T.C. İSTANBUL TİCARET ÜNİVERSİTESİ

FEN BİLİMLERİ ENSTİTÜSÜ

TELEKOMÜNİKASYON SEKTÖRÜ İÇİN VERİ MADENCİLİĞİ VE MAKİNE ÖĞRENMESİ TEKNİKLERİ İLE AYRILAN MÜŞTERİ ANALİZİ

Furkan UYANIK

Danışman
Doç. Dr. Mustafa Cem KASAPBAŞI

YÜKSEK LİSANS TEZİ
BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI
İSTANBUL - 2021

KABUL VE ONAY SAYFASI

Furkan UYANIK tarafından hazırlanan **"Telekomünikasyon Sektörü için Veri Madenciliği ve Makine Öğrenmesi Teknikleri ile Ayrılan Müşteri Analizi"** adlı tez çalışması 06/07/2021 tarihinde aşağıdaki jüri üyeleri önünde başarı ile savunularak, İstanbul Ticaret Üniversitesi Fen Bilimleri Enstitüsü **Bilgisayar Mühendisliği Anabilim Dalı**'nda **YÜKSEK LİSANS TEZİ** olarak kabul edilmiştir.

Danışman	**Doç. Dr. Mustafa Cem KASAPBAŞI** İstanbul Ticaret Üniversitesi	.....
Jüri Üyesi	**Dr. Öğr. Üyesi Arzu KAKIŞIM** İstanbul Ticaret Üniversitesi	.....
Jüri Üyesi	**Dr. Öğr. Üyesi Buket DOĞAN** Marmara Üniversitesi	.....

Onay Tarihi: 29.07.2021

İstanbul Ticaret Üniversitesi, Fen Bilimleri Enstitüsünün 29.07.2021 tarih ve 2021/317 numaralı Yönetim Kurulu Kararının 2. maddesi gereğince, ders yüklerini ve tez yükümlülüğünü yerine getirdiği belirlenen "FURKAN UYANIK" adlı öğrencinin mezun olmasına oy birliği ile karar verilmiştir.

Prof. Dr. Necip ŞİMŞEK
Enstitü Müdürü

AKADEMİK VE ETİK KURALLARA UYGUNLUK BEYANI

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

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

beyan ederim.

29.07.2021

Furkan UYANIK

İÇİNDEKİLER

	Sayfa
İÇİNDEKİLER	i
ÖZET	iii
ABSTRACT	iv
TEŞEKKÜR	v
ŞEKİLLER DİZİNİ	vi
ÇİZELGELER DİZİNİ	vii
SİMGELER VE KISALTMALAR DİZİNİ	viii
1. GİRİŞ	1
2. LİTERATÜR ÖZETİ	2
3. AYRILAN MÜŞTERİ ANALİZİ	6
3.1. Telekomünikasyon Sektöründe Ayrılan Müşteri Analizi	7
3.2. Farklı Sektörlerde Ayrılan Müşteri Analizi	7
4. VERİ MADENCİLİĞİ	8
4.1. Veri Madenciliği Süreci	9
4.1.1. Problemin tanımlanması	9
4.1.2. Verilerin hazırlanması	9
4.1.3. Modelin kurulması ve değerlendirilmesi	10
4.1.4. Modelin kullanılması	10
4.1.5. Modelin izlenmesi	10
4.2. Veri Madenciliğinde Kullanılan Yöntemler	11
4.2.1. Öznitelik indirgeme (Attributes Reduction)	11
4.2.2. Veri normalizasyonu	11
4.2.2.1. Min-max normalizasyon yöntemi	12
4.2.3. Veri kodlama	12
4.2.3.1. Etiket kodlama (Label Encoding) yöntemi	12
4.2.3.2. One-Hot Encoding yöntemi	13
4.2.4. Öznitelik seçimi	14
4.2.4.1. Pearson Correlation Coefficient yöntemi	14
4.2.4.2. Univariate Feature Selection (Tek Değişkenli Öznitelik seçimi) yöntemi	15
4.2.5. Veri aşırı örnekleme	15
4.2.5.1. SMOTE (Synthetic Minority Oversampling Technique)	16
4.2.5.2. ADASYN (Adaptive Synthetic Sampling Method)	18
5. MAKİNE ÖĞRENMESİ	20
5.1. Makine Öğrenmesi Yöntemleri	21
5.1.1. Lojistik Regresyon (Logistic Regression) yöntemi	21
5.1.2. Karar ağacı (Decision Tree) yöntemi	23
5.1.3. Yapay sinir ağları (Artificial Neural Network) yöntemi	23
5.2. Makine öğrenmesinde topluluk öğrenmesi yöntemleri	25
5.2.1. Torbalama (Bagging) yöntemi	26
5.2.2. Arttırma (Boosting) yöntemi	26
5.2.2.1. Gradient Boosting yöntemi	28
5.2.2.2. Adaptive Boosting (AdaBoost) yöntemi	28
5.2.2.3. Extreme Gradient Boosting (XGBoost) yöntemi	28
5.2.2.4. Light Gradient Boosting Machine (LightGBM) yöntemi	29
5.2.2.5. CatBoost yöntemi	30

5.2.3. Rastgele orman (Random Forest) yöntemi 30
5.3. Değerlendirme Ölçütleri 31
5.3.1. Doğruluk oranı (Accuracy Rate) 32
5.3.2. Geri çağırma (Recall) 32
5.3.3. Hassasiyet (Precision) 33
5.3.4. Özgünlük (Specificity) 33
5.3.5. Dengelenmiş doğruluk oranı (Balanced Accuracy Rate) 33
5.3.6. F1 Skoru (F1 Score) 33
5.3.7. ROC eğrisinin altında kalan alan değeri (ROC-AUC) 34
6. UYGULAMA 35
6.1. Geliştirme Ortamı 35
6.2. Veri Kümesi 35
6.3. Veri Ön İşleme 41
6.4. Önerilen Modeller 42
6.4.1. SMOTE ile önerilen tahmin modeli 43
6.4.2. ADASYN ile önerilen tahmin modeli 44
7. BULGULAR VE TARTIŞMA 46
7.1. Aşırı Örnekleme Yöntemi Kullanmadan Eğitilen Tahmin Modelleri 46
7.2. Aşırı Örnekleme Yöntemi Kullanarak Eğitilen Tahmin Modelleri 47
8. SONUÇ VE ÖNERİLER 51
KAYNAKLAR 53
ÖZGEÇMİŞ 57

ÖZET

Yüksek Lisans Tezi

TELEKOMÜNİKASYON SEKTÖRÜ İÇİN VERİ MADENCİLİĞİ VE MAKİNE ÖĞRENMESİ TEKNİKLERİ İLE AYRILAN MÜŞTERİ ANALİZİ

Furkan UYANIK

İstanbul Ticaret Üniversitesi
Fen Bilimleri Enstitüsü
Bilgisayar Mühendisliği Anabilim Dalı

Danışman: Doç. Dr. Mustafa Cem KASAPBAŞI
2021, 57 sayfa

Son yıllarda şirketler arası rekabetin artmasıyla beraber aboneliğinden ayrılacak müşterilerin tahmin edilmesi oldukça önemli hale gelmiştir. Ayrılan müşteri analizi, veri madenciliği, makine öğrenmesi ve derin öğrenme gibi yapay zekâ alanlarında sıklıkla karşılaşılan analiz çeşitlerinden biridir. Özellikle telekomünikasyon, sigortacılık ve bankacılık gibi sektörlerde yaygın olarak kullanılmaktadır. Bu çalışma da veri madenciliği ve makine öğrenmesi teknikleri ile aboneliğini sonlandırma ihtimali olan müşterileri tahmin etmeyi amaçlamaktadır. Bu çalışma Lojistik Regresyon, Karar Ağacı, Yapay Sinir Ağları, Torbalama (Bagging) ve Artırma (Boosting) sınıflandırma modelleri kullanılarak arasından en iyi sonucu bulmayı önermiştir. Veri setinde sınıf dengesizliği olduğu için SMOTE (Synthetic Minority Oversampling Technique) ve ADASYN (Adaptive Synthetic Sampling Method) tekniği ile örnekleme yapılmıştır. Bu çalışmada, 2 adet tahmin modeli önerilmiştir ve tahmin modelleri Veri Seti, Veri Ön İşleme, Veri Örnekleme, Değerlendirme olarak 4 farklı aşamadan oluşmaktadır. Veri Ön İşleme aşamasında, kullanılmayan ve önemsiz özniteliklerin veri setinden çıkartılması, normalizasyon, şifreleme ve aşırı örnekleme gibi birçok yöntem kullanılmıştır. Performans ölçütü olarak Doğruluk Oranı, Geri Çağırma (Recall), Hassasiyet (Precision) ve Özgünlük (Specificity), Dengelenmiş Doğruluk Oranı gibi birçok değer kullanılmıştır. Performans ölçütlerine göre en iyi tahmin modeli, ADASYN örnekleme yöntemi kullanılan model olmuştur. Sınıflandırma yöntemi olarak en iyi sonucu veren LightGBM (Light Gradient Boosting Machine) tekniği olmuştur. Önerilen modeller arasında Veri Ön İşleme ve Veri Örnekleme aşamalarında farklılıklar bulunmaktadır. Bu çalışmada önerilen tahmin modellerinin eğitim süresi, benzer çalışmalara göre daha iyi performans sağladığı tespit edilmiştir. Ayrıca bu çalışmada, sadece 58 öznitelik kullanarak 172 öznitelik kullanan benzer çalışmaların başardığına çok yakın sonuçlar elde edilmiştir.

Anahtar Kelimeler: Ayrılan müşteri analizi, makine öğrenmesi, müşteri karmaşası tahmini, örnekleme algoritmaları, sınıflandırma, tahmin, telekomünikasyon, topluluk sınıflandırması, veri madenciliği.

ABSTRACT

M.Sc. Thesis

CHURN ANALYSIS FOR TELECOMMUNICATION SECTOR WITH DATA MINING AND MACHINE LEARNING

Furkan UYANIK

İstanbul Commerce University
Graduate School of Applied and Natural Sciences
Department of Computer Engineering

Supervisor: Assoc. Prof. Dr. Mustafa Cem KASAPBAŞI
2021, 57 pages

With the increasing competition among companies in recent years, it has become very important to estimate the customers who are churned. Churn is one of the most common types of analysis, especially in areas such as data mining, machine learning and deep learning. It is widely used in sectors such as telecommunications, insurance and banking. In this study, it purpose to predict customers who may end their subscription with data mining and machine learning techniques. This study proposed to find the best result from using Logistic Regression, Decision Tree, Artificial Neural Network, Bagging and Boosting classification models. For the data set was unstable, sampling was performed using SMOTE (Synthetic Minority Oversampling Technique) and ADASYN (Adaptive Synthetic Sampling Method) technique. In the study, 2 prediction models are proposed and the proposed prediction models consist of 4 different phases as Data Set, Data Pre-Processing, Data Sampling and Evaluation. In the Data Pre-Processing phase, many methods were used, such as removing unused and unimportant features from the data set, normalization, encoding and oversampling. Accuracy Rate, Recall, Precision and Specificity, Balanced Accuracy Rate and Area Under the ROC Curve (ROC-AUC) value were used as performance measures. Considering the performance measures, the best prediction model suggested was the model using ADASYN sampling method. As the classification method, the best success was the LightGBM (Light Gradient Boosting Machine) technique. There are differences in the Data Pre-Processing and Data Sampling stages phases the proposed models. It was determined that the prediction models proposed in this study provide better performance than similar studies. Also, in this study, results very close to those achieved by similar studies using 172 features using only 58 features were obtained.

Keywords: Churn analysis, churn prediction, classification, data mining, ensemble classification, machine learning, oversampling algorithms, telecommunication.

TEŞEKKÜR

Yüksek Lisans öğrenimim boyunca yardımlarını, fikirlerini ve tecrübelerini benden esirgemeyen değerli danışman hocam Doç. Dr. Mustafa Cem KASAPBAŞI'ya, beni bu günlere kadar getiren varlığımı borçlu olduğum anneme ve babama, bugüne kadar desteğini hiçbir zaman eksik etmeyen daima yanımda olan biricik eşime teşekkür ederim.

Ayrıca bu çalışmayı desteklediği ve finanse ettiği için TTG International Ltd.'e müteşekkirim ve veri akışı mimarisinde yardımcı olan uzmanlara minnettarım.

TTG International Ltd., devlet kurumlarına ve mobil ağ operatörü şirketlerine OSS ürün tedarikçisidir. TTG International Ltd., araştırma çalışmalarını desteklemek ve aynı zamanda Ar-Ge çalışmalarına katılım yoluyla çalışanların yenilikçiliğini teşvik etmek için çeşitli ülkelerde etkin bir şekilde faaliyet göstermektedir.

Furkan UYANIK
İSTANBUL, 2021

ŞEKİLLER

	Sayfa
Şekil 4.1. Veri madenciliğinin süreci ile bilginin keşfedilmesi	9
Şekil 5.1. Yapay Zeka ve alt alanları	21
Şekil 5.2. Karar Ağacı (Decision Tree)	23
Şekil 5.3. Yapay Sinir Ağı	24
Şekil 5.4. Bagging ve Boosting yöntemlerinin karşılaştırılması	27
Şekil 5.5. Karışıklık Matrisi	31
Şekil 5.6. ROC eğrisinin altında kalan alan değer grafiği	34
Şekil 6.1. Aldığı Hizmetin Aboneliğinden Ayrılan & Ayrılmayan Müşteri Oranı	36
Şekil 6.2. Univariate Feature Selection yöntemi ile 15 en seçici Öznitelikler	36
Şekil 6.3. Pearson Correlation Coefficient yöntemi ile özniteliklerin birbiriyle olan ilişkisi	37
Şekil 6.4. Telefonunun yenilenmesi göre ayrılan müşteri dağılımları	38
Şekil 6.5. İnternet erişimine göre ayrılan müşteri dağılımları	39
Şekil 6.6. Kredi notuna göre ayrılan müşteri dağılımları	39
Şekil 6.7. Paket aşımı miktarına göre ayrılan müşteri dağılımları	40
Şekil 6.8. Telefonunun yenilenmesi göre ayrılan müşteri dağılımları	40
Şekil 6.9. SMOTE tekniği ile Önerilen Tahmin Modeli	44
Şekil 6.10. ADASYN tekniği ile Önerilen Tahmin Modeli	45

ÇİZELGELER

Sayfa

Çizelge 4.1. Kategorik değerlerin Label Encoding yöntemi ile kodlanması.. 13
Çizelge 4.2. Kategorik değerlerin One-Hot Encoding yöntemi ile kodlanması..... 13
Çizelge 4.3. Label Encoding ve One-Hot Encoding yöntemlerinin karşılaştırılması 14
Çizelge 7.1. Aşırı Örnekleme Yöntemi Kullanmadan Eğitilen Tahmin Modelleri..... 47
Çizelge 7.2. Birinci önerilen tahmin modeline ait performans çıktıları 48
Çizelge 7.3. İkinci önerilen tahmin modeline ait performans çıktıları 48
Çizelge 7.4. Birinci tahmin modelinin performans çıktıları 49
Çizelge 7.5. İkinci tahmin modelinin performans çıktıları 50
Çizelge 7.6. Benzer çalışmalar ile karşılaştırılması 50

SİMGELER VE KISALTMALAR

ADASYN	Adaptive Synthetic Sampling Method
AI	Yapay Zekâ (Artificial Intelligence)
ANN	Yapay Sinir Ağları (Artificial Neural Network)
AUC	Receiver Operating Characteristics
AdaBoost	Uyarlanabilir Arttırma (Adaptive Boosting)
BS	Geriye Doğru Arama (Backward Search)
CFS	Korelasyon Öznitelik Seçimi (Correlation Feature Selection)
CPU	Merkezi İşlem Birimi (Central Process Unit)
DT	Karar Ağacı (Decision Tree)
FS	Öznitelik Seçimi (Feature Selection)
GBM	Gradyan Arttırma Makinesi (Gradient Boosting Machine)
GPU	Grafik İşlemci Birimi (Graphics Processing Unit)
IG	Bilgi Kazancı (Information Gain)
KNN	K-En Yakın Komşu (K-Nearest Neighbors)
LightGBM	Hafif Gradyan Arttırma Makinesi (Gradient Boosting Machine)
NB	Naive Bayes
OLS	Ordinary Least Squared
ROC	Eğri Altında Kalan Alan (Area Under The Curve)
RSFS	Kaba Ayar Özelliği Seçimi (Rough Set Feature Selection)
SMOTE	Synthetic Minority Oversampling Technique
SVM	Destek Vektör Makinesi (Support Vector Machine)
XGBoost	Aşırı Gradyan Arttırma (Extreme Gradient Boosting)

1. GİRİŞ

Günümüzde telekomünikasyon, sigortacılık ve bankacılık gibi birçok sektörde önemli düzeyde müşteri sirkülasyonu bulunmaktadır. Bu yüzden şirketler arasında rekabet ortamı oluşmaktadır. Rekabet ortamındaki şirketler ise müşteri kayıplarını en aza indirgemek istemektedir. Buna çözüm bulmak için kullanılan yöntemlerden biri de müşterilerin isteğe bağlı veya istemsiz olarak aboneliğinin sonlandırmasını tahmin etmektir. Söz konusu analizin adı Ayrılan Müşteri Analizi (Churn Analysis) olarak geçmektedir (Gold, 2020).

Ayrılan müşteri analizi terimi birçok sektörde bulunduğu için her sektör farklı tanımlama yapabilmektedir. Genel bir deyişle, mevcut bir kullanıcının veya abonenin aldığı hizmeti sonlandırmasıdır. Ayrılan Müşteri Analizi (Churn Analysis) ile ayrılacak müşteriyi izleyerek elinde tutmak, aboneliğinden ayrılma sebebini tespit etmek gibi birçok analiz çıktısı elde edilebilmektedir. Ayrılan müşteri analizini gerçekleştirmek için müşterinin aldığı hizmet sorunları, aldığı hizmetin kullanım miktarları, telekomünikasyon ağının performansı, kişisel bilgileri ve yaşadığı bölge gibi birçok bilgi göz önüne alınarak büyük veri kapsamında incelenmektedir.

Ayrılan müşteri analizinin en çok kullanıldığı sektörlerden biri de farklı hizmet sağlayıcısına geçişi kolay olması sebebiyle Telekomünikasyon sektörüdür. Telekomünikasyon sektöründe hizmet sağlayıcılar için yeni müşteri kazanmanın çok daha maliyetlidir. Bunun sebebiyle var olan müşterisini bünyesinde tutmak istemektedir. Türkiye'de 2019 yılının 4. çeyreğinde toplam mobil abone sayısı 81 milyona yaklaşmıştır. Yine 2019 yılına ait Türkiye'nin nüfusu verileri ile ilişkilendirildiğinde %98,5 oranında bir mobil hat kullanımı görülmektedir (Bilgi Teknolojileri ve İletişim Kurumu, 2019).

Bu çalışmanın amacı; müşteri bilgilerinin veri madenciliği, makine öğrenmesi ve derin öğrenme teknikleri ile anlamlandırılarak ayrılacak müşteri analizini yapmak ve ağırlıklı olarak hangi sebeplerden dolayı aboneliğini sonlandırdığını tahmin etmektir.

2. LİTERATÜR ÖZETİ

Literatür içerisinde telekomünikasyon sektörüne dair birçok müşteri ayrılma analizi çalışması bulunmaktadır. Çalışmalarda performans ölçütlerinin değerini arttırmaya yönelik; yeni yöntem oluşturma, mevcut yöntemleri geliştirme, birden fazla yöntemi sentezleyerek yeni bir model oluşturma, özellik mühendisliği yapılarak bilgi üretme ve katkıda bulunmayan özniteliklerin veri setinden çıkartılması gibi birçok yöntem kullanılmıştır.

Bu bölümde öznitelik seçimi, tek sınıflandırıcı kullanılarak geliştirilen modeller ve birden fazla tekniğin birlikte kullanılarak melez (hybrid) olarak geliştirilen modeller tartışılmaktadır.

Altibbi ve AL-Shatnwai (2020), çalışmasında telekomünikasyon şirketlerinde müşteri tutma oranını tahmin etmek için yüksek hızda örnekleme yöntemleriyle topluluk öğrenme algoritmaları olan Gradient Boosting algoritmasına dayalı bir yaklaşım önerilmiştir. Bu yaklaşımda; rastgele yüksek hızda örnekleme, SMOTE, ADASYN ve Borderline SMOTE olmak üzere dört yaygın ve iyi bilinen yüksek hızda örnekleme yöntemi kullanılır ve karşılaştırılmıştır. Deneylerin ilk bölümü, yüksek hızda örnekleme yapılmadan Gradient Boosting algoritmasının SVM, Random Forest, Logistic Regression ve SGD sınıflandırıcı yöntemleri dâhil diğer popüler sınıflandırmalardan daha iyi performans gösterdiğini gösterdi. Yapılan deneyin ikinci kısmında yüksek hızda örnekleme yöntemlerini farklı yüksek hızda aşırı örnekleme oranlarında uygulanmıştır. Deneyler, yüksek hızda örnekleme yöntemlerinin Gradient Boosting algoritmasının kayıp sınıfını tahmin etme performansını artırdığını ve en iyi F değerine yaklaşık %84 ulaşabileceğini ve SMOTE yöntemi ile %20 yüksek hızda örnekleme oranında ulaşılabileceğini ortaya koymuştur.

Safitri ve Muslim (2020), yapmış oldukları çalışmada SMOTE tekniğini ve genetik algoritma kullanarak Naive Bayes sınıflandırma algoritmasının doğruluğunu arttırmaya yönelik bir yaklaşım önermiştir. SMOTE tekniğini veri kümesinin sınıf dengesizliğine çözüm bulmak için kullanırken, genetik algoritma ise öznitelik

seçimi yapmak için kullanmışlardır. Çalışmada Naive Bayes algoritması ile sınıflandırma yaparak doğruluk başarı oranını %47.1 olarak elde etmişlerdir. SMOTE ve Naive Bayes kullanarak doğruluk oranını %78.15 olarak elde etmişlerdir ve daha iyi bir sonuç almışlardır. Çalışmanın en iyi sonucu olarak SMOTE ile aşırı örnekleme, genetik algoritması ile öznitelik seçimi ve Naive Bayes ile sınıflandırma algoritması kullanılmıştır ve başarı ölçüsü olan doğruluk oranını %78.46 olarak elde etmiştir.

Wadikar (2020), çalışmasında bir Credit Union finans kurumunun müşteri kayıp analizini kesin olarak tahmin edebilen bir makine öğrenimi modeli geliştirmeyi amaçlamıştır. Veri kümesinde bulunan sınıf dengesizliği problemini, öznitelik seçimini ve müşteri kaybını verimli bir şekilde tahmin eden denetimli bir makine öğrenimi modeli oluşturmak için nicel ve tümdengelimle araştırma stratejileri kullanılmıştır. Müşteri kayıp analizini gerçekleştirmek için denetimli makine öğrenimi yöntemleri olan Lojistik Regresyon, Rasgele Orman, Destek Vektör Makinesi (SVM) ve Sinir Ağı yöntemlerini kullanmıştır. Çalışmasında en iyi sınıflandırıcıyı belirlemek için Doğruluk Oranı, ROC eğrisi ve AUC-ROC çıktılarını başarı ölçütleri olarak kullanmıştır. Aşırı örnekleme yöntemi olarak SMOTE tekniğini kullanmıştır. Önerilen modeller arasında en iyi sınıflandırıcı olarak Random Forest (Rastgele Orman) algoritması tespit edilmiştir.

Abbasimehr vd. (2014), erişime açık Larose isimli telekomünikasyon veri kümesini kullanarak müşteri kaybı tahmin (customer churn prediction) analizi yapmıştır. Çalışmasında topluluk öğrenimi algoritmaları ve bunun yanında en yaygın olan Karar Ağaçları, Yapay Sinir Ağları, Destek Vektör Makinesi gibi birçok algoritma kullanarak en iyi sonucu almayı amaçlamıştır. Çalışmasının ilk aşamasında temel öznitelik çıkarım işlemi yaparak sembolik değerler taşıyan öznitelikleri veri kümesinden çıkarmıştır. Sonraki adımlarda SMOTE aşırı örnekleme tekniğiyle beraber başarı ölçütlerini karşılaştırarak ve değerlendirerek bir sonuç elde etmişlerdir. Başarı ölçütü olarak AUC, Sensivity, Specificity gibi kriterleri temel almışlardır. Çalışmada kullanılan yöntemler melez (hybrid) olarak birbiriyle beraber kullanılmıştır ve en iyi sonucu veren 2 adet

model önerilmiştir. Çalışmanın en iyi sonuç veren modeller Boosting+RIPPER ve Boosting+C4.5 olarak belirlenmiştir ve önerilmiştir.

J. Vijaya vd. (2018), yapmış oldukları çalışmada yüksek boyutlu müşteri verilerini işlemek için öznitelik seçimi ve grup sınıflandırmasına bütünleşmiş bir yaklaşım önermiştir. Ön işleme sürecinden sonra Rough Set Feature Selection (RSFS), Correlation Feature Selection (CFS), Information Gain (IG), Forward Search (FS), Backward Search (BS) teknikleri kullanılarak özellik seçimi yapmaktadır. Sonrasında öznitelik seçimi teknikleri beraber kNN, Decision Tree (DT), Support Vector Machine (SVM), Naive Bayes (NB), Artificial Neural Network (ANN) ve Ensemble Learning (Bagging, Boosting, Random Subspace) yöntemlerini kullanarak en iyi sonuçları elde etmeye çalışmıştır. Çalışmada en iyi sonucu, RSFS ve Boosting yöntemlerini kullanarak elde etmiştir.

Sjarif vd. (2019), yapmış oldukları çalışmada 7043 kayıt ve 21 özniteliği bulunan bir veri kümesi kullanmıştır. Ön işleme safhasında özniteliklerin seçimi için Pearson Korelasyon Katsayısı (Pearson Correlation Coefficient) olarak adlandırılan standart korelasyon yöntemi kullanılmıştır. Geliştirdikleri modelde KNN (K Nearest Neighbor), Random Forest ve Support Vector Machine yöntemlerini kullanarak en iyi sonucu elde etmeye çalışmıştır. Sonuçlara bakıldığında KNN yönteminin diğer yöntemlere göre daha iyi performans gösterdiği tespit edilmiştir. Performans ölçütü olarak doğruluk oranı temel alınmıştır.

Shuan vd. (2017), çalışmasındaki veri kümesini UCI Machine Learning Repository adlı kaynaktan sağlamıştır. Çalışmada, veri kümesini 2 tip olarak ayırmıştır. İlk veri kümesinde tüm öznitelikler bulunmaktadır. İkinci veri kümesinde ise öznitelikler için seçim yaparak sayısını azaltmıştır. Önerdikleri modelde 2 adet yöntem kullanılmıştır. Birinci yöntemde kümeleme yöntemi olan K-Means ile Naive Bayes yöntemlerini birleştirmiştir. İkinci yöntemde ise EWD yöntemini kullanmıştır. Performans çıktılarına bakıldığında K-Means ve Naive Bayes yönteminin birlikte kullanıldığı yaklaşımın, EWD yönteminden daha iyi olduğunu tespit etmişlerdir. Ayrıca küme sayısının arttırılması ile karmaşıklık matrisi

içerisinde bulunan Gerçek Pozitif sayısını iyileştirdiğini kanıtlamışlardır. Performans ölçütü olarak doğruluk oranı ve hassasiyet kullanılmıştır.

Guan Li, Koh vd. (2019), yapmış oldukları çalışmada veri kümesini Kaggle Open Datasets adlı kaynaktan elde etmiştir. Ön işleme bölümünde veri kümesine normalizasyon uygulanmıştır ve ayrıca ROSE örnekleme yöntemini kullanarak veri kümesini dengelemiştir. Çalışmasında temel sınıflandırma yöntemleri olan Naive Bayes, Decision Tree ve Artificial Neural Network yöntemlerini kullanmıştır. Ayrıca bu yöntemleri Grid Search olarak adlandırılan hiper parametre ayarlaması yöntemi ile beraber kullanılmıştır. Çıktılara bakıldığında Grid Search algoritmasıyla birlikte kullanılan temel sınıflandırma yöntemleri daha iyi sonuç vermiştir. Performans ölçütü olarak doğruluk oranını temel almıştır.

3. AYRILAN MÜŞTERİ ANALİZİ

Günümüzde işletmeler sahip oldukları müşterilerin en değerli varlıkları olduğunun bilincindedir. Ancak müşterilerin tercih edebileceği çok sayıda işletmenin bulunması, yeni müşteriler kazanmayı daha zor ve maliyetli bir hale getirmektedir. İşletmeler potansiyel risk taşıyan yeni müşteriler aramaktansa sahip oldukları müşterilerin ihtiyaçlarına odaklanarak uzun dönemli ilişkiler geliştirmeyi daha çok tercih etmektedir. Uzun dönemli ilişkiler geliştirilen müşteriler daha fazla satın alım yapmakta ve eğer memnun kalırlarsa işletmenin olumlu yönde tanıtımını yapmaya istekli olmaktadırlar. Ayrıca uzun dönemde bu tür sadık müşterilere hizmet sağlamanın maliyeti daha düşük olmaktadır. Çünkü sadık müşteriler rakiplerin pazarlama faaliyetlerine daha az duyarlı olma eğilimindedirler (Poel and Lariviére, 2004).

Ayrılan müşteri analizi genel olarak mevcut kayıpların tahminine dayanan bir analiz yöntemidir. Telekomünikasyon, bankacılık, sigorta ve oyun sektörleri gibi birçok sektörde müşteri bilgileri kullanılarak analiz yapılmaktadır. Bu sektörlerde müşteri sürekliliği, şirketlerin değerleri ile doğru orantılıdır. Bu nedenden dolayı ayrılan müşteri analizi, müşteri sürekliliğinin istendiği sektörlerde çok kritik bir analizdir. Ayrılan müşteri analizinin literatürde müşteri tükenmesi, müşteri kaçışı, müşteri çıkışı, müşteri karmaşası gibi birçok adı bulunmaktadır.

Ayrılan müşteri analizinin en çok kullanıldığı sektörlerden biri de farklı hizmet sağlayıcısına geçişi kolay olması sebebiyle Telekomünikasyon sektörüdür. Telekomünikasyon sektöründe hizmet sağlayıcılar için yeni müşteri kazanmak çok daha maliyetlidir. Bunun sebebiyle var olan müşterisini bünyesinde tutmak istemektedir. Türkiye'de 2019 yılının 4. çeyreğinde toplam mobil abone sayısı 81 milyona yaklaşmıştır. Yine 2019 yılına ait Türkiye'nin nüfusu verileri ile ilişkilendirildiğinde %98,5 oranında bir mobil hat kullanımı görülmektedir (Bilgi Teknolojileri ve İletişim Kurumu, 2019).

3.1. Telekomünikasyon Sektöründe Ayrılan Müşteri Analizi

Günümüz teknolojisinde iletişim ihtiyaçları genellikle ek bir maliyet olarak değil, bir ihtiyaç olarak kabul edilir. Telekomünikasyon sektörü, bu sektör için rekabetçi bir pazara sahip ülkelerde çoğunlukla doymuş olma eğilimindedir. Aslında, başka bir deyişle, pazarın büyüme hızı her geçen gün daha küçük bir ivmeye doğru kaymaktadır. Bu değerlendirmeler ışığında, yeni aboneler kazanmak giderek daha zor hale geliyor. Bu durum, şirketlerin ayrılan müşteri analizi gibi faaliyetler yürütmelerine ve bunları sistemde tutmak için faaliyetler yürütmelerine yol açmaktadır. Rekabet ortamındaki iptal oranının ayda %2.2 olduğu göz önüne alındığında, bu faaliyetlerin ne kadar önemli olduğunu tahmin etmek zor olmayacaktır (Çelik, 2019).

Müşterilerin abonelik ömrü ne kadar uzun olursa, şirket için o kadar fazla kâr göz önünde bulundurulur. Şirketler kısa vadeli müşteri ilişkileri yerine uzun vadeli sözleşmeleri, kampanyaları veya tarifeleri tercih etme eğilimindedir. Sadık müşterilerin değeri nedeniyle, sadakat oluşturmak için müşteri kaybı analizi uygulanmaktadır (Çelik, 2019).

3.2. Farklı Sektörlerde Ayrılan Müşteri Analizi

Literatürde kayıp müşteri analizinin iletişim başta olmak üzere birçok farklı sektörde gerçekleştirildiği görülmektedir. Bunların arasında bankacılık, sigortacılık, turizm sektörü ve marketler sayılabilmektedir (Cengizci, 2020).

Bankacılık ve sigortacılık sektörlerinde ayrılan müşterileri tanımlamada genellikle hesap kapatma durumu temel alınmaktadır. Perakende sektöründe ayrılma davranışının genellikle kısmi kayıp olarak ortaya çıktığı görülmektedir. Bu durum herhangi bir üyelik sistemi ya da sözleşme bulunmamasından dolayı müşterilerin işletmeyi ayrılma durumuyla ilgili bilgilendirmek zorunda olmayışı ve farklı işletmelerden alışveriş yaparak satın alımlarını farklı işletmeler arasında bölüştürmesinden kaynaklanmaktadır. Bu şekilde çalışılan sektörlerde müşteriler aşamalı olarak ilişkilerini azaltabilmektedir (Cengizci, 2020).

4. VERİ MADENCİLİĞİ

Veri madenciliği, çok büyük miktarda verinin depolandığı veri tabanlarından, amacımız doğrultusunda, gelecek ile ilgili tahminler yapmamızı sağlayacak, anlamlı olan veriye ulaşma ve veriyi kullanma işidir (Savaş vd., 2012).

Günümüz dünyasında pazardaki zorlu koşullar, şirketleri daha iyi rekabet etmek için yeni yollar bulmaya yöneltmektedir. Yoğun küresel rekabet ve hızla değişen teknolojik ortamlarla, müşterilerin çeşitli ihtiyaçlarını karşılamak ve kârlı müşterilerin değerini en üst düzeye çıkarmak, birçok çağdaş şirket için tek geçerli seçenek haline gelmektedir. Teknolojik gelişmelerle birlikte şirketler ve kurumlar müşteri ve satış verilerini sürekli olarak saklamaktadır. Veri madenciliği, sürekli olarak saklanan bu verilerde firmaların amaçlarına ulaşma yolunda faydalı, gizli, anlamlı, bilinmeyen ve değerli bilgilerin elde edilmesi sürecidir. Büyük veri, veri bilimi, istatistik, veritabanı teorisi ve makine öğreniminden birçok tekniği bir araya getiren bir bilgisayar bilimi alt birimidir (Ata, 2018).

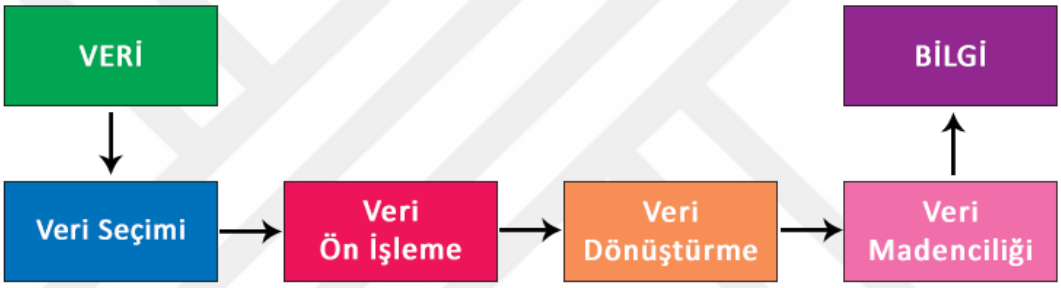
Şirketler mevcut müşterilerin davranışlarını tahmin edebilir veya analiz edebilmektedir. Ayrıca yeni ürünler/abonelikler önerebildiği gibi mevcut ve potansiyel müşterilerine kampanyalar/indirimler sunabilmektedir.

Veri madenciliği birçok sektörde kullanılabilmektedir. Veri madenciliğine aşağıdaki gibi örnekler verilmiştir:

- Müşterinin mevcut aboneliğinin sonlandırıp sonlandırmayacağını tahmin edilmesi yani Ayrılan Müşteri Analizi
- Kayıp, kâr, talep gibi risk faktörlerinin tahmin edilmesi
- Pazarlama ve satış analizi
- Piyasaya yeni çıkacak bir ürünün talebinin önceden tahmin edilmesi
- Dolandırıcılık tespiti

4.1. Veri Madenciliği Süreci

Veri madenciliğine ayrıca bir süreç denilebilmektedir. Veri madenciliği sadece veri yığınları arasında soyut biçimde kazılar yaparak veri ortaya çıkarmak değildir. Buna ek olarak bilgi keşfi sürecinde örüntüleri ayrıştırarak süzmek ve sonraki adımlara hazır hale getirmek de bu sürecin bir parçası denilebilmektedir. Veri madenciliği süreçlerinde, üzerinde inceleme yapılan veri kümesinin özniteliklerini bilmek çok önemlidir. Bilinmemesi durumunda hiçbir veri madenciliği algoritmasının fayda sağlaması mümkün olmayacaktır. Bu yüzden veri madenciliği sürecine girmeden önce veri özniteliklerinin detaylı analiz yapılması gerekmektedir. Bu süreç Şekil 4.1'de gösterilmiştir.



Şekil 4.1. Veri madenciliğinin süreci ile bilginin keşfedilmesi

Veri madenciliği sürecinde izlenen adımlar sırasıyla Problemin Tanımlanması, Verilerin Hazırlanması, Modelin Kurulması ve Değerlendirilmesi, Modelin Kullanılması, Modelin İzlenmesi olmak üzere 5 farklı adımdan oluşmaktadır.

4.1.1. Problemin yanımlanması

Veri madenciliğinde başarılı olmanın en önemli yolu verinin hangi amaç ile işletileceği ve elde edilecek sonucun başarı ölçütlerinin nasıl ölçülmesi gerektiğidir.

4.1.2. Verilerin hazırlanması

Veri madenciliği sürecinde en fazla enerji ve zaman harcatan adımı verilerin hazırlanması adımı olabilmektedir. Modelin kurulumu ve değerlendirilmesi

adımında isteğe uygun değişiklikler yapmak amacıyla “Verilerin Hazırlanması” adımına sık sık geri dönülebilmektedir.

Verilerin hazırlanması adımında veri ile ilgili “Toplama”, “Birleştirme ve Temizleme”, “Örneklemin Seçilmesi”, “Dönüştürme” gibi birçok ön işleme aşamaları bulunmaktadır.

4.1.3. Modelin kurulması ve değerlendirilmesi

Çoğu zaman optimum modelin tespit edilmesi için çok sayıda modelin kurulup test edilmesi gerekebilmektedir. Bu sebepten dolayı “Veri Hazırlama” ve “Model Kurma ve Değerlendirme” aşamaları literatürde yinelenen bir süreç olarak geçmektedir.

4.1.4. Modelin kullanılması

Kabul edilen optimum model, doğrudan bir uygulama olabileceği gibi başka uygulamaların alt parçası olarak kullanılabilmektedir.

4.1.5. Modelin izlenmesi

Tüm sistemlerin özelliklerinde ve kullanılan veriler üzerinde zaman geçtikçe herhangi bir sebepten dolayı değişiklikler yaşanabilmektedir. Haliyle kurulan modellerin sürekli olacak bir biçimde izlenmesi ve ihtiyaç oldukça yeniden düzenlenmesi gerekecektir.

Diğer yöntemler de genellikle istatistiği temel alan ama daha çok makine öğrenimi ve yapay zekâ destekli yeni nesil yöntemlerdir. Veri madenciliği modelleri, gördükleri işlevlere göre temel anlamda “Sınıflama”, “Kümeleme”, “Birliktelik Kuralları” ve “Regresyon” olarak 4 grupta toplanmaktadır (Savaş vd., 2012).

4.2. Veri Madenciliğinde Kullanılan Yöntemler

Zaman zaman verileri, kullanılacak yöntemlere göre ve istenilen veri formatına dönüştürmemiz gerekmektedir. Veri madenciliği süreçlerinden "Verilerin Hazırlanması" adımında öznitelik seçimi, normalizasyon, kodlama, aşırı örnekleme gibi birçok ön işleme safhaları bulunmaktadır.

4.2.1. Öznitelik indirgeme (Attributes Reduction)

Veri madenciliği yönteminin kullanılacağı bazı uygulamalarda veriler ham olarak toplanabilmektedir. Toplanan ham verilerin isteğe uygun bir biçime dönüştürülmesi gerekebilmektedir. Bunun için veri kümesindeki uygulanacak analize katkısı olmayan özniteliklerin performans açısından iyileştirilmesi amacı ile çıkartılması yani öznitelik indirgeme işlemi yapılması en ideal çözüm olabilmektedir.

Uygulanacak analize katkısı olmayan özniteliklere örnek olarak benzersiz (unique) değer taşıyan öznitelikler örnek olarak verilebilmektedir.

4.2.2. Veri normalizasyonu

Bir veri kümesindeki sayısal öznitelikler farklı aralıklara sahip olabilmektedir. Öznitelik değerleri, o özniteliğin minimum ve maksimum değerleri arasında herhangi bir yere düşebilmektedir. Bu değerler için hesaplama yapan bir sınıflandırıcıya iki veya daha fazla sayısal değer girildiğinde, bu sınıflandırıcıların bu değerlere dayalı çıktı yapabilmesi için matematiksel gösterimi çıkarması zorlaşmaktadır. Örneğin özniteliklerden birisi [0, 5] aralığına sahipken diğer özniteliğin [500, 1000] aralığına sahip olduğu düşünüldüğünde matematiksel olarak sınıflandırıcıların değer aralıklarını sağlıklı bir şekilde seçmesi çok zordur. Veri kümesinde öznitelik sayısının artmasıyla birlikte, oluşacak karmaşıklığın seviyesi de yüksek oranda artış gösterecektir (Quackenbush, 2002).

Normalizasyon işlemi, hesaplamaları sınıflandırıcılara göre çok daha kolay hale getirmektedir. Böylelikle daha iyi sonuçlar üretebilmektedir ve sınıflandırıcıların içindeki hesaplamayı basitleştirmektedir (Quackenbush, 2002).

4.2.2.1. Min-Max Normalizasyon yöntemi

Min-Max normalizasyon yöntemi, ham veri kümesinin minimum ve maksimum değerlerini bulur ve Denklem (4.1)'de verilen formüle göre her bir giriş değerini 0 ve 1 aralığında doğrusal olarak normalleştirir. Min-Max normalleştirme yöntemiyle ilgili temel sorun, minimum ve maksimum hesaplamada kullanılan örnek olmayan veri kümesinin değerleri bilinmemektedir (Taşdemir vd., 2018).

$$\mathbf{N_i} = \frac{\mathbf{x_i} - \mathbf{x_{min}}}{\mathbf{x_{max}} - \mathbf{x_{min}}} \tag{4.1}$$

Denklem (4.1)'de özniteliğin maksimum değerini $\boldsymbol{x_{max}}$, minimum değerini $\boldsymbol{x_{min}}$ ve öznitelikteki her bir değer $\boldsymbol{x_i}$ olarak tanımlanmaktadır.

4.2.3. Veri kodlama

Bir veri kümesindeki değerler sayısal veya kategorik değerler olabilmektedir. Ayrıca kategorik değerler ise sıralı (ordinal) veya sembolik (nominal) değer olabilmektedir. Bazı sınıflandırma teknikleri, girdi değerlerine dayalı hesaplamalar kullandığından kategorik değerleri işleme yeteneğine sahip değildir. Bu yüzden kategorik değerleri sayısal değerler ile kodlamak önemli olmaktadır. Böylece bu veri kümelerini çeşitli yöntemleri kullanarak işlemek mümkün hale gelmektedir (Kadhim, 2018).

4.2.3.1. Etiket kodlama (Label Encoding) yöntemi

Etiket kodlama, kategorik değerleri sayısal biçime dönüştürmektedir. Eğer kategorik değerler sıralı değerler ise sıralarına ve bir kategorik değer ile diğeri arasındaki mesafelere göre sayılarla atanabilmektedir. Eğer kategorik değerler sembolik (nominal) değer ise sayısal değerler rastgele olarak veya öznitelikteki

değerlerin alfabetik sıralaması gibi herhangi bir seçilen sırayla atanabilmektedir (Kadhim, 2018).

Label Encoding yönteminin kategorik değerleri kodlamasına ilişkin örnek Çizelge 4.1'de verilmiştir.

Çizelge 4.1. Kategorik değerlerin Label Encoding yöntemi ile kodlanması

Kategorik Değerler	**Label Encoded Yöntemi**
Güneşli	1
Bulutlu	2
Güneşli	1
Güneşli	1
Yağmurlu	3

4.2.3.2. One-Hot Encoding yöntemi

One-Hot Coding en yaygın kullanılan kodlama şemasıdır. Kategorik değişkenin her değerini sabit bir referans değeri ile karşılaştırmaktadır. One-Hot Encoding yönteminde, n adet gözlem içeren ve d adet farklı değerli tek bir özniteliği, her biri n adet gözlemi olan d adet ikili değişkene dönüştürmektedir. Her gözlem, ikili değişkenin varlığını veya yokluğunu sırasıyla 1 veya 0 şeklinde göstermektedir (Potdar vd., 2017). One-Hot Encoding yönteminin kategorik değerleri kodlamasına ilişkin örnek Çizelge 4.2'de verilmiştir.

Çizelge 4.2. Kategorik değerlerin One-Hot Encoding yöntemi ile kodlanması

Kategorik Değerler	**One-Hot Encoding Yöntemi**		
Güneşli	1	0	0
Bulutlu	0	1	0
Güneşli	1	0	0
Güneşli	1	0	0
Yağmurlu	0	0	1

Label Encoding yöntemi ile One-Hot Encoding yönteminin karşılaştırılması Çizelge 4.3'de örnek olarak verilmiştir.

Çizelge 4.3. Label Encoding ve One-Hot Encoding yöntemlerinin karşılaştırılması

Kategorik Değerler	**Label Encoded Yöntemi**	**One-Hot Encoding Yöntemi**		
Güneşli	1	1	0	0
Bulutlu	2	0	1	0
Güneşli	1	1	0	0
Güneşli	1	1	0	0
Yağmurlu	3	0	0	1

4.2.4. Öznitelik seçimi

Veri madenciliğinde bazı zamanlar en uygun modeli bulmak için veri kümesinde bulunan özniteliklerin birbiri arasındaki ilişki analiz edilmesi gerekebilmektedir. Ayrıca daha az öznitelik ile aynı başarıyı elde etmek için zaman maliyetini düşürmek istenebilmektedir. Buna benzer ihtiyaçlar olduğunda veri kümesinde öznitelik seçimi yapılabilmektedir.

Bunlara örnek olarak öznitelikler arasındaki ilişiyi analiz edebilmek için Pearson Correlation Coefficient gibi veri madenciliği yöntemleri kullanılabilmektedir. Ayrıca öznitelik seçimi yapmak için Univariate Feature Selection gibi veri madenciliği yöntemleri de kullanılabilmektedir.

4.2.4.1. Pearson Correlation Coefficient yöntemi

Pearson Correlation Coefficient tekniği, iki sayısal ve sürekli olan değişkenler arasındaki doğrusal bağımlılık ilişkisini istatistiksel olarak ölçmek amacı ile kullanılan en yaygın yöntemlerden birisidir. Kovaryans yöntemine dayanmaktadır.

Bu teknik ile veri kümesindeki her bir öznitelik için korelasyon katsayısı bulunursa, öznitelikler arasında bulunan doğrusal bağımlılık ilişkisi tespit edilmiş olur.

$$r = \frac{\sum(\mathrm{x}-\mathrm{m}_x)(\mathrm{y}-\mathrm{m}_y)}{\sqrt{\sum(\mathrm{x}-\mathrm{m}_x)^2 \sum(\mathrm{y}-\mathrm{m}_y)^2}} \tag{4.2}$$

Denklem (4.2)'deki gibi x ve y terimleri, n uzunluğunda iki vektördür. m_x terimi x'in ve m_y terimi y'nin ortalamasına karşılık gelmektedir. Ayrıca r korelasyon katsayısı olarak isimlendirilmektedir. Korelasyon katsayısı değer olarak -1 ile +1 arasında bir değer almaktadır.

Korelasyon katsayısının değeri 0 ise iki değişken arasında bir ilişki olmadığını, negatif (negatif korelasyon) ise birbirleri arasında zıt ilişki olduğunu, pozitif sayı (pozitif korelasyon) ise birbirleri arasında ilişki olduğunu göstermektedir. Matematiksel ifadesi Denklem (4.3) olarak gösterilmiştir.

$$\mathrm{f(r)} = \begin{cases} \text{negatif korelasyon, zıt ilişkili,} & \mathrm{r} < 0 \\ \text{pozitif korelasyon, ilişkili,} & \mathrm{r} > 0 \\ \text{ilişki yok,} & \mathrm{r} = 0 \end{cases} \tag{4.3}$$

Bu yöntemin görseli de bulunmaktadır. Bu görsele göre iki özniteliğe karşılık gelen renk; ne kadar koyu olursa o kadar ilişkilidir yani pozitif korelasyon, ne kadar açık olursa o kadar zıt ilişkilidir yani negatif korelasyon bulunmaktadır (Şekil 6.3).

4.2.4.2. Univariate Feature Selection yöntemi

Tek değişkenli, farklı istatistiksel puanlama işlevlerine dayalı olarak özelliklerin sıralı bir listesini döndüren bir özellik seçme yöntemidir. Veri kümesi özniteliklerini kullanmadan önceki bir ön işleme adımıdır (Subho vd., 2019).

4.2.5. Veri aşırı örnekleme

Veri kümesinde sınıf dengesizliği söz konusu olduğunda dengeyi sağlayabilmek için baskın olan verinin sayısını düşürmek veya azınlık olan veriyi arttırmak

gerekmektedir. Bu yöntemlerin genel adına Oversampling ve Undersampling adı verilmektedir.

Aşırı örnekleme (oversampling) yöntemi, veri kümesindeki azınlık sınıfına ait üyelerin sayısını artırmayı hedeflemektedir. Aşırı örneklemenin avantajı, tüm üyeleri azınlık ve çoğunluk sınıflarından koruduğumuz için orijinal eğitim setinden hiçbir bilginin kaybolmamasıdır (Liu, 2004).

En yaygın kullanılan veri aşırı örnekleme olarak SMOTE (Synthetic Minority Oversampling Technique) ve ADASYN (Adaptive Synthetic Sampling Method) yöntemleri örnek verilebilmektedir.

4.2.5.1. SMOTE (Synthetic Minority Oversampling Technique)

Azınlık sınıfının, değiştirme yolu ile aşırı örnekleme yerine "sentetik" örnekler oluşturarak, yüksek hızda örneklendiği bir aşırı örnekleme yaklaşımı SMOTE olarak önerilmiştir. SMOTE yaklaşımı, el yazısı ile bir yazıda karakter tanımada başarılı olan bir yöntemden esinlenerek geliştirilmiştir (Ha ve Bunke, 1997). Gerçek verilere çeşitli işlemler uygulayarak ekstra eğitim verileri oluşturmuşlardır. SMOTE yöntemi için döndürme ve çarpıklık gibi veri kümesine uygulanan işlemler, eğitim verilerini bozmanın doğal yolları olduğunu çalışmalarında belirtmişlerdir.

SMOTE yaklaşımında, sentetik örnekler "veri alanı" yerine "özellik alanı" içinde çalışarak daha az uygulamaya özgü bir şekilde oluşturulmaktadır. Veri kümesindeki her azınlık sınıf için örneklem alınarak ve en yakın komşularındaki k adet azınlık sınıfının herhangi birini/tümünü birleştiren çizgi parçaları boyunca sentetik örnekler verilerek aşırı örneklenmektedir. Gereken aşırı örnekleme miktarına bağlı olarak, kNN yönteminden gelen (en yakın k komşularından) gelen komşular rastgele seçilir (Chawla vd., 2002).

Bu işlem, iki belirli özellik arasındaki çizgi parçası boyunca rastgele bir noktanın seçilmesine neden olur. SMOTE yaklaşımı, azınlık sınıfının karar bölgesini etkili bir şekilde daha genel hale getirmeye zorlamaktadır (Chawla vd., 2002).

SMOTE tekniğinin sözde kodu (pseudo code) aşağıdaki gibidir:

Algoritma: SMOTE (T, N, k)
Girdi: Azınlık sınıf örneklerinin sayısı, T; SMOTE miktarı, N%; En yakın komşuların sayısı k
Çıktı:

```
if N < 100
        then T azınlık sınıfı örneklerini rastgele hale getirin
                T = (N/100)∗T
                N = 100
endif
N = (int)(N/100) // N değişkeni, 100'ün intergral katları olduğu varsayılır.
k = En yakın komşuların sayısı
numattrs = Öznitelik Sayısı
sample[ ][ ] // Orijinal azınlık sınıfı örneklerine ait dizi
newIndex = 0 // Oluşturulan sentetik örneklerin sayısını tutar
synthetic[ ][ ]: Sentetik örneklere ait dizi
// Yalnızca her bir azınlık sınıfı örneği için k en yakın komşuyu hesaplayın.
for i  = 1 to T
        i için k en yakın komşuyu hesaplayın
        İndisleri nnarray isimli diziye kaydedin
        Populate (N, i, nnarray)
endfor
Populate (N, i, nnarray) // Sentetik numuneler oluşturma fonksiyonu
while N != 0
        1 ile k arasında rastgele bir sayı seçin, buna nn adını verin.
        (Bu adım, i'nin en yakın komşularından birini seçer.)
        for attr = 1 to numattrs
                dif = sample[nnarray[nn]][attr] – sample[i][attr]
```

23. gap = 0 ile 1 arasında rastgele bir sayı
24. synthetic[newIndex][attr] = sample[i][attr] + gap ∗ dif
25. Endfor
26. newIndex++
27. N = N – 1
28. Endwhile

4.2.5.2. ADASYN (Adaptive Synthetic Sampling Method)

ADASYN'in temel fikri, öğrenmesi daha zor olan azınlık sınıfı örnekleri için daha fazla sentetik verinin üretildiği, öğrenme güçlüğü düzeylerine göre farklı azınlık sınıfı örnekleri için ağırlıklı bir dağılım kullanmaktır. Ek olarak ADASYN yöntemi, veri kümesinin sınıf dengesizliğini, veri dağılımının getirdiği öğrenme önyargısını azaltmaktadır. Ayrıca öğrenmesi zor olan veri örneklerine odaklanmak için karar sınırını güncellenebilir şekide değiştirebilmektedir. (He vd., 2008).

ADASYN yaklaşımı veri dağılımları ile ilgili öğrenmeyi iki şekilde geliştirmektedir. Bunlardan birincisi, sınıf dengesizliğinin getirdiği sapmanın azaltılmasıdır. İkincisi ise sınıflandırma karar sınırının zor örneklere uyarlanabilir şekilde kaydırılmasıdır (He vd., 2008).

Algoritma: ADASYN (T, N, k)
Girdi: Eğitim veri kümesi, D_{tr}; m veri örnekleri $\{x_i, y_i\}$ olmak üzere i = 1, … , m; n boyutlu X öznitelik uzayında bir örnek, x_i; azınlık sınıfı örneklerinin sayısı, m_s; çoğunluk sınıfı örneklerinin sayısı, m_l; $m_s \leq m_l$ ve $m_s + m_l = m$
Çıktı:

1. Sınıf dengesizliğinin derecesini hesaplayın. $d \in (0, 1]$ olmak üzere $d = m_s/m_l$
2. $d < d_{th}$ (Tolere edilen sınıf dengesizlik oranı için önceden belirlenen eşik) ise
    a. Azınlık sınıfı için oluşturulması gereken sentetik veri örneklerinin sayısını $G = (m_l - m_s) \times \beta$ denklemi ile hesaplayın. Burada $\beta \in [0, 1]$, sentetik verilerin oluşturulmasından sonra istenen denge seviyesini belirtmek için kullanılan bir parametredir. $\beta = 1$, genelleştirme

sürecinden sonra tamamen dengeli bir veri setinin oluşturulduğu anlamına gelir.

b. Her bir $x_i \in$ azınlık sınıfı örneği için, n boyutlu uzayda Öklid mesafesine dayalı olarak kNN'yi bulun ve aşağıdaki gibi tanımlanan r_i oranını $r_i = \Delta_i/K$, $i = 1, \ldots, m_s$ denklemi ile hesaplayın. Where Δ_i is the number of examples in the Knn of x_i that belong to the majority class, therefore $r_i \in [0, 1]$.
c. Normalize r_i according to $\hat{r}_i = r_i / \sum_{i=1}^{ms} r_i$, so that $\hat{r}_i$ is a density distribution ($\sum_{\hat{r}i=1}$)
d. Her bir azınlık örneği x_i için oluşturulması gereken sentetik veri örneklerinin sayısını $g_i = \hat{r}_i \times G$ denklemi ile hesaplayın.
e. Her bir azınlık sınıfı veri örneği x_i için, aşağıdaki adımlara göre g_i sentetik veri örnekleri oluşturun. Bunun için 1'den gi'ye kadar döngü oluşturun.
    i. x_i verisi için kNN'den rastgele bir azınlık veri örneği, x_{zi} seçin.
    ii. Sentetik veri örneğini $si = xi + (xzi - xi) \times \lambda$ denklemi ile oluşturun.
    iii. $(x_{zi} - x_i)$ n boyutlu uzaylardaki fark vektörü ve λ rastgele bir sayıdır: $\lambda \in [0, 1]$

Döngü Sonu

5. MAKİNE ÖĞRENMESİ

İstatistik veriden öğrenme sanatıdır. Veriden öğrenme konusunda yaşanan zorluklar sonucunda istatistik biliminde köklü gelişmeler ortaya çıkmıştır. Veriden anlam çıkarma ve öğrenme sürecinde matematiksel hesaplamalar ve bilgisayar kodları hayati önem taşıdığından bu konudaki en büyük katkılar bilgisayar bilimciler tarafından sağlanmıştır (Friedman vd., 2009).

Makine öğrenmesi genellikle verilerin yapısını anlamaya yarayan yapay zekânın bir alt alanıdır. Ayrıca verileri insanların anlayabileceği ve kullanabileceği modellere sığdırmaya çalışmaktadır (Mitchell, 2006).

Şekil 5.1'de, yapay zeka alanının alt bölümleri gösterilmektedir.

Makine öğrenmesi bilgisayar bilimi alanında olmasına rağmen, geleneksel hesaplama yaklaşımlarından biraz farklıdır. Algoritmalar, bilgisayarların geleneksel hesaplamadaki problemleri hesaplamak veya çözmek için kullandıkları açıkça programlanmış talimat türleridir. Bunun yerine, makine öğrenmesi yöntemleri, bilgisayarların veri girdilerini eğitmesine ve belirli bir aralıkta değerler üretmek için istatistiksel analizler kullanmasına izin vermektedir. Makine öğrenmesi, bilgisayarların veri girişine dayalı karar verme süreçlerini otomatize etmek için örnek verilerden modeller oluşturmasını sağlamaktadır (Mitchell, 2006).

Günümüzde çoğu teknoloji kullanıcısı makine öğreniminden yararlanmaktadır. Makine öğrenimi günlük hayatta birçok yerde kullanılmaktadır. Kullanım alanlarına örnek olarak yüz tanıma, sahtekârlık tespiti, görsel ve işitsel tanıma, trafik tahmini gibi örnekler verilebilmektedir (Müller, 2016).

Makine öğrenimi; hangi filmi seyredeceğimize, hangi yemeği sipariş edeceğimize ya da hangi ürünü satın alacağımıza yönelik otomatik öneriler, kişiselleştirilmiş çevrimiçi radyo, fotoğraflarımızdaki arkadaşlarımızı tanıma vb. teknikler kullanan çoğu modern internet sitesi ve uygulamada kullanılmaktadır. Facebook,

Amazon ve Netflix gibi karmaşık internet siteleri ziyaret edildiğinde göz gezdirilen hemen her noktada birden fazla makine öğrenimi modeli çalışmaktadır (Müller, 2016).

Şekil 5.1. Yapay Zeka ve Alt Alanları

5.1. Makine Öğrenmesi Yöntemleri

Makine öğrenmesi için çeşitli amaçta kullanılan Lojistik Regresyon (Logistic Regression), Rastgele Orman (Random Forest), Karar Ağacı (Decision Tree), Yapay Sinir Ağları (Artificial Neural Network), Torbalama (Bagging), Arttırma (Boosting) gibi birçok yöntem bulunmaktadır.

5.1.1. Lojistik Regresyon (Logistic Regression) yöntemi

Lojistik regresyon ilk olarak 1940'larda, ikili sonuçları ele almada sıradan en küçük kareler (OLS) regresyonunun sınırlamalarının üstesinden gelmek için alternatif bir teknik olarak önerildi ve 1980'lerin başlarında istatistiksel paketlerde yerini aldı. O zamandan beri, sosyal bilimlerde, yükseköğretimde eğitim araştırmalarında, epidemiyolojide, medikal uygulamalarda, tıp alanında, ekonomide, tarımda, veterinerlik ve taşıma sahalarında lojistik regresyon yaygın olarak kullanılmaktadır (Soyer, 2020).

Lojistik regresyon, bağımlı değişkenin ikili (binary), sıralı (ordinal) ve çoklu (multinominal) kategorilerde bulunduğu durumlarda bağımsız değişken veya değişkenler ile olan ilişkisinin belirlenmesinde kullanılan parametrik olmayan istatistiksel bir yöntemdir. Bir başka ifadeyle, bağımlı değişken üzerinde etki sahibi olan bağımsız değişkenlerin etki derecesini belirlemeyi sağlayan bir yöntemdir. Bağımsız değişkenlerin aldığı değerlere göre bağımlı değişkenlerin beklenen değerinin olasılık olarak elde edildiği lojistik regresyon, bağımlı değişkenin değerlerinin sınıflama ve atama işlemiyle en uygun gruba atanmasına yardımcı olan bir yöntemdir (Soyer, 2020).

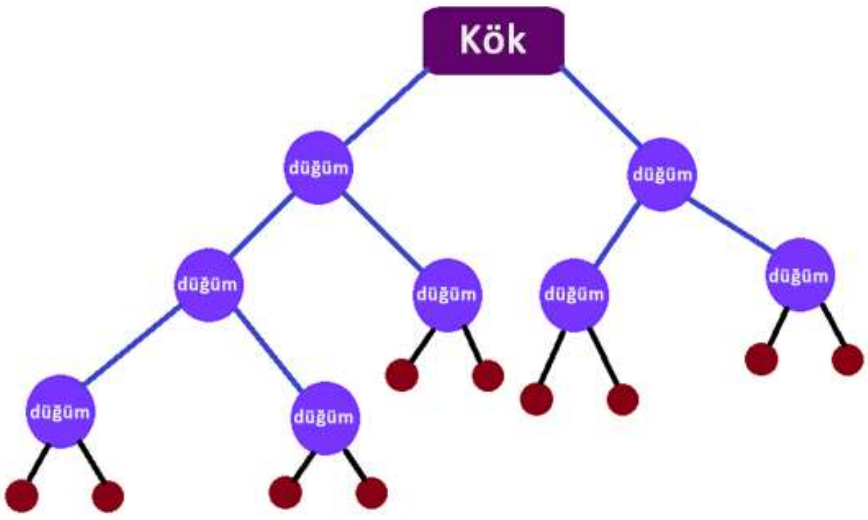
Lojistik regresyon algoritmasında İkili Lojistik Regresyon (Binary Logistic Regression), Sıralı Lojistik Regresyon (Ordinal Logistic Regression) ve Çok Terimli Lojistik Regresyon (Multinomial Logistic Regression) olmak üzere 3 farklı yöntem bulunmaktadır. Bu yöntemlerden hangisinin kullanılması gerektiği bağımlı olan değişkenin kategori sayısına ve ölçeğine göre belirlenmektedir. Lojistik regresyonda bağımlı değişken olarak var/yok, evet/hayır, başarılı/başarısız gibi örnekler verilebilmektedir. Sıralı değişken olarak az/orta/çok, 14-18 yaş aralığı/19-26 yaş aralığı gibi örnekler verilebilmektedir. Çok terimli değişken olarak çocuk-yetişkin-yaşlı, ortaokul-lise-üniversite, çocuk-yetişkin-yaşlı gibi örnekler verilebilmektedir.

Bağımlı değişkenin yapısından kaynaklı olarak varsayımların sağlanamadığı durumlarda gözlemleri verilerin yapısında bulunan olası gruplara atamak için kullanılan lojistik regresyon algoritması, diskriminant (ayırma) algoritması ve kümeleme algoritmalarına alternatif olarak uygulanmaktadır. Lojistik regresyonda bağımlı değişken kategorik olduğundan dolayı çoklu doğrusal regresyon algoritmasında geçerli olan varsayımlar burada geçerliliğini yitirmektedir. Bağımsız değişkenlerin normal dağılım göstermesi gerektiği ve çoklu doğrusal bağlantı içermemesi gibi varsayımların sağlanamadığı durumlarda da lojistik regresyon algoritması kullanılabilmektedir. Yani, lojistik regresyon algoritması için herhangi bir varsayımı sağlama gerekliliği bulunmamaktadır (Soyer, 2020).

5.1.2. Karar ağacı (Decision Tree) yöntemi

Karar ağaçları denetimli bir sınıflandırma yaklaşımı içerir. Fikir, bir kök ve düğümlerden (dalların bölündüğü konumlar), dallardan ve yapraklardan oluşan sıradan ağaç yapısından geldi. Benzer şekilde, daireleri temsil eden düğümlerden bir karar ağacı oluşturulmaktadır. Devamında dallar, düğümleri birbirine bağlayan bölümler tarafından temsil edilmektedir. Karar ağaçları kökten başlayarak, aşağı doğru hareket eder. Genellikle soldan sağa doğru çizilmektedir. Karar ağacının başlangıç düğümüne kök düğüm adı verilmektedir. Zincirin bittiği düğüm ise yaprak düğüm olarak adlandırılmaktadır. Karar ağacının içindeki her düğümden yani yaprak düğüm olmayan bir düğümden toplamda iki veya daha fazla dal uzatılabilir. Bir düğüm belirli bir özelliği temsil etmektedir. Dallar bir dizi değeri temsil etmektedir. Bu değer aralıkları, verilen karakteristiğin değerler kümesi için bir bölme noktası görevi görmektedir (Ali vd., 2012).

Şekil 5.2'de bir karar ağacının yapısını açıklamaktadır.



Şekil 5.2. Karar Ağacı (Decision Tree)

5.1.3. Yapay sinir ağları (Artificial Neural Network) yöntemi

İnsan beyninde tüm kararlar, vücudumuzda doğal olarak bulunan ve temel yapı taşı olan nörondan oluşan sinir ağları aracılığıyla alınır. Biyolojik nöron, diğer

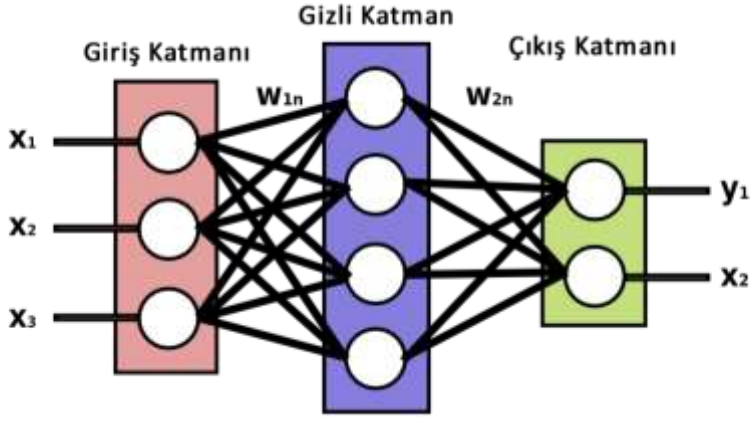
nöronlardan veri almaktan sorumlu olan dendritlerden oluşur, hücre gövdesi içeriden alınan tüm girdileri toplar ve verileri nöronun dışındaki akson aracılığıyla verir. Tüm iletişim ve işlemler, önceki nörondan gelen dendritler ve akson arasında bir bağlantı noktası olan sinapslar aracılığıyla elektrik sinyallerinde gerçekleştirilir (Shafiq vd., 2019).

Benzer şekilde, yapay nöron girdilerinde $x_1\ x_2\ \dots\ x_n$ her nöron tarafından alınır ve karar verme için toplama ve aktivasyon / transfer fonksiyonu için eklenir. Çıktı katmanı, sinir ağı tarafından alınan ortak karar temelinde dışarıdan alınır (Shafiq vd., 2019). Çıktı katmanı Şekil 5.3'te verilen yapay sinir ağının en sağ bölümünde gösterilmektedir.

Benzer şekilde, yapay nöron girdilerinde $x_1, x_2 \dots x_n$ her nöron tarafından alınır ve karar verme için toplama ve aktivasyon / transfer fonksiyonu için eklenir. Çıktı, tüm sinir ağı tarafından alınan ortak karar temelinde dışarıdan alınır. Nöron; giriş katmanı, gizli katmanlar ve çıkış katmanı olarak 3 ana parçadan oluşur. Giriş değeri x_i nörona uygulandığında, ağırlık eklenir ve sonuçlanır:

$$o_k = f(\sum w_i x_i + b_j) \tag{5.1}$$

Denklem (5.1)'te bir yapay sinir ağının çıktı değerini formüle etmek için tanımlanmaktadır. $\boldsymbol{w_i}$ her girdi verisi ($\boldsymbol{x_i}$) için ağırlığı ve $\boldsymbol{b_j}$ her algılayıcı için sapmayı temsil etmektedir. Yapay sinir ağının çıktısı ise $\boldsymbol{o_k}$ olarak tanımlanmaktadır.



Şekil 5.3. Yapay Sinir Ağı

5.2. Makine Öğrenmesinde Topluluk Öğrenmesi Yöntemleri

Son yıllarda makine öğrenmeye artan ilgi ile birlikte çoklu sınıflandırma sistemleri olarak adlandırılan topluluk sistemleri popülerlik kazanmaya başlamıştır. Makine öğrenmesinde topluluk öğrenmesi (Ensemble Learning), elde edilen sonuçlar neticesinde performansın arttırılması amacıyla farklı öğrenme algoritmalarının birlikte kullanılmasını ifade etmektedir. Topluluk öğrenmesi yöntemlerinin gerçek hayattaki uygulamaları, problemleri çok yönlü ele almaktadır ve problem çözümüne ulaştırması ile ön plana çıkmaktadır. Bu konudaki çalışmalar; bir hastalığın teşhisinde doktorların ortak tanı koyması, bir kişinin kendi hastalığı ile ilgili birden fazla doktordan görüş alması, bir ürünün satın alınmadan önce kullanıcı yorumlarının okunması, bir firmanın stratejik karar alma aşamasında kurullarına danışması ve sonuçların ortak oylama ile alınması örnek olarak gösterilebilir. Sonuç olarak, topluluk öğrenimi, doğru kararı alabilmek için, karar verme sürecinde farklı düşünceleri de değerlendirerek nihai düşünceyi karara bağlamaktadır (Yangın, 2019).

Topluluk öğrenmesi yöntemleri, aynı problemi çözmek için çoklu öğrenicileri eğitmektedir. Bir öğreniciyi eğitim verisinden kurmaya çalışan sıradan öğrenici yaklaşımlarının aksine, topluluk öğrenmesi yöntemleri bir dizi öğreniciyi oluşturmayı ve onları birleştirmeyi amaçlamaktadır (Zhou, 2012).

Bir topluluk yapısı, temel öğreniciler olarak adlandırılan bir grup öğreniciyi içermektedir. Temel öğreniciler genellikle eğitim verilerinden karar ağaçları, sinir ağları veya diğer tür öğrenme algoritmaları olabilen temel öğrenme algoritmaları tarafından oluşturulmaktadır. Çoğu topluluk metotları, homojen temel öğreniciler üretmek için temel bir öğrenici algoritması kullanır, diğer bir ifadeyle, homojen topluluklara giden aynı tür öğrenicilerden oluşur. Benzer şekilde farklı tür öğrenicilerin oluşturmuş olduğu heterojen öğreniciler de mevcuttur. Yani, heterojen öğrenicileri üretmek için çoklu öğrenme algoritmaları kullanılmakta ve bu durum tek bir öğrenme algoritması ile olmamaktadır. Bir topluluğun genelleme yeteneği 15 genellikle temel öğreniciden daha güçlüdür. Topluluk yöntemleri, rastgele bir tahmine göre, daha iyi olan zayıf

öğrenicilerinden net tahminlemeler yapabilen güçlü öğrenicilere yükseltebilme özelliğine sahiptir. Bu sebeple, temel öğreniciler zayıf öğreniciler olarak da adlandırılır. Araştırmacılar genellikle zayıf öğreniciler üzerinde çalışmakta ve performansı zayıftan güçlüye yükseltmek için güçlü algoritmalar tasarlamaktadırlar. Bunun sonucu olarak Adaboost, Bagging, vb. gibi topluluk yöntemlerinin doğuşuna ve güçlü bir öğrenici olmak için, zayıf öğrenicilerin neden ve nasıl arttırabileceği hakkında teorik anlayış geliştirilmesine yol açmıştır (Zhou, 2012).

5.2.1. Torbalama (Bagging) yöntemi

Bagging (Torbalama) algoritması, L. Breiman tarafından önerilen topluluk sınıflandırması için bir yöntemdir (Breiman, 1994). Bagging terimi, Bootstrap (Önyükleme) ve Aggregating (Toplama/Birleştirme) teriminin birleşiminden oluşan kısaltılmış bir ifadedir.

Bagging, bir grup tahminciyi kullanarak ve birleştirerek daha iyi bir başarı elde etmeyi amaçlamaktadır. Bagging yönteminde, bazı bağımsız hatalar yapan bazı tahmincilere ihtiyaç vardır. Bagging algoritmasına göre Bootstrap yöntemini kullanarak örnekleme yoluyla eğitim veri kümesinden k adet alt küme oluşturur. Daha sonra her bir alt kümeye olacak şekilde k adet sınıflandırıcı oluşturulur. Sınıflandırıcılar en son toplanarak tek bir yerde birleştirilir. Algoritmanın tahmin adımı ise farklı k öğrenici için çoğunluk oylamasına göre tahmin edilmektedir.

$$\widehat{f_{bag}} = \widehat{f_1}(x) + \widehat{f_2}(x) + \cdots + \widehat{f_b}(x) \tag{5.2}$$

Denklem (5.2)'de sol taraftaki terim toplu bir tahminciyi temsil etmektedir. Sağ taraftaki terimler ise bireysel tahmincilerdir.

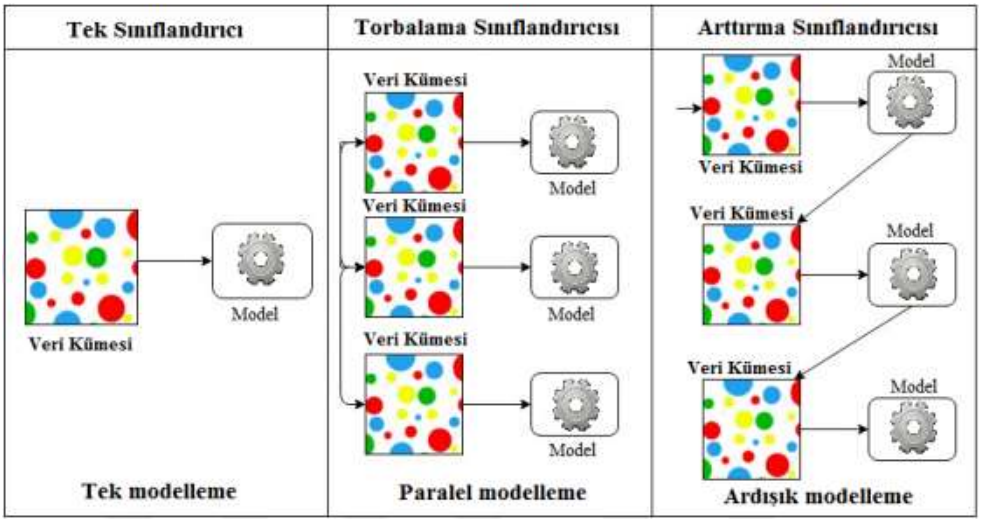
5.2.2. Arttırma (Boosting) yöntemi

Arttırma (Boosting), Schapire tarafından 1989 yılında geliştirilen bir yöntemdir. Freud ve Schapire ve Friedman son çalışmalarında bu algoritmayı daha da ileriye

götürmüşlerdir. Arttırma yöntemi, yavaş öğrenmeye dayalı sıralı bir yöntemdir ve hatadan öğrenmeyi amaçlamaktadır. Gradient Boosting Machine, Adaptive Boosting (AdaBoost), Extreme Gradient Boosting (XGBoost), Kategorik Arrtırma (CatBoost) gibi birçok farklı topluluk öğrenmesi yöntemini kapsamaktadır. Bu algoritmalar, yüksek hassasiyetli modeller oluşturmak için düşük hassasiyetli birkaç modeli birleştirmeyi amaçlamaktadır. Genel prensip olarak her iterasyonda elde edilen modeller, belirli kurallar çerçevesinde birleştirerek güçlü bir model oluşturmaktır (Coşkun, 2020).

Arttırma (Boosting) yöntemi sürecinde; ilk önce eğitim veri kümesinden rastgele örneklemler oluşturulur. Bu örneklem için bir sınıflandırıcı eğitilir ve tüm eğitim veri kümesi test edilir. Her bir örneklem tahmini için hata hesaplanır. Eğer örneklem yanlış sınıflandırılmışsa, o örneklem için ağırlık arttırılır ve başka bir örneklem oluşturulur. Sistemden yüksek doğruluk elde edilene kadar bu işlemler tekrarlanır (Coşkun, 2020).

Torbalama ve arttırma algoritmalarının karşılaştırılması Şekil 5.4'te verilmiştir (Yangın, 2019). Eğitim esnasında Torbalama (Bagging) algoritmasında veri kümesi ile model oluşturulurken paralel, Arttırma (Boosting) algoritmasında ise sıralı bir biçimde gerçekleştirilmektedir.



Şekil 5.4. Bagging ve Boosting yöntemlerinin karşılaştırılması (Yangın, 2019)

5.2.2.1. Gradient Boosting yöntemi

Gradient Boosting, Friedman tarafından 2001 yılında tanıtılan güçlü bir makine öğrenme tekniğidir. Gradient Boosting, bir tahmin modelinde, genellikle karar ağaçları gibi zayıf tahmini modellerin topluluk formunu üreten, regresyon ve sınıflandırma problemleri için bir makine öğrenme tekniği olarak tanımlanmaktadır. Gradient Boosting, temeli boosting tekniklerine dayanmaktadır. Boosting yöntemine dayandığından sırayla çok sayıda zayıf öğreniciyi inşa etmek ve onları karmaşık bir modele dâhil etmeyi amaçlamaktadır (Yangın, 2019).

5.2.2.2. Adaptive Boosting (AdaBoost) yöntemi

Freund ve Schapire tarafından 1995 yılında geliştirilen Arttırma yöntemi, arttırma yöntemini kullanarak zayıf öğrenme algoritmalarının performanslarının arttırılmasını amaçlayan bir sınıflandırma algoritmasıdır. Öznitelik uzayındaki her öznitelik üzerinde bir zayıf sınıflandırıcıyı eğitmektedir ve doğrusal olarak birleştirerek güçlü sınıflandırıcıların elde edilmesini sağlamaktadır (Coşkun, 2020).

Öğrenme veri kümesi üzerinde her modelin bir ağırlık katsayısı bulunmaktadır. Her öznitelik için zayıf olan sınıflandırıcı eğitilmektedir. Sınıflandırıcı eğitildikten sonra yanlış sınıflandırılan modelin ağırlık katsayıları arttırılmaktadır. Doğru sınıflandırılan örneklerin ağırlık katsayıları ise azaltılmaktadır. Aşırı Arttırma algoritması, optimizasyon işlemlerini içermektedir. Optimizasyon işlemleri sonrasında oluşan son sınıflandırma, zayıf sınıflandırma algoritmalarının bir lineer kombinasyonudur. Bu algoritmanın diğer arttırma algoritmalarından farkı büyük veri kümesine ihtiyaç duymamaktadır (Coşkun, 2020).

5.2.2.3. Extreme Gradient Boosting (XGBoost) yöntemi

Uzun ismi Extreme Gradient Boosting olarak adlandırılmaktadır. XGBoost yöntemi gradyan arttırma ve karar ağacı tekniklerine dayanan bir makine

öğrenme yöntemidir. XGBoost yönteminin orijinal hali Friedman tarafından ilk olarak 2002 yılında geliştirilmiştir (Yangın, 2019).

XGBoost yöntemi, hem sınıflandırma hem de regresyon modelleri için yüksek doğruluk oranına sahip bir yöntemdir. Ayrıca diğer algoritmalardan 10 kat daha hızlı olup performansı iyileştiren ve aşırı uyum ya da aşırı öğrenmeyi azaltan bir dizi düzenleme içerir. Bu sayede daha iyi performans elde etmeyi başarmaktadır. XGBoost yöntemi, parametre almadan kendi içerisinde çapraz doğrulama yaparak modelin doğruluğunun maksimum olmasını sağlayan bir yöntemdir (Ekiz, 2019).

XGBoostalgoritması yüksek oranda esneklik sağlamaktadır. Bu esneklikle birlikte modele yeni bir boyut katan kriterin değerlendirilmesi mümkün hale gelmektedir. Ayrıca en uygun amaç fonksiyonun tanımlanmasını da mümkün hale getirmektedir. Ek olarak XGBoost algoritması ağaç budaması (tree pruning) yaparken, belirtilen "max_depth" parametre değerine kadar dallara ayırma işlemini yapmaktadır. Bu işlemden sonra ağacı geriye doğru budamaya başlamaktadır. Budama esnasında olumlu yönde katkısı olmayan dallanmaları kaldırır (Yangın, 2019). XGBoost algoritması bu özellikleri sayesinde karar ağacı çok büyümemektedir. Bu yüzden algoritmanın performansı diğer arttırma yöntemlerine göre çok daha yüksek olmaktadır (Coşkun, 2020).

5.2.2.4. Light Gradient Boosting Machine (LightGBM) yöntemi

LightGBM, 2017 yılında Microsoft'un düzenlemiş olduğu DMTK projesi kapsamında geliştirilen bir gradyan arttırma algoritmasıdır. Hafif Gradyan Arttırma Makinesi (LightGBM); regresyon, sınıflandırma ve sıralama amacı ile kullanılmaktadır. Dağıtık sistemlerle çalışan, özel tanımlanmış hata fonksiyonunu desteklemektedir. Sürekli değer içeren öznitelikleri kategorik değere dönüştürerek modelin eğitim süresini kısaltmakta ve kaynak kullanımını azaltmaktadır. LightGBM yaprak tabanlı bir yapıda çalışmaktadır. Bu sayede ağaç öncelikli bir şekilde yatay büyüme sağladığından ve ağaç derinliğinin büyük

sayılara çıkmadığından dolayı aşırı öğrenmenin önüne geçebilmektedir (Gümüştaş, 2019).

5.2.2.5. CatBoost yöntemi

CatBoost, kategorik özellikleri desteklemektedir ve gradyan arttırma algoritmalarından bir tanesidir. Hem GPU (Grafik İşlemci Birimi) hem de CPU (Merkezi İşlem Birimi) uygulamalarında kullanılabilmektedir. CatBoost, kategorik özniteliklere sahip verileri sınıflandırmak için kullanılmaktadır. (Tozlu, 2019). CatBoost yöntemi Yandex firması tarafından geliştirilen açık kaynaklı (open-source) yazılım kütüphanesidir.

5.2.3. Rastgele orman (Random Forest) yöntemi

Rastgele Orman yöntemi (Breiman, 2001), eğitim veri setinin önyükleme örnekleri üzerinde büyüyen ve ağaç yapımı sürecinde rastgele özellik seçimini içeren bir karar ağaçları topluluğu algoritmasıdır. Yapılan tahminler, tek tek tüm ağaçların tahminlerinin toplanmasıyla yapılır. Rastgele Orman, karar ağaçlarının toplulukları olduğu için, tek ağaca dayalı sınıflandırıcılara göre kesinlikle önemli performans artışı sergiler. Rastgele Orman, büyük boyutlu verilerin işlenmesinde iyi bir seçim olarak kabul edilmektedir ancak dengesiz eğitim veri kümesi durumunda da zarar görmektedir. Rastgele Orman, genel hata oranını en aza indirir ve bu nedenle dengesiz veri kümesi durumunda daha yüksek toplam doğruluk, bazen azınlık sınıfının gerçek tahminini zayıflatır (Idris ve Khan, 2012).

Rastgele Orman yöntemi, işlemsel olarak regresyon ve sınıflandırmayı kolaylıkla ele almaktadır. Eğitme ve tahminleme işlemlerinde nispeten daha hızlıdır. Sadece birkaç parametre ayarına bağlı olması, çok boyutlu problemlerde rahatlıkla kullanılabilmektedir. Bu sebeplerden dolayı oldukça fazla tercih edilmektedir.

Rastgele Orman yöntemi, istatistiksel olarak özniteliklerin önem seviyeleri, sınıf ağırlığı, görselleştirme, aykırı değerlerin tespiti, denetimsiz öğrenme gibi birçok ek özelliği bulunmaktadır. Bu ek özellikler Rastgele Orman algoritmasını tercih

edilebilir yapmaktadır. Ayrıca bu yöntem tek bir veri kümesi içerisinde en iyi özniteliği aramaktadır. Bu yüzden düşük varyans ve daha yüksek sapma değerinin oluşmasına sebep olmaktadır ve bunun sonucunda ağaç çeşitliliğinin daha büyük olmasına imkân sağlamaktadır (Yangın, 2019).

Rastgele Orman yöntemi (Breiman, 2001), Torbalama (Bagging) yönteminin genişletilmiş haldir. Rastgele Orman algoritması öznitelik seçimi işlemini rastgele yapmaktadır. Torbalama ise rastgele yapmamaktadır. İki algoritma arasındaki en büyük fark olarak öznitelik seçimi denilebilmektedir (Coşkun, 2020).

Telekomünikasyon veri kümeleri normalde daha yüksek derecede çarpıklıktan mustariptir, bu nedenle Rastgele Orman bazen kayda değer bir performans sergilemekten mustariptir (Idris ve Khan, 2012).

5.3. Değerlendirme Ölçütleri

Ayrılan Müşteri Analizi gibi çeşitli Veri Madenciliği ve Makine Öğrenimi yöntemlerinin kullanıldığı analizlerde önerilen sistem için değerlendirilmesi gereken bir veya birden fazla ölçüt belirlenmesi gerekmektedir.

Doğruluk (accuracy) ölçüsü her sınıfı eşit derecede önemli gördüğü için, nadir sınıfın çoğunluk sınıfından daha ilginç olduğu düşünülen dengesiz veri kümelerini analiz etmek için uygun olmayabilir. İkili sınıflandırma için, nadir sınıf genellikle pozitif sınıf olarak, çoğunluk sınıfı ise negatif sınıf olarak ifade edilir. Bir sınıflandırma modeli tarafından doğru veya yanlış tahmin edilen örnek sayısını özetleyen matrise, karışıklık matrisi denir (Tan vd., 2014). Karışık matrisi Şekil 5.5'te gösterilmektedir.

Şekil 5.5. Karışıklık Matrisi

Değerlendirme ölçütü olarak Doğruluk Oranı (Accuracy Rate), Geri Çağırma (Recall), Hassasiyet (Precision), Özgünlük (Specificity), Dengelenmiş Doğruluk Oranı (Balanced Accuracy Rate), F1 Skoru (F1 Score), ROC Eğrisinin Altında Kalan Alan Değeri (ROC-AUC) gibi birçok ölçüt kullanılabilmektedir.

5.3.1. Doğruluk oranı (Accuracy Rate)

Doğruluk oranı; kurulan modelin ne kadar doğru çalıştığını, doğru pozitif ve doğru negatifin diğer değerlere oranını temsil etmektedir. Doğruluk oranı, Denklem (5.3)'teki gibi hesaplanarak bulunmaktadır.

$$\text{Doğruluk Oranı} = \frac{\text{TP+TN}}{\text{TP+FP+FN+TN}} \tag{5.3}$$

5.3.2. Geri çağırma (Recall)

Geri çağırma ölçütü, duyarlılık (sensivity) olarakta bilinmektedir. Doğru pozitif tahmin edilen ilgili örneklerin, gerçek tüm pozitif miktarına oranıdır. Sonuçlar ne kadar eksiksiz sorusuna cevap verir. Denklem (5.4)'teki gibi hesaplanarak bulunmaktadır.

$$\text{Geri Çağırma} = \frac{\text{TP}}{\text{TP+FN}} \tag{5.4}$$

5.3.3. Hassasiyet (Precision)

Hassasiyet, bazen pozitif tahmin değeri (positive predictive value) olarakta bilinir, doğru pozitif tahminlerin, pozitif tahminlere oranıdır. Arama sonuçları ne kadar geçerli sorusuna cevap verir. Hassasiyet, Denklem (5.5)'teki gibi hesaplanarak bulunmaktadır.

$$\text{Hassasiyet} = \frac{\text{TP}}{\text{TP+FP}} \tag{5.5}$$

5.3.4. Özgünlük (Specificity)

Özgünlük, negatif olarak sınıflandırılan gerçekten olumsuz durumların oranıdır; dolayısıyla, sınıflandırıcınızın olumsuz durumları ne kadar iyi tanımladığının bir ölçüsüdür. Aynı zamanda gerçek negatif oran (true negatif rate) olarak da bilinir. Özgünlük, Denklem (5.6)'daki gibi hesaplanarak bulunmaktadır.

$$\text{Özgünlük} = \frac{\text{TN}}{\text{FP+TN}} \tag{5.6}$$

5.3.5. Dengelenmiş doğruluk oranı (Balanced Accuracy Rate)

Dengelenmiş doğruluk oranı, bir ikili sınıflandırıcının ne kadar iyi olduğunu değerlendirirken kullanabileceği bir ölçüdür. Özellikle sınıflar dengesiz olduğunda, yani iki sınıftan biri diğerinden çok daha sık göründüğünde kullanışlıdır. Denklem (5.7)'deki gibi hesaplanarak bulunmaktadır.

$$\text{Dengelenmiş Doğruluk Oranı} = \frac{\text{Geri Çağırma+Özgünlük}}{2} \tag{5.7}$$

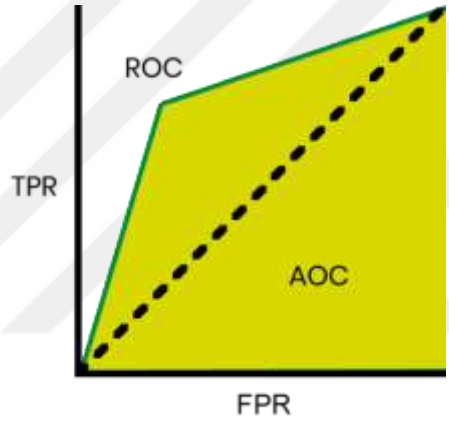
5.3.6. F1 Skoru (F1 Score)

F1 skoru, hassasiyet ve geri çağırmanın harmonik ortalamasıdır. Testlerin doğruluğunun bir ölçütüdür. Denklem (5.8)'deki gibi hesaplanarak bulunmaktadır.

$$\text{F1 Skoru} = 2x\left(\frac{\text{Hassasiyet x Geri Çağırma}}{\text{Hassasiyet+Geri Çağırma}}\right) \quad (5.8)$$

5.3.7. ROC eğrisinin altında kalan alan değeri (ROC-AUC)

Bir alıcının çalışma karakteristiği (ROC) eğrisi, bir sınıflandırıcının gerçek pozitif oranı ile yanlış pozitif oranı arasındaki dengeyi göstermek için grafiksel bir yaklaşımdır. Bir ROC eğrisinde, doğru pozitif oranı y ekseni boyunca çizilir ve yanlış pozitif oranı x ekseninde gösterilir (Tan vd., 2014). ROC-AUC ölçüsü ise ROC eğrisinin altında kalan alanı temsil etmektedir. ROC-AUC ölçüsü Şekil 5.6'da gösterilmektedir.



Şekil 5.6. ROC eğrisinin altında kalan alan değer grafiği

6. UYGULAMA

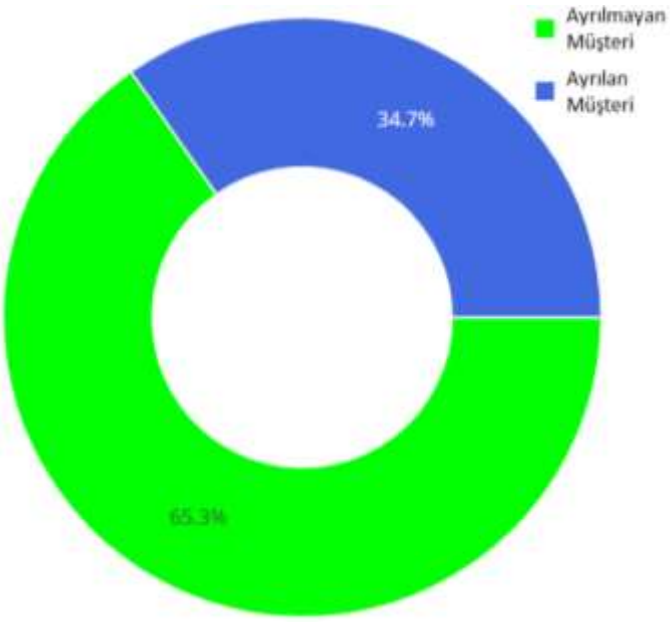
Bu bölümde, Bölüm 2'de anlatılan ayrılan müşteri analizi için Bölüm 3 ve 4'de anlatılan veri madenciliği ve makine öğrenmesi yöntemleri kullanılarak cell2cell veri kümesi üzerine uygulanmasıyla elde edilen sonuçların analizi anlatılmaktadır.

6.1. Geliştirme Ortamı

Bu çalışmaya ait önerilen modeller, 2.60 GHz CPU, 16 GB RAM özelliklerini taşıyan Windows 10 işletim sistemine ait bir bilgisayarda geliştirilmiştir. Ayrıca çalışma ortamında Python yazılım dili ile geliştirilmiştir ve Python 3.8.6 versiyonu kullanılmıştır.

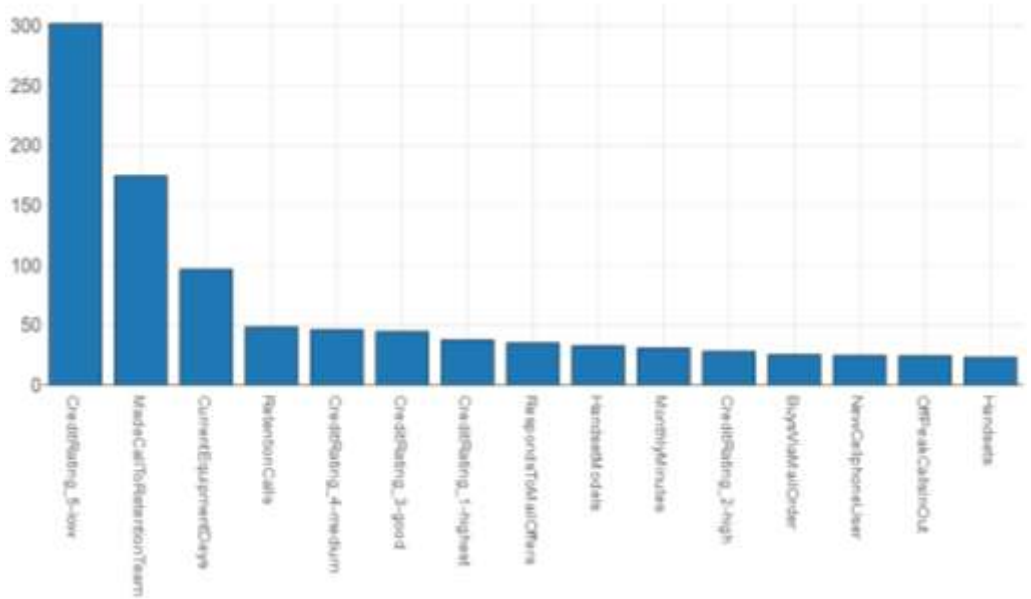
6.2. Veri Kümesi

Bu çalışmada cell2cell telekomünikasyon firması tarafından oluşturulan veri kümesi kullanılmıştır (Kaggle, 2018). Veri kümesinde 51.047 abonenin bilgisi bulunmaktadır. Ayrıca abonelerin 33.335'i yani %65.3'ü hizmet almaya devam ederken 17.712'i yani %34.7'si aldığı hizmeti sonlandırmıştır. Bu yüzden veri kümesinde bir sınıf dengesizliği bulunmaktadır. Bu konu "Veri Aşırı Örnekleme" başlığı altında detaylı olarak ele alınmıştır. Veri kümesinde, aboneliğinden ayrılan ve ayrılmayan müşteri oranı Şekil 6.1'de grafik üzerinde gösterilmiştir.



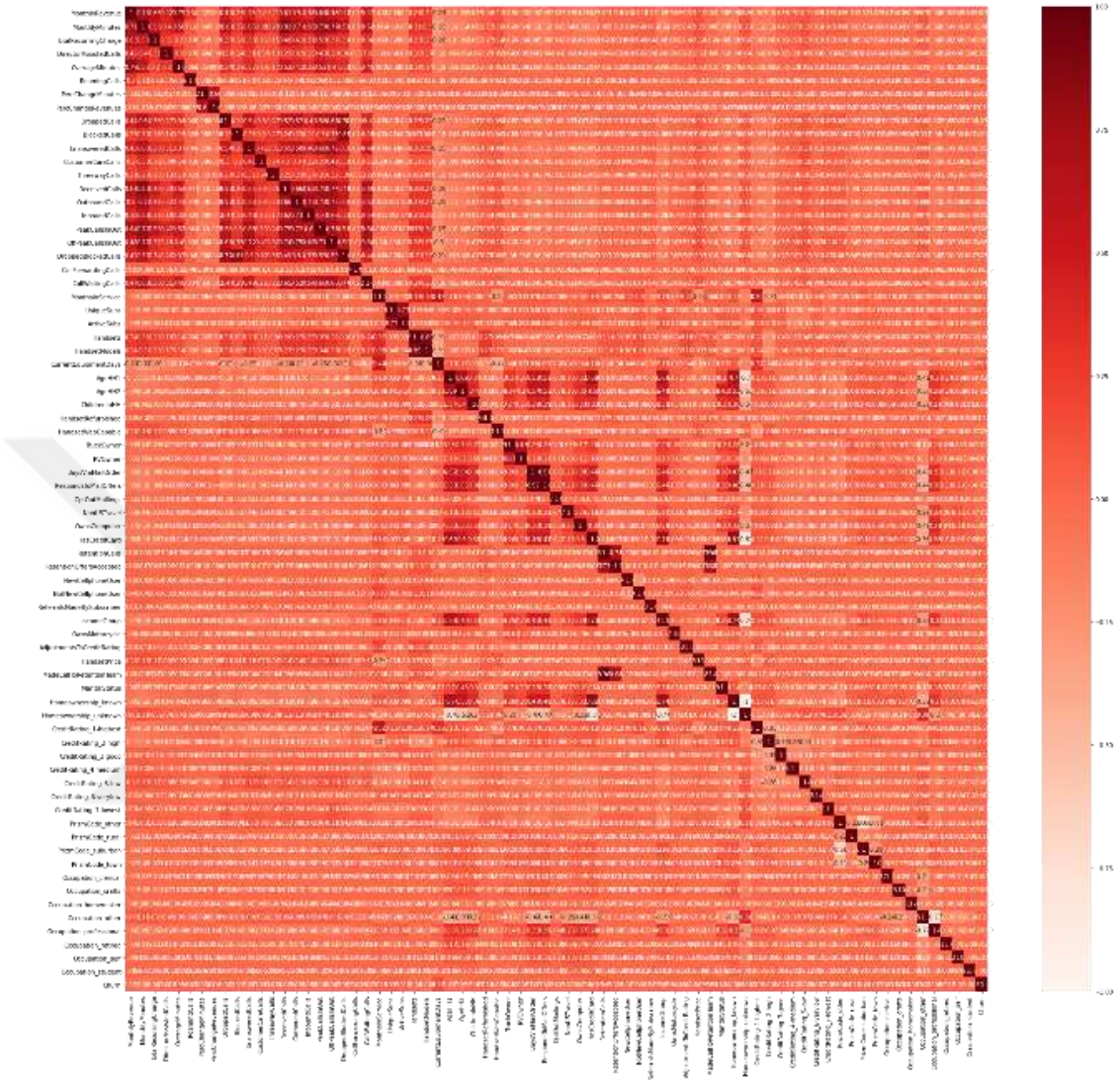
Şekil 6.1. Aldığı hizmetin aboneliğinden ayrılan & ayrılmayan müşteri Oranı

Her abonenin 35 sayısal, 23 kategorik özniteliği olmak üzere toplamda 58 adet öznitelik vardır ve eksik veriler bulunmaktadır. Sayısal özniteliklerin tümü oransal ölçek türündedir. Kategorik öznitelikler ise sembolik (nominal) ölçek türündedir. Veri kümesinin en seçici 15 özniteliği Univariate Feature Selection yöntemi kullanılarak Şekil 6.2'de grafiksel gösterilmiştir.



Şekil 6.2. Univariate Feature Selection yöntemi ile 15 en seçici öznitelikler

Ayrıca özniteliklerin arasındaki ilişkiyi tespit edebilmek için Pearson Correlation Coefficient yöntemi kullanılarak analiz edilmiştir ve Şekil 6.3'te gösterilmiştir.



Şekil 6.3. Pearson Correlation Coefficient yöntemi ile özniteliklerin birbiriyle olan ilişkisi

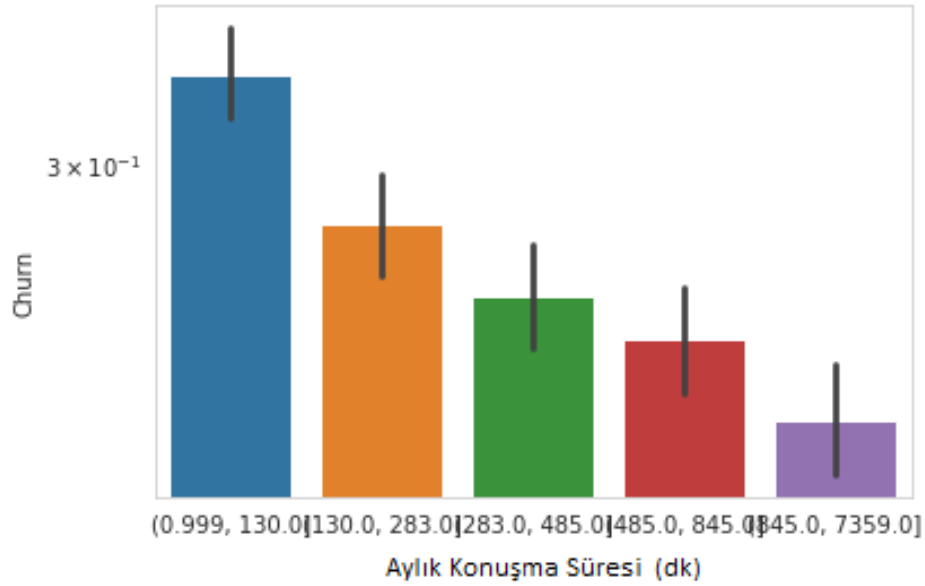
Veri kümesinde müşterilere ait birçok veri bulunmaktadır.

Müşterinin almış olduğu hizmete dair veriler olarak; müşteri hizmetleri ile konuşma miktarı, toplam paket aşımı miktarı, kllanım dakikasındaki yüzdelik değişim, ortalama cevapsız, reddedilen ve konferans arama sayıları, aylık konuşma süresi gibi birçok veri bulunmaktadır.

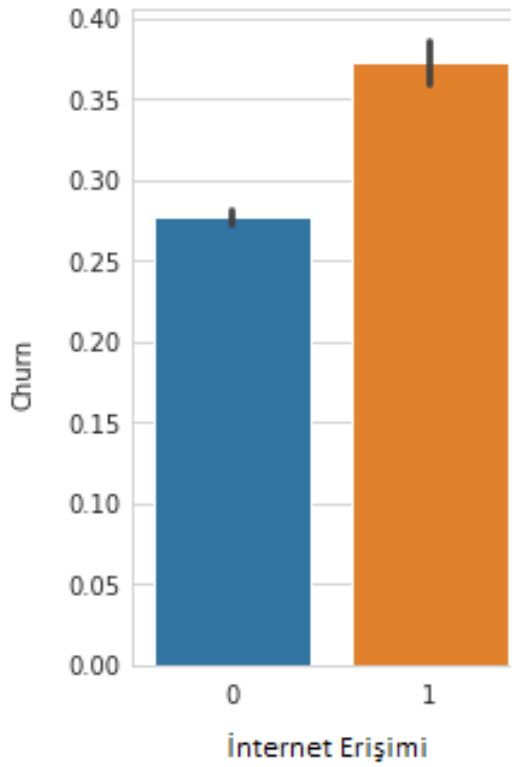
Müşterinin kişisel bilgileri olarak; müşterinin yaşadığı bölge, evli olup olmadığı, varsa çocuk sayısı ve çocuklarının yaş aralıkları, kredi notu gibi birçok veri bulunmaktadır.

Müşterinin maddi bilgileri olarak; aylık geliri, telefonunun fiyatı, kaç adet arabasının olduğu, kaç adet evi olduğu gibi birçok veri bulunmaktadır.

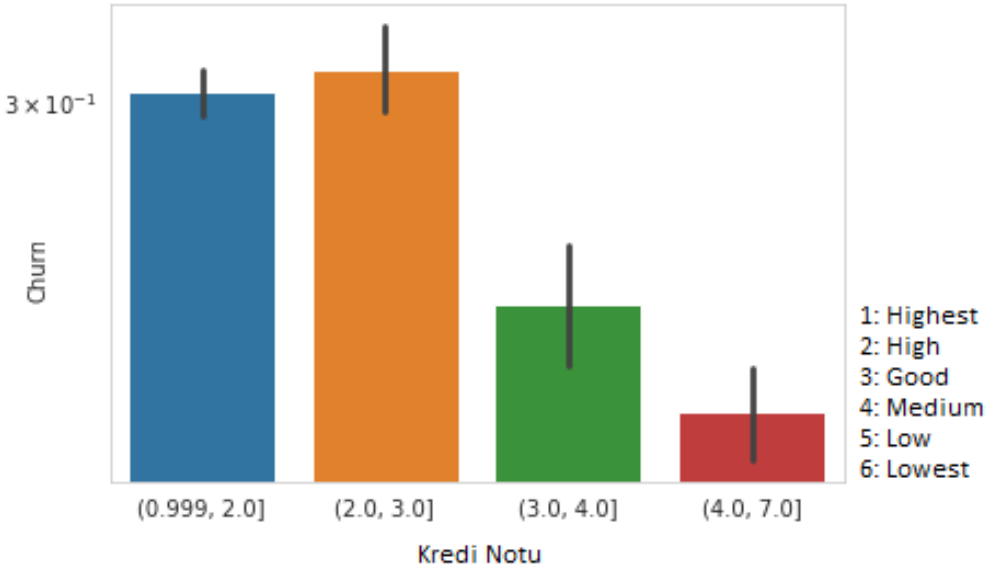
Veri kümesinde ayrılan müşterilerin verilerine göre bir inceleme yapılmıştır. Aylık Konuşma Süresi, İnternet Erişiminin Olup Olmaması, Kredi Notu, Paket Aşımı Miktarı ve Telefonunun Yenilenip Yenilenmemesi gibi birçok öznitelik için Churn olma istatistikleri görsel bir şekilde ortaya konulmuştur. Bunlar sırasıyla Şekil 6.4., Şekil 6.5., Şekil 6.6., Şekil 6.7. ve Şekil 6.8. olarak verilmiştir.



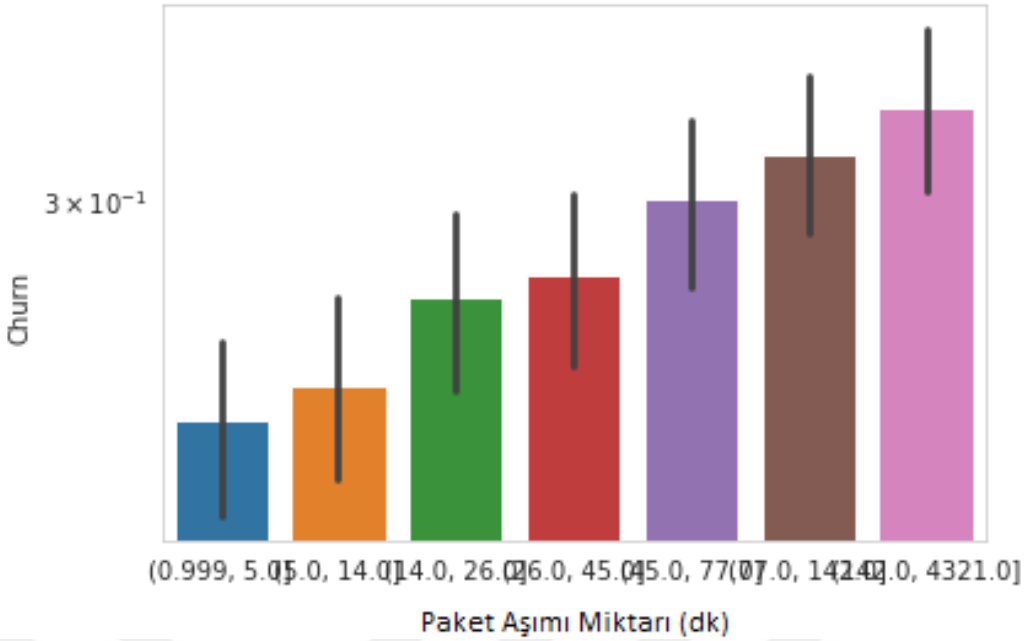
Şekil 6.4. Aylık konuşma süresine göre ayrılan müşteri dağılımları



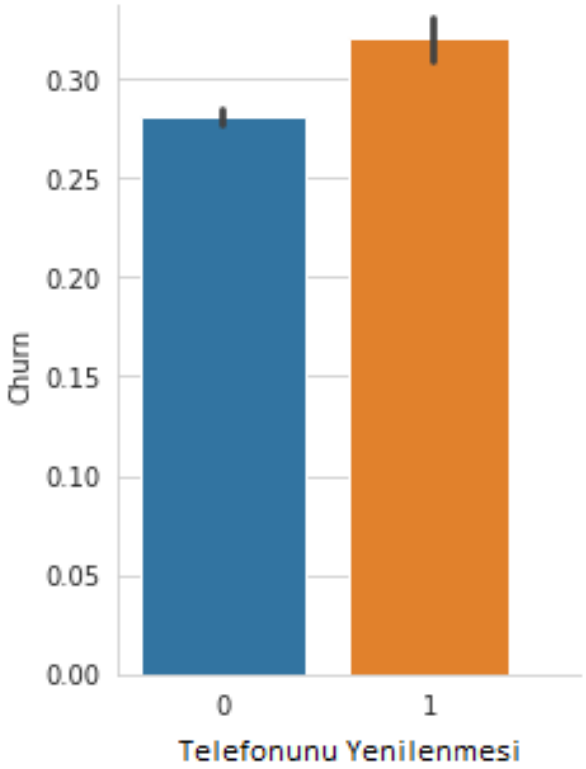
Şekil 6.5. İnternet erişimine göre ayrılan müşteri dağılımları



Şekil 6.6. Kredi notuna göre ayrılan müşteri dağılımları



Şekil 6.7. Paket aşımı miktarına göre ayrılan müşteri dağılımları



Şekil 6.8. Telefonunun yenilenmesi göre ayrılan müşteri dağılımları

6.3. Veri Ön İşleme

Veri kümesinde benzersiz (unique) değer taşıyan öznitelikler, analize katkıda bulunamayan öznitelikler bulunmaktadır. Analize katkıda bulunamayan öznitelikler için 4.2.1.'de anlatılan öznitelik indirgeme yöntemi uygulanmıştır. Dolayısıyla ilk adımda CustomerID, ServiceArea ve Handsets olan öznitelikler veri setinden çıkartılmıştır.

Veri kümesinin %2'si kayıp veya bilinmeyen değerlerden oluşmaktadır. Bu değerlerin tümü sayısal özniteliklerin içerisinde bulunmaktadır. Eksik verilerin tamamlanması için özniteliklerin aritmetik ortalamaları alınarak bahsi geçen eksik değerlere atanmıştır.

Diğer bir ön işleme safhası olan kısım ise kategorik değerlerin sayısallaştırılmasıdır. Çoğu makine öğrenme algoritması kategorik verileri kullanmamaktadır. Bazı öznitelikler "Yes" ve "No" değerlerinden oluşmaktadır. Buna benzer öznitelikler 4.2.3.1.'de anlatılan Label Encoding yöntemi ile kodlanarak 1 ve 0 olarak değiştirilmiştir.

Ayrıca ikiden fazla değere sahip olan kategorik öznitelikler için 4.2.3.2.'de anlatılan One-Hot Encoding yöntemi ile kullanarak kodlanmıştır. Bu yöntem sonunda veri kümesinde bulunan tüm öznitelikler sayısal özniteliklere dönüştürülmüştür.

Bir sonraki adım olarak veri kümesine, normalizasyon yöntemi uygulanmıştır. Bahsi geçen normalizasyon yöntemi, sayısal değerler arasındaki yüksek varyansların hesaplarda birbirlerini etkilememesi için kullanılmıştır. Ayrıca sayısal öznitelikler birbirleriyle karşılaştırılmak istendiğinde ortak bir sayı sisteminde bulunması son derece önemlidir. Normalizasyon için en yaygın olan ve 4.2.2.1.'de anlatılan Min-Max Normalizasyon yöntemi kullanılmıştır.

Son olarak ise veri kümesinde bir sınıf dengesizliği mevcuttur. Veri kümesindeki sınıf dengesizliğini ortadan kaldırmak amacıyla 4.2.5.1.'de anlatılan SMOTE

(Synthetic Minority Oversampling Technique) ve 4.2.5.2.'de anlatılan ADASYN (Adaptive Synthetic Sampling Method) aşırı örnekleme yöntemleri kullanılmıştır. Aşırı örnekleme yöntemleri kullanıldığında toplam müşteri sayısı artarak 66.670'e yükselmiştir.

Bu çalışmada azınlık olan sınıfın verilerini, aboneliğini sonlandırmış olan müşterilerin, sayısı arttırılmıştır ve veri kümesini dengeye getirmek amacıyla SMOTE ve ADASYN olmak üzere toplam da 2 adet aşırı örnekleme tekniği kullanılmıştır.

6.4. Önerilen Modeller

Bu çalışmada, toplamda 2 adet tahmin modeli önerilmiştir. Yüksek boyutlu veri kümesinde müşteri verilerine; normal ön işleme, öznitelik seçimi, aşırı örnekleme ve normalizasyon yöntemleri uygulanan 2 adet tahmin modeli önerilmiştir. Önerilen modellerin arasındaki fark; aşırı örnekleme yöntemlerinin farklı olmasından ve öznitelik seçimi yöntemini sadece ADASYN aşırı örnekleme yöntemi ile kullanılmasından dolayıdır.

ADASYN ile önerilen tahmin modeli, Univariate Feature Selection tekniğiyle birlikte çalıştığında daha doğru sonuçlar alındığı gözlendiği için ADASYN ile Önerilen Tahmin Modeli'ne eklenmiştir. SMOTE ile önerilen tahmin modelinde ise aynı işlem yapılmıştır fakat Univariate Feature Selection tekniği ile belirgin bir iyileştirme elde edilemediği için SMOTE ile önerilen tahmin modelinden çıkarılmıştır.

Her iki model içinde ön işleme safhasının ilk adımı olan gereksiz, önemsiz öznitelikler veri kümesinden çıkartılmıştır. Çıkartılan öznitelikler için kullanıcının kimlik numarası, müşterinin hizmet alanı gibi örnekler verilebilir. Bu ön işleme safhası, Univariate Feature Selection öznitelik seçimi ile karıştırılmamalıdır.

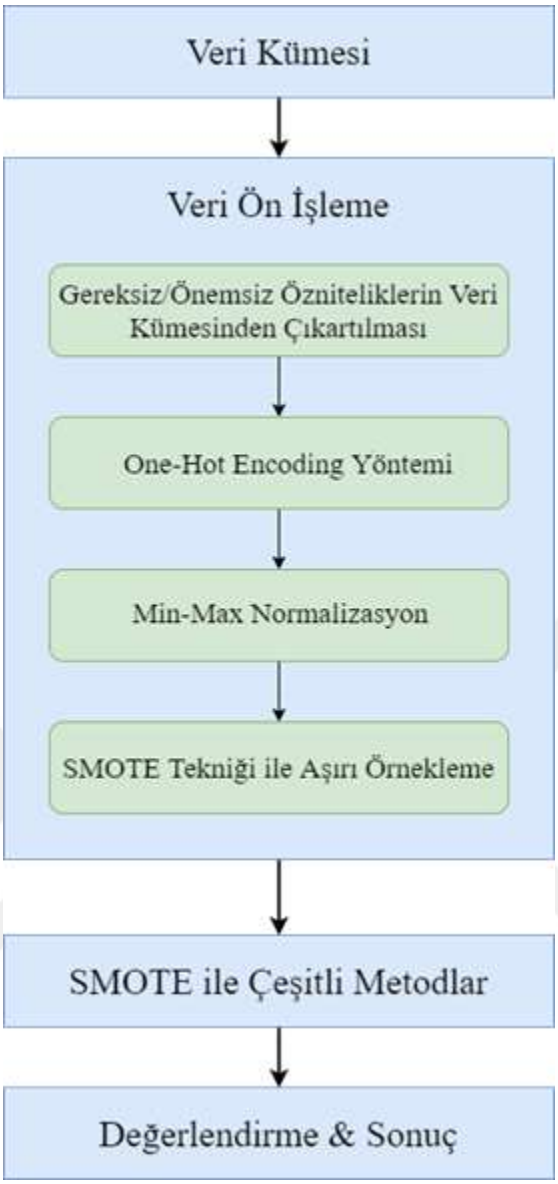
Birinci tahmin modelinde Univariate Feature Selection öznitelik seçimi yapılmamakla beraber SMOTE aşırı örnekleme yöntemi kullanılmıştır. İkinci tahmin modelinde ise Univariate Feature Selection yöntemi ile birbiriyle ilişkisi olmayan öznitelikler veri kümesinden çıkartılmıştır ve aşırı örnekleme yöntemi olarak ADASYN yöntemi kullanılmıştır.

ADASYN ve SMOTE algoritmalarının temel farkı; ADASYN algoritmasının ana fikri yoğunluk dağılımını kullanmaktır. SMOTE algoritması ise n adet orijinal azınlık sınıfı için aynı sayıda sentetik numune üretmektedir (He vd., 2008).

Önerilen modeller eğitilmeden önce K-Fold Cross Validation tekniği kullanılarak veri kümesi 10 farklı alt kümeye bölünmüştür. Böylelikle veri kümesi eğitim ve test verileri olarak ayrılmıştır.

6.4.1. SMOTE ile önerilen tahmin modeli

SMOTE algoritması aşırı örnekleme algoritmalarına göre en yaygın olarak kullanılan algoritmalardan biridir. Bu modelde, öznitelik seçimi yapılmamış olup Min-Max Normalizasyonu ve SMOTE aşırı örnekleme yöntemi ve yöntemi kullanılmıştır. Bu çalışmada, önerilen diğer tahmin modeline göre doğruluk oranı (accuracy rate) biraz daha iyi performans vermiştir. İlgili model için en iyi sonucu Light Gradient Boosting Machine algoritması vermiştir. Bu model Şekil 6.9. olarak tanımlanabilir.



Şekil 6.9. SMOTE tekniği ile önerilen tahmin modeli

6.4.2. ADASYN ile önerilen tahmin modeli

Bu modelde ise Min-Max Normalizasyonu ve ADASYN aşırı örnekleme yöntemi kullanılmasının yanı sıra Univariate Feature Selection algoritması yardımıyla birbiri ile ilişkili olan öznitelikler kullanılarak analiz edilmiştir ve toplamda 56 adet öznitelik kullanılmıştır. Bu çalışmada, önerilen diğer tahmin modeline göre özgünlük (specificity) ve hassasiyet (precision) daha iyi performans vermiştir. Bu model Şekil 6.10. olarak tanımlanabilir.

Şekil 6.10. ADASYN tekniği ile önerilen tahmin modeli

7. BULGULAR VE TARTIŞMA

Bu çalışmada önerilen tahmin modelini bulmak için makine öğrenmesi algoritmaları, derin öğrenme algoritmaları ve topluluk sınıflandırma teknikleri kullanılmış ve her biri için eğitim yapılmıştır. Bunlar sırasıyla Decision Tree, Artificial Neural Network, Logistic Regression, Bagging ve Boosting algoritmalarıdır.

Çalışmada aşırı örnekleme yöntemlerini kullanan ve kullanmayan tahmin modelleri olmak üzere 2 farklı model tipi oluşturulmuştur. Aşırı örnekleme yöntemlerini kullanan tahmin modelleri, kullanmayan modellere göre çok daha iyi bir başarı göstermiştir. Bu yüzden çalışmada önerilen modellere Önerilen Tahmin Modeller'i denmiştir. Önerilen tahmin modellerinde SMOTE ve ADASYN aşırı örnekleme yöntemleri kullanılmıştır.

7.1. Aşırı Örnekleme Yöntemi Kullanmadan Eğitilen Tahmin Modelleri

Bu bölümde veri kümesine SMOTE, ADASYN gibi aşırı örnekleme yöntemleri kullanılmadan sadece sınıflandırma algoritmaları kullanarak eğitilen modellerdir. Aşırı örnekleme kullanılmadan eğitilen modellerin başarısı, aşırı örnekleme yapılan modellere göre başarı elde edememiştir. Bu yüzden çalışmada ilgili tahmin modellerine yer verilmemiştir. Ayrıca bu modellere ait başarı oranları Çizelge 7.1' de gösterilmektedir.

Çizelge 7.1. Aşırı örnekleme yöntemi kullanmadan eğitilen tahmin modelleri

	Doğruluk Oranı (%)	Geri Çağırma (%)	Hassasiyet (%)	Özgünlük (%)	Dengelenmiş Doğruluk Oranı (%)	F1 Skoru (%)	ROC-AUC (%)
LightGBM	72.3	12.8	59.2	96.4	54.6	21.1	68.2
Hist Gradient Boosting	72.4	12.7	59.8	96.6	54.6	20.9	68.2
CatBoost	72.6	15.7	59.4	95.6	55.7	24.9	68.6
XGBoost	71.6	19.7	52.1	92.6	56.2	28.6	66.5
AdaBoost	71.6	9.9	54.5	96.6	53.2	16.7	66.1
Bagging	70.1	17.4	45.3	91.4	54.4	25.2	61.9
Logistic Regression	71.1	2.6	47.8	98.8	50.7	4.9	61.4
Decision Tree	62.0	35.7	35.0	72.8	54.2	35.3	54.2
Artificial Neural Network	69.4	7.2	40.3	95.4	51.3	12.2	61.2
Random Forest	65.4	37.9	39.5	76.5	57.2	38.7	61.9

7.2. Aşırı Örnekleme Yöntemi Kullanarak Eğitilen Tahmin Modelleri

Bu bölümde önerilen tahmin modelleri anlatılmaktadır. Veri kümesine SMOTE ve ADASYN aşırı örnekleme yöntemleri kullandıktan sonra sınıflandırma algoritmaları ile eğitilen modellerdir.

İki tahmin modeli içinde en iyi sonucu veren algoritmalar birbirine çok yakın sonuçlardan oluşmaktadır ve en iyi sonucu veren topluluk sınıflandırma tekniği olan Light Gradient Boosting Machine (LightGBM) olarak tespit edilmiştir.

İlgili veri kümesi için performans ölçütlerinin çıktılarına bakıldığında, topluluk sınıflandırma tekniklerinin temel sınıflandırma algoritmalarına göre daha iyi çalıştığı incelenmiştir. Genellikle her iki önerilen model için en iyi sonucu veren topluluk sınıflandırma algoritmaları Gradient Boosting teknikleri olmuştur.

SMOTE ile önerilen tahmin modelinde doğruluk oranı (accuracy rate), hassasiyet (precision), özgünlük (specificity), dengelenmiş doğruluk oranı (balanced accuracy rate) performans ölçütleri için en iyi sonucu veren LightGBM tekniğidir. Geri çağırma (recall) performans ölçütü için ise Decision Tree en iyi sonucu vermektedir. Performans ölçülerinin çıktıları Çizelge 7.2'de gösterilmektedir.

ADASYN ile önerilen tahmin modelinde doğruluk oranı (accuracy rate), hassasiyet (precision), özgünlük (specificity), dengelenmiş doğruluk oranı (balanced accuracy rate) performans ölçütleri için en iyi sonucu veren LightGBM tekniğidir. Geri çağırma (recall) performans ölçütü için ise Artificial Neural Network en iyi sonucu vermektedir. Performans ölçülerinin çıktıları Çizelge 7.3'te gösterilmektedir.

Çizelge 7.2. Birinci önerilen tahmin modeline ait performans çıktıları

	Doğruluk Oranı (%)	**Geri Çağırma (%)**	**Hassasiyet (%)**	**Özgünlük (%)**	**Dengelenmiş Doğruluk Oranı (%)**	**F1 Skoru (%)**	**ROC-AUC (%)**
LightGBM	80.2	64.9	93.5	95.5	80.2	76.6	86.9
Hist Gradient Boosting	80.0	64.8	93.0	95.1	80.0	76.4	86.8
CatBoost	79.2	66.8	88.9	91.7	79.2	76.3	86.1
XGBoost	78.4	63.5	90.6	93.4	78.4	74.6	85.2
AdaBoost	73.6	67.7	76.7	79.4	73.6	71.9	81.3
Bagging	77.7	65.2	87.0	90.5	77.8	74.6	82.3
Logistic Regression	59.2	62.0	58.6	56.3	59.1	60.2	62.6
Decision Tree	69.4	70.9	69.0	67.9	69.4	69.9	69.4
Artificial Neural Network	65.6	70.6	64.3	60.7	65.6	67.2	71.5
Random Forest	76.9	66.2	84.2	87.6	76.9	74.1	84.0

Çizelge 7.3. İkinci önerilen tahmin modeline ait performans çıktıları

	Doğruluk Oranı (%)	**Geri Çağırma (%)**	**Hassasiyet (%)**	**Özgünlük (%)**	**Dengelenmiş Doğruluk Oranı (%)**	**F1 Skoru (%)**	**ROC-AUC (%)**
LightGBM	80.3	65.0	93.9	95.7	80.4	76.8	87.1
Hist Gradient Boosting	80.0	64.9	93.2	95.2	80.1	76.5	86.9
CatBoost	79.3	67.1	89.0	91.6	79.4	76.5	86.2
XGBoost	78.6	63.6	91.0	93.6	78.6	74.9	86.3
AdaBoost	73.2	67.9	76.2	78.5	73.2	71.8	81.0
Bagging	77.9	65.4	87.3	90.2	77.8	74.8	82.3
Logistic Regression	57.8	58.5	52.9	57.2	57.8	55.5	60.7
Decision Tree	69.9	71.0	69.9	68.7	69.9	70.5	69.9
Artificial Neural Network	64.7	75.3	62.3	54.1	64.7	68.2	70.5
Random Forest	77.0	66.4	84.5	87.8	77.1	74.4	84.0

Bu çalışmada, oluşturulan model eğitilmeden önce cell2cell veri kümesine ait abone bilgileri veri ön işleme yöntemleri ile işlenmiştir. Veri kümesi büyük miktarda özniteliğe sahip olduğu için normalizasyon ve öznitelik seçimi yöntemleri uygulanmıştır. Ardından veri kümesinin sınıf dengesizliğini yok etmek için aşırı örnekleme yöntemleri kullanılmıştır ve dengelenmiş veri kümesi eğitilmeye hazır hale gelmiştir.

Eğitilmeye hazır olan veri kümesi için temel sınıflandırma ve topluluk sınıflandırma teknikleri kullanılarak en iyi sonucu veren model, önerilen tahmin modeli olarak seçilmiştir. Bu çalışma için çıktılar birbirine çok yakın çıkmıştır. Bu yüzden toplamda 2 adet önerilen tahmin modeli seçilmiştir.

Geliştirilen modeller arasında en iyi modeli değerlendirmek amacıyla doğruluk oranı, geri çağırma, hassasiyet, özgünlük, dengelenmiş doğruluk oranı, F1 skoru ve ROC AUC olmak üzere toplamda 7 adet performans ölçütü kullanılmıştır.

Birinci tahmin modeli için en iyi performans çıktıları Çizelge 7.4'te verilmiştir. İkinci tahmin modeli için en iyi performans çıktıları ise Çizelge 7.5'te verilmiştir.

Çizelge 7.4. Birinci tahmin modelinin performans çıktıları

Performans Ölçütü	**Performans Değeri**
Doğruluk Oranı	% 80.2
Geri Çağırma	% 70.9
Hassasiyet	% 93.5
Özgünlük	% 95.5
Dengelenmiş Doğruluk Oranı	% 80.2
F1 Skoru	% 76.6
ROC-AUC	% 86.9
Çalışma Zamanı	14.75 saniye

Çizelge 7.5. İkinci tahmin modelinin performans çıktıları

Performans Ölçütü	Performans Değeri
Doğruluk Oranı	% 80.3
Geri Çağırma	% 75.3
Hassasiyet	% 93.9
Özgünlük	% 95.7
Dengelenmiş Doğruluk Oranı	% 80.4
F1 Skoru	% 76.8
ROC-AUC	% 87.1
Çalışma Zamanı	15.34 saniye

Çalışmada kullanılan veri kümesi, erişimi herkese açık olan bir veri kümesidir. İlgili veri kümesi ile birden fazla çalışma bulunmaktadır. Bu çalışmayı benzer çalışmalarla karşılaştırmak amacıyla bir çizelge hazırlanmıştır. Bu karşılaştırma Çizelge 7.6'da gösterilmiştir.

Çizelge 7.6. Benzer çalışmalar ile karşılaştırılması

	Idris ve Khan, 2012	**Yıldız ve Albayrak, 2015**	**Jamil ve Khan, 2016**	**Amin ve diğerleri, 2019**	**Vijaya ve Sivasankar, 2018**	**İlgili Çalışma**
Öznitelik	76	172	78	100	172	58
Geri Çağırma	% 76.5	% 60.6	% 94.5	-	-	% 75.0
Hassasiyet	-	% 60.4	% 94.5	-	% 95.2	% 93.9
Özgünlük	% 74.6	-	-	-	% 98.3	% 95.7
Doğruluk Oranı	-	-	-	% 57.0	-	% 80.3
ROC-AUC Değeri	% 81.6	-	-	-	-	% 87.1
Eğitim Zamanı	-	-	60.6 saniye	-	-	15.34 saniye
Yöntem	mRMR, RotBoost	Information Gain, Chi Square, OneR, Naive Bayes, Logistic Regression	SMOTE, Grid Search, SVM, DT, RIP with Active Learning Based Approach	Distance based Sampling, Naive Bayes	RSFS, Bagging, Boosting, Random Subspace	Univariate Selection, Min-Max Normalization, One-Hot Encoding, SMOTE, ADASYN, Bagging, Boosting, Random Forest

8. SONUÇ VE ÖNERİLER

Kullanılan veri kümesinde ilk olarak analize katkısı olmayan öznitelikler çıkartılmıştır ve tanımsız olan bazı kayıtlar veri kümesinden kullanılmamak üzere kaldırılmıştır. İkinci olarak normalizasyon ve kodlama yöntemleri kullanılarak veri kümesi eğitilmek üzere belirli bir formata alınmıştır. Veri kümesine genel itibariyle bakıldığında sınıf dengesizliği mevcuttur. Şekil 1 (Abonelerin Bulunduğu Hizmetten Ayrılma Oranı)'de ki gibi gösterilmiştir. Veri kümesini dengeye getirmek amacıyla SMOTE ve ADASYN aşırı örnekleme teknikleri kullanılmıştır. Ardından belirli sınıflandırıcı yöntemleri ile veri kümesi eğitilmiştir. Eğitim aşamasında toplamda 56 adet öznitelik kullanılmıştır.

Geliştirilen tahmin modelleri arasından en iyi sonuç verenler belirlenmiştir. Toplamda 2 adet tahmin modeli önerilmiştir ve tahmin modelleri sırasıyla "ADASYN ile Önerilen Tahmin Modeli" ve "SMOTE ile Önerilen Tahmin Modeli" olarak adlandırılmıştır. Birinci tahmin modeli, ikinci tahmin modelinin performans çıktılarına göre daha iyi olduğu belirlenmiştir. Önerilen modeller, eğitim süresi olarak karşılaştırıldığında ise birbirine yakın sonuçlar vermiştir ve en iyi performansı "SMOTE ile Önerilen Tahmin Modeli" sağlamıştır.

ADASYN ve SMOTE ile önerilen tahmin modellerinin, çalışmada kullanılan veri kümesi gibi dengede olmayan (veri sınıflarının denk sayıda olmaması) veri kümelerinde başarılı olduğu (Chawla vd., 2002)'de ve (He vd., 2008)'de ifade edilmiştir.

Modellerin karşılaştırılması kullanılırken belirlenmiş olan Karmaşıklık Matrisi göz önüne alınmıştır ve performans ölçütü olarak Doğruluk Oranı (%80.3), Geri Çağırma (%75.3), Hassasiyet (%93.9), Özgünlük (%95.7), Dengelenmiş Doğruluk Oranı (%80.4), F1 Skoru (%76.8), ROC-AUC Oranı (%87.1) ve Çalışma Zamanı (14.75 saniye) kullanılarak iyi sonuçlar elde edilmiştir. İki tahmin modeli de incelendiğinde benzer sonuçlar verdiğini tespit edilmiştir. İki tahmin modelinde en iyi sonucu veren Light Gradient Boosting Machine sınıflandırıcısı olmuştur.

Ayrıca, iki modelde de Geri Çağırma performansı ölçütünün %75.3 ile en iyi sonucu verdiği yöntem Yapay Sinir Ağları (Artificial Neural Network) olmuştur.

Benzer çalışmalar ile karşılaştırıldığında 58 öznitelik kullanılarak diğerlerinin 172 öznitelikle başardığına çok yakın sonuçlar elde edildiği kanıtlanmıştır. Ayrıca eğitim süresi olarak karşılaştırıldığında eğitim süresini göze alan çalışmalardan 4 kata yakın daha performanslı olduğu tespit edilmiştir. 58 adet öznitelik kullanarak elde edilen bazı başarı ölçütlerinin, benzer çalışmalara göre çok daha iyi olduğu tespit edilmiştir.

Gelecekte, veri kümesi ham halinde sınıf dengesizliği olduğu için aşırı örnekleme yöntemlerini kullanarak daha dengeli bir hale getirilmesi sağlanabilir. Ayrıca öznitelik seçimi algoritmaları ve normalizasyon algoritmaları kullanılarak performans karşılaştırılması yapılabilir.

KAYNAKLAR

Abbasimehr H., Setak M., Tarokh M.J., 2014. A Comparative Assessment of the Performance of Ensemble Learning in Customer Churn Prediction, The International Arab Journal of Information Technology, 11(6), 599-606.

Ali J., Khan R., Ahmad N., Maqsood I., 2012. Random Forests and Decision Trees, IJCSI International Journal of Computer Science Issues International Journal of Computer Science Issues, 9(3).

AL-Shatnwai A.M., Altibbi M.F., 2020. Predicting Customer Retention using XGBoost and Balancing Methods, International Journal of Advanced Computer Science and Applications, 11(7), 704-712. 10.14569/IJACSA.2020.0110785.

Amin A., Obeidat F., Shah B., Adnan A., Loo J., Anwar S., 2019. Customer Churn Prediction in Telecommunication Industry Using Data Certainty, Journal of Business Research, 94, 290-301. 10.1016/j.jbusres.2018.03.003.

Ata F., 2018. Understanding Customer Value Using Data Mining Applications: A Case Study of An Insurance Broker, İstanbul Arel Üniversitesi, Fen Bilimleri Enstitüsü. Yüksek Lisans Tezi, İstanbul.

Bilgi Teknolojileri ve İletişim Kurumu, 2019. İletişim Hizmetleri İstatistikleri. Erişim Tarihi: 01.09.2020. https://www.btk.gov.tr/uploads/pages/iletisim-hizmetleri-istatistikleri/istatistik-2019-4-5ec51cf389753.pdf.

Breiman L., 1994. Bagging Predictors, Department of Statistics, University of California Berkeley, Technical Report No. 421.

Breiman L., 2001. Random Forests, Machine Learning, 45(1), 5-32.

Cell2Cell Dataset: Teradata Center For Customer Relationship Management at Duke University, Dec. 2018. Erişim Tarihi: 15.10.2020. https://www.kaggle.com/Jpacse/Datasets-for-Churn-Telecom.

Cengizci A.D., 2020. Otel İşletmelerinde Kayıp Müşteri Tahminlemesi, Akdeniz Üniversitesi, Sosyal Bilimler Enstitüsü. Yüksek Lisans Tezi, Antalya.

Chawla N.V., Bowyer K.W., Hall L.O., Kegelmeyer W.P., 2002. SMOTE: Synthetic Minority Over-Sampling Technique, Journal of Artificial Intelligence Research, 16, 321-357. 10.1613/jair.953.

Coşkun K., 2020. Ağ Saldırılarının Sınıflandırılmasında Karar Ağaçlarına Dayalı Arttırma (Boosting) Algoritmalarının Karşılaştırılması, Muğla Sıtkı Koçman Üniversitesi, Fen Bilimleri Enstitüsü, Yüksek Lisans Tezi, Muğla.

Çelik E., 2019. Customer Churn Analysis Based on Machine Learning by Using Data Mining Techniques in Telecommunication Sector, Yeditepe Üniversitesi, Sosyal Bilimler Enstitüsü. Yüksek Lisans Tezi, İstanbul.

Ekiz E., 2019. Makine Öğrenmesi Teknikleri ile Tahsilat Davranışı Tahmini: Telekomünikasyon Sektörü Örneği, İstanbul Teknik Üniversitesi, Fen Bilimleri Enstitüsü, Yüksek Lisans Tezi, İstanbul.

Freund Y., Schapire R.E., 1997. A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting, Journal of Computer and System Sciences, 55(1), 119-139. 10.1006/jcss.1997.1504.

Gold C., 2020. Fighting Churn With Data, O'reilly Media, 1. Baskı, Kaliforniya, Sebastopol.

Gümüştaş E., 2019, Kayıp Gözlem İçeren Dengesiz Veri Setlerinin Topluluk Öğrenme Algoritmaları İle Sınıflandırılması, Mimar Sinan Güzel Sanatlar Üniversitesi, İstanbul. - Cilt Yüksek Lisans Tezi.

Ha T.M., Bunke H., 1997. Off-line, Handwritten Numeral Recognition by Perturbation Method, Pattern Analysis and Machine Intelligence, 19(5), 535–539.

Hastie T., Tibshirani R., Friedman J., 2009. The Elements of Statistical Learning: Data Mining, Inference, and Prediction., Springer Series in Statistics, Stanford Üniversitesi, Kaliforniya.

He H., Bai Y., Garcia E.A., Li S., 2008. ADASYN: Adaptive Synthetic Sampling Approach for Imbalanced Learning, 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence), 1322-1328. 10.1109/ijcnn.2008.4633969.

Hosmer D.W., Lemeshow S., Sturdivant R.X., 2013. Applied Logistic Regression, 3. Baskı, WILEY.

Idris A., Khan A., 2012. Customer Churn Prediction for Telecommunication: Employing Various Various Features Selection Techniques and Tree Based Ensemble Classifiers, 15th International Multitopic Conference (INMIC), 23-27. 10.1109/inmic.2012.6511498.

Jamil S., Khan A., Churn Comprehension Analysis for Telecommunication Industry Using ALBA, 2016 International Conference on Emerging Technologies (ICET), 1-5. 10.1109/icet.2016.7813259.

Kadhim, K.A., 2018. Intrusion Detection Model Based on Data Mining and Machine Learning, Altınbaş Üniversitesi, Fen Bilimleri Enstitüsü. Yüksek Lisans Tezi.

Khan Y., Shafiq S., Abid A., Ahmed S., Safwan N., Hussain S., 2019. Customers Churn Prediction using Artificial Neural Networks (ANN) in Telecom Industry, International Journal of Advanced Computer Science and Applications, 10(9), 132-142. 10.14569/IJACSA.2019.0100918.

Li K.G., Marikannan B.P., 2020. Hyperparameters Tuning and Model Comparison for Telecommunication Customer Churn Predictive Models, 3rd Global Conference on Computing & Media Technology, 475-83.

Liu A.Y., 2004. The Effect of Oversampling and Undersampling on Classifying Imbalanced Text Datasets, University of Texas at Austin, USA.

Mitchell T.M., 2006. The discipline of machine learning, 9. Baskı, Carnegie Mellon University, School of Computer Science, Machine Learning.

Müller A.C., Guido S., 2016. Introduction to Machine Learning with Python: A Guide for Data Scientists, O'Reilly Media, Kaliforniya, Sebastopol.

Poel D.V.D., Lariviére B., 2004. Customer Attrition Analysis for Financial Services Using Proportional Hazard Models, European Journal of Operational Research 157, 196-217. European Journal of Operational Research 157.

Potdar K., Pardawala T., Pai C., 2017. A Comparative Study of Categorical Variable Encoding Techniques for Neural Network Classifiers, International Journal of Computer Applications, 175(4), 7-9. 10.5120/ijca2017915495.

Quackenbush J., 2002. Microarray data normalization and transformation, Nature genetics, 32, 496.

Safitri A.R., Muslim M.A., 2020. Improved Accuracy of Naive Bayes Classifier for Determination of Customer Churn Uses SMOTE and Genetic Algorithms, JOSCEX Journal of Soft Computing Exploration 1(1), 70-75.

Savaş S., Topaloğlu N., Yılmaz M., 2012. Veri Madenciliği ve Türkiye'deki Uygulama Örnekleri, İstanbul Ticaret Üniversitesi Fen Bilimleri Dergisi, 21, 1-23.

Sjarif N.N.A., Yusof M.R.M., Wong D.H., Yaakob S., Ibrahim R., Osman M.Z., 2019. A Customer Churn Prediction using Pearson Correlation Function and K Nearest Neighbor Algorithm for Telecommunication Industry, International Journal of Advances in Soft Computing & Its Applications, 11(2), 46-59.

Soyer M., 2020. Küresel Rekabet ile Ekonomik Özgürlük Arasındaki İlişkinin Lojistik Regresyon Yöntemiyle Analizi, Uşak Üniversitesi, Lisansüstü Eğitim Enstitüsü, Temmuz. - Cilt Yüksek Lisans Tezi

Subho M.R.H., Chowdhury M.R., Chaki D., Islam S., Rahman M.M., 2019. A Univariate Feature Selection Approach for Finding Key Factors of

Restaurant Business, 2019 IEEE Region 10 Symposium (TENSYMP), 605-610, Kolkata, India. 10.1109/TENSYMP46218.2019.8971127.

Tan P., Steinbach M., Kumar V., 2014. Introduction to Data Mining, Pearson Education Limited, UK.

Tan Y., Shuan L.H., Yan L.J., Guo X., 2017. Prediction on Customer Churn in the Telecommunications Sector Using Discretization and Naïve Bayes Classifier, International Journal of Advances in Soft Computing and its Applications, 9(3), 23-35.

Taşdemir Ş., Yanıktepe B., Güher A.B., 2018. The Effect on the Wind Power Performance of Different Normalization Methods by Using Multilayer Feed-Forward Backpropagation Neural Network, International Journal of Energy Applications and Technologies, 5, 131-139. 10.31593/ijeat.464210.

Tozlu İ., 2019. Simplifying Balance Sheet Adjustment Process in Commercial Loan Applications Using Machine Learning Methods, İstanbul Teknik Üniversitesi, İstanbul. - Cilt Yüksek Lisans Tezi

Vijaya J., Sivasankar E., 2018. Computing Efficient Features Using Rough Set Theory Combined with Ensemble Classification Techniques to Improve the Customer Churn Prediction in Telecommunication Sector, Computing, 100(8), 839–860. 10.1007/s00607-018-0633-6.

Wadikar D., 2020. Customer Churn Prediction, Technological University Dublin. 10.21427/kpsz-x829.

Yangın G., 2019. Xgboost ve Karar Ağacı Tabanlı Algoritmaların Diyabet Veri Setleri Üzerine Uygulaması, Mimar Sinan Güzel Sanatlar Üniversitesi, Fen Bilimleri Enstitüsü. - Cilt Yüksek Lisans Tezi.

Yıldız M., Albayrak S., 2015. Customer Churn Prediction in Telecommunication, 23nd Signal Processing and Communications Applications Conference (SIU), 256-259. 10.1109/siu.2015.7129808.

Zhou, Z.-H, 2012. Ensemble Methods Foundations and Algorithms, Chapman & Hall / CRC Press, U.S.

ÖZGEÇMİŞ

Adı Soyadı : FURKAN UYANIK

Eğitim Durumu

Lise : Ümraniye Anadolu Meslek Lisesi

Lisans : Sakarya Üniversitesi,
Bilgisayar ve Bilişim Bilimleri Fakültesi,
Bilgisayar Mühendisliği Bölümü

Lisans : Anadolu Üniversitesi,
Açıköğretim Fakültesi,
Yönetim Bilişim Sistemleri Bölümü

Yüksek Lisans : İstanbul Ticaret Üniversitesi,
Fen Bilimleri Enstitüsü,
Bilgisayar Mühendisliği Anabilim Dalı

Yayınları

Uyanık, F., Kasapbaşı, M.C., 2021. Telekomünikasyon Sektörü için Veri Madenciliği ve Makine Öğrenmesi Teknikleri ile Ayrılan Müşteri Analizi İstanbul Ticaret Üniversitesi, Fen Bilimleri Enstitüsü. 9(3), 172-191, 10.29130/dubited.807922