşteri kayıp tahmin modelinin performans üzerinde önemli bir etkisi olduğu ortaya koyulmuştur. Nath(2014) tarafından naive bayes yönteminin kullanıldığı bir diğer çalışmada ise; bir telekominasyon şirketine ait veriler kullanılmış. 50000 gerçek müşteriden oluşan veri seti ile analiz yapılmıştır. Analiz sonucunda naive bayes modelinin % 68 doğruluk oranı gösterdiği elde edilmiştir.

Karar ağacı, yapay sinir ağları, k-nn ve destek vektör makinesi modellerinin performanslarını karşılaştırmak için bir çalışma yapılmış ve bu çalışmada İran'da bulunan bir Telekom şirketinin verileri kullanılmıştır. Modellerin davranışları analiz edilerek ve uzmanlıkları öğrenilerek değerlendirme ölçütlerinin sonucunu iyileştiren karma bir metodoloji sunulmuştur. Sunulan bu metodoloji de % 95 oranında bir doğruluk oranı elde edilmiştir (Keramati et al., 2014). Tuncay Özcan ve arkadaşları

2000 yılından 2016 yılına kadar olan müşteri kayıp analizi ile ilgili yapılmış olan ve çeşitli alan indekslerine giren uluslararası 100 makaleyi inceleyerek; sektör, veri seti, performans değerlendirme ölçütü ve kullanılan yöntemlerini ortaya koymuştur. Sektör olarak en fazla makale telekominasyon sektöründe, kullanılan yöntem karar ağaçları ve performans ölçütü ise doğruluk olarak açıklanmıştır (Ozcan et al., 2016). Müşteri kayıp analizi çalışmalarında kullanılan yöntemlerinden olan lojistik regresyon ve rastgele orman yöntemi kullanılarak 2009 yılında yapılan çalışmada veri seti 6 farklı kategoride ele alınmış sınıf dengesizliği ile nasıl başa çıkılacağı araştırılmıştır. Kullanılan 2 yöntemde karşılaştırılmıştır (Burez & Van den Poel, 2009).

Çin'de gerçek bir banka verileri kullanılarak müşteri kayıp analizi çalışması yapılmıştır. Rastgele orman, yapay sinir ağları, karar ağaçları ve destek vektör makineleri yöntemleri kullanılarak çalışma gerçekleştirilmiş ve rastgele orman yöntemin müşteri kaybını tahmin etmedeki etkinliği araştırılmıştır (Xie et al., 2009).

Yapmış olduğumuz literatür taramasına baktığımızda hizmet sektörü olan yazılım ve yazılım destek alanlarında çok az sayıda çalışma yapılmıştır. 2009 yılında yapılan bir çalışmada bu sektör ve telekominasyon şirketi analizi arasında birçok benzerlik olmasına rağmen bazı farklılıkların olduğu ortaya koyulmuştur. Makale 4 deney sunmakta ve gelecekte yapılacak olan çalışmalara fırsatlar vermektedir. 600 den fazla müşteri kayıtları ve 8 farklı özellik incelenerek çalışma gerçekleştirilmiştir (Frank & Pittges, 2009). Tayland'da yüksek bir kayıp oranıyla karşı karşıya olan bir yazılım şirketi için en iyi müşteri kayıp analizi modelini veren bir çalışma yapılmıştır. Çalışmada dört farklı yöntem kullanılmış bunlar; lojistik regresyon, karar ağacı, destek vektör makinesi ve rastgele orman yöntemleridir. Rastgele orman yöntemi yapılan analizler sonucunda en az hata oranına sahip yöntem olarak belirlenmiştir (Amornvetchayakul & Phumchusri, 2020).

Yüksek boyutlu Telekom verilenin kullanıldığı bir çalışmada veri setlerinin dengesiz yapısının kayıp tahmini modeli performansını etkilediği üzerinde durulmuştur. Bileşen azaltma teknikleri kullanılarak Parçacık sürü optimizasyonu örnekleme yöntemi araştırılmış. Elde edilen yeni veri setine K-nn ve Rastgele orman modelleri

uygulanarak çalışma yapılmıştır. Elde edilen sonuçlara göre Rastgele orman modeli, K-nn modeline göre daha başarılıdır (Idris et al., 2012). 2019 yılında Güney Asya'daki bir Telekom şirketine ait veriler kullanılarak bir çalışma yapılmış. Bu çalışma birçok analiz yöntemi kullanılmış; Random Tree (RT), J48, Rastgele orman, Decision Stump, AdaboostM1 + Decision Stump, Bagging + Random Tree, Naive Bayes, Multilayer Per- ceptron (MLP), lojistik regresyon, IBKandLW. Çalışmadaki veri seti cross validation yapıldıktan sonra tekrar analiz edilmiş ve iki analiz sonucunda da Rastgele orman yönteminin en başarılı sonucu olduğu gözlenmiştir (Ullah et al., 2019).

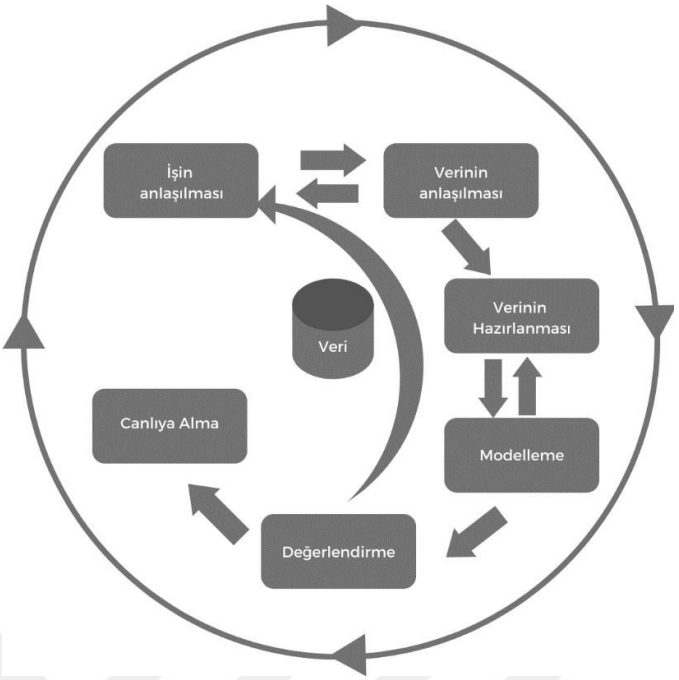
Abonelik sistemlerine dayanan meslek gruplarından olan yazılım sektöründe de müşteri kaybı yaşanmaktadır. Müşteri kaybına sebep olan faktörler tespit edilmeli ve bunlar için bir çözüm yolu geliştirmelidir. Yazılım (SaaS) sektörlerinde ki müşteri kaybını belirlemek için yapılan bir çalışmada yapay sinir ağları, destek vektör makinesi ve random forest modeli kullanılmış. Sonuca etki eden unsurlar belirlenmiş ve yapılan analizler sonucunda destek vektör makinesi modeli en başarılı sonucu vermiştir (Rautio, 2019). Yazılım(SaaS) sektöründe yapılan diğer bir çalışmada ise müşterinin önümüzdeki 3 ayda çalışmayı bırakıp bırakmayacağı tahmin edilmiştir. Çalışma Dört sınıflandırma algoritmasını kullanarak gerçekleştirilmiş. XGBoost modeli, en önemli yazılım kullanım özelliklerini belirlemek ve müşterileri kayıp türü veya riskli olmayan tür olarak sınıflandırmak için en iyi sonuçları vermiştir (Ge et al., 2017).

BÖLÜM 4. YAZILIM SEKTÖRÜNDE VERİ MADENCİLİĞİ YÖNTEMLERİYLE MÜŞTERİ KAYBI ANALİZİ

Hizmet olarak yazılım bilgisayarımıza kurduğumuz uygulamalardan farklı olarak internet tarayıcı üzerinden kullanabildiğimiz bulut tabanlı bir hizmettir. Aboneliğe bağlı lisanslama olarak da bilinmektedir. Uygulama kullan öde ve kiralama mantığına dayanmaktadır. İş hayatında ki kullanım alanları; kurumsal kaynak planlama(erp), insan kaynakları, finans, e-ticaret, müşteri ilişkileri yönetimi gibi alanlardır.

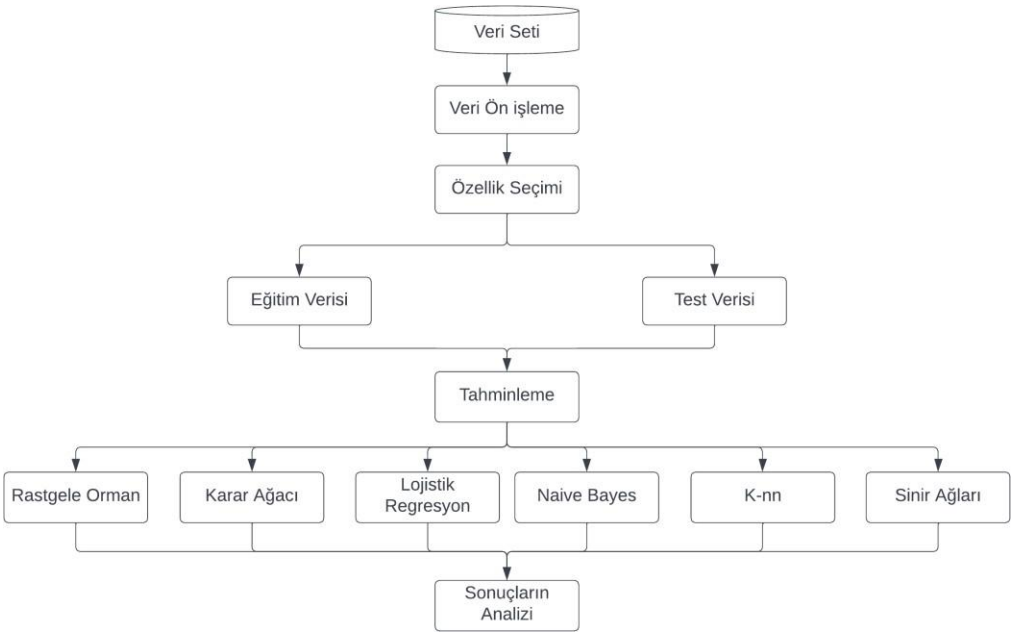
Veri madenciliği çalışmaları çeşitli aşamalardan oluşmaktadır. Çalışmaların çoğunluğunda verilerin toplanması, temizlenmesi, modelin kurulması ve sonuçlar. CRISP-DM (Cross Industry Standard Process Model for Data Mining) veri madenciliği modellerinden yaygın olarak kullanılan endüstri ve kullanılan yazılımdan bağımsız bir veri madenciliği süreç modelidir (Şekeroğlu, 2010).

En temel adımı ile 6 farklı süreçten oluşan CRISP-DM modeli; iş süreçlerinin anlaşılması, verinin anlaşılması, verinin işleme aşaması, model aşaması, değerlendirme aşaması ve ürün aşamaları süreçlerinden oluşur. Aşağıdaki şekilde akış gösterilmektedir.



Şekil 4.1. CRISP-DM Adımları Ve Akışı

Bu çalışma CRISP-DM adımlarına uygun olarak gerçekleştirilmiştir. Çalışmanın akış diyagramı Şekil 4.2.'de gösterilmiştir.



Şekil 4.2. Akış Diyagramı

4.1. Veri Setinin Oluşturulması Ve Önişleme

Çalışma için kullanacağımız veri seti özel bir yazılım firmasından temin edilmiştir. Veri seti 17 özellikten oluşmaktadır. Bu özellikler ve açıklamaları Tablo 4.1.'de yer almaktadır.

Tablo 4.1. Özellikler Ve Açıklamaları

Ürün Sayısı	Müşterinin programda ki ürün sayısı
Müşteri Sayısı	Müşterinin programdaki cari sayısı
Teklif Sayısı	Müşterinin programdaki teklif sayısı
Sipariş Sayısı	Müşterinin programdaki sipariş sayısı
Fatura Sayısı	Müşterinin programdaki fatura sayısı
Pazaryeri sayısı	Müşterinin programdaki aktif pazaryeri bağlantı sayısı
Ödeme Belgesi Sayısı	Müşterinin programdaki ödeme belgesi sayısı
Kargo Kullanımı	Müşterinin programdaki kargo kullanımı
Kasa Bağlantı Sayısı	Müşterinin programdaki kasa bağlantı sayısı
Mail Bağlantısı	Müşterinin programdaki mail bağlantı durumu
Kasa Belge Sayısı	Müşterinin programdaki kasa fişi sayısı
Özel Rapor Kullanımı	Müşterinin programdaki özel rapor kullanım durumu
Üretim Siparişi Sayısı	Müşterinin programdaki üretim için oluşturduğu belge sayısı
Kullanıcı Sayısı	Müşterinin programdaki kullanıcı sayısı
Destek Sayısı	Müşterinin firmadan aldığı ettiği destek sayısı
Müşteri Grubu	Müşteri, hangi şubenin müşterisidir.
Müşteri Kaybı(churn)	Müşterinin programı kullanıp kullanmama durumu

Veri setinde 1951 müşteri verisi, 17 farklı öznitelik bulunmaktadır. Orange veri analizi programı kullanılarak, özellik öncelikli özellikler belirlenmiştir. 16 bağımsız özelliğin seçimi için özellik seçimi yöntemlerinden ki- kare yöntemi, bilgi kazancı yöntemi, kazanç oranı yöntemi ve gini indeksi yöntemi veri setine uygulanarak 4 yöntem içinde en yüksek değeri almış 10 özellik belirlenmiştir ve bu özellikler kullanılarak müşteri kayıp analizi yapılmıştır. Analiz sonucu Tablo 4.2.'de gösterilmektedir.

Tablo 4.2. Orange Analiz Sonuçları Ve Sıralaması

	Ki-Kare	Gini Index	Info. Gain	Gain Ratio
1	Fatura Sayısı	Cari Sayısı	Cari Sayısı	Kasa Bağlantı Sayısı
2	Teklif Sayısı	Fatura Sayısı	Fatura Sayısı	Cari Sayısı
3	Destek Sayısı	Ürün Sayısı	Teklif Sayısı	Özel Rapor Kullanımı
4	Cari Sayısı	Teklif Sayısı	Destek Sayısı	Mail Bağlantısı
5	Ürün Sayısı	Destek Sayısı	Ürün Sayısı	Teklif Sayısı
6	Kasa Bağlantı Sayısı	Kullanıcı Sayısı	Kullanıcı Sayısı	Kullanıcı Sayısı
7	Kargo Kullanımı	Kargo Kullanımı	Kasa Bağlantı Sayısı	Fatura Sayısı
8	Ödeme Belge Sayısı	Kasa Bağlantı Sayısı	Kargo Kullanımı	Destek Sayısı
9	Özel Rapor Kullanımı	Özel Rapor Kullanımı	Özel Rapor Kullanımı	Kargo Kullanımı
10	Mail Bağlantısı	Ödeme Belge Sayısı	Ödeme Belge Sayısı	Ödeme Belge Sayısı
11	Kasa Belge Sayısı	Sipariş Sayısı	Mail Bağlantısı	Ürün Sayısı
12	Kullanıcı Sayısı	Kasa Belge Sayısı	Sipariş Sayısı	Kasa Belge Sayısı
13	Sipariş Sayısı	Mail Kullanımı	Kasa Belge Sayısı	Üretim Sipariş Sayısı
14	Müşteri Grubu	Müşteri Grubu	Müşteri Grubu	Sipariş Sayısı
15	Üretim Sipariş Sayısı	Üretim Sipariş Sayısı	Üretim Sipariş Sayısı	Müşteri Grubu
16	Pazaryeri Sayısı	Pazaryeri Sayısı	Pazaryeri Sayısı	Pazaryeri Sayısı

Müşteri kayıp özelliği ile karşılaştırılan 16 özellikten; ürün sayısı, cari sayısı, sipariş sayısı, teklif sayısı, fatura sayısı, kasa kullanıcı sayısı, kargo kullanımı, mail kullanımı, özel rapor kullanımı ve kullanıcı sayısının anlamlı bir ilişkisinin olduğu belirlenmiştir. Sonuçlar Tablo 4.3.'de gösterilmektedir.

Tablo 4.3. Öncelikli Özellikler

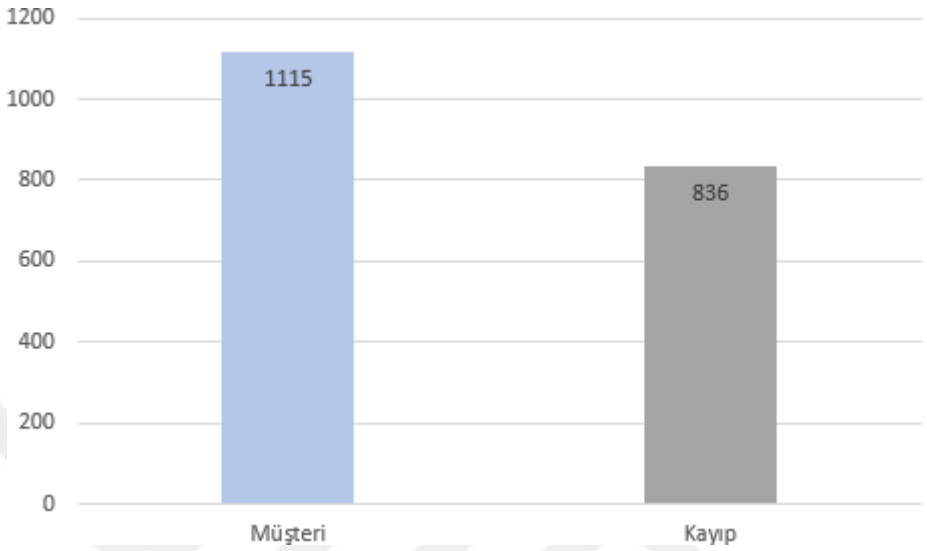
Model Kapsamında Değerlendirilen Özellikler	**Analiz Dışı Özellikler**
Ürün Sayısı	Pazaryeri sayısı
Müşteri Sayısı	Ödeme Belgesi Sayısı
Teklif Sayısı	Destek Sayısı
Sipariş Sayısı	Müşteri Grubu
Fatura Sayısı	Üretim Siparişi Sayısı
Kargo Kullanımı	Kasa Bağlantı Sayısı
Kullanıcı Sayısı	
Özel Rapor Kullanımı	
Kasa Belge Sayısı	
Mail Bağlantısı	

Öncelikli olarak belirlediğimiz 10 özellik ile python programlama dilinde müşteri kayıp analizi(churn) yapılmıştır. Veri setinin istatistiksel sonuçları aşağıdaki Tablo 4.4.'te gösterilmiştir.

Tablo 4.4. Verilerin İstatistiksel Sonuçları

	Ürün Sayısı	**Cari Sayısı**	**Sipariş Sayısı**	**Teklif Sayısı**	**Fatura Sayısı**	**Kasa Belge Sayısı**	**Mail Bağlantısı**	**Özel Rapor Kullanımı**	**Kargo Kullanımı**	**Kullanıcı Sayısı**	**Müşteri Kayıp**
Sayı	1951	1951	1951	1951	1951	1951	1951	1951	1951	1951	1951
Ortalama	1014.04	1550.39	469.36	12.51	673.55	0.03	0.02	0.07	0.16	1.24	0.57
Std	10450.94	10069.71	3490.81	280.05	4890.94	0.41	0.42	1.26	0.61	1.22	0.50
Min	1	0	0	0	0	0	0	0	0	0	0
25%	4	4	0	0	0	0	0	0	0	1	0
50%	6	4	0	0	0	0	0	0	0	1	1
75%	54	12	0	0	0	0	0	0	0	1	1
Max	368883	184842	59001	10888	107190	12	12	49	10	17	1

Veri setinde kullanacağımız 1951 verinin 1115 tanesi firmanın müşterisi olup, 836 tanesi müşteri değildir. Arada ki farka bakıldığında veri setinin dengeli olduğu gözlemlenmiştir. Şekil 4.3.'te grafiksel olarak gösterimi vardır.



Şekil 4.3. Veri Seti Dağılımı

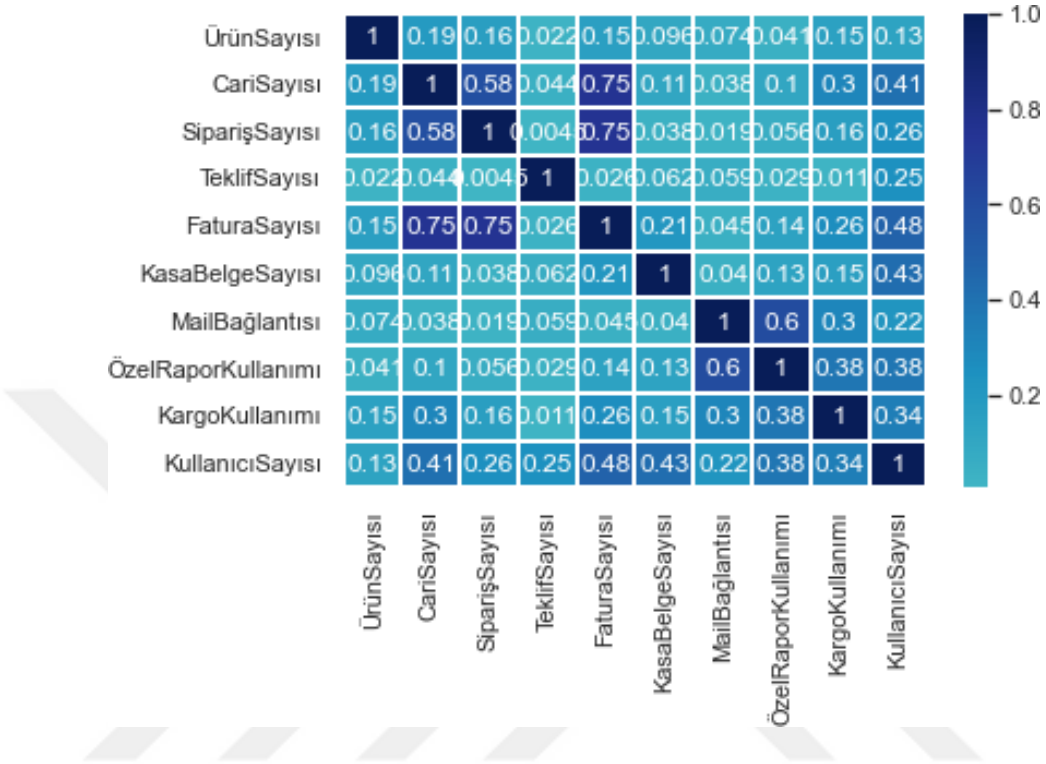
Veri setimizde bulunan ve skalası değişkenlik gösteren özelliklerimize değerleri 0 ve 1 arasında olması için normalizasyon işlemi uygulanır. Normalizasyon adımından sonra verilerimizin ön adımları gerçekleşmiş olur. Veriler iki farklı şekilde gruplara ayrılır. İlk Grup %25 test %75 eğitim verisi olacak şekilde 2 parçaya ayrılır. 1463 satır eğitim verisi, kalan 488 satır test verisidir. İkinci grupta ise %30 test %70 eğitim verisi olacak şekilde 2 parçaya ayrılır. Analizler iki grup için gerçekleştirilecek ve doğruluk oranı iyi olan grup ile çalışmaya devam edilecektir. Grup dağılımları Tablo 4.5.'te gösterilmektedir.

Tablo 4.5. Eğitim Ve Test Veri Sayısı

	Eğitim Verisi Sayısı	**Test Verisi Sayısı**
%75-%25	1463	488
%70-30	1365	586

Değişkenlerimizin kendi aralarında ki ilişkisi olup olmadığını korelasyon ile hesaplayabiliriz. Korelasyon: iki değişken arasında ilişki olup olmadığını ölçmeye yarayan bir yöntemdir.-1 ile 1 arasında değer alan korelasyon sonucu, -1 ve 1 e yakın

olduğunda değişkenler arasında güçlü bir ilişki olduğu söylenebilir. Şekil 4.4.'de değişkenlerin korelasyon matrisi yer almaktadır.



Şekil 4.4. Korelasyon Matrisi

Şekil 4.4.'ye göre değişkenler arasından pozitif yönlü ilişki olduğu yorumu yapılabilir. Koyu renkler birbiri ile ilişkisi olan değişkenleri ifade eder. 1'e en yakın olan değerler koyu renkten açık renge doğru azalmaktadır.

Ön hazırlığı tamamlanan verilere algoritmalar uygulanarak analiz yapılmıştır. Analiz sonuçları Tablo 4.6.'da gösterilmiştir.

Tablo 4.6. Analiz Sonuçları

		%75-%25	%70-%30
	Algoritma	**Doğruluk Oranı**	**Doğruluk Oranı**
0	Karar Ağacı(Decision Tree)	74.59%	73.54%
1	Lojistik Regresyon(Logistic Regression)	57.58%	57.84%
2	Naive Bayes	47.95%	48.12%
3	K-nn (K-Nearest Neihbors)	65.36%	74.23%
4	Rastgele Orman (Random Forest)	78.27%	77.47%
5	Yapay Sinir Ağları (Neural Network)	57.99%	58.19%

Yukarıda ki tabloyu incelediğimizde müşteri kayıp durumunu en iyi analiz eden algoritma Rastgele Orman(Random Forest) algoritmasıdır. Çapraz doğrulama(Cross validation) yöntemi kullanılarak uygulama tekrar yapılmış ve analiz sonuçları Tablo 4.7.'de gösterilmiştir. Çapraz doğrulama için 10 katlı olacak şekilde gerçekleştirilmiştir.

Tablo 4.7. İkinci Analiz Sonuçları

		%75-%25	%70-%30
	Algoritma	**Doğruluk Oranı**	**Doğruluk Oranı**
0	Lojistik Regresyon(Logistic Regression)	%57.09	%57.09
1	K-nn (K-Nearest Neihbors)	%64.32	%64.32
2	Karar Ağacı(Decision Tree)	%73.55	%73.55
3	Rastgele Orman (Random Forest)	%76.42	%76.32
4	Naive bayes	%47.60	%47.60
5	Yapay Sinir Ağları (Neural Network)	%57.04	%57.04

Elde edilen sonuçlar incelendiğinde, karar ağaçları ve rastgele orman algoritmalarının görece yüksek sonuç verdiği görülmektedir. Bu bağlamda her iki algoritmada kullanılan bağımsız değişkenlerin özellikleri ve seçilen kriter yöntemleri bir sonra ki bölümde detaylandırılmıştır.

4.2. Etkili Sonuç Veren Algoritmaların İncelenmesi

4.2.1. Rastgele Orman(Random Forest) algoritması parametreleri

- n_estimators: Oluşturulmak istenen ağaç sayısıdır. Sayı ne kadar büyürse çalışma hızı o kadar yavaşlar. Default değer 10'dur.
- min_sample_leaf : Her ağacın son düğümünün minimum boyutunu belirlememizi sağlar. Varsayılan değer 1'dir.

- min_sample_split: Bölünme için gereken minimum örnek sayısıdır. Varsayılan değer 2'dir.
- max_depth: ağacın maksimum derinliğini belirtir. Varsayılan değer None'dur.
- bootstrap Var olan veri setindeki örneklerinin kullanılıp kullanılmadığı kontrol edilir.
- class_weight: İlişkili ağırlıkları belirlenir.
- criterion: Bölünmede kullanılan ölçüttür. Gini ve entropi değerleri kullanılmaktadır.
- Varsayılan değer gini'dir.
- max_features: Düğüm ayrılırken dikkate alınacak maksimum özellik sayısıdır. Varsayılan değer auto'dur.
- max_leaf_nodes: Maksimum yaprak sayısını ifade eder. Default değer None'dur.
- min_impurity_decrease: Kirlilik değerlerini ifade eder. Varsayılan değer 0'dır.
- min_weight_fraction_leaf: Yaprak düğümde olması gereken ağırlıklar toplamının minimum ağırlıklı değerini belirlemede kullanılır. Varsayılan değer 0'dır.
- n_jobs Paralel çalıştırılan iş sayısını gösterir. Varsayılan 1'dir.
- oob_score: Çapraz doğrulama yöntemizdir.
- verbose: Ağacın ayrıntı düzeyini ayarlar. Varsayılan değer 0'dır.
- warm_start: Önceki uygun çözümü yeniden kullanmak yerine, yeni bir orman oluşturulmasını sağlar. Varsayılan değer false'tur (sklearn.ensemble.RandomForestClassifier, n.d.).

Çalışmada tüm özellikler için varsayılan değerler kullanılmıştır.

4.2.2. Karar Ağaçları (Decision Tree) algoritması parametreleri

- criterion: Bölünmenin kalitesini ölçen işlevdir. Varsayılan değeri gini'dir.
- splitter: Her düğümde bölmeyi seçmek için kullanılan yöntemdir. Varsayılan değer best'tir.

- max_depth: Ağacın maximum derinliğidir. Varsayılan değer none'dur.
- min_samples_split: Bir düğümü bölmek için gereken minumum örnek sayısıdır. Varsayılan 2'dir.
- min_samples_leaf: Her ağacın son düğümünün minimum boyutunu belirlememizi sağlar. Varsayılan değer 1'dir.
- min_weight_fraction_leaf: Yaprak düğümde olması gereken ağırlıklar toplamının minimum ağırlıklı değerini belirlemede kullanılır. Varsayılan değer 0'dır.
- max_features: En iyi bölünmeyi ararken göz önünde bulundurulması gereken özelliklerin sayısıdır. Varsayılan none'dır.
- random_state: Tahminleyicinin rastgeleliğini kontrol eder. Varsayılan None'dır.
- max_leaf_nodes: Maksimum yaprak sayısını ifade eder. Default değer None'dur.
- min_impurity_decrease: Kirlilik değerlerini ifade eder. Varsayılan değer 0'dır.
- class_weight: ilişkili ağırlıkları belirlenir (sklearn.tree.DecisionTreeClassifier, n.d.).

Çalışmada tüm özellikler için varsayılan değerler kullanılmıştır.

BÖLÜM 5. ARAŞTIRMA BULGULARI VE SONUÇLAR

Bu bölümde değerlendirme ölçütleriyle birlikte uygulamanın sonuçlarından bahsedilmektedir. Makine öğrenmesi çalışmalarında modellerin performanslarını ölçmek için bazı ölçütler yer almaktadır. Ölçütlerin hesaplanmasında karışıklık(Confusion Matrix) kullanılmaktadır (Akyiğit, 2021). Tablo 5.1'de karışıklık matrisinin gösterimi bulunmaktadır. Satırlarda yer alan değerler veri setimizde ki gerçek değerler, sütunlarda yer alan değerler ise çalışma sonrası tahmin ettiğimiz değerleri içerir.

Aynı zamanda doğruluk oranı %70' in üzerinde olan algoritmaların önemli özelliklerinin sıralanması için bir çalışma yapılmış ve sonuçlar Şekil 5.1. ve Şekil 5.2.'de gösterilmiştir.

Tablo 5.1. Karışıklık Matrisi (Confusion Matrix)

		Tahmin Edilen Sınıf(Predicted Class)	
		Sınıf=1	Sınıf=2
Gerçek Sınıf Değeri (Actual Class)	Sınıf = 1	True positive(TP)	False Negative (FN)
	Sınıf = 2	False Positive(FP)	True Negative(TN)

Doğruluk (Accuracy): Doğru tahmin edilen verilen, veri setindeki tüm örnek verilerin sayısına oranı olarak bulunur. Denklem 5.1'de gösterilir.

$$\text{Doğruluk} = \frac{TP+TN}{TP+TN+FP+FN} \tag{5.1}$$

5.1. Araştırma Bulguları

5.1.1. Karar Ağacı(Decision Tree)

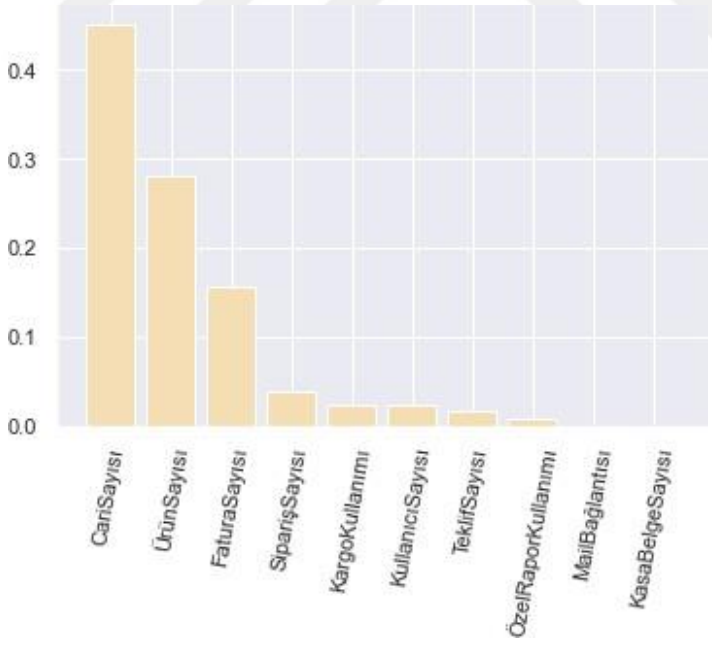
Tablo 5.2. Karar Ağacı Algoritması Karışıklık Matrisi

		Tahmin Edilen Sınıf(Predicted Class)	
		0	**1**
Müşteri Durumu (Churn)	**Kayıp 0**	172	35
	1	89	192

5.1 denklemine göre hesaplama yaptığımızda elde ettiğimi sonuç aşağıdaki gibi olmaktadır.

$$\text{Doğruluk} = \frac{172+192}{172+35+89+192} = \% 74.59$$

Karar ağaçları algoritmasında özniteliklerin önem sırası Şekil 5.1.'de gösterilmiştir.



Şekil 5.1. Karar Ağacı Algoritmasına Göre Önemli Özelliklerin Sıralanması

5.1.2. Rastgele Orman(Random Forest)

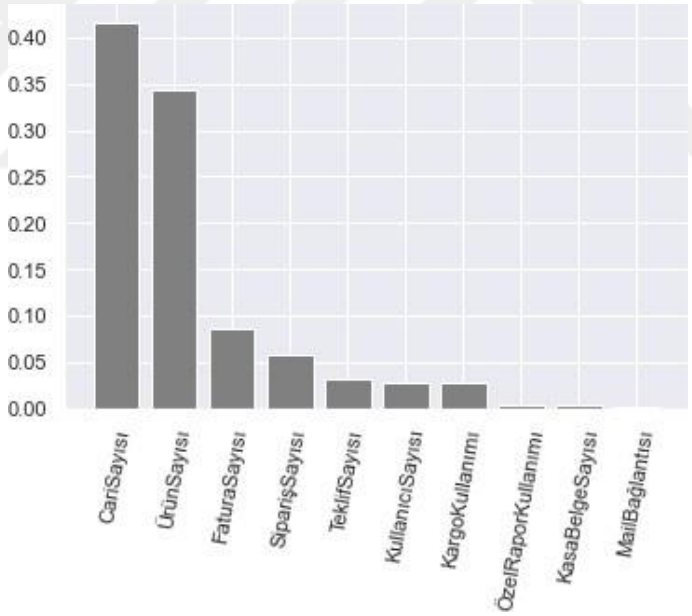
Tablo 5.3. Rastgele Orman Algoritması Karışıklık Matrisi

		Tahmin Edilen Sınıf(Predicted Class)	
		0	**1**
Müşteri Durumu (Churn)	**Kayıp 0**	173	39
	1	67	209

5.1 denklemine göre hesaplama yaptığımızda elde ettiğimi sonuç aşağıdaki gibi olmaktadır.

$$\text{Doğruluk} = \frac{173+209}{173+39+67+209} = \%78.27$$

Rastgele Orman algoritmasında özniteliklerin önem sırası Şekil 5.2.'de gösterilmiştir.



Şekil 5.2. Rastgele Orman Algoritmasına Göre Önemli Özelliklerin Sıralanması

5.2. Sonuçlar Ve Öneriler

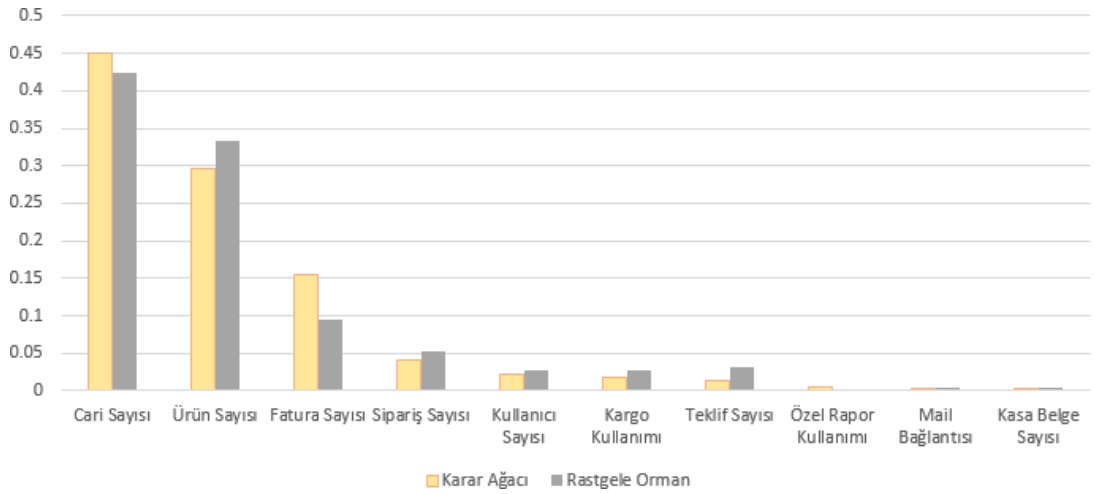
Rekabet gücünün arttığı ve hızlı büyüyen sektörlerde devamlılığın sağlanması için sadık müşterilerin artırılması önemli bir konudur. Yeni müşteri kazanmak için harcanan maliyet mevcut müşteriyi elde tutmak için harcanan maliyetten çok daha

yüksektir ve bu açıdan mevcut müşterilerin davranışları incelenerek, firmayı bırakma ihtimali olan müşteriler belirlenip memnuniyet artırmaya yönelik çalışmaların yapılması gereklidir.

Bu çalışmada 16 adet değişken için özellik seçimi yapılmış ve yazılım sektöründe, kullandıkları hizmeti bırakma olasılığı olan müşterileri tahmin etmek için 6 farklı makine öğrenmesi algoritması kullanılmıştır. Bu algoritmalar Tablo 3.1.'de ki literatür özetinde kullanılan makine algoritmalarından en sık kullanılanlar olarak belirlenmiştir. Toplamda 1951 müşteri verisi incelenerek çalışma tamamlanmış.

Çalışmada veriler %75 ve %70 eğitim verisi olacak şekilde iki kez gruplandırılmış en iyi sonucu analiz sonucu veren grup %75 olduğundan çalışmaya bu grup ile devam edilmiştir. Çalışma analiz yapıldıktan sonra veriler optimize (cross validation) edilerek tekrarlanmış ve Tablo 4.5. ve 4.6.'da gözüktüğü gibi her iki analiz sonunda da Rastgele Orman(Random Forest) algoritması başarılı olmuştur.

Tablo 3.1. ile karşılaştırıldığında özelliklerden kullanıcı sayısı, kullanım metrikleri benzerlik göstermektedir. Çalışmamızda %70 başarının üstünde sonuç gösteren algoritmaların önemli özelliklerinin sıralaması da Şekil 5.1. ve Şekil 5.2.'de gösterilmiştir. Özelliklerin karşılaştırılması Şekil 5.3.'de paylaşılmıştır.



Şekil 5.3. Öznitelik önemlerinin karşılaştırılması

Tablo 3.1. ile karşılaştırıldığında başarılı sonuç veren algoritmaların, çalışmada başarılı sonuç veren algoritmalarla birçoğunun aynı olduğu gözlemlenmiştir. Rastgele Orman algoritması literatürdeki çalışmalarda olduğu gibi bu çalışmada da başarılı olmuştur.

Çalışmada Orange yazılımı ve Python programlama dili kullanılarak tüm süreçler gerçekleştirilmiştir. Python programlama dili ve kütüphaneleri ile başarılı sonuçlar elde edildiği gözlemlenmiştir.

Sektörel olarak analiz sonuçları değerlendirildiğinde, bir müşterinin temel iş süreçlerinde kullanmış olduğu özelliklerin ve kullanıcı sayısı özelliğinin müşterinin kayıp veya devam durumuyla anlamlı ilişkisinin olduğu gözlemlenmiştir. Temel olmayan özelliklerin kullanımının aslında müşterin devam durumuyla ilişkisi yoktur. Yapılan analizler sonucunda yazılım sektörü için öncelikli özellikler belirlenmiş ve bu özelliklerle yapılan analizler sonucunda firmayı terk etme ihtimali olan müşteriler tahmin edilmiştir. Müşteri kaybı analizi ile kaybın minimalize edilmesi ve müşteri memnuniyetinin arttırılması amaçlanmıştır.

Gelecekte ki çalışmalarda kullanılan öz niteliklerin sayısı ve veri sayısı arttırılarak daha etkili sonuç veren analizler yapılabilir. Farklı tahminleme algoritmaları kullanılarak performans karşılaştırması yapılabilir.

KAYNAKLAR

Akyiğit, H. E. (2021). Sigortacılık sektöründe makine öğrenmesi ile müşteri kaybı analizi. Sakarya Üniversitesi Fen Bilimleri Enstitüsü.

Amornvetchayakul, P., & Phumchusri, N. (2020). Customer Churn Prediction for a Software-as-a-Service Inventory Management Software Company: A Case Study in Thailand. 2020 IEEE 7th International Conference on Industrial Engineering and Applications, ICIEA 2020, 514–518. https://doi.org/10.1109/ICIEA49774.2020.9102099, Erişim Tarihi: 01.06.2022.

Burez, J., & Van den Poel, D. (2009). Handling class imbalance in customer churn prediction. Expert Systems with Applications, 36(3 PART 1), 4626–4636. https://doi.org/10.1016/j.eswa.2008.05.027, Erişim Tarihi: 01.06.2022.

Çelik, M. (2009). Veri Madenciliğinde Kullanılan Sınıflandırma Yöntemleri ve Bir Uygulama.

Çelik, U. (2018). Veri Madenciliği İle Müşteri Kaybı Analizi. International Conference on Applied Economics and Finance, 1(February), 185–204.

Dahiya, K., & Bhatia, S. (2015). Customer churn analysis in telecom industry. 2015 4th International Conference on Reliability, Infocom Technologies and Optimization: Trends and Future Directions, ICRITO 2015, 1–6. https://doi.org/10.1109/ICRITO.2015.7359318, Erişim Tarihi: 01.06.2022.

Demir, M. (2021). Özellik Seçim Yöntemleri Kullanılarak Sınıflandırma Algoritmalarının Performanslarının Karşılaştırılması. 6.

Emhan, Ö., & Akın, M. (2019). Filtreleme Tabanlı Öznitelik Seçme Yöntemlerinin Anomali Tabanlı Ağ Saldırısı Tespit Sistemlerine Etkisi. DÜMF Mühendislik Dergisi, 10(2), 549–559. https://doi.org/10.24012/dumf.565842, Erişim Tarihi: 01.06.2022.

Frank, B., & Pittges, J. (2009). Analyzing Customer Churn in the Software as a Service (SaaS) Industry. Southeastern InfORMS Conference Proceedings, 481–488.

Ge, Y., He, S., Xiong, J., & Brown, D. E. (2017). Customer Churn Analysis for a Software-as-a-service Company. 106–111.

Gürsoy, Ş. U. T. (2009). Customer churn analysis in telecommunication sector. İstanbul Üniversitesi İşletme Fakültesi Dergisi, 39(1), 35–49.

Idris, A., Rizwan, M., & Khan, A. (2012). Churn prediction in telecom using Random Forest and PSO based data balancing in combination with various feature selection strategies. Computers and Electrical Engineering, 38(6), 1808–1819. https://doi.org/10.1016/j.compeleceng.2012.09.001, Erişim Tarihi: 01.06.2022.

Kaptan, F. (2019). Müşteri Kayıp Analizi: Hava Yolu Sektöründe Bir Uygulama. İstanbul Teknik Üniversitesi Fen Bilimleri Enstitüsü, 8(5), 55.

Kaynar, O., Tuna, M. F., Görmez, Y., & Deveci, M. A. (2017). Makine Öğrenmesi Yöntemleriyle Müşteri Kaybı Analizi. C.Ü. İktisadi ve İdari Bilimler Dergisi, 18(1), 1–14.

Keramati, A., Jafari-Marandi, R., Aliannejadi, M., Ahmadian, I., Mozaffari, M., & Abbasi, U. (2014). Improved churn prediction in telecommunication industry using data mining techniques. Applied Soft Computing Journal, 24, 994–1012. https://doi.org/10.1016/j.asoc.2014.08.041, Erişim Tarihi: 01.06.2022.

Mutanen, T. (2006). Customer churn analysis: a case study: research report. Technical Research Centre of Finland (VTT), 1–19. http://www.vtt.fi/inf/julkaisut/muut/2006/customer_churn_case_study.pdf, Erişim Tarihi: 01.06.2022.

Nath, S. V. (2014). Customer Churn Analysis in the Wireless Industry: A Data Mining Approach. Florida Atlantic University, July.

Orange3 for Windows, Version 3.25. (2020).

Ozcan, T., Önay Koçoğlu, F., & Baray, Ş. A. (2016). Veri Madenciliğinde Ayrılan Müşteri Analizi Problemi Üzerine Bir Literatür Araştırması. Uluslararası Katılımlı 16. Üretim Araştırmaları Sempozyumu, February 2019.

Rautio, A. (2019). Churn Prediction In SaaS Using. May.

Seker, S. E. (2016). Müşteri Kayıp Analizi (Customer Churn Analysis). YBS Ansiklopedi, 1–4.

Şekeroğlu, S. (2010). Hizmet Sektöründe Bir Veri Madenciliği Uygulaması. Interagir: Pensando a Extensão, 0(15), 1–9. https://www.golder.com/insights/block-caving-a-viable-alternative/, Erişim Tarihi: 01.06.2022.

https://scikitlearn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html, Erişim Tarihi: 01.06.2022.

https://scikitlearn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html, Erişim Tarihi: 01.06.2022.

Sukow, A. E. R., & Grant, R. (2013). Forecasting and the Role of Churn in Software-as-a-Service Business Models. IBusiness, 05(01), 49–57. https://doi.org/10.4236/ib.2013.51a006, Erişim Tarihi: 01.06.2022.

Ullah, I., Raza, B., Malik, A. K., Imran, M., Islam, S. U., & Kim, S. W. (2019). A Churn Prediction Model Using Random Forest: Analysis of Machine Learning Techniques for Churn Prediction and Factor Identification in Telecom Sector. IEEE Access, 7, 60134–60149. https://doi.org/10.1109/ACCESS.2019.2914999, Erişim Tarihi: 01.06.2022.

Verbeke, W., Martens, D., & Baesens, B. (2014). Social network analysis for customer churn prediction. Applied Soft Computing Journal, 14(PART C), 431–446. https://doi.org/10.1016/j.asoc.2013.09.017, Erişim Tarihi: 01.06.2022.

Xie, Y., Li, X., Ngai, E. W. T., & Ying, W. (2009). Customer churn prediction using improved balanced random forests. Expert Systems with Applications, 36(3 PART 1), 5445–5449. https://doi.org/10.1016/j.eswa.2008.06.121. Erişim Tarihi: 01.06.2022.

Yazıcı, B., Yaslı, F., Gürleyik, H. Y., & Turgut, U. O. (n.d.). Veri Madenciliğinde Özellik Seçim Tekniklerinin Bankacılık Verisine Uygulanması Üzerine Araştırma ve Karşılaştırmalı Uygulama. 72–83.

ÖZGEÇMİŞ

Adı Soyadı **:** Sena KASIM

ÖĞRENİM DURUMU

Derece	**Eğitim Birimi**	**Mezuniyet Yılı**
Yüksek Lisans	Sakarya Üniversitesi / Fen Bilimleri Enstitüsü / Bilişim Sistemleri Mühendisliği	Devam Ediyor
Lisans	Sakarya Üniversitesi / Bilgisayar ve Bilişim Bilimleri Fakültesi / Bilgisayar Mühendisliği	2016
Lise	Şehit Üsteğmen Selçuk Esedoğlu Anadolu Lisesi	2012

İŞ DENEYİMİ

Yıl	**Yer**	**Görev**
2020-Halen	Edit Ar-Ge Yazılım(reybex)	Erp Danışmanı
2017-2020	Bym Yazılım Eğitim ve Danışmanlık	Yazılım Geliştirme

YABANCI DİL

İngilizce

ESERLER (makale, bildiri, proje vb.)

1. Konfeksiyon Sektöründe FP-Growth Algoritmasıyla Sepet Analizi (Erkek Giyim Örneği),
Uluslararası Marmara Fen ve Sosyal Bilimler Kongresi bildiri-2020,
2. Veri Madenciliği Yöntemleriyle Müşteri Kayıp Analizi: Yazılım Sektörü İçin Öncelikli Özelliklerin Belirlenmesi, Uluslararası Marmara Fen ve Sosyal Bilimler Kongresi bildiri-2022,