Düzce Üniversitesi Bilim ve Teknoloji Dergisi

Araştırma Makalesi

Telekomünikasyon Sektörü için Veri Madenciliği ve Makine Öğrenmesi Teknikleri ile Ayrılan Müşteri Analizi

Furkan UYANIK [a,*], Mustafa Cem KASAPBAŞI [b]

[a] *Bilgisayar Mühendisliği Bölümü, Fen Bilimleri Enstitüsü, İstanbul Ticaret Üniversitesi, İstanbul, TÜRKİYE*
[b] *Bilgisayar Mühendisliği Bölümü, Fen Bilimleri Enstitüsü, İstanbul Ticaret Üniversitesi, İstanbul, TÜRKİYE*
** Sorumlu yazarın e-posta adresi: furkanuyanikk@hotmail.com*
DOI: **10.29130/dubited.807922**

ÖZET

Son yıllarda şirketler arası rekabetin artmasıyla beraber aboneliğinden ayrılacak müşterilerin tahmin edilmesi oldukça önemli hale gelmiştir. Müşteri karmaşası analizi, veri madenciliği, makine öğrenmesi ve derin öğrenme gibi alanlarda 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. Çalışma Lojistik Regresyon (Logistic Regression), Karar Ağacı (Decision Tree), Yapay Sinir Ağları (Artificial Neural Network), 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. Çalışmada, 2 adet tahmin modeli önerilmiştir ve önerilen 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 (encoding) ve aşırı örnekleme gibi birçok yöntem kullanılmıştır. Performans ölçütü olarak Doğruluk Oranı (Accuracy Rate), Geri Çağırma (Recall), Hassasiyet (Precision) ve Özgünlük (Specificity), Dengelenmiş Doğruluk Oranı ve ROC Eğrisi Altındaki Alan (ROC-AUC) değeri kullanılmıştır. Performans ölçütlerine bakıldığında önerilen 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, Müşteri Karmaşası Tahmini, Veri Madenciliği, Makine Öğrenmesi, Tahmin, Örnekleme Algoritmaları, Sınıflandırma, Topluluk Sınıflandırması, Telekomünikasyon*

Churn Analysis for Telecommunication Sector with Data Mining and Machine Learning

ABSTRACT

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

Geliş: 08/10/2020, Düzeltme: 25/02/2013, Kabul: 04/03/2021

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, Data Mining, Machine Learning, Churn Prediction, Oversampling Algorithms, Classification, Ensemble Classification, Telecommunication*

I. 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 Müşteri Karmaşası Analizi (Churn Analysis) olarak geçmektedir [1].

Müşteri karmaşası 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.

Müşteri karmaşası 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 [2].

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.

II. LİTERATÜR ÇALIŞMASI

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.

A. M. AL-Shatnwai ve arkadaşı, 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 dahil diğer popüler sınıflandırmalardan daha iyi performans gösterdiğini gösterdi. Deneylerin ikinci bölümünde, yüksek hızda örnekleme yöntemleri farklı yüksek hızda 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 [3].

A. R. Safitri ve arkadaşı, 2020 yılında 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 [4].

D. Wadikar, çalışmasında bir Credit Union finans kurumunun müşteri kayıp analizini kesin olarak tahmin edebilen bir makine öğrenimi modeli geliştirmeyi amaçlamıştır. Sınıf dengesizliği sorununu, özellik seçimini ve müşteri kaybını verimli bir şekilde tahmin eden denetimli bir makine öğrenimi modeli oluşturmak için nicel ve tümdengelimle araştırma stratejileri kullanmıştır. 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 [5].

H. Abbasimehr ve diğerleri, 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 baz 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 [6].

J. Vijaya ve diğerleri, 2018 yılında 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 [7].
N. N. A Sjarif ve diğerleri, 2019 yılında 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 [8].

L.H. Shuan ve diğerleri, 2017 yılında yapmış oldukları çalışmada veri kümesini UCI Machine Learning Repository adlı kaynaktan elde etmiştir. 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 [9].

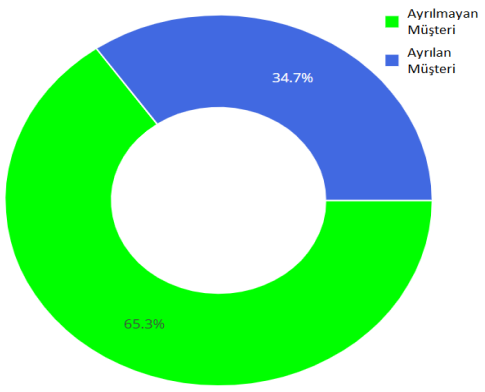
Guan Li, Koh ve diğerleri 2019 yılında yapmış oldukları çalışmada veri kümesini Kaggle Open Datasets adlı kaynaktan elde etmiştir. Ön işleme bölümünde veri kümesine normalizasyon uygulanmıştır ve ayrıca ROSE örnekleme yöntemini kullanarak veri kümesini dengelemiştir. Çalışmasında temel sınıflandırma 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 [10].

III. METODOLOJİ

Bu çalışmanın önerilen tahmin modelleri 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.

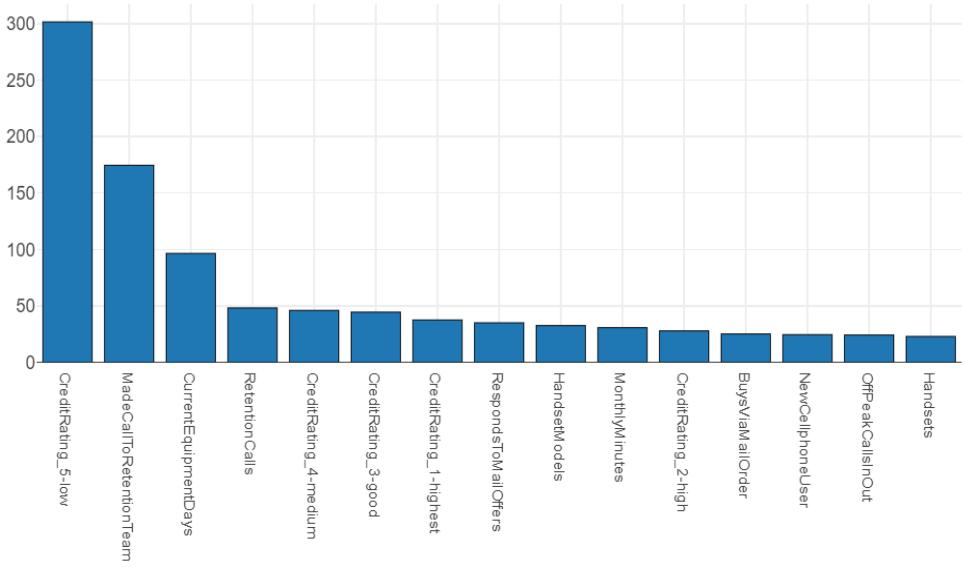
A. VERİ KÜMESİ

Bu çalışmada cell2cell veri kümesi kullanılmıştır [11]. 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.



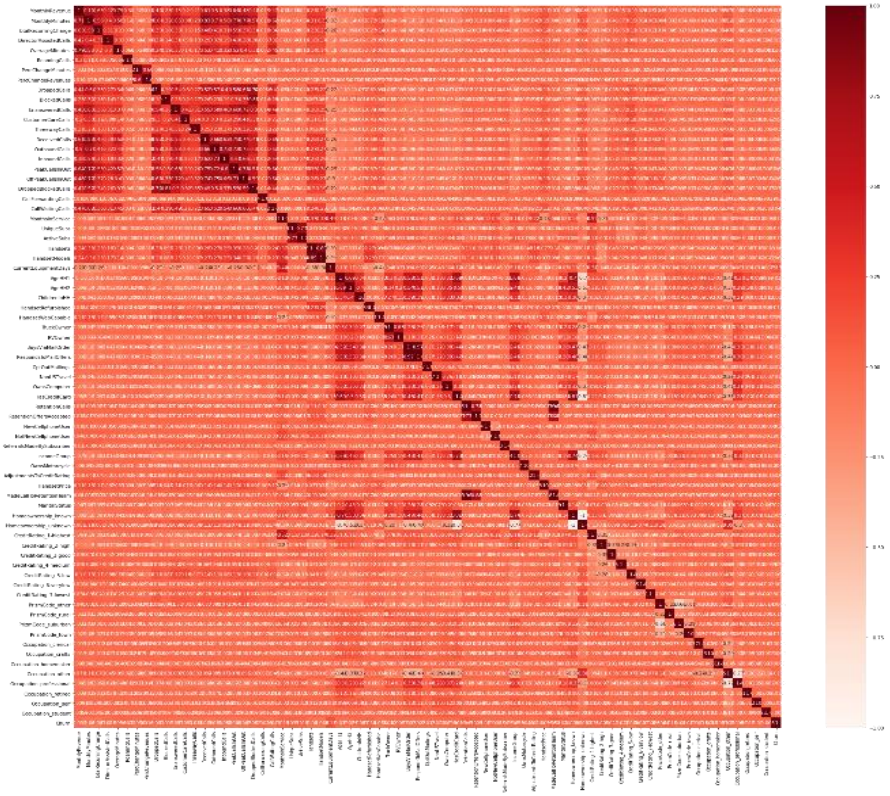
Şekil 1. *Aldığı Hizmetin Aboneliğinden Ayrılan & Ayrılmayan Müşteri Oranı*

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 nominal ölçek türündedir. Veri kümesinin en seçici 10 özniteliği Univariate Feature Selection yöntemi kullanılarak bulunmuştur ve Şekil 2'de gösterilmektedir.



Şekil 2. *Univariate Feature Selection yöntemi ile en seçici öznitelikler*

Ayrıca özniteliklerin arasındaki ilişkiyi tespit edebilmek için Pearson Correlation Coefficient yöntemi kullanılarak analiz edilmiştir ve bu analize göre renklendirilmiştir. Bu renklendirme Şekil 3'te gösterilmektedir.



Şekil 3. *Pearson Correlation Coefficient yöntemi ile özniteliklerin birbiriyle olan ilişkisi*

B. VERİ ÖN İŞLEME

Veri kümesinde benzersiz (unique) değer taşıyan öznitelikler, analize katkıda bulunamayan öznitelikler bulunmaktadır. 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 1 ve 0 olarak değiştirilmiştir.

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 Min-Max Normalization yöntemi kullanılmıştır.

Ayrıca ikiden fazla değere sahip olan kategorik öznitelikler için One-Hot Encoding yöntemi kullanılmıştır [12]. Bu yöntem sonunda veri kümesinde bulunan tüm öznitelikler sayısal özniteliklere dönüştürülmüştür.

Son olarak ise veri kümesinde bir sınıf dengesizliği mevcuttur. Veriyi dengeye almak amacıyla aşırı örnekleme algoritmalarından SMOTE ve ADASYN teknikleri kullanılmıştır.

B. 1. Min-Max Normalizasyon

Min-Max normalizasyon yöntemi, ham veri kümesinin minimum ve maksimum değerlerini bulur ve Denklem (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 [13].

$$N_i = \frac{x_i - x_{min}}{x_{max} - x_{min}} \quad (1)$$

B. 2. One-Hot Encoding

One-Hot Coding en yaygın kullanılan kodlama şemasıdır. Kategorik değişkenin her seviyesini sabit bir referans seviyesiyle karşılaştırır. Bir sıcak kodlama, n gözlem içeren ve d farklı değerli tek bir değişkeni, her biri n gözlemi olan d ikili değişkene dönüştürür. Her gözlem, ikili değişkenin varlığını (1) veya yokluğunu (0) gösterir [12].

C. VERİ AŞIRI 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, eğitim setindeki azınlık sınıfı üyelerinin sayısını artırmayı amaçlamaktadır. Aşırı örneklemenin avantajı, tüm üyeleri azınlık ve çoğunluk sınıflarından koruduğumuz için orijinal eğitim setinden hiçbir bilginin kaybolmamasıdır [14].

Bu çalışmada ise 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.

C. 1. SMOTE (Synthetic Minority Oversampling Technique)

SMOTE yaklaşımı, el yazısı karakter tanımada başarılı olan bir teknikten esinlenmiştir. "Veri alanı" yerine "özellik alanı" içinde çalışarak sentetik örnekleri daha az uygulamaya özgü bir şekilde üretiyoruz [15].

Azınlık sınıfı, her azınlık sınıfı örneği alınarak ve en yakın komşulardaki k azınlık sınıfının herhangi birini / tümünü birleştiren çizgi parçaları boyunca sentetik örnekler verilerek aşırı örneklenir. Gereken aşırı örnekleme miktarına bağlı olarak, en yakın k komşularından gelen komşular rastgele seçilir [15].

C. 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 [16].

ADASYN yaklaşımı veri dağılımları ile ilgili öğrenmeyi iki şekilde geliştirir: (1) sınıf dengesizliğinin getirdiği sapmanın azaltılması ve (2) sınıflandırma karar sınırının zor örneklere uyarlanabilir şekilde kaydırılması [16].

D. YÖNTEMLER

D. 1. Bagging (Torbalama) Algoritması

Bagging (Torbalama) algoritması, L. Breiman tarafından önerilen topluluk sınıflandırması için bir yöntemdir [17]. 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.

$$\widehat{f_{bag}} = \widehat{f_1}(x) + \widehat{f_2}(x) + \cdots + \widehat{f_b}(x) \quad (2)$$

Denklem (2)'de sol taraftaki terim toplu bir tahminciyi temsil etmektedir. Sağ taraftaki terimler ise bireysel tahmincilerdir.

D. 2. Boosting (Güçlendirme) Algoritması

Boosting, bir dizi "zayıf" sınıflandırıcıyı yüksek doğrulukla "güçlü" olacak şekilde birleştirmek için etkili bir yöntemdir. Boosting sadece deneysel olarak güçlü bir öğrenme algoritması değildir, aynı zamanda sınıflandırma hatasının üst sınırını en aza indirmede optimal olduğu da kanıtlanmıştır [18].

İlk ve en yaygın kullanılan Boosting algoritması Adaptive Boosting algoritması olarak literatüre geçmektedir.

D. 3. Pearson Correlation Coefficient (Pearson Korelasyon Katsayısı) 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.

$$r = \frac{\sum(\mathrm{x}-\mathrm{m}_x)(\mathrm{y}-\mathrm{m}_y)}{\sqrt{\sum(\mathrm{x}-\mathrm{m}_x)^2 \sum(\mathrm{y}-\mathrm{m}_y)^2}} \tag{3}$$

Denklem 3'teki 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 -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) olarak gösterilmiştir.

$$f(r) = \begin{cases} negatif\ korelasyon, zıt\ ilişkili,\ r < 0 \\ pozitif\ korelasyon, ilişkili,\ r > 0 \\ ilişki\ yok, r = 0 \end{cases} \tag{4}$$

Bu yöntemin görsel hali Şekil 3'te gösterilmektedir. Şekil 3'te bulunan 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.

D. 4. Univariate Feature Selection (Tek Değişkenli Öznitelik Seçimi) 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 [19].

Çalışmada öznitelik seçimi yöntemi olarak bu yöntem kullanılmaktadır. Ayriyeten özniteliklerin arasındaki ilişki hakkında bilgi almak amacıyla Pearson Korelasyon Katsayısı yöntemi kullanılmaktadır ve bu çalışmaya etki etmemektedir.

D. 5. Logistic Regression (Lojistik Regresyon) Yöntemi

Regresyon yöntemleri, bir sınıf özniteliği ile bir veya daha fazla öznitelik arasındaki ilişkiyi açıklamakla ilgili herhangi bir veri analizinin ayrılmaz bir bileşeni haline gelmiştir. Lojistik Regresyon algoritmasında çoğunlukla sınıf özniteliği ayrık tipte değer taşır ve iki veya daha fazla olası değeri alır. Lojistik regresyon modeli, bu verilerin analizi için en sık kullanılan regresyon modelidir [20].

D. 6. Random Forest (Rastgele Orman) Yöntemi

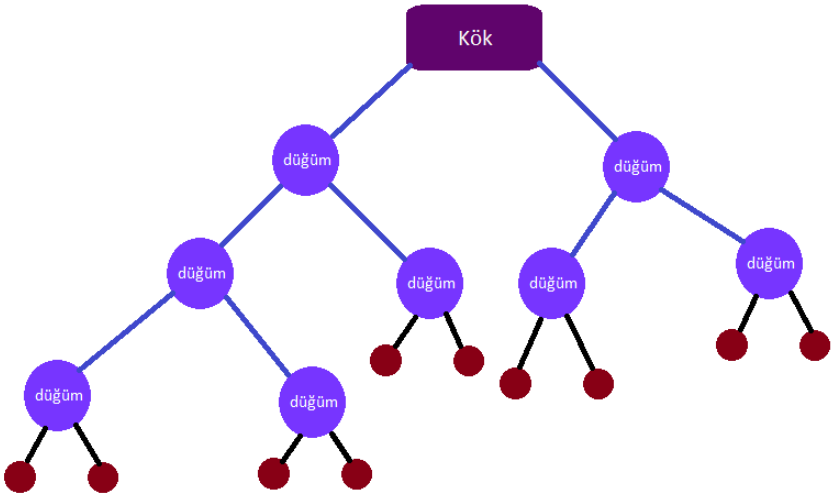
Rastgele Orman yöntemi [21], eğitim veri setinin önyükleme örnekleri üzerinde büyüyen ve ağaç yapımı sürecinde rastgele özellik seçimini içeren bir karar ağaçları topluluğu algoritmasıdır. Yapılan tahminler, tek tek tüm ağaçların tahminlerinin toplanmasıyla yapılır. Rastgele Orman, karar ağaçlarının toplulukları olduğu için, tek ağaca dayalı sınıflandırıcılara göre kesinlikle önemli performans artışı sergiler. Rastgele Orman, büyük boyutlu verilerin işlenmesinde iyi bir seçim olarak kabul edilmesine rağmen, dengesiz eğitim veri kümesi durumunda da zarar görür. 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 [22].

Telekomünikasyon veri kümeleri normalde daha yüksek derecede çarpıklıktan mustariptir, bu nedenle Rastgele Orman bazen kayda değer bir performans sergilemekten mustariptir [22].

D. 7. Decision Tree (Karar Ağacı) 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 oluşturulur ve dallar, düğümleri birbirine bağlayan bölümler tarafından temsil edilir. Bir Karar Ağacı kökten başlar, aşağı doğru hareket eder ve genellikle soldan sağa doğru çizilir. Ağacın başladığı yerdeki düğüme kök düğüm denir. Zincirin bittiği düğüm, "yaprak" düğüm olarak bilinir. Her iç düğümden, yani yaprak düğüm olmayan bir düğümden iki veya daha fazla dal uzatılabilir. Bir düğüm belirli bir özelliği temsil ederken, dallar bir dizi değeri temsil eder. Bu değer aralıkları, verilen karakteristiğin değerler kümesi için bir bölme noktası görevi görür. Şekil 4, bir ağacın yapısını açıklamaktadır [23].

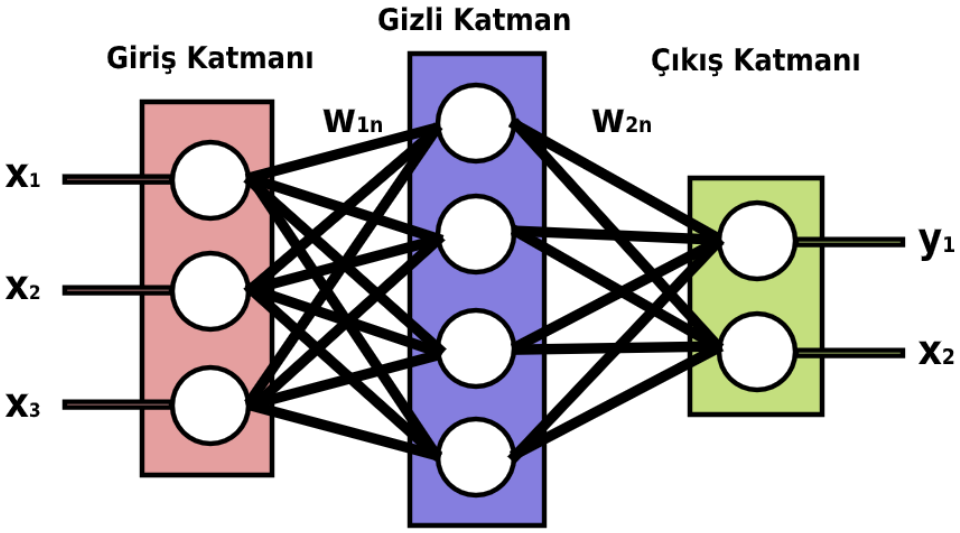


Şekil 4. *Decision Tree (Karar Ağaçları) Yöntemi Ağaç Yapısı*

D. 8. Artificial Neural Network (Yapay Sinir Ağları) Yöntemi

İnsan beyninde tüm kararlar, vücudumuzda doğal olarak bulunan ve temel yapı taşı olan "nöron" dan 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 [24].

Benzer şekilde, yapay nöron girdilerinde $x_1\ x_2\ \ldots\ x_n$ her nöron tarafından alınır ve karar verme için toplama ve aktivasyon / transfer fonksiyonu için eklenir. Çıktı, Şekil 5'te verilen tüm sinir ağı tarafından alınan ortak karar temelinde dışarıdan alınır [24].



Şekil 5. *Artificial Neural Network Yöntemi Ağaç Yapısı*

Benzer şekilde, yapay nöron girdilerinde x_1, $x_2 \ldots x_n$ 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, üç ana katmandan (giriş, gizli ve çıkış katmanları) oluşur. Giriş değeri x_i nörona uygulandığında, ağırlık eklenir ve sonuçlanır:

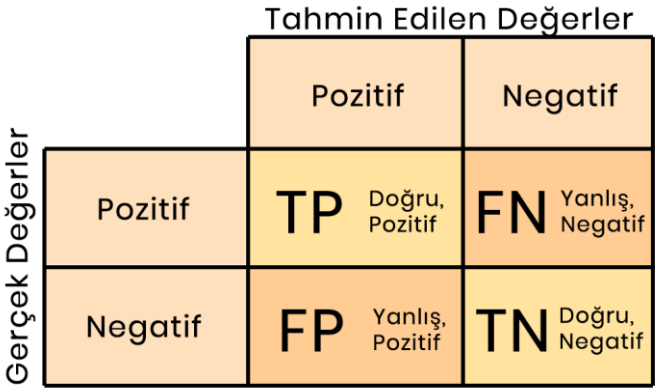
$$o_k = f(\sum w_i x_i + b_j) \tag{5}$$

Denklem (5)'te bir yapay sinir ağının çıktı değerini formüle etmek için tanımlanmaktadır. w_i her girdi verisi (x_i) için ağırlığı ve b_j her algılayıcı için sapmayı temsil etmektedir. Yapay sinir ağının çıktısı ise o_k olarak tanımlanmaktadır.

E. PERFORMANS ÖLÇÜTLERİ

Çalışmanın amacı, derin öğrenme ve makine öğrenmesi algoritmaları ile en iyi performansı veren tahmin modelini oluşturmaktır.

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 [25]. Şekil 6'da gösterilmektedir.



Şekil 6. *Karışıklık Matrisi*

Bunun için Karışıklık Matrisi (Confusion Matrix) ile performans ölçütleri yardımıyla değerlendirilmiştir. Performans ölçütü olarak aşağıda verilen metrikler kullanılmıştır.

E. 1. Doğruluk Oranı (Accuracy Rate)

Bu oran ile verilen modelin ne kadar doğru çalıştığı, doğru pozitif ve doğru negatifin diğer değerlere oranı Denklem (6)'daki gibi hesaplanarak bulunur.

$$Doğruluk\ Oranı = \frac{\mathrm{TP+TN}}{\mathrm{TP+FP+FN+TN}} \tag{6}$$

E. 2. Geri Çağırma (Recall)

Duyarlılık (Sensitivity) olarakta bilinir, doğru pozitif tahmin edilen ilgili örneklerin, gerçek tüm pozitif miktarına oranıdır. Sonuçlar ne kadar eksiksiz sorusuna cevap verir. Denklem (7)'deki gibi hesaplanır.

$$\text{Geri Çağırma} = \frac{\mathrm{TP}}{\mathrm{TP+FN}} \tag{7}$$

E. 3. Hassasiyet (Precision)

Hassasiyet, bazen pozitif tahmin değeri (positive predictive value) olarak da bilinir, doğru pozitif tahminlerin, pozitif tahminlere oranıdır. Arama sonuçları ne kadar geçerli sorusuna cevap verir. Denklem (8)'deki gibi hesaplanarak bulunur.

$$Hassasiyet = \frac{\mathrm{TP}}{\mathrm{TP+FP}} \tag{8}$$

E. 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. Denklem (9)'daki gibi hesaplanarak bulunur.

$$\text{Özgünlük} = \frac{\mathrm{TN}}{\mathrm{FP+TN}} \tag{9}$$

E. 5. Dengelenmiş Doğruluk Oranı (Balanced Accuracy Rate)

Dengeli doğruluk, 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 (10)'daki gibi hesaplanır.

$$Dengelenmiş\ Doğruluk\ Oranı = \frac{\text{Geri Çağırma+Özgünlük}}{2} \tag{10}$$

E. 6. F1 Skoru (F1-Score)

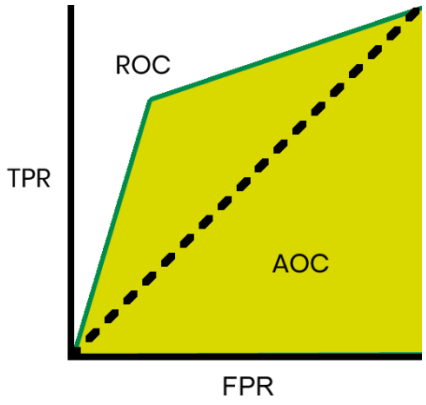
Hassasiyet ve geri çağırmanın harmonik ortalamasıdır. Testlerin doğruluğunun bir ölçütüdür. Denklem (11)'deki gibi hesaplanır.

$$F1\ Skoru = 2x\left(\frac{\text{Hassasiyet x Geri Çağırma}}{\text{Hassasiyet+Geri Çağırma}}\right) \tag{11}$$

E. 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 [25].

ROC-AUC ölçüsü ise ROC eğrisinin altında kalan alanı temsil etmektedir.



Şekil 7. *ROC eğrisinin altında kalan alan değer grafiği*

F. ÖNERİLEN TAHMİN MODELLERİ

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 tahmin modellerinin 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 Modeli'nden 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.

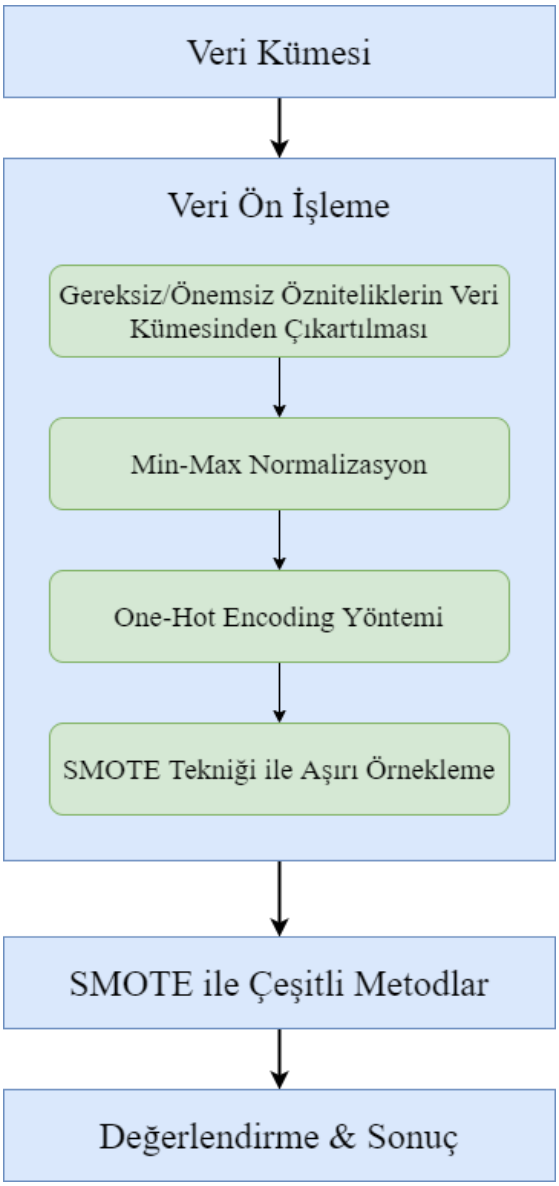
Birinci tahmin modelinde Univariate Feature Selection öznitelik seçimi yapılmamakla beraber SMOTE aşırı örnekleme yöntemi kullanılmıştır. İkinci tahmin modelinde ise Univariate Feature Selection yöntemi ile birbiriyle ilişkisi olmayan öznitelikler veri kümesinden çıkartılmıştır ve aşırı örnekleme yöntemi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

ADASYN ve SMOTE algoritmalarının temel farkı; ADASYN algoritmasının ana fikri yoğunluk dağılımını kullanmaktır. SMOTE algoritması ise her orijinal azınlık örneği için aynı sayıda sentetik numune üretir [16].

F. 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. 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 8 olarak tanımlanabilir.

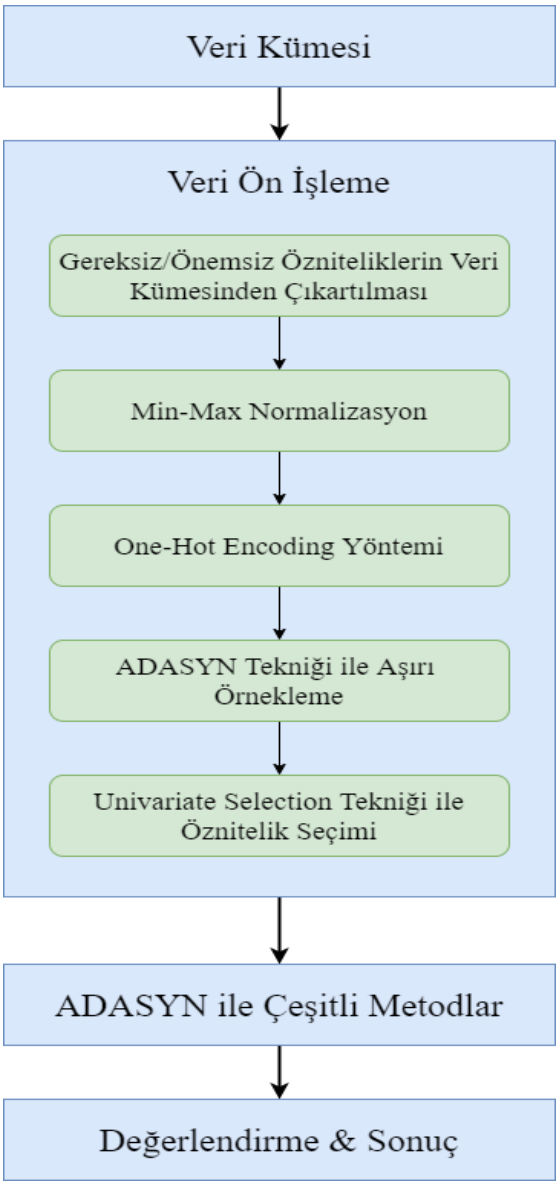


Şekil 8. *SMOTE tekniği ile Önerilen Tahmin Modeli*

F. 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. Çalışmada, önerilen diğer tahmin modeline göre özgünlük (specificity) ve hassasiyet (precision) daha iyi performans vermiştir.

Bu model Şekil 9 olarak tanımlanabilir.



Şekil 9. *ADASYN tekniği ile Önerilen Tahmin Modeli*

G. DEĞERLENDİRME & 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.

G. 1. Aşırı Örnekleme Yöntemi Kullanmadan Eğitilen Tahmin Modelleri

Bu bölümde önerilmeyen tahmin modelleri anlatılmaktadır. 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 önerilmeyen tahmin modellerine ait başarı oranları Tablo 1' de gösterilmektedir.

Tablo 1. *Önerilmeyen tahmin modellerinin başarı oranları*

	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*

G. 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 Tablo 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 Tablo 3'te gösterilmektedir.

Tablo 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*

Tablo 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 seçilen modelin performans çıktıları Tablo 4'te verilmiştir. İkinci tahmin modeli için seçilen modelin performans çıktıları ise Tablo 5'te verilmiştir.

Tablo 4. *Birinci tahmin modelinin performans çıktıları*

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 Değeri	*% 86.9*
Çalışma Zamanı	*14.75 saniye*

Tablo 5. *İkinci tahmin modelinin performans çıktıları*

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 Değeri	*% 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 tablo hazırlanmıştır. Bu karşılaştırma tablosu Tablo 6'da gösterilmiştir.

Tablo 6. *Benzer çalışmalar ile karşılaştırılması*

	[22]	**[26]**	**[27]**	**[28]**	**[7]**	**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	*Random Forest, Rotation Forest, RotBoost*	*Naive Bayes, Logistic Regression, AdaBoost, OneR..*	*SVM, DT, RIP with Active Learning Based Approach*	*Distance based Sampling*	*Bagging, Boosting, SVM, ANN..*	*Gradient Boosting, Bagging, ANN, Logistic Regression..*

IV. SONUÇ

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 58 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 [15]'da ve [16]'da ifade edilmiştir.

Modellerin karşılaştırılması kullanılırken belirlenmiş olan Karmaşıklık Matrisi göz önüne alınmıştır ve performans ölçütü olarak Doğruluk Oranı (%80.3), Geri Çağırma (%75.0), Hassasiyet (%93.9), Özgünlük (%95.7), Dengelenmiş Doğruluk Oranı (%80.4), F1 Skoru (%76.8) ve ROC-AUC Oranı (%87.1), Çalışma Zamanı (15.34 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.

TEŞEKKÜR: Bu çalışmayı desteklediği ve finanse ettiği için TTG International Ltd. 'ye müteşekkirim ve veri akışı mimarisinde bize 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.

V. KAYNAKLAR

[1] C. Gold, "*What this book is about*" in Fighting Churn With Data, 1. Baskı, O'reilly Media, 2020.

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

[3] A. M. AL-Shatnwai, M. F. Altibbi, "Predicting Customer Retention using XGBoost and Balancing Methods," *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 7, pp. 704- 712, 2020.

[4] A. R. Safitri, M. A. Muslim, "Improved Accuracy of Naive Bayes Classifier for Determination of Customer Churn Uses SMOTE and Genetic Algorithms," *JOSCEX Journal of Soft Computing Exploration*, vol. 1, no. 1, pp. 70-75, 2020.

[5] D. Wadikar, "Customer Churn Prediction," *Yüksek Lisans Tezi, Technological University Dublin*, 2020.

[6] H. Abbasimehr, M. Setak, M. J. Tarokh, "A Comparative Assessment of the Performance of Ensemble Learning in Customer Churn Prediction," *The International Arab Journal of Information Technology*, vol. 11, no. 6, pp. 599-606, 2014.

[7] J. Vijaya ve E. Sivasankar, "Computing Efficient Features Using Rough Set Theory Combined with Ensemble Classification Techniques to Improve the Customer Churn Prediction in Telecommunication Sector," *Computing*, vol. 100, no. 8, pp. 839–860, 2018.

[8] N.N.A. Sjarif, M.R.M. Yusof, D.H. Wong, S. Yaakob, R. Ibrahim ve M.Z. Osman, "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*, c. 11, s. 2, ss. 46-59, 2019.

[9] Y. Tan, L.H. Shuan, L.J. Yan ve X. Guo, "Prediction on Customer Churn in the Telecommunications Sector Using Discretization and Naïve Bayes Classifier," *International Journal of Advances in Soft Computing and its Applications*, c. 9, s. 3, ss. 23-35, 2017.

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

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

[12] K. Potdar, T. Pardawala ve C. Pai "A Comparative Study of Categorical Variable Encoding Techniques for Neural Network Classifiers," *International Journal of Computer Applications*, c. 175, s. 4, ss. 7–9, 2017.

[13] Ş. Taşdemir, B. Yanıktepe ve A.B. Güher, "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*, c. 5, ss. 131–139, 2018.

[14] A.Y. Liu, "The Effect of Oversampling and Undersampling on Classifying Imbalanced Text Datasets," *Yüksek Lisans Tezi, University of Texas at Austin*, USA, 2004.

[15] N.V. Chawla, K.W. Bowyer, L.O. Hall ve W.P. Kegelmeyer, "SMOTE: Synthetic Minority Over-Sampling Technique," *Journal of Artificial Intelligence Research*, c. 16, ss. 321–357, 2002.

[16] H. He, Y. Bai, E.A. Garcia ve S. Li, "ADASYN: Adaptive Synthetic Sampling Approach for Imbalanced Learning," *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, ss. 1322-1328, 2008.

[17] L. Breiman, "Bagging Predictors," *Department of Statistics, University of California Berkeley*, Technical Report No. 421, 1994. Retrieved 2019-07-28.

[18] Y. Freund ve R. E. Schapire, "A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting," *Journal of Computer and System Sciences*, c. 55, s. 1, ss. 119-139, 1997.

[19] M. R. H. Subho, M. R. Chowdhury, D. Chaki, S. Islam and M. M. Rahman, "A Univariate Feature Selection Approach for Finding Key Factors of Restaurant Business," *2019 IEEE Region 10 Symposium (TENSYMP)*, Kolkata, India, 2019, pp. 605-610.

[20] D. W. Hosmer, S. Lemeshow ve R. X. Sturdivant, "*Introduction*" in Applied Logistic Regression, 3. Baskı, WILEY, 2013.

[21] L. Breiman, "Random Forests," *Machine Learning*, c. 45, s. 1, ss. 5-32, 2001.

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

[23] J. Ali, R. Khan, N. Ahmad ve I. Maqsood, "Random Forests and Decision Trees," *IJCSI International Journal of Computer Science Issues International Journal of Computer Science Issues*, c. 9, s. 3, 2012.

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

[25] P. Tan, M. Steinbach, V. Kumar, "*Performance Measure*" in Introduction to Data Mining, Pearson Education Limited (UK), 2014.

[26] M. Yıldız ve S. Albayrak, "Customer Churn Prediction in Telecommunication," *2015 23nd Signal Processing and Communications Applications Conference (SIU)*, ss. 256-259, 2015.

[27] S. Jamil ve A. Khan. "Churn Comprehension Analysis for Telecommunication Industry Using ALBA," *2016 International Conference on Emerging Technologies (ICET)*, ss. 1-5, 2016.

[28] A. Amin, F. Obeidat, B. Shah, A. Adnan, J. Loo ve S. Anwar, "Customer Churn Prediction in Telecommunication Industry Using Data Certainty," *Journal of Business Research*, c. 94, ss. 290–301, 2019.