MAKİNE ÖĞRENMESİ ALGORİTMALARI KULLANILARAK MODEM VERİSİ ÜZERİNDEN MÜŞTERİ MEMNUNİYETİNİN TAHMİNİ

ESTIMATION OF CUSTOMER SATISFACTION USING MACHINE LEARNING ALGORITHMS OVER MODEM

YASİN SARI

Dr. Öğr. Üyesi İbrahim Zor

Tez Danışmanı

Hacettepe Üniversitesi

Lisansüstü Eğitim-Öğretim ve Sınav Yönetmeliğinin

İstatistik Anabilim Dalı İçin Öngördüğü

Yüksek Lisans Tezi Olarak hazırlanmıştır.

2024

ÖZET

MAKİNE ÖĞRENMESİ ALGORİTMALARI KULLANILARAK MODEM VERİSİ ÜZERİNDEN MÜŞTERİ MEMNUNİYETİNİN TAHMİNİ

YASİN SARI

Yüksek Lisans, İstatistik Bölümü

Tez Danışmanı: Dr. Öğr. Üyesi İbrahim Zor

Ağustos 2024, 70 Sayfa

Verinin aktarılması, ayrıştırılması ve algoritmaların test edilmesi gibi tüm süreçlerde KNIME Analitik Platformu kullanılmıştır. Modem verileri haftalık bazda ele alınmış ve indirme-yükleme verileri kategorize edilerek 6 farklı saat diliminde değerlendirilmiştir. Sınıflandırma analizinde ise AutoML üzerinden çalışmalar yapılmıştır. Burada değerlendirilen algoritmalar; Sade Bayes, Lojistik Regresyon, Sinir Ağları, Gradyan Artırma Ağaçları, Karar Ağacı, Rastgele Orman ve XGBoost'dur. Kullanılan kütüphane ve platformlar ise; Genelleştirilmiş Doğrusal Modeller için H2O yazılımı, Derin Öğrenme için Keras Kütüphanesi ve diğer birçok algoritma için H2O AutoML (Otomatik Edilmiş Makine Öğrenmesi) çatısıdır.

Bu çalışmada amaç memnuniyetsiz müşterileri belirlemektir. Dengesiz bir veri kümesi ile çalışma yapıldığından dolayı farklı veri çoklama teknikleri kullanılmıştır. Memnuniyetsiz müşterileri bulmak için modemlere ait sorunlu sinyal bilgisi ve arıza kaydı bırakan abonelere ait veriler üzerinden işaretleme yapılmıştır. Uygulama

sonucunda, veriye Temel Bileşenler Analizi yöntemi ile 4 Temel Bileşen indirgemesi yapılması ve sonrasında verinin rastgele seçilerek SMOTE (Synthetic Minority Over-sampling) tekniği ile verinin zenginleştirilmesi modeli iyileştirmiştir. Genel olarak ağaç algoritmaları dengesiz veri üzerinden sınıflandırma problemini çözmede daha iyi sonuçlar vermiştir. Algoritmalar arasında seçim yapmak için Duyarlılık (TPR) ve Seçicilik (TNR) değerlerinin geometrik ortalaması (GBA), ağırlıklı ortalaması (WPN) ve Bookmaker Informedness (BM) kriterleri kullanılmıştır. Bu ölçütlerdeki sonuçların birbirine yakın olmasından dolayı, FP (False Positive – Yanlış Pozitif) oranı daha düşük üzerinden değerlendirme yapılmıştır. Bunun seçilmesinin nedeni memnuniyetsiz müşteriye yapılacak olan yatırım maliyetini düşürmektir. Burada uygulanan 10 adet algoritma ve kütüphane içerisinden en iyi sonuçları XGBoost algoritmasının verdiği görülmüştür.

Anahtar Kelimeler: Makine Öğrenimi, AutoML, Müşteri Memnuniyet Tahmini, Büyük Veri, Sınıflandırma

ABSTRACT

ESTIMATION OF CUSTOMER SATISFACTION USING MACHINE LEARNING ALGORITHMS OVER MODEM

YASİN SARI

M. Sc Thesis, Department of Statistics

Supervisor: Asst. Prof. Dr. İbrahim Zor

August 2024, 70 pages

The KNIME Analytics Platform was used throughout all processes, including data transfer, parsing, and algorithm testing. Modem data was analyzed weekly, and download-upload data was categorized and evaluated across six different time slots. For classification analysis, AutoML was utilized, assessing algorithms such as Naive Bayes, Logistic Regression, Neural Networks, Gradient Boosted Trees, Decision Trees, Random Forest, and XGBoost. The libraries and platforms used include H2O software for Generalized Linear Models, the Keras library for Deep Learning, and H2O AutoML for various other algorithms.

The aim of this study is to identify dissatisfied customers. Different sampling methods were used due to working with an unbalanced dataset. Data from modems with faulty signal information and data from subscribers who left a service complaint were used for labeling. The model was improved by reducing the data to four Principal Components using Principal Component Analysis (PCA) and then enriching it with the SMOTE (Synthetic Minority Over-sampling) technique. Tree-based algorithms yielded better

results in solving the classification problem on imbalanced data. Algorithms were evaluated based on the geometric mean of Sensitivity (TPR) and Specificity (TNR), weighted average (WPN), and Bookmaker Informedness (BM) criteria. Due to the closeness of the results, the False Positive (FP) rate was chosen as the final criterion to minimize the investment cost in dissatisfied customers. XGBoost provided the best results among the ten algorithms applied.

Keywords: Machine Learning, Classification, AutoML, Customer Satisfaction Estimation, Big Data

TEŞEKKÜR

Bilimsel Hazırlık ve Yüksek Lisans eğitim boyunca bilgi ve birikimlerini hiç eksik etmeyen İstatistik Anabilim dalındaki öğretmenlerime ve danışmanım Dr. Öğr. Üyesi İbrahim Zor'a,

Çalıştığım firmada bilgi ve tecrübelerini benimle paylaşan çalışma arkadaşlarıma ve bana verileri inceleme imkanı sağlayan kurumuma,

Bilimsel Hazırlık ve Yüksek Lisans eğitimim boyunca beni sürekli destekleyen değerli eşime ve çocuklarıma teşekkür ederim.

Yasin SARI

Ağustos 2024, Ankara

İÇİNDEKİLER

ŞEKİLLER DİZİNİ

ÇİZELGELER DİZİNİ

SİMGELER VE KISALTMALAR

Kısaltmalar

KNIME	Konstanz Information Miner - KNIME Analitik Platformu
TB	Terabyte
FTP	File Transfer Protocol - Dosya Transfer Protokolü
AutoML	Automated Machine Learning - Otomatik Makine Öğrenimi
IoT	Internet Of Things - İnternetin Şeyleri
CRM	Customer Relationship Management – Müşteri İlişkileri Yönetimi
JSON	JavaScript Object Notation – JavaScript Nesne Notasyonu
H2O	Distributed In-Memory Machine Learning Platform – Dağıtık Yapıda Hafıza Üzerinden Makine Öğrenimi Platformu
TP	True Positive - Doğru Pozitif
TN	True Negative - Doğru Negatif
FP	False Positive - Yanlış Pozitif
FN	False Negative - Yanlış Negatif
CK	Cohen’s Kappa
MCC	Matthews Correlation Coefficient - Mathews İlişki Katsayısı
GBA	The G- Mean Of TPR and TNR - TPR Ve TNR’nin Geometrik Ort.
PPV	Positive Predictive Value - Pozitif Tahmin Değeri
TPR	True Positive Rate - Doğru Pozitif Oranı
TNR	True Negative Rate - Doğru Negatif Oranı
BM	Bookmaker Informedness
WPN	Weighted TPR-TNR – Ağırlıklı TPR ve TNR
MAC	Media Access Control Address – Modeme Ait Özel Adres Alanı
SQL	Structured Query Language - Yapısal Sorgu Dili

PostgreSQL	Object-Relational Database Management System – Nesne İlişkisel Veritabanı Yönetim Sistemi
SingleStore	A Database Designed For Data-Intensive Applications – Yoğun Veri Uygulamaları Tasarlanmış Bir Veritabanı
OLAP	Online Analytical Processing – Online Analitik İşleme
Lateral	SQL Subqueries With Single Join - SQL Alt Sorgularının Tek Bir Birleştirme İle Yapılması
CS	Customer Satisfaction - Müşteri Memnuniyeti
CL	Customer Loyalty - Müşteri Sadakati
CSI	Customer Satisfaction Index - Müşteri Sadakat İndeksi
MLE	Machine Learning Engineer-Makine Öğrenimi Mühendisi
MSE	Mean Squared Error - Ortalama Kare Hata
Apache Hive	Distributed, Fault-Tolerant Data Warehouse System That Enables Analytics At A Massive Scale - Büyük Ölçekli Analizlere Olanak Sağlayan, Dağıtık Ve Hata Toleranslı Veri Ambarı Sistemi
Apache Impala	Open Source, Native Analytic Database For Open Data And Table Formats – Açık Veri Ve Tablo Formatları İçin Açık Kaynaklı, Yerel Analitik Veri Tabanı
MapReduce	Programming Model Within The Hadoop Framework That Is Used To Access Big Data Stored In The Hadoop File System (HDFS) – HDFS'de Depolanan Büyük Verilere Erişmek İçin Kullanılan Hadoop Çerçevesi İçinde Yer Alan Bir Programlama Modeli Veya Deseni
Radoop	An Graphical Interface For Analyzing Data On A Hadoop Cluster With A Running Hive Server - Hadoop Kümesinde Yer Alan Hive Sunucusuna Erişerek, Grafik Arayüzü İle Veri Analizi Yapılmasını Sağlar
ETL	Extract Transform Load - Çıkarma Dönüştürme Yükleme

1. GİRİŞ

Günümüzde, rakip firmalara odaklanarak gelişim planlaması yapmak, hem mevcut piyasanın gerisinde kalınmasına hem de rekabet ortamından uzaklaşılmasına neden olabilmektedir. Eldeki verilerin aslında araştırmacıyı ileriye taşıyacak işaretleri barındırdığı bilgi dahilindedir. Diğer yandan, dijital dünyada müşteri memnuniyetsizlikleri, hem parasal hem de prestij kayıplarına yol açabilir. Bu nedenle, eldeki bulunan verilerden memnuniyetsiz müşterilerin profillerini ortaya çıkarmak, aslında daha fazla memnun müşteriye ulaşılmasına yardımcı olabilecek bulgular içerebilir. Zıtlıklardan yola çıkarak yapılacak incelemelerin, başarıya götürecek anahtarları içerdiği söylenebilir.

Bugün geçmişte de olduğu gibi müşterisini önceden anlama becerisini gösteren şirketler bir adım daha öne çıkacaktır. Google veya Apple uygulama havuzundan bedava bir ürün indirildiğinde verinin kendiniz olduğu bilinmektedir. Servis hizmeti satan firmalar günümüz dünyasında çok önemli verilere müşterileri tarafından ulaşmakta ve belki de milyon dolarla alınamayacak kıymetteki bilgilere erişmektedirler. Bu veriler sadece o ürüne ait değil aynı zamanda pazardaki müşterilerin profillerini anlama adına da ayrı bir önem arz etmektedir.

Verinin meydan okuması bağlamında, büyük verinin nasıl saklanacağı ve bunun nasıl değere dönüştürüleceği önemlidir ve akan bir verinin önünde durmak oldukça zordur. Dolayısıyla firmaların stratejik olarak rekabet edebilmeleri için verinin içindeki nitelikleri doğru ve anlamlı şekilde bir araya getirerek geleceği inşa edecek iyi veri mühendislerine sahip olmaları gerekmektedir.

Normalde 10 bin sene sürecek bir hesaplamayı Google makineleri kuantum mekaniklerini kullanarak saniyeler içerisinde gerçekleştirmişlerdir. 1980'li yıllarda sadece teorik seviyede yer alan bilgi, 2020'li yıllarda uygulama aşamasına gelmiştir. Q-Günü isimli güne gelindiğinde yeni bir çığır açılacak ve şifreler/şifreleme algoritmaları işlevsiz hale gelecektir. Ford ve Microsoft'taki araştırmacılar kuantum hesaplama yöntemlerini kullanarak yoğun trafik saatlerinde trafik akışını kolaylaştıracak model üzerinde çalışmalar yapmaktadırlar (Suleyman, 2023).

Özet olarak, bu yeni çağda Yapay Zeka, Biyoteknoloji ve Kuantum kavramları, her zamankinden daha etkin bir rol alacak faktörler olarak çok yakın bir gelecekte herkesi etkileyecektir. Hızla büyüyen teknoloji ve verilerle birlikte yeni meydan okumalara karşı karşıya kalınması durumu ve bu hızla büyüme karşısında etkin çözümler ve adaptasyonlar sağlanması kaçınılmazdır. Hızlı teknoloji dönüşümüne hazır olanlar, büyük veri ve hesaplamaları daha hızlı bir şekilde stratejik öngörülere dönüştürebileceklerdir.

Bu çalışmada, hızla büyüyen bir sektörde büyük verilerin nasıl analiz edilebileceği ve nasıl hızlı okumalar sağlanabileceği konularına da değinilecektir. Bu çalışma, internet kullanan abonelere ait sorunlu modem sinyal değerleri ve çağrı merkezine arıza bırakan abonelerin verisi üzerinden müşteri memnuniyetini tespit etmeyi amaçlamaktadır.

1.1. Literatürde İlgili Çalışmalar

Veri tiplerinin ve modellerinin farklılığından dolayı birçok farklı yöntem ve yaklaşım mevcuttur. Genel olarak hedef müşteri memnuniyetini artırmaya yönelik çalışmalardır. Bu çalışmalarda farklı makine öğrenmesi algoritmaları kullanılmıştır. Müşteri memnuniyetini sağlamak adına öncesinde beklentilerin belirlenmesi ve sonrasında doğru niteliklerin seçilmesi gerektiğidir.

Aktepe ve ark. (2014) yaptıkları çalışmada müşteri memnuniyetini (CS) artırmak ve sonrasında müşteri sadakatini (CL) elde etmek için müşterileri 4 farklı grupta ele almışlardır. Burada sınıflandırma algoritmaları için WEKA yazılımı ve LISREL aracı ile Yapısal Eşitlik Modeli (SEM) üzerinden yapılan analizler ile her bir grup için memnuniyet ve sadakat kriterlerine ait etkiler memnuniyet-sadakat matrisi üzerinden incelenmiştir. Değerlendirme yapmak için 4 farklı grup belirlenmiş ve toplamda 15 adet kriter üzerinden 200 müşteri ile anket yapılmıştır. Sınıflandırma algoritmaları hem grupların belirlenmesinde hem de CS-CL matrislerinin belirlenmesinde kullanılmıştır. Müşteri Memnuniyeti ve sadakati grupları belirlenmiş, sonraki aşamada literatür ve alan uzmanları üzerinden değerlendirme yapılmıştır.

Lee ve ark. (2016) Tayvan Müşteri Memnuniyeti İndeksini (TCSI) kullanarak turizm firmalarındaki Müşteri Memnuniyeti ve Müşteri Sadakati üzerine analizler yapmışlardır. Turizm firmasından hizmet alan 242 müşteri ile anket yapılmıştır. Burada Kısmi En Küçük Kareler yöntemi uygulanarak teorik modelin test ve analizleri yapılmıştır. 14 değişken ve 6 temel madde üzerinden analiz yapılmıştır: imaj, müşteri beklentisi, alınan kalite, alınan değer, müşteri memnuniyeti ve sadakat. Sonuç olarak Müşteri Memnuniyeti ve Sadakati için iki tane Müşteri Memnuniyeti İndeksi (CSI) modeli önerilmiştir. Diğer elde edilen bulgu ise firmaların Müşteri Sadakatini elde etmede kullandıkları servis ve ürünün sahip olduğu imajdan çok Müşteri Memnuniyeti ile sağlanmasıdır.

Meinzer ve ark. (2017) yaptıkları çalışmada araba endüstrisinde müşteri memnuniyetsizliğini tespit etmede 2 temel soru üzerinden ilerlemiştir. Bunlardan birincisi memnuniyetsiz müşteri arabayı servise her getirdiği zaman oluşan veriden mi tespit edilecek? Yoksa servise gelen müşteri ile birlikte arkada işlenen diğer veriler üzerinden mi tespit edilmelidir? Burada memnuniyetsizliği tespit etmede AdaBoost, kNN, SVM(doğrusal), SVM(RBF) ve Rastgele Orman makine öğrenmesi teknikleri uygulamıştır. Analiz edilen verinin boyutu saatlik 1 TB civarındadır. 5 sınıflandırma metodu ile 105 tane nitelik üzerinden çalışma yapılmıştır. Burada en iyi sonucu % 88.8'lik skor ile Destek Vektör Makineleri (SVM) RBF çekirdeği vermiştir.

Tong ve ark. (2017) telekomünikasyon sektöründe Müşteri Sadakatini daha iyi yapmak için NPS (Net Promoter Score – Net Taraftar Skoru) veri madenciliği tekniğinden faydalanmışlardır. Net Promoter konsepti Fred Reicheld (2003) tarihinde ortaya atılmıştır. Ölçümleme şu soruya göre yapılır: "Bir arkadaşınıza veya meslektaşınıza X şirketini tavsiye etme olasılığınız nedir?". Puanlama sisteminde 9 veya 10 verenler "Promoter" (taraftar) olarak isimlendirilir. Aradaki kullanıcılar "Passive" (Pasif) ve "Detractor" (Karşıt) olarak nitelendirilmektedir. Bunlar arasındaki yüzdelik farkı NPS skoru ölçütünü belirler. 35 tane nitelik üzerinden eldeki bilginin analiz edilmesi için XGBoost algoritması tercih edilmiştir. XGBoost algoritmasının tercih edilmesinin nedeni yüksek Doğruluk ve hız; aynı zamanda karar ağacında budama ihtiyacı gerektirmemesidir.

Kumar ve Zymbler (2019) havayolu endüstrisindeki müşteri deneyimini tweetler üzerinden incelemişlerdir. Makine Öğrenmesi algoritmalarını kullanarak Tweetler üzerinden sınıflandırma (pozitif/negatif) yapmışlardır. Konvansiyonel Sinir Ağları (CNN) modeli, Destek Vektör Makineleri (SVM) ve Yapay Sinir Ağları (ANN) modellerine göre daha iyi performans göstermiştir. Uygulamanın ikinci aşamasında tweetlere birliktelik kuralları madenciliği uygulanarak duygu kategorizasyon haritalaması yapılmıştır.

Wei ve ark. (2020) yaptıkları çalışmada müşteri memnuniyetini anlamak için müşteri yorumları (CC) ve ajans notları (AN) üzerinden eşleştirme yaklaşımını kullanmışlardır. İlk aşamada CC üzerinden duygu analizi yapılmıştır; sonrasında Konvansiyonel Gizli Semantik Model (CLSM) üzerinden CC ve AN eşleştirmesi gerçekleştirilmiştir. Mesafe hesaplaması üzerinden uyumsuzluk skoru elde edilmiştir. Özellik çıkarmada Bayes sınıflandırması kullanılmıştır. Duygu analizi ve uyumsuzluk skoru ile NPS değeri tahmin edilmesi için CNN kullanılmıştır. Sonuç olarak yüksek uyumsuzluk skoruna sahip olan kişiler için müşteri servisinde iyileştirme yapılması gerektiği ortaya konmuştur.

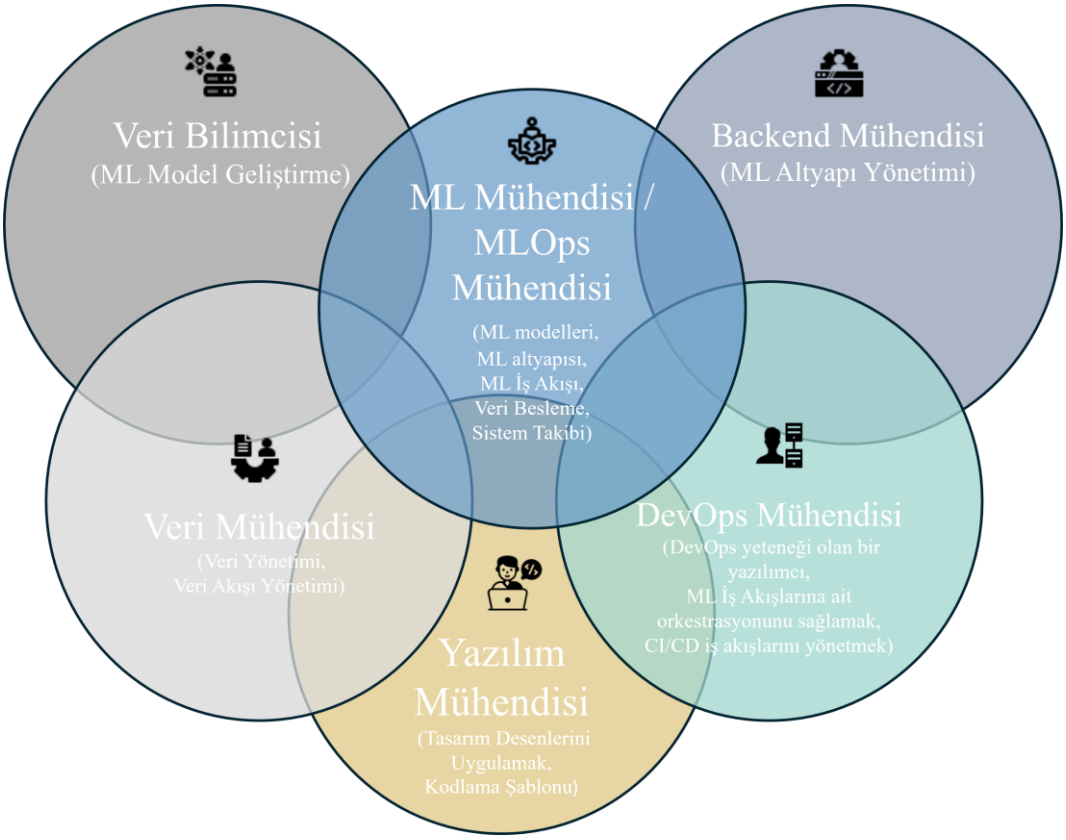
Wong ve ark. (2020) e-ticaret alanında müşteri memnuniyetini geliştirmek için müşteri içgörülerini incelemiştir. Burada 4 farklı algoritma ile sınıflama yapmışlardır: Karar Ağacı, Rastgele Orman, Yapay Sinir Ağları ve Destek Vektör Makineleridir. Karşılaştıkları en büyük zorluk verinin dengesiz olması, çarpıklık ve eksik değerlerdir. Rastgele Orman en yüksek doğruluk ve kabul edilebilir işleme süresinden dolayı tercih edilmiştir.

2. VERİ YÖNETİMİ VE MAKİNE ÖĞRENMESİ YÖNTEMLERİ

2.1. Büyük Verinin Tanımı

Büyük veri nedir sorusunun yanıtına McAfee ve Brynjolfsson verdikleri cevap dikkate değerdir. Onlara göre büyük veri tanımı şudur: "Akıllı liderler endüstri üzerinden büyük verinin kullanımını tariflemektedir. Bu da: bir yönetim devrimidir." (McAfee, 2012). Büyük verinin ne kadar büyük olduğundan ziyade nasıl yönetilmesi gerektiği ile ilgili yöntemler konuşulmaktadır.

Yapay zeka ve makine öğrenmesine gelene kadar işleyen bir süreç altyapısı ve burada çalışan organizasyonel yapılar ve meslek tanımları da belirmiştir. Şekil 2.1 de Makine Öğrenmesi Operasyonları adı altında değerlendirilen bu roller; kapsamları, tanımları ve mimari yönden değerlendirilmiştir (Kreuzberger, 2023):



Şekil 2.1. Roller ve MLOps paradigmaları arasındaki katkıların geçişleri

Geleneksel bakış açısından farklı olarak "büyük verinin 7-V'si": Hacim (Volume), Hız (Velocity), Kaynak Çeşitliliği (Variety), Doğruluk (Veracity), Değer (Value), Değişkenlik (Variability), ve Görselleştirme (Visualization) adı altında tanımlanmıştır (Mishra, 2017).

1. Hacim: Genel kavram açısından ele alındığında terabyte (TB) seviyelerini aştığı ifade edilebilir. Dünyadaki verileri incelediğimizde bilgisayar, mobil cihazlar ve makinalar (IoT) günlük 2.5 exabyte'ı (1 Exabyte = 1 milyon TB) geçmektedir. Sadece Walmart ve Amazon tüketicilerinden elde ettiği veriler devasa boyutlara ulaşmış durumdadır. Bu tezde işlenen veri, an itibariyle aylık olarak 1 TB düzeyini geçmektedir ve büyümeye devam etmektedir.

2. Hız: Verinin üretilme hızını ve cinsini ifade etmektedir. Örnek olarak Facebook sadece metinsel değil görsel veri de üretmektedir. Aynı zamanda sürekli güncellenen verileri ve farklı yorumları içermektedir. Buradaki müşterilere ait İndirme-Yükleme verileri saatlik bazda yaklaşık 1milyar satırlık veriye karşılık gelmektedir.

3. Kaynak Çeşitliliği: Farklı veri kaynaklarının ürettiği veriler bu tanımda ele alınmaktadır. Örnek olarak sosyal medya verisi, dijital işlemler. Çalışmada kullanılan veriler 3 farklı veri kaynağından beslenmektedir.

4. Doğruluk: Verinin doğruluğundan emin olunmasını kapsamaktadır. Doğruluğu teyit edilmeyen veriler güvenilmez sonuçlar doğuracaktır. Modem verisi ilk aşamada modem MAC (medya erişim kontrol adresi) değerine göre tutulmaktadır. Modem MAC Adresi her modem de bulunan eşsiz bir değerdir. Bir MAC adresi, ağa bağlı her cihaza atanan 12 basamaklı bir 16 tabanlı sayı sistemidir. Buna göre örnek olarak bir müşterinin internette bıraktığı izler modem mac adresi üzerinden takip edilmektedir.

5. Değer: Verinin ekonomik ve stratejik olarak sahip olduğu değerlerdir. Modem değerleri ve sahadaki diğer cihazlar üzerinden birçok stratejik çalışma yapılmaktadır.

6. Değişkenlik: Bu başlık altında verinin akış ve işleniş hızındaki değişimler ve miktarında artışın farklı değerlendirilmesi konuşulmaktadır. Bu tez kapsamında ele alınacak verinin oluşmasında çalışan sistemler 3 saat geriden saatlik veriyi güncel olarak saklayabilmektedir. Ancak son 5 günlüğe ait saatlik veriler günlük ortalamalara dönüştürülüp silinmektedir.

7. Görselleştirme: Büyük verinin görselleştirilmesi için kullanılan yapılardır. Microsoft TB seviyesindeki Telemetri datalarını görselleştirmek için CitusData isimli analitik ve verinin dağıtık tutulduğu bir mimaride inceleyebilmektedir (Cubukcu, 2021). Bu kapsamda verilerde PostgreSQL'de veri bölümleme (partition) yapısında ve SingleStore (Yüksek performans, gerçek zamanlı analiz ve operasyonel işlemler için tasarlanmış dağıtık mimaride çalışan SQL veri tabanı) veri tabanlarından analiz edilmiştir.

Çalışmada kullanılan veriler Dosya Transfer Protokolü (FTP) ile sunuculara aktarıldıktan sonra KNIME Analitik Platformu Aracı ile işlenmektedir. Makine öğrenmesine uygun hale getirilmek istenen veriler ise yüksek RAM ve CPU hacmine sahip bir sunucuda Müşteri İlişkisi Yönetimi (CRM) verileri ve arıza verileri ile birleştirilerek uygun nitelik ve veri tiplerine dönüştürülmektedir. Veriyi işlemek ve dönüştürmenin yanında nasıl okunacağı da bir problem olarak durmaktadır.

Örneklem verisinin tüm Türkiye'yi temsil etmesi nedeniyle İstanbul bölgesine ait abonelerin verisi üzerinden çalışma yapılmıştır. Bu kapsamda haftalık olarak çekilen verinin büyüklüğü yaklaşık olarak 35 GB'tır. Tez kapsamındaki verilerde saatlik veri daha anlamlı analiz yapabilmek için günün 6 dilimine ayrılmış ve aynı zamanda bu veri hem yükleme hem indirme bazında sütunlara ayrılmıştır. Burada önemli olan soru zaman dilimlerinin neye göre ayrılması gerektiğidir. Veri modelleme aşamasında; hangi nitelikler alınmalı, sınıflandırılması ve ayrıştırılması gibi konular için şirket tecrübelerinden yola çıkılmıştır. Burada Silveira ve ark. (2022) yazdığı makalede bu fikri destekler niteliktedir. Yani erken saatler 4-8, sabah saatleri 8-12 ve bu şekilde devam etmektedir.

2.2. Büyük Veri İşleme Teknolojileri

Büyük veri işleme sürecinde yüksek performanslı ve ölçeklenebilir teknolojilerin kullanılması gereklidir. Bu bağlamda bazı ileri seviye ürünler ve ETL (Extract Load Transform - Çıkarma Dönüştürme Yükleme) araçları aşağıda ele alınmıştır.

Hadoop, büyük veri analitiği için bulut tabanlı ve açık kaynaklı bir çözüm olarak geliştirilmiştir. Apache (açık kaynaklı yazılım projeleri geliştiren vakıf)'nin üst düzey projelerinden biri olan Hadoop, Java programlama dili ile yazılmıştır ve bağımsız bilgisayarlarda büyük veri kümeleri ile çalışmayı mümkün kılar. Hadoop'un iki ana bileşeni vardır: Hadoop Dağıtılmış Dosya Sistemi (HDFS) ve MapReduce programlama modeli. HDFS, büyük dosyaları birden fazla makineye dağıtarak saklar ve veri güvenilirliğini, veriyi farklı makinelere (host) kopyalayarak sağlar. MapReduce ise veri işleme görevlerini paralel olarak dağıtır ve yürütür. Bu model, özellikle veri madenciliği, sosyal ağ analizi ve makine öğrenmesi gibi işlemler için idealdir (Bagheri, 2015).

Zaharia ve ark. (2010) yaptıkları çalışmada, tekrar eden makine öğrenmesi algoritmaları ve etkileşimli veri analiz araçları gibi, bir veri kümesini birden çok paralel işlemde yeniden kullanan uygulamalara odaklanılmıştır. Spark, MapReduce'un ölçeklenebilirlik ve hata toleransını korurken, dayanıklı dağıtık veri kümeleri (RDD) adı verilen ve kaybolan bölümlerin yeniden oluşturulmasını sağlayan bir soyutlama kullanarak, yinelemeli makine öğrenmesi işlerinde Hadoop'dan 10 kat daha hızlı çalışmış ve 39 GB'lık bir veri kümesini saniyenin altında etkileşimli olarak sorgulama imkanı sağlamıştır.

Google Cloud Platform (GCP), altyapı, depolama, veri tabanları, makine öğrenmesi ve analizi dahil olmak üzere ölçeklenebilir bulut bilişim çözümleri sunarak işletmelerin dijital çağda verimli bir şekilde yenilik yapmalarını ve büyümelerini sağlar. Mage AI, verilerin akıllı karar alma ve öngörüler için kullanılmasını sağlamak üzere gelişmiş veri hazırlama ve analitik yetenekler sunan yenilikçi bir platformdur. Bu platform, GCP'den verileri alarak gerekli dönüşümleri yapar ve verileri görselleştirme araçlarına aktarır. Bu entegrasyon, kullanıcılara stratejik bilgiler sunarak karar verme süreçlerini iyileştirir ve kazançlarını artırır (Balaji, 2024).

RapidMiner, bir veri madenciliği aracıdır ve büyük veri ortamında analiz yapılmasını sağlayan Radoop adında bir büyük veri uzantısı sunar. Radoop, ileri analiz süreçlerinin tasarımı için kodsuz bir ortam sağlar ve Hadoop kümesinde çalışan veri dönüşümleri ve gelişmiş tahmin modellemeleri için 60'tan fazla operatör içerir (Oliveira, 2019).

KNIME Analitik Platformu, Apache Spark ve Apache Hadoop'un gücünü entegre eden büyük veri uzantıları sunar. KNIME, HDFS verilerini okuma/yazma ve Hive (Apache tarafından geliştirilen veri ambarı) ile Impala (Apache Hadoop üzerinde kullanılan SQL sorgulama motoru) içinde analiz yapma imkanı sağlar. Hem KNIME hem de RapidMiner, kodsuz bir ortam ve grafiksel kullanıcı arayüzü sunar (Oliveira, 2019).

Oliveira ve ark. (2019), hem RapidMiner hem de KNIME'ın büyük veri ortamında iyi performans sergilediğini göstermiştir. MapReduce, HDFS'de depolanan büyük miktarda yapılandırılmış ve yapılandırılmamış veriyi işlemek için uygulamalar yazmak amacıyla kullanılırken; Apache Tez, yüksek performanslı toplu ve etkileşimli veri işleme uygulamaları oluşturmak için tasarlanmıştır (Koetter, 2015). Apache Tez, Apache MapReduce'a kıyasla daha hızlı yanıt süresi sunarak veri setlerini sürekli yeniden kullanan uygulamalar için daha verimli bir alternatif teşkil etmektedir. Apache MapReduce, aynı veri setini sürekli yeniden kullanması gereken uygulamalar için verimsizdir (Torres, 2018). Radoop, veri boyutu ve düğüm sayısı arttıkça iyi performans göstermiş, ancak küçük veri setlerinde tek bir makine daha iyi performans sergilemiştir.

2.3. AutoML

KNIME Analitik Platformu, sınıflandırma problemlerini çözmek için çeşitli makine öğrenmesi yöntemlerini kullanabilme kapasitesine sahiptir. AutoML (Otomatik Edilmiş Makine Öğrenmesi) düğümü ile birlikte Sade Bayes, Lojistik Regresyon, Sinir Ağları, Gradyan Artırma Ağaçları, Karar Ağacı, Rastgele Orman ve XGBoost Ağaçları gibi algoritmalar; ayrıca Genelleştirilmiş Doğrusal Modeller, Derin Öğrenme ve H2O AutoML çerçevesi tek bir düğüm üzerinden yönetilebilir. Bu düğüm, makine öğrenmesi algoritmaları, kütüphaneleri ve çerçeveleri üzerinde detaylı işlemler yapmaya izin verir. Asıl amaç, büyük firma ve organizasyonlar için uçtan uca çözümler sunmaktır. AutoML mekanizması ile sınıflama veya regresyon için kullanılan algoritma ve kütüphanelere, hızlı ve pratik biçimde ulaşılabilir. AutoML düğümü, verinin hazırlanması, özellik

seçimi, model seçimi, hiperparametre optimizasyonu ve süreçlerin gözlenmesi gibi birçok faktörü içermektedir. Bu mekanizma, makine öğrenmesi uzmanlarının hızla çözüm üretebilmesini sağlar (Singh, 2022).

AutoML düğümünü kullanmadan önce, özellik mühendisliği ile doğru değişkenler seçilir ve aykırı değer analizleri yapılır. AutoML düğümü ile hangi modellerin kullanılacağına, belirleme ölçütü olarak hangi metriklerin değerlendirileceğine, Kategorik Değişkenlere Dönüştürme (One-Hot Encoding) yapılıp yapılmayacağına, Çapraz Doğrulama (n-Folds) seçeneğine ve eğitim kümesi için ne kadar veri kullanılacağına karar verilir.

2.3.1 AutoML Aşamaları

AutoML işleminde süreçlerin akışında 4 aşama ele alınmıştır. Birinci aşamada temel ayarlar gelmektedir Algoritma seçimi, k katman sayısı, eğitim ve test oranı vs.

İkinci aşamada ise Veri Ön İşleme süreçleri işletilmektedir. Bu kapsamda eşsiz değer seçimi ile kategori seçiminde kullanılacak sayının kısıtlanması, normalleştirme, sayısal değerlerinin ondalıklı sayıya dönüştürülmesi, eksik değerlerin atanması, kategorik değişkenlere dönüştürme seçilmişse uygulanması, eğitim-doğrulama ve test verisinin ayrıştırılması gibi süreçleri kapsamaktadır (nodepit.com, 2024).

Üçüncü aşamada Eğitim ve Skorlama yapılan bölümdür. Test kümeleri üzerinden tahmin etme ve performans ölçümlemesi yapılmaktadır.

Dördüncü ve son aşamada ise en iyi model seçilmesi işlemidir. Burada model seçilen metriğe göre sıralanmakta ve en iyi model seçilmektedir.

2.4 Sade Bayes

Sade Bayes algoritması, veri sınıflama önemli bir yere sahiptir. Bu önem, algoritmanın basitliği ve doğruluğundan kaynaklanmaktadır. Adından da anlaşılacağı üzere, Sade Bayes yaygın olması Bayes teoremine dayanmaktadır ve makine öğrenmesi ile veri analitiğinde kullanılan önde gelen olasılıksal sınıflandırma tekniklerinden biridir. Sade Bayes'in yaygın kullanılması sadece basitliğinden değil, aynı zamanda algoritmanın

etkinliği ve dayanıklılığından da kaynaklanmaktadır (Arar, 2017). Literatüre göre, Sade Bayes, veri madenciliğinde en iyi performans (doğruluk, dayanıklılık ve verimlilik) gösteren algoritmalardan biridir (Wu, 2008; Settouti, 2016) (Kalutarage, 2017). Bayes formülü aşağıdaki gibi ifade edilir:

$$\Pr(X = x \mid Y = y) = \frac{\Pr\,(Y=y|X=x)\Pr(X=x)}{\Pr(Y = y)} \tag{2.1}$$

Eşitlik 2.1'de Pr ifadesi olasılık anlamına gelir. Başka bir rastgele Y rastlantı değişkeninin belirli bir y değerine sahip olması koşulunda, X rastlantı değişkenin belirli bir x değerine sahip olması olasılığıdır.

2.5 Lojistik Regresyon

Lojistik regresyon, aynı zamanda bir sınıflama öğrenme algoritması olarak kullanılmaktadır. Temel amacı, ikili veya çok sınıflı sınıflandırma problemlerde, verilerin doğrusal bir fonksiyonla modellenmesidir. Bu model, iki olası değeri olan bağımlı değişkenlerin (etiketlerin) olasılıklarını hesaplar. Lojistik regresyon, doğrusal regresyonun aksine, model çıktısını 0 ile 1 arasında sınırlayan sigmoid (lojistik) fonksiyonunu kullanır. Bu fonksiyon, modelin çıktısını olasılık olarak yorumlamamızı sağlar. Model, parametrelerini optimize ederek, verilen bir veri noktasının belirli bir sınıfa ait olma olasılığını tahmin eder. Bu olasılık, belirli bir eşik değerine göre pozitif veya negatif sınıf olarak yorumlanır. Lojistik regresyonda optimizasyon kriteri maksimum olabilirlik olarak adlandırılır ve bu, modele göre eğitim verilerinin olabilirliğini maksimize etmek anlamına gelir. Maksimum olabilirlik yöntemi, modelin parametrelerini, verilen verilerin gözlemlerine en iyi uyacak şekilde ayarlar. Bu amaçla, gradyan inişi gibi sayısal optimizasyon teknikleri kullanılır (Burkov, 2019).

Standart lojistik fonksiyon (sigmoid fonksiyonu olarak da bilinir) şu şekildedir:

$$f(x) = \frac{1}{1+ e^{-x}} \tag{2.2}$$

Eşitlik 2.2'de, doğal logaritmanın tabanı (Euler sayısı olarak da bilinir; e^x programlama dillerinde exp(x) fonksiyonu olarak da bilinir). Lojistik regresyon modeli şu şekildedir:

$$f_{w\,b}(x) = \frac{1}{1+ e^{-(wx+b)}} \tag{2.3}$$

(Burkov, 2019).

2.7 Karar Ağacı

Bir karar ağacı, her dallanmada yer alan düğüm bir özelliğe ait değer yer alır. O noktada belirlenen eşik noktasına göre ağacın soldan ya da sağdan gitmesini belirler. Tüm bu dal yapısı yaprak düğüme geldiğinde hangi sınıfa dahil olacağına karar verilir (Burkov, 2019).

Karar ağacı parametrik olmayan hem regresyon hem sınıflama problemlerinde kullanılan bir denetimli öğrenme algoritmasıdır. Hunt Algoritması 1960 yıllarında psikolojide insan ögrenmesi için geliştirilmiş ve sonrasında karar ağacı algoritmalarının temelini oluşturmuştur (Hassani, 2016).

2.8 Rastgele Orman

Leo Breiman ve Adele Cutler imzasını taşıyan algoritmanın amacı fazla karar ağacının çıktısını değerlendirerek tek bir çıktı üretmektedir. Uygulaması kolay ve esnek yapısı nedeniyle hem sınıflama hem de regresyon problemlerinde kullanılmaktadır (Yalovetzky, 2024).

Rastgele Orman algoritması topluluk öğrenmesi (ensemble) yöntemlerinden biri olan torbalama (bagging) metodunu kullanmaktadır. Topluluk öğrenmesi yönteminde oylama, ortalama gibi yöntemlerle farklı doğruluk sonuçları üretilmektedir. Böylelikle birden fazla sınıflayıcı ile daha doğru bir sonuca varılmaktadır. Torbalama ile farklı alt kümeler rastgele seçilir. Sonrasında veri, eğitim ve test olarak ayrılır. Farklı eğitim kümelerinin seçilmesindeki amaç farklı kararlar ortaya koymaktır ve bu kararlar ağırlıklı oylama ile birleştirilmektedir (Cutler, 2011).

2.9 Gradyan Artırma Ağaçları

Artırma (Boosting), makine öğrenmesinde bilinen ve topluluk öğrenmesi algoritmasına dayanan bir yöntemdir. Hem sınıflandırma hem de regresyon problemlerinde yaygın olarak kullanılmaktadır. Boosting teknolojisi, son 30 yılda önemli bir gelişme göstermiştir. Bu yöntemle, zayıf tahmin modelleri kullanılarak güçlü bir topluluk öğrenmesi tahmin edicisi elde edilir. Rastgele Orman yönteminde tahmin modelleri arasında seçim yapılırken, boosting, ileri doğru hareket stratejisiyle ardışık olarak yeni tahmin modelleri eklenir. Bu yöntem, zayıf modelleri sekanslar halinde birleştirerek tahmin gücünü artırır (Guillen, 2023).

İlk aşamada kayıp fonksiyonu tanımlanır. Bu fonksiyon, modelin tahminlerinin ne kadar yanlış olduğunu ölçer. Ardından, modelin ilk tahmini yapılır ve bu tahmin üzerinden hatalar hesaplanır. Gradyan artırma (Gradient Boosting) algoritmasında, her adımda bir ağaç eklenerek bu hatalar minimize edilmeye çalışılır. Bu süreçte her yeni ağaç, bir önceki ağacın hatalarını düzeltmeye çalışır ve tüm ağaçlar sıralı şekilde birbirine bağlıdır. Bu da boosting algoritmalarının yavaş çalışmasına neden olur, ancak sonuçta daha yüksek doğruluk sağlar. Gradyan Artırma öğrenmesinde öğrenme oranı (Learning Rate) ve ağaç sayısı (n_estimators) iki kritik hiperparametredir. Öğrenme oranı, modelin ne kadar hızlı öğrendiğini belirlerken, ağaç sayısı modelin karmaşıklığını ve doğruluğunu etkiler (He, 2019).

2.10 XGBoosting Algoritması

Gradyan Artırma algoritması üzerine inşa edilmiştir. Açılımı eXtreme Gradient Boosting olan bu algoritmayı Chen ve Guestrin büyük ölçekli makine öğrenme yöntemlerinde nasıl etkili çalışacağını anlatmışlardır. Buna göre çok az kaynak kullanarak gerçek dünyaya ait problemleri çözdüğü görülmüştür (Chen, 2016). Bu göstergelerin bazıları da internet üzerinden yapılan Makine Öğrenmesi müsabakalarında ilk ona giren projelerin XGBoost algoritmasını kullanmalarıdır.

XGBoost'u öne çıkaran bazı durumlar vardır. Örnek L1 (Lasso olarak bilinir. Bazı Parametrelerin aldığı değerleri sıfır yapar.) ve L2 (Ridge olarak bilinir. Parametrelerin bazılarını sıfıra yakınsamaya zorlar.) kullanabilme, Ağırlıklı Çeyrek Sıkıştırma tekniğini

kullanarak ağırlıklı verinin daha efektif ele alabilmesi, bloklar şeklindeki yapısından dolayı çoklu işlemciler ile çalışabilme, ön bellek mekanizmasını her bir iş birimi kapsamında ele aldığından dolayı donanımı etkin kullanabilmesi, RAM'e sığmayan büyük veri kümelerinde verileri optimize bir şekilde disk'e yazması nedeniyle çok ciddi bir hız ve model performansı sağlamaktadır. XGBoost sahip olduğu avantajlardan yola çıkarak farklı boosting teknikleri kullanarak daha etkin çözüm arayışları da devam etmektedir (Cui, 2023).

2.11 Genelleştirilmiş Doğrusal Modeller

Genelleştirilmiş Doğrusal Modeller (GLM) geleneksel doğrusal modellerin bir uzantısıdır. Bu modelin istatiksel veri analizinde yaygınlaşmasının sebebi doğrusal regresyon, ikili sınıflama için kullanılan lojistik regresyon gibi tipik modelleri birleştirmesi, modele uyacak yazılımın olması ve büyük veri kümelerine ölçeklenebilir olmalarıdır (Nykodym, 2015).

H2O açık-kaynaklı (Open Source) makine öğrenmesi algoritmalarını içeren bir platformdur. Sigorta, sağlık ve finans kuruluşlarının yapay zeka ve derin öğrenme algoritmalarını kullanarak kompleks problemleri çözmelerine olanak sağlamaktadır. H2O platformu altında GLM'de kullanılmaktadır.

2.12 H2O AutoML Çatısı

H2O AutoML H2O.ai tarafından tasarlanmış bir otomatik makine öğrenmesi çatısıdır. Anlaması ve uygulaması yönünden kolay olması ve yüksek kalitede modeller üretmesi sebebiyle şirketlerin ekosistemine uygundur. H2O platformu ikili sınıflama, çoklu sınıflama ve regresyon problemleri gibi birden fazla probleme cevap vermektedir. Diğer bir önemli özelliği ise hızlı skorlama yapmasıdır, böylelikle kısa sürede tahmin alınabilmektedir. Aynı zamanda sunduğu API (Application Programming Interface: Yazılım uygulamalarının birbirleriyle iletişim kurmasını sağlayan arabirim) ile KNIME gibi farklı platformlardan destek alınabilmekte ve uygulamaya konulabilmektedir. Açık kaynaklı olması, dağıtık yapıda çalışması ve ölçeklenebilir olması sebebiyle endüstri, akademi ve farklı uygulamalarda yaygın biçimde kullanılmaktadır (Madni, 2023).

H2O AutoML aynı zamanda Pyhton, R, Scala ve Java üzerinden de kullanılabilmektedir. Veri ön işleme sürecinde veri tamamlama (imputation), gerekliyse normalleştirme uygulanması ve kategorik değişkenlere dönüştürme teknikleri yer almaktadır. H2O karar ağaçları tekniklerinden Gradyan Makineleri ve Rastgele Orman kullandığından dolayı kategorik veriyi de işleyebilmektedir. H2O AutoML geliştirme yol haritasında boyut indirgeme yöntemi de eklenmiştir. 2017 Yılından beri kullanılmakta olan bu platform, yeni geliştirmeler ile büyümektedir. H2O içinde kullanılan algoritmalar: XGBoost Gradyan Artırımı Makineleri, H2O Gradyan Artırımı Makineleri, Rastgele Orman (hem varsayılan hem de aşırı rastgele versiyonu), Derin Yapay Ağlar ve Genelleştirilmiş Doğrusal Modeller (LeDell, 2020).

2.13 Sinir Ağı

Bir sinir ağı (NN), tıpkı regresyon veya SVM modeli gibi, matematiksel bir fonksiyondur:

$$y = f_{NN}(x) \tag{2.4}$$

Eşitlik 2.4'te yer alan f_{NN}, iç içe geçmiş bir fonksiyon yapısına sahiptir. Üç katmanlı bir sinir ağı için f_{NN} şu şekilde ifade edilir:

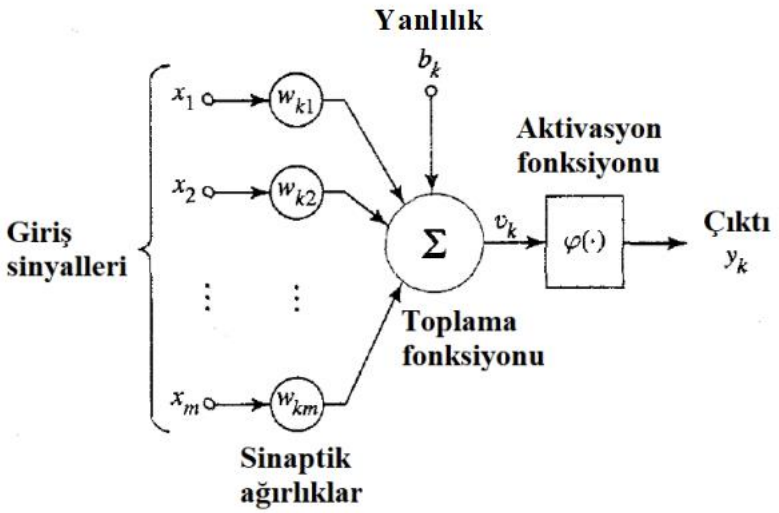
$$y = f_{NN}(x) = f_3(f_2(f_1(x))) \tag{2.5}$$

Eşitlik 2.5'te yer alan f_1 ve f_2 vektör fonksiyonlarıdır.

$$f_1(z) = g_1(W_l z + b_1) \tag{2.6}$$

Eşitlik 2.6'da yer alan l, katman indeksidir ve 1'den başlayarak herhangi bir katman sayısına kadar uzanabilir. g_1, aktivasyon fonksiyonu olarak adlandırılır. Aktivasyon fonksiyonları arasında en yaygın olarak kullanılanlar Tanh (hiperbolik tanjant) ve ReLU (Düzeltilmiş Doğrusal Birim) 'dur. Tanh fonksiyonu çıktıyı -1 ile 1 arasında sınırlandırır. ReLU fonksiyonu, negatif girişleri 0'a dönüştürürken, pozitif girişleri olduğu gibi bırakır. Her katmanın parametreleri, W_l (bir matris) ve b_1 (bir vektör), gradyan iniş yöntemi kullanılarak öğrenilir. Bu öğrenme süreci, belirli bir maliyet fonksiyonunu (örneğin, MSE

(Ortalama Kare Hatası)) optimize etmeye dayanır. Eşitlik 2.5'te yer alan fonksiyon f_3, regresyon görevi için bir skalar fonksiyon olabilir, ancak probleme bağlı olarak bir vektör fonksiyonu da olabilir (Burkov, 2019).



Şekil 2.2. Sinir Ağı (Haykin, 1999)

2.13 Derin Öğrenme

Derin öğrenme, iki veya daha fazla ara katmana sahip sinir ağlarının eğitilmesini ifade eder. Geçmişte, bu tür ağları eğitmek katman sayısı arttıkça daha zor hale gelmişti. İki büyük zorluk, patlayan gradyan ve kaybolan gradyan sorunları olarak adlandırılmıştır. Patlayan gradyan problemi, gradyan kesme ve L1 veya L2 düzenleme gibi tekniklerle ele alınabilirken, kaybolan gradyan problemi uzun yıllar çözümsüz kalmıştı. Kaybolan gradyan, sinir ağlarının parametrelerini güncellemek için kullanılan geri yayılım algoritmasının, bazı durumlarda gradyanın çok küçük olması nedeniyle parametre değerlerini etkili bir şekilde değiştirmekte zorlanmasıyla ortaya çıkar. Bu durum, ağın eğitimini tamamen durdurabilir (Burkov, 2019).

Geleneksel aktivasyon fonksiyonları, örneğin Tanh gibi, (0, 1) aralığında gradyanlara sahiptir ve geri yayılım sırasında gradyanların üssel olarak azalmasına neden olur. Bu, erken katmanların yavaş veya hiç eğitilmemesiyle sonuçlanabilir. Ancak, modern sinir ağı öğrenme algoritmaları, ReLU ve LSTM (Uzun-Kısa Süreli Bellek) gibi iyileştirmeler ve artık sinir ağlarındaki atlama bağlantıları gibi tekniklerle çok derin sinir ağlarını (yüzlerce katmana kadar) etkili bir şekilde eğitmenizi sağlar. Bu nedenle, günümüzde

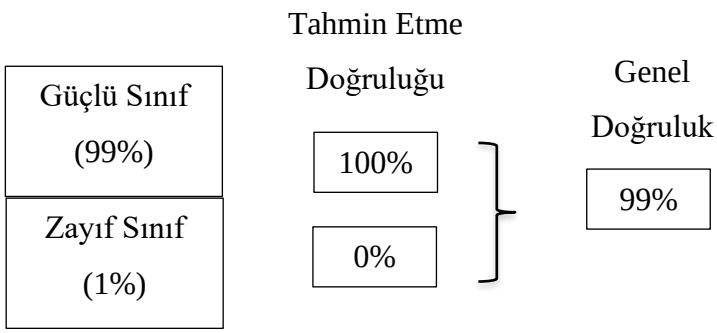
patlayan ve kaybolan gradyan problemleri büyük ölçüde çözülmüş veya etkileri azaltılmıştır (Burkov, 2019).

Konvolüsyonel Sinir Ağları (CNN) ve Tekrarlayan Sinir Ağı (RNN), derin öğrenme alanındaki bu zorlukları aşmak için geliştirilmiş özel sinir ağı türleridir. CNN, görüntü ve metin işleme gibi uygulamalarda başarılı sonuçlar verir. Görüntülerde yerel özellikleri tanımak için küçük kare pencereler (filtreler) kullanarak veriyi tarar ve bu filtreler belirli desenleri tespit etmek üzere eğitilir. Ayrıca, havuzlama (pooling) teknikleriyle modelin karmaşıklığı azaltılır ve hesaplama yükü hafifletilir (Burkov, 2019).

RNN'ler ise sıralı verilerle çalışmak için tasarlanmıştır ve metin ve konuşma işleme gibi alanlarda kullanılır. Her adımda hem mevcut giriş vektörünü hem de önceki adımlardaki gizli durumu (bellek) kullanarak veriyi işler. Bu yapısıyla zaman içindeki bağımlılıkları modelleyebilir. Ancak uzun dizilerle çalışırken uzun dönem bağımlılıklarını öğrenme konusunda zorluk yaşar, bu sorunları aşmak için LSTM ve GRU (Kapılı Yinelemeli Birim) gibi kapalı birimler geliştirilmiştir. Bu birimler, bilgiyi daha uzun süre saklayabilir ve önemli bilgilerin korunmasını sağlar (Burkov, 2019).

2.14 Dengesiz Sınıflama

Dengesizlik bir veya birden fazla sınıfın, diğer sınıf öğelerine göre eğitim setinde çok düşük ya da çok büyük oranda olmasıdır (Kuhn, 2012). İncelenen veride ikili sınıflama durumu vardır. Memnuniyetsiz müşteri oranı memnun olan müşterilerin sayısına oranla çok düşük kalmaktadır. Bu nedenle Dengesiz Sınıflama problemine yönelik kullanılan teknikler ve ölçümleme yöntemleri incelenecektir.



Şekil 2.3. Dengesiz Veri Probleminin Görselleştirilmesi

Şekil 2.3'te gösterildiği üzere zayıf sınıf toplam veride %1'lik bir oranda temsil edilmektedir. Buradaki sorun model eğitildiğinde doğruluk oranının hep %99 civarında olduğudur. Yani modeli hiçbir zaman doğru bir öğrenme gerçekleştirememektedir.

Dengesiz sınıflamada karşılaşılan en büyük sorun Doğruluk değerinin yüksek gelmesidir. Dengesiz sınıflama işleminde en basit yaklaşım modele ince ayar yapmaktır. Bu şekilde düşük olan duyarlılık değerini artırma yoluna gidilir. Örnek olarak daha az sayıda değişkenle eğitmek olabilir (Kuhn, 2012).

Birçok makine öğrenmesi algoritması, Rastgele Orman, Geri Beslemeli Sinir Ağları (BPNN), ve Destek Vektör Makineleri (SVM) gibi algoritmalar ön şart olarak sınıfların eşit dağıldığını kabul eder. Bu şekilde çok değerli olan ve az sayıda olan bir sınıf göz ardı edilmiş olmaktadır (Basha, 2022).

2.14.1 Oversampling, Undersampling ve SMOTE Teknikleri

Oversampling, undersampling ve SMOTE (Synthetic minority over-sampling) teknikleri yardımıyla, azınlık sınıfının verileri çoğunluk sınıfının verileriyle dengelenir. Oversampling ile azınlık sınıfındaki veriler artırılırken, undersampling ile çoğunluk sınıfındaki veriler azaltılır.

En popüler oversampling yöntemlerinden biri, Chawla ve ark. (2002) tarafından geliştirilen SMOTE'dur. SMOTE yöntemi, azınlık sınıfı noktasını ve en yakın κ komşularından birini kullanarak doğrusal interpolasyon uygulayarak sentetik veriler oluşturur. SMOTE, birçok uygulamada yaygın olarak benimsenmiş güçlü bir oversampling yöntemidir (Fernandez, 2018; Ahsan, 2018; Kishor, 2021). Ayrıca, Borderline SMOTE (Han, 2005), Safe-level SMOTE (Bunkhumpornpat, 2009), ADASYN (He, 2008), SVM SMOTE (Nguyen, 2011), Localized Random Affine Shadowsampling (LoRAS) (Bej, 2021), CDSMOTE (Elyan, 2021) ve Deep SMOTE (Dablain, 2022) gibi birçok SMOTE genişletmesi geliştirilmiştir (Bolívar, 2022).

Öğrenme algoritması sınıflara düzgün bir şekilde ağırlık verme imkanı sunmadığında, oversampling tekniği denenebilir. Bu yöntemde, bazı sınıflarda yer alan önemli örneklem verileri birçok kez çoğaltılır. Bunun tam tersi yönteme ise undersampling denir. Bu

yöntemde, eğitim kümesindeki çoğunluk sınıfa ait bazı örneklem kayıtları rastgele olarak çıkartılır (Burkov, 2019).

Diğer bir yaklaşım ise, azınlık sınıfına ait sütunlardan rastgele değerler alarak sentetik örneklem verisi oluşturulması ve sonrasında bu sentetik veri ile asıl verinin birleştirilerek yeni bir örneklem kümesi oluşturulmasıdır. Bu alanda en popüler kullanılan iki teknik SMOTE ve ADASYN (Adaptive Synthetic Sampling Method) yöntemleridir. Bu iki yöntem ile azınlık sınıfa ait sentetik veri üretimi yapılır (Burkov, 2019).

Azınlık sınıfına ait verilen x_i örnekleri için k-en yakın komşular ile örnekler seçilir. (k kadar alınan örnek veri kümesi S_k ile gösterilsin) Buradan yola çıkarak x_{yeni} isminde yeni bir sentetik örnek kümesi, Eşitlik 2.7'de yer alan formül ile oluşturulur:

$$x_i + \lambda(x_{zi} - x_i) \tag{2.7}$$

Eşitlik 2.7'de x_{zi} azınlık sınıfına ait S_k kümesinden rastgele seçilmiş örneklemdir. λ ara değer hiperparametresi 0 ile 1 arasından rastgele seçilen bir sayıdır. Bu yöntemle veri kümesindeki tüm olası x_i'ler seçilerek çok sayıda sentetik veri üretilir. Böylece azınlık sınıfına ait nadir bulunan örnekler çoğaltılmış olur (Burkov, 2019).

2.14.2 Alıcı İşletim Karakteristik Eğrisi (ROC)

ROC eğrisi genel olarak sınıflandırma modellerinin değerlendirilmesi için kullanılan bir metottur. ROC eğrisi, TPR (True Positive Rate - Doğru Pozitif Oranı) ile FPR (False Positive Rate - Yanlış Pozitif Oranı) kombinasyonunu kullanarak sınıflama performansına dair bir özet bir resim çıkarır. ROC eğrisi sadece belli bir güven rakamı veya olasılığa sahip olan sınıflayıcıları değerlendirmek için kullanılabilir. ROC eğrisinin yaygın kullanılmasının nedeni anlaşılır olmasıdır (Burkov, 2019).

$$\text{TPR} = \frac{\text{TP}}{(\text{TP}+\text{FN})} \tag{2.8}$$

$$\text{Seçicilik} = \frac{\text{TN}}{(\text{TN}+\text{FP})} \tag{2.9}$$

$$\text{FPR} = 1 - \text{Seçicilik} = \frac{\text{FP}}{(\text{TN}+\text{FP})} \quad (2.10)$$

Kesinlik-Duyarlılık (Precision-Recall (PR)) Grafiği, dengesiz veri kümeleri üzerinde ikili sınıflayıcıları değerlendirirken ROC Grafiğinden daha bilgilendiricidir. PR eğrisi, azınlık sınıfına odaklanarak ve modelin pozitif tahminler arasındaki gerçek pozitifleri ne kadar iyi ayırt ettiğini daha açık bir şekilde göstererek, yüksek derecede dengesiz veri kümelerinde model performansını değerlendirmede ROC eğrisinden daha bilgilendirici olmaktadır (Saito, 2015).

2.14.3 F-Ölçüsü

F-Ölçüsü ağırlıklı harmonik ortalama hesaplamasını Kesinlik ve Duyarlılık değerleri üzerinden yapar. Beta kullanıcının kesinlik ve duyarlılıktan hangisine daha fazla önem verdiğini belirlemekte kullanılan bir parametredir. Beta'nın sonsuz olması kesinliğin herhangi bir önemi olmadığını ifade ederken beta'nın 0 olması duyarlığın öneme sahip olmadığı anlamına gelir. Eğer beta 1'e eşitlenirse her ikisine verilen önemin aynı derecede olduğunu gösterir (Christen, 2023).

$$\text{F}_\beta = (1 + \beta^2) * \frac{\text{kesinlik} * \text{duyarlılık}}{(\beta^2 * \text{kesinlik}) + \text{duyarlılık}} \quad (2.11)$$

Dengesiz verinin değerlendirilmesinde kullanılan bir performans ölçütüdür (Minaee, 2021).

2.14.4 Kesinlik

Kesinlik bir makine öğrenmesi algoritmasının hedef sınıfını ne kadar doğru tahmin ettiği hakkında bilgi vermektedir. Diğer bir ifadeyle Kesinlik, bize pozitif tahminlerin hangi oranda doğru olduğunu gösterir:

$$\text{Kesinlik} = \frac{\text{TP}}{\text{TP}+\text{FP}} \quad (2.12)$$

Kesinlik'in sağladığı iki ana fayda vardır:

- Dengesiz veride hedef sınıfı tespit etmede modelin doğruluğu hakkında bilgi verir.

- Diğer taraftan FP (False Positive – Yanlış Pozitif) in maliyeti (işlem süresi, hesaplama kaynakları) yüksek ise yani bedeli ağır olacaksa Kesinlik kullanımı hedef sınıf tespit etmede daha güvenilir bir sonuç verecektir.

2.14.5 Duyarlılık

TPR, modelin doğru tahmin ettiği pozitif verilerin oranını ölçen bir metriktir. Denetimli öğrenme altındaki ikili sınıfların anlaşılmasında kullanılan bir metriktir. FN (False Negative – Yanlış Negatif) kümesinde yer alanlar aslında pozitif olduğu halde makine öğrenmesinin yanlış tahmin ettikleridir. Buna göre formül şu şekildedir:

$$\text{Duyarlılık} = \frac{\text{TP}}{\text{TP+FN}} \tag{2.13}$$

Dengesiz öğrenmede Duyarlılık ölçütü azınlık sınıfını anlamada kullanılan tipik bir ölçüttür (He, 2013).

2.14.6 Matthews Korelasyon Katsayısı (MCC)

1975 yılında Brian Matthews tarafından bulunmuş ve MCC olarak kısaltılmıştır. İkili sınıf problemlerine ait model değerlendirmelerinde kullanılan istatistiksel bir araçtır. Sınıflama probleminde kullanılan karışıklık matrisindeki TP (True Positive - Doğru Pozitif), TN (True Negative - Doğru Negatif), FP, FN çıktılarını kullanarak 0 ile 1 arasında bir değer üretir. 1 değeri tahmin edilen değer ile gerçek arasındaki yüksek uyumluluğa işaret ederken 0 değeri hiç uyum olmadığı anlamına gelmektedir. Bunun anlamı ise tahmin edilen değerin gerçek değerler açısından raslantısal olduğunu gösterir.

$$\text{MCC} = \frac{\text{TP}*\text{TN}-\text{FP}*\text{FN}}{\sqrt{(\text{TP+FP})*(\text{TP+FN})*(\text{TN+FP})*(\text{TN+FN})}} \tag{2.14}$$

Dengesiz verinin değerlendirilmesinde kullanılan Doğruluk ve Kesinlik, üzerinden hesaplanan geometrik ortalama, Duyarlılık ve F1 gibi ölçütlerin uygun olmadığını gösteren ölçümler mevcuttur. Bilimsel makalelerde MCC ölçütünün dengesiz veride kullanılıp kullanılmayacağı ile fikir ayrılığı ve farklı yorumlar mevcuttur. Yapılan test sonuçlarına göre veri daha çok dengesiz oldukça MCC ve Cohen's Kappa (CK) değerlerinin çarpıklığının arttığı görülmektedir. Bu sebepten dolayı MCC'de F1, MK ve CK ile beraber tam olarak dengesiz verinin değerlendirilmesinde doğrudan uygulanamaz. BM ve GBA ölçütleri dengesiz veride hep tutarlı kalmıştır (Zhu, 2020).

2.14.7 Duyarlılık (TPR) ve Seçicilik (TNR) üzerinden kullanılan ölçütler

Duyarlılık ile gerçekte memnun olmayanların arasında testin pozitif sonuç verme oranı ölçülür ve Seçicilik ile gerçekte memnun olanların arasında testin negatif sonuç verme oranı üzerinden değerlendirme yapılır. Burada bu iki ölçütün geometrik ortalamasını alan ölçüt, GBA olarak kısaltıldı (Zhu, 2020). Geometrik ortalama sınıf büyüklüklerinden etkilenmeyeceği için dengesiz sınıfa ait problemlerden bağımsızdır. Bu sebepten dolayı daha tutarlı sonuçlar vermektedir (Zhu, 2020).

$$\text{TPR ve TNR'nin Geometrik Ortalaması} = \sqrt{\text{TPR} * \text{TNR}} \quad (2.15)$$

Kullanılacak başka bir ölçüt olan ağırlıklı TPR – TNR ölçütü de değerlendirmeye dahil edilecektir. Dengesiz veride kullanılan performans ölçütlerinde sınıfa ait dengesizliklerin dikkate alınıp alınmayacağı önemli ayrım noktalarından birisidir. Buradan yola çıkarak Duyarlılık ve Seçicilik ölçütlerini kullanarak ağırlıklı bir formül kullanılacaktır. Formül kısaltması WPN olarak şeklinde kullanılacaktır (Jadhav, 2020).

$$\text{Ağırlıklı TPR} - \text{TNR} = \left(\text{TPR}\frac{\text{N}}{\text{P+N}}\right) + \left(\text{TNR}\frac{\text{P}}{\text{P+N}}\right) \quad (2.16)$$

Formül 2.15'te yer alan P değeri toplam pozitif durumların sayısını, N değeri ise toplam negatif durumların sayısını vermektedir.

Bookmaker Informedness (BM) ölçütü aynı zamanda Youden indeksi veya Jouden'ın J istatistiği olarak da bilinmektedir ve BM şeklinde kısaltılmıştır. Dengesiz ve dengeli veri kümelerinde BM ölçütü daha doğrusal ve oldukça tutarlı sonuçlar vermiştir (Zhu, 2020).

$$\mathrm{BM} = \mathrm{TPR} + \mathrm{TNR} - 1 \tag{2.17}$$

2.15 Temel Bileşenler Analizi

Temel bileşenler analizi (TBA), modern veri analizinde (nörobilimden bilgisayar grafiklerine kadar çeşitli alanlarda) standart bir araçtır. TBA, karışık veri kümelerinden önemli bilgileri çıkarsamada yaygın olarak kullanılan basit ve açıklaması kolay bir çok değişkenli analiz yöntemidir. Minimum çabayla, karmaşık veri setlerini daha düşük boyuta indirger. Bu sayede, verinin altında yatan gizli ve basitleştirilmiş yapıları ortaya çıkarır (Shlens, 2014).

TBA'nın temel avantajı ve veri indirgeme yöntemi olarak popülerliğini sürdürmesinin nedeni, bileşenlerin birbiriyle ilişkisiz olmasıdır. Bazı modeller, tahmin edicilerin ilişkisiz (veya en azından düşük korelasyonlu) olmasını çözüm bulma ve modelin sayısal kararlılığını artırma açısından tercih eder (Kuhn, 2012).

Belirgin çarpıklık veya basıklık gösteren değişkenler, normal dağılıma daha iyi yaklaşmak için dönüştürülebilir veya normalize edilebilir (Rosenbaum, 2010). Bu nedenle, normalizasyon adımı, değişkenler normal dağılıma sahip olmadığında önemli bir ön işleme adımıdır. Verilerin normalize edilmesi, TBA'nın sonucunu iyileştirmektedir (Salama, 2010).

3. MODEM VERİSİ VE UYGULAMA ÇALIŞMALARI

3.1. Veri Analizi İçin Gerekli Olan Ortamın ve Verinin Hazırlanması

Mevcutta kullanılan modemlerin verileri saatlik olarak okunup JSON adı verilen veri formatında saklanmaktadır. Bu veri daha sonra FTP ile verinin işleneceği sunuculara aktarılmaktadır. Sunuculara gelen veriler analiz edilip okunabilecek formata dönüştürülerek PostgreSQL veri tabanına yazılmaktadır. Saatlik verilerin son 5 gün, günlük ortalama verilerin ise 40 gün geriye yönelik tutulmasından dolayı veri analizi için kullanılması düşünülen verilerin alınıp saklanması gerekmektedir.

Sadece İstanbul bölgesine ait son 1 haftalık veriler (2,3 milyon) KNIME Analitik Platformu kullanılarak veri tabanından okunmuş ve veri analizi çalışmalarının yapılacağı sunucuya parçalı olarak (saatlik bazdaki veriler) aktarılmıştır. Parça parça yapılmasının sebebi veri okuma ve yazma sürelerinin uzun sürmesidir.

Bu veride yer alan yükleme ve indirme bilgileri 4 saatlik aralıklarla, sorunlu modem sinyallerini gösterebilmek için PIVOT (Değişkenin sütunlara dönüşmesi) yöntemi ile veri tabanı bölümleme yapıları kullanılarak sütunlara yazılmıştır. Veri tabanı bölümlemesi verilerin günlük olarak veri tabanında ayrı tablolarda tutulmasını sağlar ve bu sayede okuma işleminde hız kazandırır. PIVOT işlemi ise download ve upload bilgilerinde yer alan sorunlu sinyal bilgilerinin tek sütunda değil de günlük 4 saatlik dilimlerde gösterebilmek için KNIME Analitik Platformu kullanılarak yapılmıştır.

Bu çalışma tamamlandıktan sonra müşterilere ait diğer internet ile alakalı verileri CRM veri tabanından alınmıştır. Buradan müşteriye ait internet yükleme hızı, indirme hızı, ortalama indirme miktarları, internet ile alakalı açtığı arıza kayıt adeti, eski arıza kayıt adeti, bina bilgisi, il bilgisi, ilçe bilgisi, bölge bilgisi, binaya ait homepass (binadaki potansiyel abone sayısı) gibi diğer sütunlar da veriye eklenerek veri zenginleştirilmiştir.

Müşteri memnuniyetini tespit etmek için bağımlı değişken oluştururken iki ana kriterden faydalanılmıştır. Birincisi, abonenin kullandığı modeme ait sorunlu sinyalin varlığı, ikincisi ise arıza kaydının bırakılması durumudur. Bu iki kriter kullanılarak işaretleme yapılmıştır. İşaretleme yapılırken sahada yer alan diğer cihazlar (ana dağıtım kutuları)

değerlendirmeye alınmamıştır. Sebebi o veriler için mevcutta bir yazılım altyapısının olmaması ve verilerin toplanıp bir yerde kaydedilmemiş olmasıdır. Kaydedilse bile verilerin veri analizi yapabilmek için veri işleme süreçlerine tabi tutulması gerekmektedir.

Veriler hazırlandıktan sonra KNIME aracılığı ile AutoML düğümünü kullanarak veri analizi aşamalarına geçilmiştir. Burada diğer bir konu ise performans ve metriklerin kaydedilmesi için ek geliştirme gerektirdiğidir. Ayrıca değerlendirme metriklerinde Doğruluk, CK, MCC, F1 gibi dengeli veride kullanılabilecek ölçütlerin haricinde başka ölçütlerin yer almamasıdır. Hem çalışmada kullanılan GBA, BM ve WPN ölçütleri için hem de doğrulama/test sonuçlarını kaydetmek için KNIME Analitik Platformunda makine öğrenmesi dosyası üzerinde ek geliştirmeler yapılmıştır.

3.2. Verilerin Seçilmesi

Çalışmaların tamamı verinin büyük olması ve işlenmesindeki zorluklar nedeniyle 256 GB Ram ve 64 CPU'luk bir makine ile yapılmıştır. Diğer bir zorluk milyonlarca verinin içindeki bazı verilerin PIVOT'a dönüştürme ve JSON ayrıştırma işleminde çok yüksek CPU ve RAM kullanımı gerektirmesidir. Veriyi daha iyi okuyabilmek için farklı yöntemler denedikten sonra son olarak verinin günler bazında parçalanmış tablo yapısında tutulmasına karar verilmiştir.

Çalışma kapsamında değerlendirilecek veri için yansız, bağımsız, eşit ve Türkiye genelini en iyi temsil edeceği düşünülen İstanbul verisi seçilmiştir. JSON veri içeren bu veriyi sunuculara kaydetme, uygun işleme yönteminin tespit edilmesi ve dönüştürme süreçleri bir aylık bir çalışmayı kapsamaktadır. Bu veri ilk önce PostgreSQL veri tabanında oluşturulan bölümlere kaydedilmiştir. Buradaki bölümleme yapısı günlük bazda çalışmaktadır. Verilerde müşteri ile ilişkilendirebilecek veri kullanılmamıştır. Veriler anonimleştirilmiş ve karıştırma (shuffle) teknikleri ile veri karıştırılmıştır. Ayrıca internet abonelerine ait modem verilerinden işleme izni veren abonelere ait veriler kullanılmıştır.

Sonrasında İndirme-Yükleme verileri günün dilimlerine bölünerek hem indirme hem de yükleme bazında pivot uygulanarak sütunlara dönüştürülmüştür:

- 0-4 arası saat 24'ten saat 03:59:59'a kadar olan veriyi kapsar.

- 4-8 arası saat 4'ten saat 07:59:59'a kadar olan veriyi kapsar.
- 8-12 arası saat 8'den saat 11:59:59'a kadar olan veriyi kapsar.
- 12-16 arası saat 12'den saat 15:59:59'a kadar olan veriyi kapsar.
- 16-20 arası saat 16'dan saat 19:59:59'a kadar olan veriyi kapsar.
- 20-24 arası saat 20'den saat 23:59:59'a kadar olan veriyi kapsar.

"Parallel Chunk Start" ve "Parallel Chunk End" düğümleri arasında yer alan süreçler birden fazla iş parçacığı (7 thread) ile işlenmektedir. Her bir gözlem verisi döngüye girerek; bulunduğu saate ait JSON verisinde yer alan ve sorunlu olarak işaretlenen İndirme-Yükleme değerlerine göre okuma yapılmaktadır. Tüm veriler tamamlandığında PIVOT işlemi esnasında her bir abonenin 24 satırlık verisi yukarıda yer alan saat dilimlerine göre pivota dönüştürülmektedir.

Detaylı bir SQL sorgusu ile veri tabanına giderek günlük olarak parçalanmış tablolardan okuma yapılır. Bu şekilde tüm veriye gitmeyerek hız kazanımı elde edilmiştir. Aynı zamanda sorgu yapısında JSON ve Lateral (veri tabanı sorgu çekme tekniği) yapısı kullanılarak pratik ve hızlı bir şekilde veriler içinden sorunlu sinyal değerleri okunmuştur. Son olarak ise okuma zamanında yer alan saat bilgisi çıkartılarak gruplama işlemleri yapılmıştır. Bu işlemlerin veri tabanı seviyesinde yapılması da hız kazandırmıştır. Tüm hizmetlere ait verilerde saatlik dönüşüm tamamlandıktan sonra KNIME pivot düğüm ile satırdan sütuna geçiş yapılmıştır.

Homepass yani binadan alınabilecek hizmet sayısı veriye eklenmiştir. Bunun yanında abonelerin aylık kota verileri incelenmiş ve indirme-yükleme sayıları da veriye eklenmiştir. İncelenen veri bir haftalık olduğundan geçmişe yönelik olarak toplam kaç tane arıza bıraktığı da arıza alışkanlığı olarak veri kümesine eklenmiştir. Ayrıca o hafta arıza bırakan abonelerin verisi de alınmıştır. Bağlantı hızı, yükleme hızı ve santralden verinin okunduğu tarih bilgileri de eklenmiştir.

3.2.1. Büyük Veriyi İşleme Süreçleri

Modem verileri yapısal olmayan veri şeklinde sisteme gelmektedir. Daha önce yapılan Veri Mühendisliği işlemleriyle bu veriler veri tabanına öncelikle saatlik olarak kaydedilmiş, ardından günlük ortalamalar şeklinde depolanmıştır.

İkinci aşamada, modem verilerinde bulunan İndirme-Yükleme bilgileri haftanın günlerine göre ayrıştırılmış ve hata alınan sinyal değerleri günün altı dilimine bölünerek sütunlara dönüştürülmüştür.

Üçüncü aşamada ise, kota verileri ve CRM verileri alınarak birleştirilmiş ve tez kapsamında kullanılmak üzere hazır hale getirilmiştir.

Veri işleme sürecinde diğer zorluk JSON verisinin ayrıştırılması işlemidir. Burada modeme ait her gün için 24 saat veri kaydı bulunmaktadır. Her saatlik veride ise 24 tane port için yükleme ve indirmeye ait detaylı sinyal verisi içermektedir. Buradaki diğer zorluk ise toplama ve gruplama (aggregation) işleminin yüksek donanım istemesidir. JSON verisinin ayrıştırılması işleminin KNIME ile yapılması da ayrı bir zorluk çıkardığı deneysel olarak tespit edilmiştir. Bundan dolayı KNIME geliştirilen paket ile PostgreSQL'e ait JSON kütüphaneleri kullanılmış ve aynı zamanda KNIME veri tabanı düğümleri kullanılarak işlem yükü veri tabanı seviyesine alınmıştır.

Örnek JSON verisi modemlerden okunan sinyal değerlerini port bazında vermektedir. Port yapısı abonenin internete çıkmasını sağlanan kanallardır. Bu kanallar bölgedeki kapasiteye göre belirlenerek modemlere tanımlanmaktadır. Şöyle ki o bölgede aşırı bir internet yoğunluğu olduğunda aboneler internete yavaş erişme veya hiç erişememe durumu yaşamamaları için o bölgedeki modemlere ek kanallar tanımlanmaktadır. Genel olarak modemlerde indirme portlarında 2, yükleme portlarında ise 4-8-12-16-24 vs. adet kanallar tanımlanmaktadır. Aşağıda yer alan örnek veride interface yazan kısımda /19 veya /4 portun sıra numarasını ifade etmektedir. Bu portların içinde yer alan cer_dn, ccer_dn, snr_dn ve rxpower_dn tanımları yer almaktadır. Bu tanımların sahip olduğu belli aralıklar bulunmaktadır. Eğer o kanalda yer alan değer belirlenen aralığın içinde değilse o kanal için sorunlu sinyale sahip denilmektedir. Müşteri memnuniyetini belirlerken kriter olarak kullanılan sorunlu sinyal değişken bilgisinin içeriği buradan elde

edilmektedir. Bir aboneye ait saatlik bazdaki verinin metin olarak büyüklüğü ise bir A4 sayfası kadardır:

[{"interface" : "11/1/19", "modem_port" : 19, "port_type" : "DS", "sorunlu" : false, "cer_dn" : 0.00, "ccer_dn" : 0.00, "snr_dn" : 39.50, "rxpower_dn" : 9.50}, {"interface" : "11/1/4", "modem_port" : 4, "port_type" : "DS", "sorunlu" : false, "cer_dn" : 0.00, "ccer_dn" : 0.00, "snr_dn" : 38.00, "rxpower_dn" : 8.30}]

Tez kapsamında kullanılan değişkenler:

Kısaltmalar: [sps: sorunlu port sayısı] [d: download -indirme][u: upload - yükleme]
sps_d_0_4 [0-4][00:00 ile 03:59 arası]
sps_d_4_8 [4-8][04:00 ile 07:59 arası]
sps_d_8_12 [8-12][08:00 ile 11:59 arası]
sps_d_12_16 [12-16][12:00 ile 15:59 arası]
sps_d_16_20 [16-20][16:00 ile 19:59 arası]
sps_d_20_24 [20-24][20:00 ile 23:59 arası]
sps_u_0_4 [0-4][00:00 ile 03:59 arası]
sps_u_4_8 [4-8][04:00 ile 07:59 arası]
sps_u_8_12 [8-12][08:00 ile 11:59 arası]
sps_u_12_16 [12-16][12:00 ile 15:59 arası]
sps_u_16_20 [16-20][16:00 ile 19:59 arası]
sps_u_20_24 [20-24][20:00 ile 23:59 arası]
homepass: bina içinde yer alan potansiyel abone sayısı
ariza_adet: abonenin açtığı arızanın sayısı
baglanti_hizi: abonenin kullandığı internetin hızı
yükleme_hizi: abonenin kullandığı yükleme hızı
ariza_aliskanliği: geçmişe yönelik abonenin açtığı arızaların toplam sayısı
last_ds_bytes: indirme miktarı
last_us_bytes: yükleme miktarı

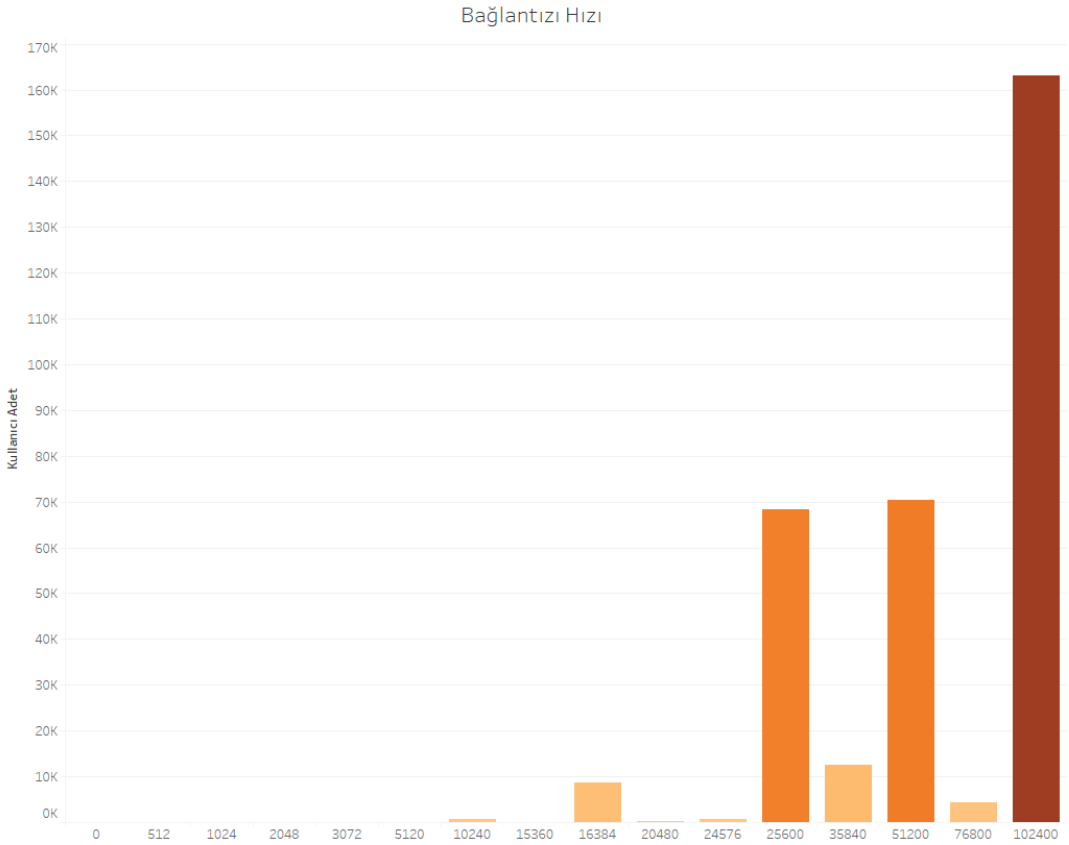
Bir haftalık verinin sayısal değişkenlere ait betimsel istatistiği Tablo 3.1'de verilmiştir:

Sütun Adı	En Düşük	En Yüksek	Ortalama	Standart Sapma	Varyans	Çarpıklık	Basıklık	Eksik Bilgi
0-4 arası Sorunlu Port Sayısı (Down.)	0	192,00	5,67	23,97	574,67	5,63	34,54	40417
0-4 arası Sorunlu Port Sayısı (Up.)	0	256,00	22,51	46,49	2161,64	2,39	5,11	40417
4-8 arası Sorunlu Port Sayısı (Down.)	0	192,00	5,53	23,64	558,67	5,72	35,84	124124
4-8 arası Sorunlu Port Sayısı (Up.)	0	192,00	19,40	43,01	1849,48	2,74	7,27	124124
8-12 arası Sorunlu Port Sayısı (Down.)	0	192,00	5,55	23,78	565,70	5,71	35,51	76194
8-12 arası Sorunlu Port Sayısı (Up.)	0	256,00	17,69	42,25	1785,08	2,90	8,11	76194
12-16 arası Sorunlu Port Sayısı (Down.)	0	192,00	5,54	23,45	550,08	5,68	35,33	52074
12-16 arası Sorunlu Port Sayısı (Up.)	0	192,00	18,57	42,20	1780,80	2,76	7,31	52074
16-20 arası Sorunlu Port Sayısı (Down.)	0	192,00	5,70	24,00	576,16	5,62	34,44	45835
16-20 arası Sorunlu Port Sayısı (Up.)	0	192,00	19,54	43,90	1926,81	2,68	6,77	45835
20-24 arası Sorunlu Port Sayısı (Down.)	0	192,00	5,78	24,17	584,03	5,57	33,88	25938
20-24 arası Sorunlu Port Sayısı (Up.)	0	192,00	24,34	46,88	2198,02	2,25	4,50	25938
Homepass	0	2564,00	17,90	33,43	1117,64	42,35	2995,48	0
Arıza Adeti	1	3,00	1,07	0,26	0,07	4,04	16,72	2264005
Bağlantı Hızı	0	102400,00	69988,32	33768,51	1140312551,44	-0,23	-1,70	0
Yükleme Hızı	128	20480,00	4680,15	2564,85	6578442,45	4,21	24,77	21
Arıza Alışkanlığı	0	2201,00	6,30	9,48	89,89	48,22	9758,55	0
Son Aya Ait Download	0	17776,51	240,47	236,40	55887,02	6,09	179,39	0
Son Aya Ait Upload	0	4513,58	16,23	33,33	1110,67	35,24	2788,17	0

Tablo 3.1. Sütunlara Ait Betimsel İstatistik

3.2.2. Verinin İncelenmesi

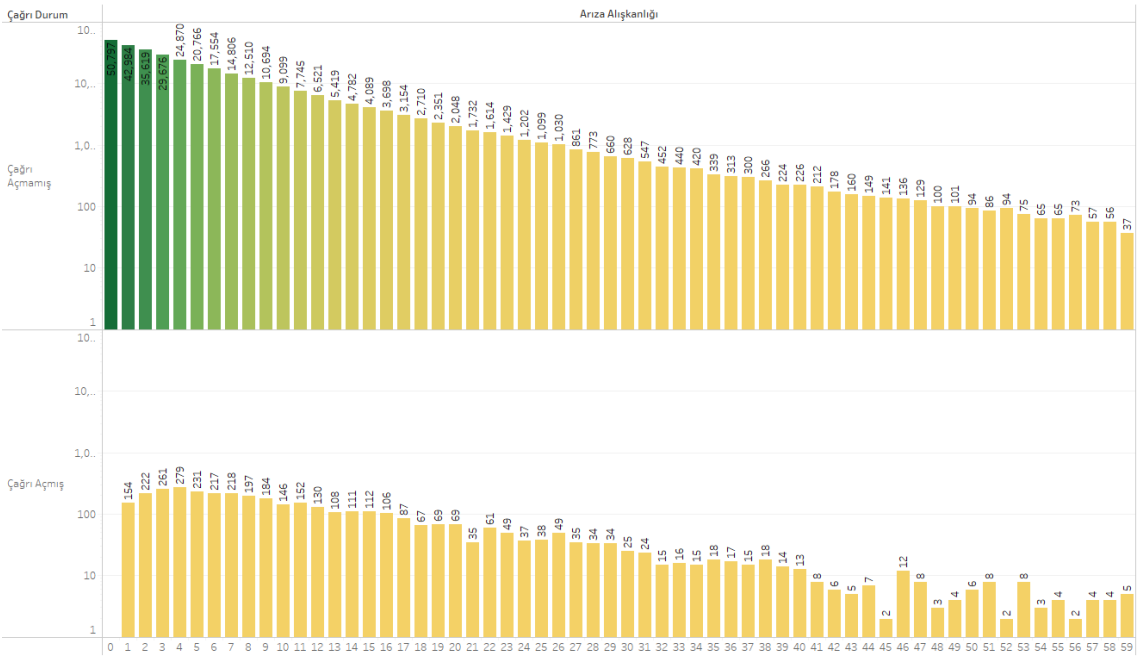
Aşağıda yer alan çubuk grafiği üzerinden farklı Bağlantı Hızı kullanan abonelerin toplam abone sayısına oranları incelenecektir.



Şekil 3.1. Internet Hız İstatistikleri

Şekil 3.1'e göre bağlantı hızında ortalamanın yüksek olması, internet kullanıcılarının yüksek hızda internet kullandıklarını göstermektedir. Abonelerin çoğunluğunun 25 Mbit, 50 Mbit ve 100 Mbit hızlarını kullandığı görülmektedir.

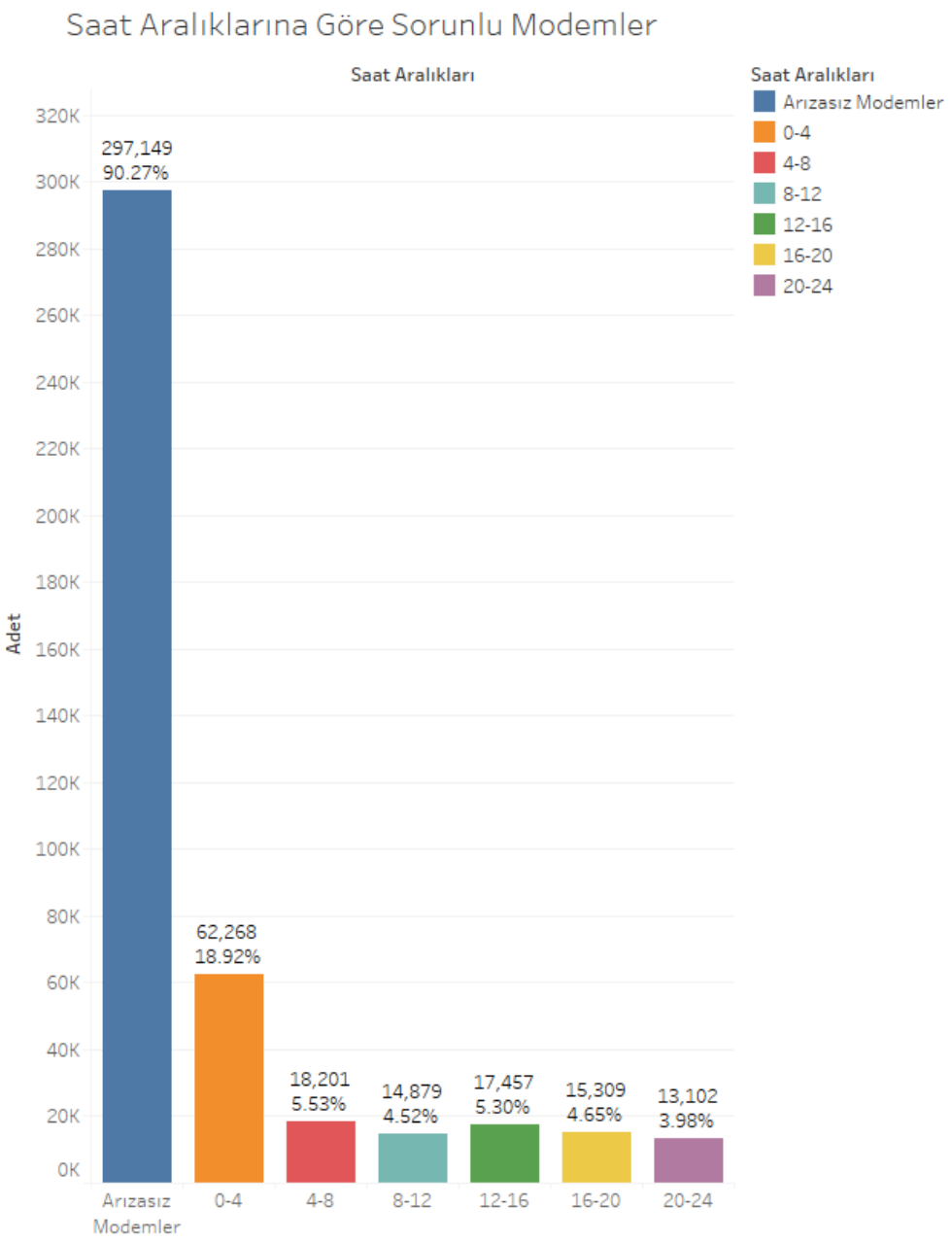
Arıza kaydı (Çağrı Açan) açan abonelerin sayıları arıza alışkanlıkları sayıları karşılaştırılmıştır:



Şekil 3.2. Çağrı Durumu ve Arıza Alışkanlığı Sayıları

Şekil 3.2'de dikeyde veri çağrı açmış/çağrı açmamış şeklinde gruplanmıştır. Yatay da ise arıza alışkanlıklarının sayıları verilmiştir. Arıza alışkanlıkları ve çağrı kaydı bırakma arasındaki ilişki incelendiğinde, örneğin, arıza alışkanlığı 3 tane olan grubun toplamda açtığı arıza kaydı 261 iken, açmayanların sayısı ise 30 bin'dir. Çağrı açmamış olanlarda sol tarafta bir yoğunluk gözlenmektedir. Çağrı açmayanları gösterebilmek için logaritmik gösterim tercih edilmiştir. Bu tercihin sebebi, çağrı açanların sayısının çok az olmasıdır. Müşteri memnuniyetsizliğini belirlemede, arıza kaydı bırakanlar yanında sorunlu modem değerleri de işaretleme için kullanılmıştır. Verideki dengesizlikten kurtulmak amacıyla; oversampling, undersampling ve SMOTE teknikleri kullanılmıştır.

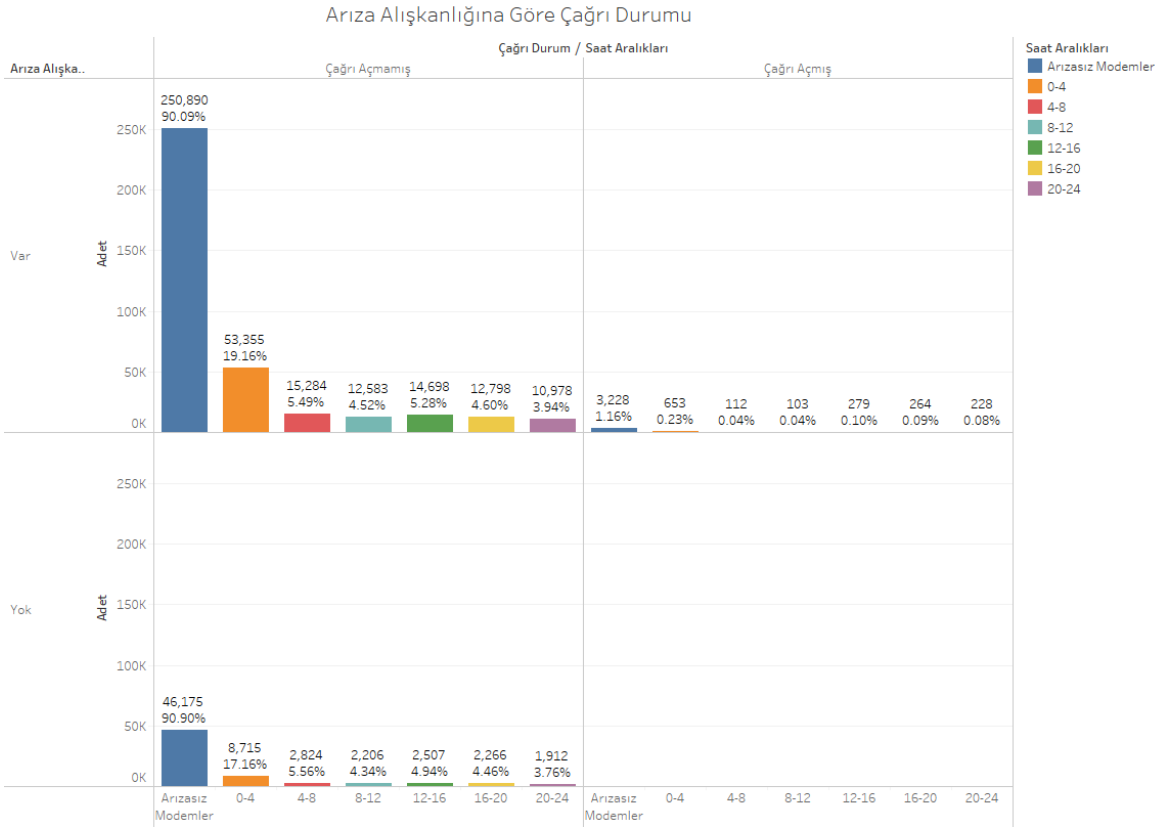
Saat aralıklarına göre modemlere ait sorunlu sinyal miktarları oranlanmıştır.



Şekil 3.3. Saat Aralıklarına Göre Sorun Yaşama Oranları

Şekil 3.3'te yer alan grafikte sorunlu sinyal üreten modemlerin değerleri incelendiğinde, modemlerin çoğunun sorunsuz çalıştığı görülmektedir. Genel olarak modemlerin gün içinde sorun çıkarma oranlarının birbirine yakın olduğu gözlemlenmektedir. Gece sorun çıkaran modemlerin sayısının gündüze göre daha çok olduğunu söylenebilir.

Arıza alışkanlıkları olan abonelerin çağrı açma durumları ve açtıkları saatler karşılaştırılacaktır. Maksadımız aralarında bir ilişki olup olmadığını gözlemlemektir.

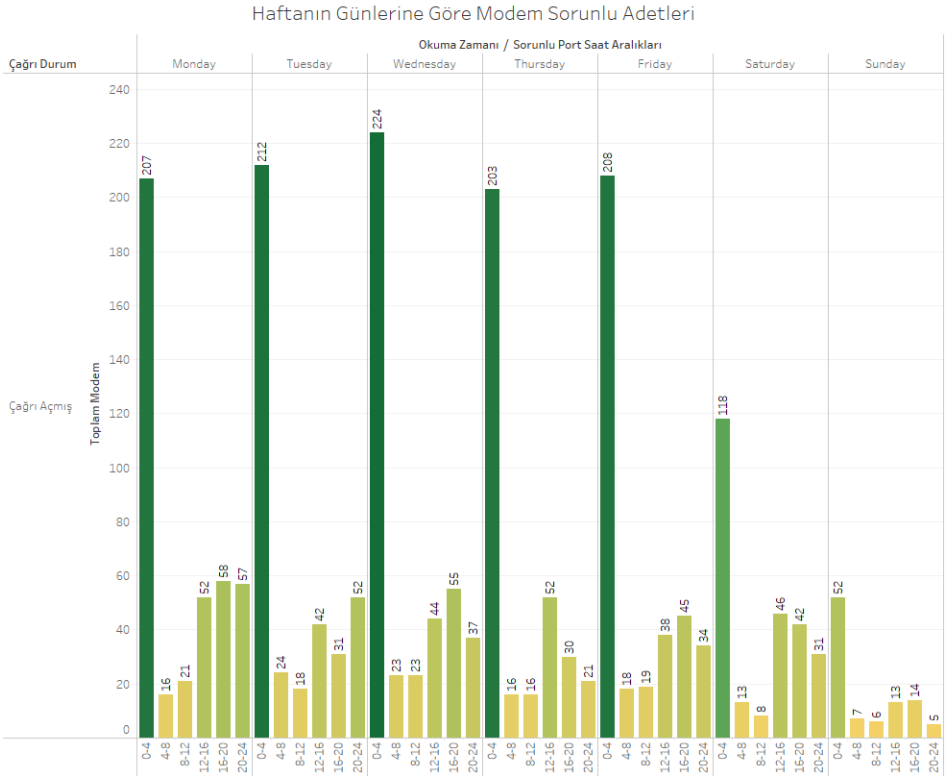


Şekil 3.4. Çağrı Durumu/Arıza Alışkanlığı ve Sorunlu Modemler

Şekil 3.4'te yer alan grafikte önceden arıza alışkanlığı olan aboneler en az bir çağrı açmış ise arıza alışkanlığı var olarak işaretlenmiştir. Burada arıza alışkanlığı olduğu halde çağrı açmamış abonelerin olduğu görülmektedir. Ayrıca bu grafikten çağrı açanların aynı zamanda eskiden arıza açma alışkanlıkları olanlar olduğunu söyleyebiliriz. Arıza kaydı bırakanlar aslında genel olarak problemlerini hemen dile getiren, şikayet açma konusunda kararlı kişiler olabilmektedir.

Genel olarak aboneler memnuniyetsizliklerini hemen bildirmeyip bir süre daha beklemeyi tercih edebilmektedirler. Diğer bir konu ise saha çalışmaları, altyapı çalışmaları ve sahadaki diğer cihazlardan kaynaklı problemlerin de var olmasıdır. Bu değerler bu veride bulunmamaktadır. Genel olarak arıza alışkanlığı ile arıza açma arasında bir ilişki veya memnuniyetsizlikle ilişkilendirme tespit edilememiştir.

Abonelerin arıza açtıkları modemlerin haftanın günlerine göre sorunlu adetleri incelenmektedir.



Şekil 3.5. Haftanın Günlerine Göre Sorunlu Modem Sayıları

Şekil 3.5 ile haftanın günlerine ve modemlerin saat aralıklarına göre arıza bırakma sayıları gösterilmiştir. Burada arıza bırakmayan modemler grafikten çıkarılmıştır. Buna göre Cumartesi ve Pazar günleri internet kullanımının azlığına bağlı olarak modemlerin daha az sorun bıraktıklarını görülmektedir. Ayrıca hafta içi Çarşamba gününe doğru sorunlu sinyal oranları yükseliş göstermektedir. Gün içindeki saat aralıklarına bakarak internet kullanımının artması ile modemlerin sorunlu aralıkta olması arasında bir oran olduğu da gözlemlenmektedir.

Verinin işaretlenmesi aşamasında değerlendirilen kriterler abonen arıza bırakması ve modemin de en az bir kere sorunlu değer bırakması dikkate alınmıştır. Memnuniyetsizliğini belli etmeyen müşteriler çoğunluktadır. Bu da verideki azınlık sınıfın oranını oldukça düşürmektedir. Memnuniyet durumuna göre arıza kaydı bırakan aboneler ile sorunlu modemlerin adetleri karşılaştırılmıştır.

Memnuniyet Durumu	Arıza Bırakan/ Bırakmayan Müşteri	Sorunlu Port Saat Aralıkları	Sorunlu Modem Ort.	Arıza Kaydı Bırakan	Toplam Gözlem	Gözlem Sayısı Yüzdelik
Memnun Müşteri	**Arıza Bırakmayan Müşteri**	Arızasız Modemler	0		1,899,142	83.516%
		0-4			16,670	0.733%
		4-8			750	0.033%
		8-12			3,675	0.162%
		12-16			3,986	0.175%
		16-20			2,811	0.124%
		20-24			2,840	0.125%
	Toplam				**1,929,874**	**84.868%**
Memnuniyetsiz Müşteri	**Arıza Bırakan Müşteri**	Arızasız Modemler	0	8,200	7,711	0.339%
		0-4	260	1,300	1,224	0.054%
		4-8	47	123	117	0.005%
		8-12	29	121	111	0.005%
		12-16	40	308	287	0.013%
		16-20	36	313	275	0.012%
		20-24	27	275	244	0.011%
		Ara Toplam	**439**	**10,640**	**9,969**	**0.438%**
	Arıza Bırakmayan Müşteri	0-4	273		253,535	11.149%
		4-8	34		21,264	0.935%
		8-12	20		15,057	0.662%
		12-16	22		18,151	0.798%
		16-20	17		14,421	0.634%
		20-24	8		11,703	0.515%
	Toplam		**814**	**10,640**	**344,100**	**15.132%**

Tablo 3.2. Sorunlu Modem Sayıları ve Arıza Bırakanlar

Tablo 3.2'de sorun bırakan/bırakmayan modemlere ait veriler ile arıza bırakan/bırakmayan abonelere ait veriler üzerinden müşteri memnuniyetine ait tablo verilmiştir. Genel olarak, 1 haftalık dönemde sorunlu sinyal değerlerine sahip modemlerin sayısı azdır. Ayrıca, arıza bildiren abone sayısı da oldukça azdır. Yalnızca arıza bildirenlerin oranı tüm kümede %1'in altındadır. Sorunlu modem sinyaline sahip olanlarla birlikte memnuniyetsiz müşteri oranı %15 civarındadır. Müşteri memnuniyetsizliğini etkileyen faktörler arasında bölgesel arızalar, altyapı çalışmalarından kaynaklı kesintiler, internet bant genişliğinden kaynaklı hız düşüşleri ve internet sinyalini aktaran cihazlardan kaynaklı aksaklıklar bulunmaktadır. Tablo 3.2'den

görüleceği üzere, modeminde sinyal sorunu olmayan ancak arıza kaydı bırakan aboneler bulunmaktadır.

3.3. İstatistiksel Analizler

3.3.1. Doğrusal Korelasyon Katsayıları

Büyük veride değişken sayısını az tutmak ve normalleştirme kullanmak verinin eğitim aşamasında hız kazandırmaktadır (Burkov, 2019). Bu nedenle, doğrusal korelasyon analizi gibi yöntemlerle değişken seçimi yapmak, hem performansı artırmak hem de modelin genel başarısını iyileştirmek açısından önemlidir. AutoML, eksik değerleri otomatik olarak düzelterek ve gerekli gördüğü yerde normalleştirmeye başvurarak ve bazı modeller için sadece sayısal değerler kullanarak model geliştirmektedir. Müşteri Memnuniyet Durumu üzerinden sayısal değerler arasındaki doğrusal korelasyon analizi için KNIME üzerinden aşağıdaki yapı kullanılmıştır.

Sütun Adı	Çağrı Açma Durumu
Download - 0 ile 4 Arası Sorunlu Port Sayısı	-0.02492
Upload - 0 ile 4 Arası Sorunlu Port Sayısı	-0.04725
Download - 4 ile 8 Arası Sorunlu Port Sayısı	-0.04507
Upload - 4 ile 8 Arası Sorunlu Port Sayısı	-0.08818
Download - 8 ile 12 Arası Sorunlu Port Sayısı	-0.03864
Upload - 8 ile 12 Arası Sorunlu Port Sayısı	**-0.00086**
Download - 12 ile 16 Arası Sorunlu Port Sayısı	-0.01952
Upload - 12 ile 16 Arası Sorunlu Port Sayısı	**0.00702**
Download - 16 ile 20 Arası Sorunlu Port Sayısı	-0.01567
Upload - 16 ile 20 Arası Sorunlu Port Sayısı	**-0.00585**
Download - 20 ile 24 Arası Sorunlu Port Sayısı	-0.00801
Upload - 20 ile 24 Arası Sorunlu Port Sayısı	**0.00336**
Homepass	**0.00338**
Arıza Adeti	0.22159
Bağlantı Hızı	-0.06580
Yükleme Hızı	-0.04486
Arıza Alışkanlığı	**0.00047**
En Son Aya Ait Download Miktarı	-0.07802
En Son Aya Ait Upload Miktarı	-0.05017

Tablo 3.3. Korelasyon Katsayısı

Tablo 3.3'te yer alan korelasyon katsayıları incelendiğinde homepass, arıza alışkanlığı ve özellikle yüklemeye ait saat 08:00 ile 24:00 arasındaki değerler memnuniyetsizliği anlamada etkisiz oldukları görülmektedir. Bunlar tabloda koyu renkle verilmiştir.

3.3.2. Özelliklerin Değerlendirilmesi

AutoML üzerinden aykırı değer ve eksik değer analizleri yapılsa da veriyi daha doğru anlama adına bazı işlemlerden geçirildi. Buna göre eksik değer olanları veriden çıkarılmadı. Veri AutoML ile modellendiğinde verilere normalleştirme uygulanmaktadır. Bazı durumlarda abonenin verisinin okunmaması arıza yaşadığını veya modemin kapalı olduğu anlamına gelebilir. Homepass, arıza alışkanlığı ve yüklemeye ait 08:00-24:00 arası değerler düşük korelasyon nedeniyle veriden çıkartılmıştır. İlk denemelerde sadece sayısal özellikler ile model geliştirilmiştir.

3.4. AutoML Detayları

Verinin en doğru eğitim kümesi ile modellemek adına kullanılan bir yöntemdir. Parametre optimizasyonun çapraz doğrulama ile yapıldığı yerdir. Veri dengesizliğinden kaynaklı olarak doğru metrikleri içermeyen küme ile modele gidilmesi istenen bir durum değildir. Küme 5 eşit rastgele alt kümelere bölümlenerek her bir kümeye katman adı verilir. Bu biçimde K-katmanlı çapraz doğrulama yöntemi ile her bir parça ile doğrulama verisi kullanılarak modelin genel performansının daha doğru değerlendirilmesi sağlanmış olur. Genel görüşe göre 5-katmanlı çapraz doğrulama tercih edilmektedir (Burkov, 2019). Veri 5 kere döngüye girerek ilgilenilen metrik değerine göre değerlendirilir. En doğru veri seti bulunmuş olur. Aynı zamanda kıyaslama (benchmarking) ile modelin çalışma zamanı ölçülerek eğitim süreleri kayıt altına alınmıştır. Bir sonraki aşamada eğitilen modeller ile öğrenme gerçekleşir.

Modeller tahmin edildikten sonra istatiksel sonuçlar üretilir. Model sonuçları toplandıktan sonra başlangıçta seçilen değerlendirme metriğine göre en iyi modelin oylaması yapılır ve en iyi model kaydedilir.

En iyi model seçildikten sonra gerçek veri üzerinden skorlama yapılmaktadır. Burada AutoML kısmında tüm algoritma, kütüphane ve çatı yapıları belirlenen kritere göre

sıralandıktan sonra en yüksek ölçüt değerine sahip model otomatik olarak test verisi için kullanıma hazır hale getirilmektedir. Sonrasında bu modelin ölçütleri test verisi üzerinden değerlendirmek üzere skorlanmaktadır.

3.5. Dengesiz Veri ile Yapılan Tahminler

Veride tahmin edilmek istenen sınıf, veri setinde düşük seviyede temsil edildiği için dengesiz veri kapsamına girmektedir. Yaklaşık olarak temsil oranı %15 civarındadır. Verinin eğitilebilmesi için oversampling, undersampling ve SMOTE yapılabilir. KNIME üzerinde yer alan "Bootstrap" düğümü oversampling için kullanılacaktır. Bu yaklaşımın tam tersi olan undersampling ve rastgele yöntemi kullanarak baskın sınıftan bazı örnekler çıkarılabilmektedir. KNIME üzerinden undersampling için "equal size sampling" düğümü kullanılacaktır. Ayrıca bu iki yöntemin dışında sentetik veri üretme teknikleri de bulunmaktadır. Bunlardan SMOTE ve ADASYN algoritmaları ile zayıf sınıfa ait örnekler çoğaltılabilmektedir (Burkov, 2019).

Veri kümelerinin seçiminde optimum olarak kabul edilebilecek bir kural bulunmamaktadır. Genel olarak veriler üç kategoriye ayrılmaktadır: eğitim, doğrulama ve test verisi. Geçmiş zamanlarda kabul gören kanaat verinin %70'ini eğitim setine, geri kalan %15 doğrulama ve %15 ise teste ayrılmaktaydı. Ancak günümüzde büyük verilerle çalışıldığı için eğitim kümesine verinin %95'ini, geri kalan %2.5 test ve %2.5 ise doğrulamaya bırakılabilir (Burkov, 2019). Buna göre ilk uygulamalar %80 ile yapılmış ve sonrasında eğitim kümesi için %95 ayarlanarak da uygulama yapılmıştır.

Genel olarak algoritmaların dengesiz veri eğitme konusunda hassasiyetleri düşüktür. Karar Ağaçları, Rastgele Orman ve Gradyan Artırma algoritmaları dengesiz veride iyi sonuç verebilmektedir (Burkov, 2019). SMOTE ve undersampling kombinasyonun genel olarak modellerde daha iyi sonuçlar verdiği söylenmektedir (Chawla, 2002).

Büyük veride değişken sayısını az tutmak ve normalleştirme kullanmak verinin eğitim aşamasında hız kazandırmaktadır. Burada min-max normalleştirmesi tercih edilmektedir. Z-score normalleştirme, standart normal dağılıma benzeyen dağılımlar için tercih edilmektedir. Verilerin yaklaşık olarak oldukça küçük bir aralıkta olması; bilgisayarın çok büyük veya çok küçük değerleri hesaplamasında yaşanan aritmetik taşma problemini

de engellenmiş olmaktadır (Burkov, 2019). Büyük verinin eğitilmesinde değişken sayısının az olması hem işlem hızı hem de daha etkin çalışma adına önemlidir. Bu nedenle ilk denemeler, sayısal özelliklerin korelasyon analizi sonuçları da dikkate alınarak: indirme-8-24 saat aralıkları, yükleme-0-8 saat aralıkları, bağlantı hızı, yükleme hızı, arıza adet, indirme toplam, yükleme toplam özellikleri ile toplamda 13 değişken ile model eğitim aşamasına geçilmiştir. K-Katmanlı (k-Fold) olarak 5 ve 10 seçilmiş, veri seti %80 Doğrulama-Eğitim, %20 Test şeklinde ayrıştırılmıştır. Toplamda 2.3 milyon gözlem verisi bulunmaktadır. Eğitim tarafındaki veriyi kaynak ve süre yönünden kısıtlamak adına rastgele örnekleme ile 100 bin kayıt seçilerek devam edilmiştir.

3.5.1. Veri Çoklama Teknikleri Kullanılmadan Yapılan Çalışma

AutoML Parametreleri olarak k-Katmanlı k=10, verinin %80'ni Eğitim ve %20'si Test olarak seçilmiştir. Sonuçlar ölçüt bazında ısı haritası ile renklendirilmiş ve sütun içerisinde max-min değerlerine göre en yüksek yeşil en düşük kırmızı olacak şekilde belirlenmiştir.

Min Max

Algoritma	Doğruluk	F-Ölçütü	GBA	BM	WPN
XGBoost Ağaçları	0,9994	0,9981	0,9997	0,9993	0,9999
Karar Ağacı	0,9987	0,9957	0,9986	0,9971	0,9984
Gradyan Artırma Ağaçları	0,9973	0,9912	0,9984	0,9968	0,9995
H2O AutoML	0,9982	0,9942	0,998	0,9961	0,9978
Rastgele Orman	0,9891	0,9637	0,9738	0,9481	0,9587
Derin Öğrenme	0,9777	0,9285	0,968	0,9362	0,9584
Sinir Ağları	0,9821	0,9413	0,9666	0,9338	0,9514
GLM	0,9272	0,7119	0,7647	0,5795	0,6306
Lojistik Regresyon	0,9219	0,6715	0,7225	0,5185	0,5662
Sade Bayes	0,9063	0,5691	0,6373	0,4034	0,4481

Tablo 3.4. Model Doğrulama Sonuçları

Model dengesiz veri içerdiği için doğruluk değerlerinin %98'in üzerinde olduğu görülmektedir. Ayrıca burada MCC ölçütü ile de sonuçlar değerlendirilmiş ve genel olarak tüm olasılıkların 1'e çok yakın olduğu görülmüştür. Bundan dolayı bu ölçüt ile de sonuca varmakta zorluk çekilmiştir.

Zhu tarafından yazılan makalede tavsiye edilen GBA, BM ve Jadhav tarafından önerilen WPN ölçütleri ile değerlendirilmesi yapılacaktır. CK, MCC ve F1 ölçütleri, dengesiz verinin değerlendirilmesinde yetersiz kaldıkları için değerlendirmeye alınmayacaktır. Modellerin değerlendirilmesinde ROC eğrisi yerine PR eğrisi tercih edilmiştir. Bunun sebebi, Saito tarafından yazılan makaleye göre, özellikle azınlık sınıfının temsil oranı %15 olan dengesiz veri kümelerinde, PR eğrisinin ikili sınıflayıcıların değerlendirilmesinde ROC eğrisinden daha bilgilendirici olmasıdır.

Modellerin Duyarlılık değerlerine bakıldığında genel olarak hepsinin FP oranını oldukça düşük seviyede tuttukları söylenebilir. AutoML düğümü, verilere normalleştirme yapmakta ve eksik değerleri kendisi doldurmaktadır.

Algoritma	Kullanılan Parametreler
Gradyan Artırma Ağaçları	nrModels = 90
Sade Bayes	threshold = 0.01
Lojistik Regresyon	stepSize = 0.1
Karar Ağacı	minNumberRecordsperNode = 8
Rastgele Orman	maxLevels = 8, minNodesize = 20, nrModels = 100
Sinir Ağları	hiddenlayer = 2, nrhiddenneurons = 75
GLM	CFG_FAMILY = Binomial, CFG_ALPHA = 0.0, CFG_LAMBDA = 1.0
XGBoost Ağaçları	eta = 0.2, Max_depth = 5
Derin Öğrenme	m_units = 64, batch_size = 512, epochs = 100
H2O AutoML	Bilinmiyor ve H2O AutoML Öğrenme ile yönetiliyor.

Tablo 3.5. Algoritmalara ait Parametreler

3.5.2. Oversampling Yöntemi Kullanılarak

AutoML Parametreler olarak k-Katmanlı 10 seçilmiş, verinin %80'ni Eğitim %20'si Test olarak seçilmiştir. Dengesizliği gidermek adına doğrulama için kullanılan örneklem verisine oversampling tekniği uygulanmıştır. Oversampling tekniği, rastgele örnekleme ile elde edilen veriye uygulanmıştır.

Oversampling ile model eğitilmiş ve doğrulama sonuçları aşağıdaki gibi elde edilmiştir:

Min Max

Algoritma	Doğruluk	F-Ölçütü	GBA	BM	WPN
XGBoost Ağaçları	0,9999	0,9999	0,9999	0,9998	0,9999
Karar Ağacı	0,9994	0,9994	0,9994	0,9987	0,9994
H2O AutoML	0,9993	0,9993	0,9993	0,9985	0,9993
Gradyan Artırma Ağaçları	0,9979	0,9979	0,9979	0,9958	0,9979
Rastgele Orman	0,9966	0,9966	0,9966	0,9933	0,9966
Derin Öğrenme	0,9897	0,9897	0,9897	0,9794	0,9897
Sinir Ağları	0,9712	0,9706	0,9710	0,9424	0,9708
GLM	0,9468	0,9449	0,9461	0,8936	0,9455
Lojistik Regresyon	0,8256	0,8456	0,8155	0,6513	0,8054
Sade Bayes	0,7091	0,7732	0,6504	0,4182	0,5966

Tablo 3.6. Oversampling Doğrulama Sonuçları

Oversampling ile yapılan tahminlerde bir önceki tahminlere göre FN ve FP değerlerinde 0 yer almamıştır. Ayrıca bu modelde H2O AutoML çatısı Gradyan Artırma algoritmasına göre daha iyi sonuçlar vermiştir. Oversampling sonrası Sade Bayes haricindeki tüm modeller birbirine yakın sonuçlar vermiştir. GLM'de iyileşme olduğu söylenebilir.

Algoritma	Kullanılan Parametreler
Gradyan Artırma Ağaçları	nrModels = 90
Sade Bayes	threshold = 0.0044
Lojistik Regresyon	stepSize = 0.05
Karar Ağacı	minNumberRecordsperNode = 8
Rastgele Orman	maxLevels = 8, minNodesize = 10, nrModels = 100
Sinir Ağları	hiddenlayer = 1, nrhiddenneurons = 75
GLM	CFG_FAMILY = Binomial, CFG_ALPHA = 0.2, CFG_LAMBDA = 0.8
XGBoost Ağaçları	eta = 0.2, Max_depth = 10
Derin Öğrenme	m_units = 64, batch_size = 512, epochs = 100
H2O AutoML	Bilinmiyor ve H2O AutoML Öğrenme ile yönetiliyor.

Tablo 3.7. Oversampling Algoritmalarına ait Parametreler

Oversampling ile yapılan modellere ait test sonuçları aşağıdaki gibidir:

Min Max

Algoritma	Doğruluk	F-Ölçütü	GBA	BM	WPN
XGBoost Ağaçları	0,9995	0,9982	0,9995	0,999	0,9995
H2O AutoML	0,9989	0,9962	0,9991	0,9983	0,9994
Karar Ağacı	0,9989	0,9963	0,9988	0,9977	0,9988
Gradyan Artırma Ağaçları	0,9959	0,9867	0,9973	0,9946	0,9987
Rastgele Orman	0,9952	0,9842	0,9962	0,9925	0,9973
Derin Öğrenme	0,9847	0,9515	0,9878	0,9757	0,991
Sinir Ağları	0,9573	0,8754	0,971	0,9423	0,9848
GLM	0,9233	0,795	0,9467	0,8947	0,9707
Lojistik Regresyon	0,917	0,7184	0,8174	0,655	0,7287
Sade Bayes	0,9068	0,5834	0,6536	0,4226	0,4712

Tablo 3.8. Oversampling Test Sonuçları

Test ile Doğrulama sonuçları arasında farklar bulunmamaktadır. H2O AutoML çatısı Karar Ağacına göre daha iyi sonuç vermiştir. Algoritmalar en düşük FP'ye göre sıralıdır:

FP-FN Max Min
TP-TN Min Max

	Doğrulama				Test			
Algoritmalar	TP	FP	TN	FN	TP	FP	TN	FN
XGBoost Ağaçları	17368	0	17372	4	68832	217	385716	30
H2O AutoML	17349	3	17369	23	68832	493	385440	30
Karar Ağaçları	17354	4	17368	18	68777	420	385513	85
Gradyan Artırma Ağaçları	17304	5	17367	68	68813	1810	384123	49
Rastgele Orman	17284	29	17343	88	68711	2054	383879	151
Derin Öğrenme	17100	85	17287	272	68340	6452	379481	522
Sinir Ağları	16515	143	17229	857	68249	18820	367113	613
GLM	15846	323	17049	1526	67607	33613	352320	1255
Lojistik Regresyon	16589	5275	12097	783	48137	17013	368920	20725
Sade Bayes	17226	9961	7411	146	29666	3181	382752	39196

Tablo 3.9. Oversampling Karışıklık Matrisi

XGBoost algoritması Doğrulama aşamasında 0 FP vermiştir. Burada algoritmanın hiç hata vermemesi aşırı öğrenme eğilimini göstermektedir.

3.5.3. Undersampling Yöntemi Kullanılarak

AutoML Parametreler olarak k-Katmanlı 10 seçilmiş, verinin %80'ni Eğitim %20'si Test olarak ayrılmıştır. Verideki dengesizliği gidermek adına Doğrulama için kullanılan örneklem verisi Eşit Sayılı Örneklem düğümü ile undersampling yapılmıştır. Undersampling tekniği rastgele örnekleme ile elde edilen veriye uygulanmıştır.

Min Max

Algoritma	Doğruluk	F-Ölçütü	GBA	BM	WPN
XGBoost Ağaçları	0,9989	0,9989	0,9989	0,9977	0,9989
H2O AutoML	0,9984	0,9984	0,9984	0,9968	0,9984
Karar Ağacı	0,9977	0,9977	0,9977	0,9955	0,9977
Gradyan Artırma Ağaçları	0,9963	0,9963	0,9963	0,9926	0,9963
Rastgele Orman	0,9892	0,9893	0,9892	0,9784	0,9892
Sinir Ağları	0,964	0,9628	0,9635	0,9279	0,963
Derin Öğrenme	0,9472	0,9482	0,9471	0,8945	0,9469
GLM	0,9411	0,9397	0,9408	0,8822	0,9406
Lojistik Regresyon	0,8298	0,8492	0,8197	0,6596	0,8097
Sade Bayes	0,7111	0,7748	0,6524	0,4221	0,5986

Tablo 3.10. Undersampling Doğrulama Sonuçları

Undersampling ile Oversampling tekniklerine ait sonuçların benzerlik taşıdığı söylenebilir. Burada H2O AutoML çatısı seçimini, "GBM" yerine "StackedEnsemble" olarak yapmıştır. Ayrıca Sinir Ağları, Derin Öğrenmenin önüne geçmiştir. Genel olarak GLM, Lojistik Regresyon ve Sade Bayes ile alakalı bir değişim gözükmemektedir.

Algoritmalar	Kullanılan Parametreler
Gradyan Artırma Ağaçları	nrModels = 90
Sade Bayes	threshold = 0.0044
Lojistik Regresyon	stepSize = 0.1
Karar Ağacı	minNumberRecordsperNode = 8
Rastgele Orman	maxLevels = 8, minNodesize = 20, nrModels = 100
Sinir Ağları	hiddenlayer = 1, nrhiddenneurons = 75
GLM	CFG_FAMILY = Binomial, CFG_ALPHA = 0.2, CFG_LAMBDA = 0.8
XGBoost Ağaçları	eta = 0.2, Max_depth = 5
Derin Öğrenme	m_units = 64, batch_size = 512, epochs = 100
H2O AutoML	Bilinmiyor ve H2O AutoML Öğrenme ile yönetiliyor.

Tablo 3.11. Undersampling Algoritmalarına ait Parametreler

Min Max

Algoritma	Doğruluk	F-Ölçütü	GBA	BM	WPN
XGBoost Ağaçları	0,9983	0,9944	0,9989	0,9978	0,9995
H2O AutoML	0,9979	0,993	0,9984	0,9967	0,9988
Karar Ağacı	0,9973	0,991	0,9973	0,9946	0,9973
Gradyan Artırma Ağaçları	0,994	0,9804	0,9962	0,9925	0,9985
Rastgele Orman	0,9924	0,9752	0,9903	0,9805	0,9881
Sinir Ağları	0,9398	0,8336	0,9622	0,9255	0,9851
Derin Öğrenme	0,9587	0,8724	0,9481	0,8964	0,9376
GLM	0,9211	0,7876	0,9395	0,8797	0,9582
Lojistik Regresyon	0,9183	0,722	0,8189	0,6577	0,7302
Sade Bayes	0,9071	0,5825	0,6518	0,4206	0,4684

Tablo 3.12. Undersampling Test Sonuçları

Doğrulama ile test sonuçları arasında yakın bir ilişki olduğu görülmektedir.

	FP-FN	Max		Min
	TP-TN	Min		Max

	Doğrulama				Test			
Algoritma	**TP**	**FP**	**TN**	**FN**	**TP**	**FP**	**TN**	**FN**
Gradyan Artırma Ağaçları	3085	0	3108	23	68827	2714	383219	35
XGBoost Ağaçları	3102	1	3107	6	68848	764	385169	14
H2O AutoML	3103	5	3103	5	68797	905	385028	65
Karar Ağacı	3101	7	3101	7	68677	1060	384873	185
Sinir Ağları	2899	15	3093	209	68565	27086	358847	297
Rastgele Orman	3084	43	3065	24	67980	2569	383364	882
GLM	2854	112	2996	254	66575	33613	352320	2287
Derin Öğrenme	2999	219	2889	109	64263	14203	371730	4599
Lojistik Regresyon	2980	930	2178	128	48242	16524	369409	20620
Sade Bayes	3089	1777	1331	19	29476	2863	383070	39386

Tablo 3.13. Undersampling Karışıklık Matrisi

Bu örnekleme ile yapılan işlemde Gradyan Artırma algoritması aşırı öğrenme eğilimindedir. Diğer taraftan XGBoost yine iyi algoritmalardan biri olarak öne çıkmaktadır. Lojistik regresyon ve Sade Bayes algoritmaları her iki örneklemede en düşük olan algoritmalar olmuştur. XGBoost, H2O AutoML Çatısı ve Karar Ağacı algoritmaları her iki yöntemde de ilk 3 sıraya girmişlerdir.

3.5.4. SMOTE - Azınlık Sınıfı Örnekleme Yöntemi Kullanılarak

AutoML Parametreleri olarak k-Katmanlı 5 ile 10 seçilmiş, verinin %80'ni Eğitim %20'si Test olarak ayrılmıştır. SMOTE tekniği, rastgele örnekleme ile elde edilen veriye uygulanmıştır. KNIME üzerinden SMOTE' a ait azınlık sınıf yöntemi kullanarak sentetik veri üretilmiş ve k-5 ile k-10 arasındaki farklar doğrulama verileri üzerinden gösterilmiştir. GBA – BM – WPN arasındaki değerleri anlayabilmek için tüm tabloya ısı haritası uygulanmıştır.

Min Max

Doğrulama	GBA			BM			WPN		
Algoritma	k-5	k-10	Fark	k-5	k-10	Fark	k-5	k-10	Fark
XGBoost Ağaçları	0.997	0.9962	0.0008	0.994	0.9923	0.0017	0.997	0.9962	0.0008
Karar Ağacı	0.9931	0.9922	0.0009	0.9862	0.9844	0.0018	0.9931	0.9922	0.0009
H2O AutoML	0.9909	0.9908	0.0001	0.9818	0.9817	0.0001	0.9909	0.9908	0.0001
Gradyan Artırma Ağaçları	0.9879	0.9884	-0.0005	0.9758	0.9767	-0.0009	0.9879	0.9884	-0.0005
Rastgele Orman	0.9824	0.9837	-0.0013	0.9648	0.9674	-0.0026	0.9824	0.9837	-0.0013
Derin Öğrenme	0.979	0.9777	0.0013	0.9581	0.9555	0.0026	0.979	0.9777	0.0013
Sinir Ağları	0.9627	0.9779	-0.0152	0.9265	0.9558	-0.0293	0.9621	0.9778	-0.0157
GLM	0.9504	0.9458	0.0046	0.9021	0.893	0.0091	0.9497	0.9451	0.0046
Lojistik Regresyon	0.8079	0.8128	-0.0049	0.6378	0.6459	-0.0081	0.797	0.8028	-0.0058
Sade Bayes	0.6227	0.6284	-0.0057	0.3804	0.3879	-0.0075	0.5619	0.5691	-0.0072

Tablo 3.14. SMOTE - MC - Doğrulama Ölçütler k-5 ve k-10

Doğrulama sonuçları izlendiğinde k-5 ile k-10 arasında farklara göre k-5 daha iyi sonuç vermiştir. Burada da XGBoost en iyi algoritma olarak sonuçlanmıştır.

Min Max

Test	GBA			BM			WPN		
Algoritma	**k-5**	**k-10**	**Fark**	**k-5**	**k-10**	**Fark**	**k-5**	**k-10**	**Fark**
XGBoost Ağaçları	0.9975	0.9973	0.0002	0.9949	0.995	0.0003	0.998	0.998	0.0001
Karar Ağacı	0.9947	0.9946	0.0001	0.9893	0.989	0.0001	0.996	0.996	0.0003
H2O AutoML	0.9923	0.9899	0.0024	0.9846	0.98	0.0047	0.996	0.99	0.0058
Gradyan Artırma Ağaçları	0.9921	0.9923	-0.0002	0.9843	0.985	-0.0004	0.997	0.997	0.0001
Derin Öğrenme	0.9795	0.9792	0.0003	0.9591	0.959	0.0006	0.986	0.986	0.0000
Rastgele Orman	0.9792	0.9795	-0.0003	0.9584	0.959	-0.0007	0.976	0.976	-0.0007
Sinir Ağları	0.962	0.9683	-0.0063	0.9251	0.937	-0.0116	0.985	0.965	0.0194
GLM	0.9465	0.9464	0.0001	0.8942	0.894	0.0002	0.97	0.97	0.0002
Lojistik Regresyon	0.8229	0.8233	-0.0004	0.6641	0.665	-0.0008	0.738	0.739	-0.0008
Sade Bayes	0.6537	0.6506	0.0031	0.4213	0.417	0.0040	0.472	0.468	0.0042

Tablo 3.15. SMOTE - MC - Test Ölçütler k-5 ve k-10

Test ile Doğrulama arasında farklar bulunmamaktadır. Test sonuçlarında da XGBoost en iyi sırayı alırken, H2O AutoML Çatısı, Karar Ağacı ve Gradyan Artırma ilk sıraları paylaşmaktadır.

FP-FN Max Min
TP-TN Min Max

Test	TP			FP			TN			FN		
Algoritma	**k-5**	**k-10**	**Fark**	**k-5**	**k-10**	**Fark**	**k-5**	**k-10**	**Fark**	**k-5**	**k-10**	**Fark**
XGBoost Ağaçları	68779	68775	4	1492	1598	-106	384441	384335	106	83	87	-4
Karar Ağacı	68668	68643	25	3036	2953	83	382897	382980	-83	194	219	-25
Sade Bayes	29739	29452	287	4096	3994	102	381837	381939	-102	39123	39410	-287
H2O AutoML	68691	68186	505	4972	3970	1002	380961	381963	-1002	171	676	-505
Gradyan Artırma Ağaçları	68829	68817	12	5876	5635	241	380057	380298	-241	33	45	-12
Rastgele Orman	67087	67148	-61	6099	6195	-96	379834	379738	96	1775	1714	61
Derin Öğrenme	68134	68146	-12	11688	11985	-297	374245	373948	297	728	716	12
Lojistik Regresyon	48815	48878	-63	17285	17344	-59	368648	368589	59	20047	19984	63
Sinir Ağları	68538	66391	2147	27086	10595	16491	358847	375338	-16491	324	2471	-2147
GLM	67572	67558	14	33613	33613	0	352320	352320	0	1290	1304	-14

Tablo 3.16. SMOTE - MC - Karışıklık Matrisi

Karışıklık matrisi incelendiğinde test sonuçlarında k-5 sonuçları k-10'a göre daha iyi sonuçlar vermiştir. Burada tüm kategorilerde değerlendirme yapılırsa XGBoost algoritmasının en iyi sonuca sahip olduğunu söylenebilir.

Algoritmalar	Kullanılan Parametreler
XGBoost Ağaçları	eta = 0.2, Max_depth = 10
Karar Ağacı	minNumberRecordsperNode = 8
H2O AutoML	Bilinmiyor ve H2O AutoML Öğrenme ile yönetiliyor.
Gradyan Artırma Ağaçları	nrModels = 90
Rastgele Orman	maxLevels = 8, minNodesize = 15, nrModels = 100
Derin Öğrenme	m_units = 64, batch_size = 512, epochs = 100
Sinir Ağları	hiddenlayer = 1, nrhiddenneurons = 45
GLM	CFG_FAMILY = Binomial, CFG_ALPHA = 0.2, CFG_LAMBDA = 0.8
Lojistik Regresyon	stepSize = 0.05
Sade Bayes	threshold = 0.0044

Tablo 3.17. SMOTE - MC Örnekleme Algoritmalarına ait Parametreler (k-5)

3.5.5. Temel Bileşenler Analizi ve SMOTE

SMOTE ile sentetik veri üreterek dengesiz veriye uygulandığında modelin performansını artırıldığı bilinmektedir. Bununla birlikte TBA ile beraber uygulandığında daha da iyi performans verdiği görülmüştür (Mulla, 2021).

Buraya kadar uygulanan yöntemler sonucunda modelde belirgin bir iyileşme sağlanmamıştır. Bu nedenle, modelde yer alan verilere boyut indirgeme yöntemi olarak TBA uygulanacaktır. Bu kapsamda, internet kullanımına ait 0-24 saat arasındaki 6 sütunluk saatlik veriler, 0-8 saat arasındaki 2 sütunluk yükleme verileri, bağlantı hızı, yükleme hızı, yükleme miktarı ve indirme miktarı gibi veriler dikkate alınacaktır. KNIME AutoML düğümü TBA düğümünü içermemektedir. Bu nedenle, AutoML düğümünden önce verilerin normalleştirilmesi ve ardından TBA'nın uygulanması gerekecektir. Normalleştirme, verilerin çarpıklığı nedeniyle yapılacaktır. Ayrıca, aykırı olarak değerlendirilen bazı değerlerin azınlık sınıflara ait değerli bilgiler içerebileceği düşünülerek, bu değerler veriden çıkarılmamıştır.

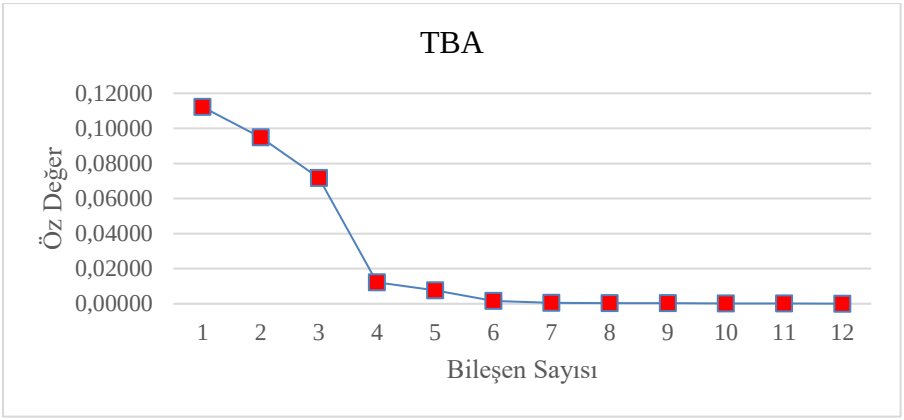
Kategorik değerlere, Kategorik Değişkenlere Dönüştürme tekniğini uygulayarak değişken sayısını arttıracağı için değişiklik yapılmamıştır. Bunun nedeni değişken sayısının artırmasıdır. Bu da eğitim süresini uzatmakta ve makineye ek yük getirmektedir. TBA'da 3, 4, 5 TB ile çalışma yapılmış ve çıkan sonuçlar değerlendirilmiştir. Normalleştirme yapılan ve TBA uygulanan veriler %95'i doğrulama-öğrenme %5'i test için ayrılmıştır. Sonrasında verinin çok büyük olmasından dolayı doğrulama tarafında

rastgele örnekleme ile 102400 kayıt alınmıştır. Bundan sonra bir önceki aşamalarda iyileştirme gösteren SMOTE – MC sentetik veri üretme yöntemi ile veri zenginleştirilmiştir. AutoML tarafında k-Katmanlı 5 seçilmiş ve eğitim seti için verinin %95'i ayrılmıştır. Bundan sonraki aşamada bir önceki denemelerde başarılı olan XGBoost Ağaçları, Gradyan Artırma Ağaçları, Karar Ağacı ve H2O AutoML çatısı ile model geliştirilecektir. TBA'ya eklenen 12 özelliğe ait sütunlar ve ne kadar bileşen kullanılması gerektiğinin analiz edildiği değerler aşağıdaki tabloda verilmiştir.

Bileşen Sayısı	Öz Vektör	Öz Değer	Oran	Varyans Açıklama Oranı
1	0	0.11216	0.3712684	37.1268388
2	1	0.09492	0.3142179	68.54862431
3	2	0.07173	0.2374403	92.29265827
4	3	0.01224	0.0405033	96.34298427
5	4	0.00762	0.0252291	98.86589792
6	5	0.00164	0.0054318	99.40907612
7	6	0.00062	0.0020359	99.61266743
8	7	0.00043	0.0014309	99.75575475
9	8	0.00029	0.000969	99.8526587
10	9	0.00022	0.0007382	99.92647679
11	10	0.00018	0.0005827	99.98474407
12	11	0.00005	0.0001526	100
	Toplam Varyans	0.30210		

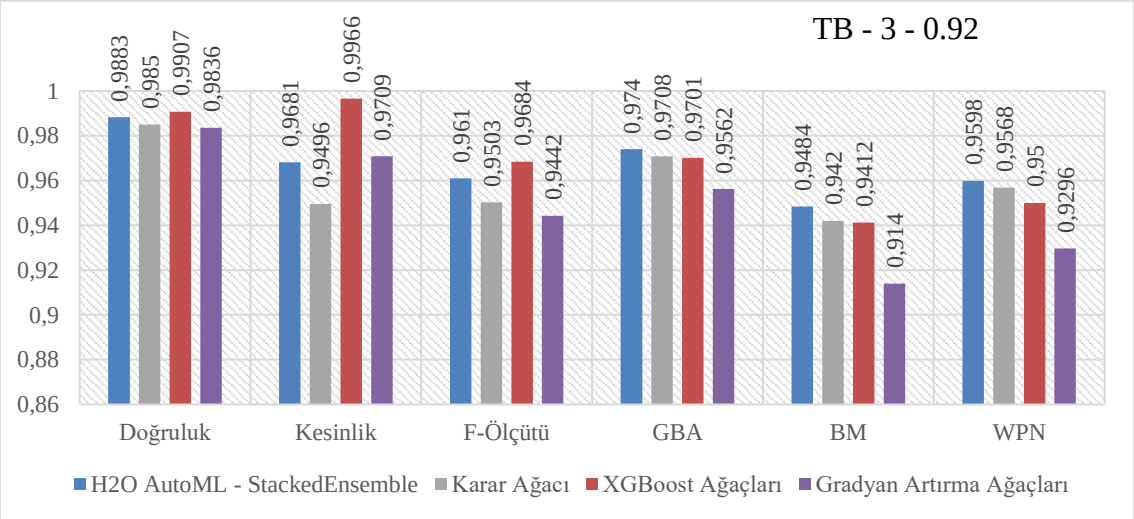
Tablo 3.18. TBA Tablosu

Tablo 3.18 incelendiğinde ilk 3 TB varyansın %92'sini açıkladığı görülmektedir. Bu da kabul edilebilir bir orandır. TB sayısını bulmak için kullanılan bir diğer yöntem, yamaç eğim grafiği ile önemli değişimin yaşandığı noktayı tespit etmektir. Burada öz değerler ile bileşen sayısı arasındaki değişimi gösteren grafik incelenecektir:

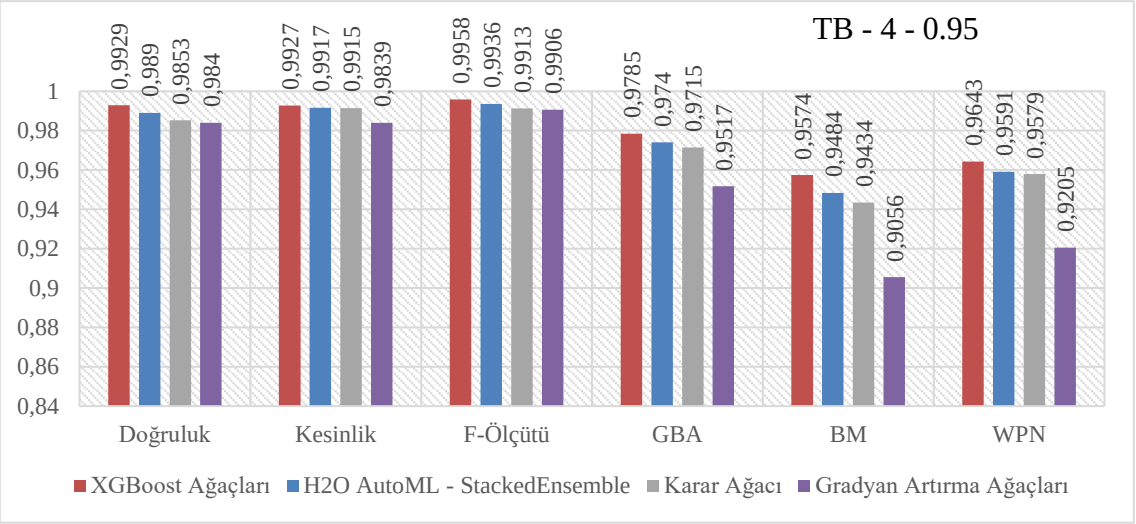


Şekil 3.6. Yamaç Eğim Grafiği ile Öz Değer Değişimi

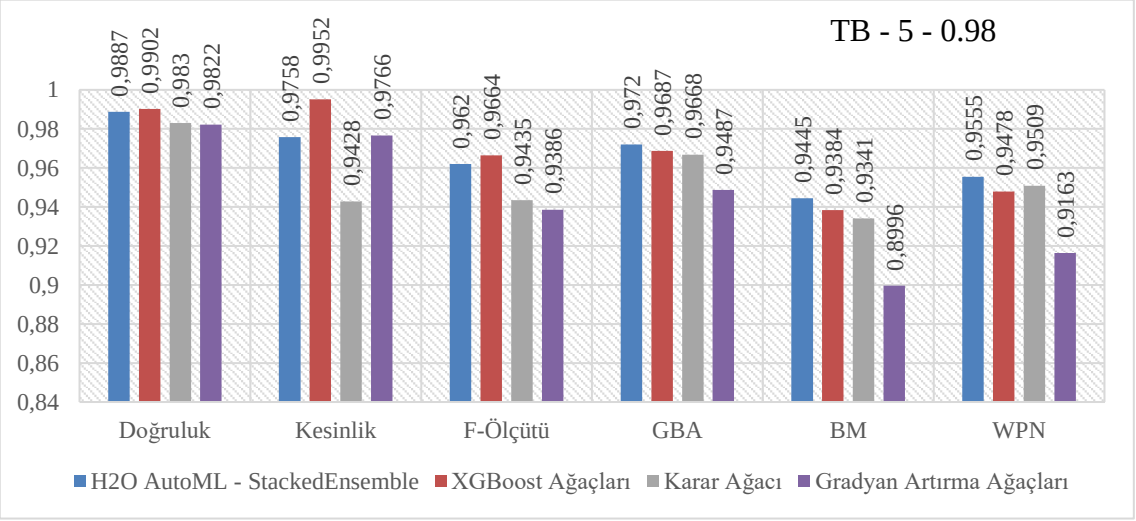
Şekil 3.6'daki grafik incelendiğinde TB 3 ile 4 arasında önemli bir düşüş görülmektedir. 4 ve sonrasındaki bileşenler değişkenliğin çok küçük bir kısmını açıklamaktadır. Sıfıra yakın olmaları nedeniyle önemsiz oldukları söylenebilir. Grafik ve TBA incelendiğinde TB olarak 3-4-5 bileşen ile çalışma yapılmasının daha iyi sonuçlar vereceği görülmektedir. TBA analizinde %92'lik açıklama ile yamaç eğim grafiği birbirini desteklemektedir. TB olarak 3-4-5 seçimlerine ait sonuçlar:



Şekil 3.7. TBA - 3 TB'ye ait Sonuçlar

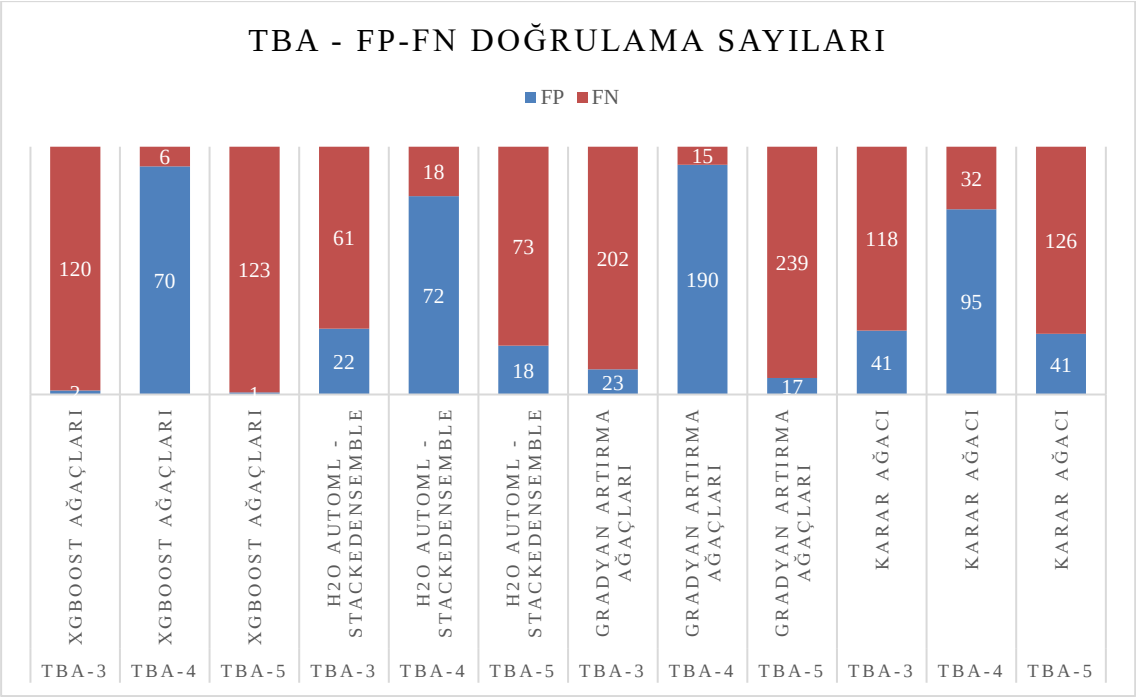


Şekil 3.8. TBA - 4 TB'ye ait Sonuçlar

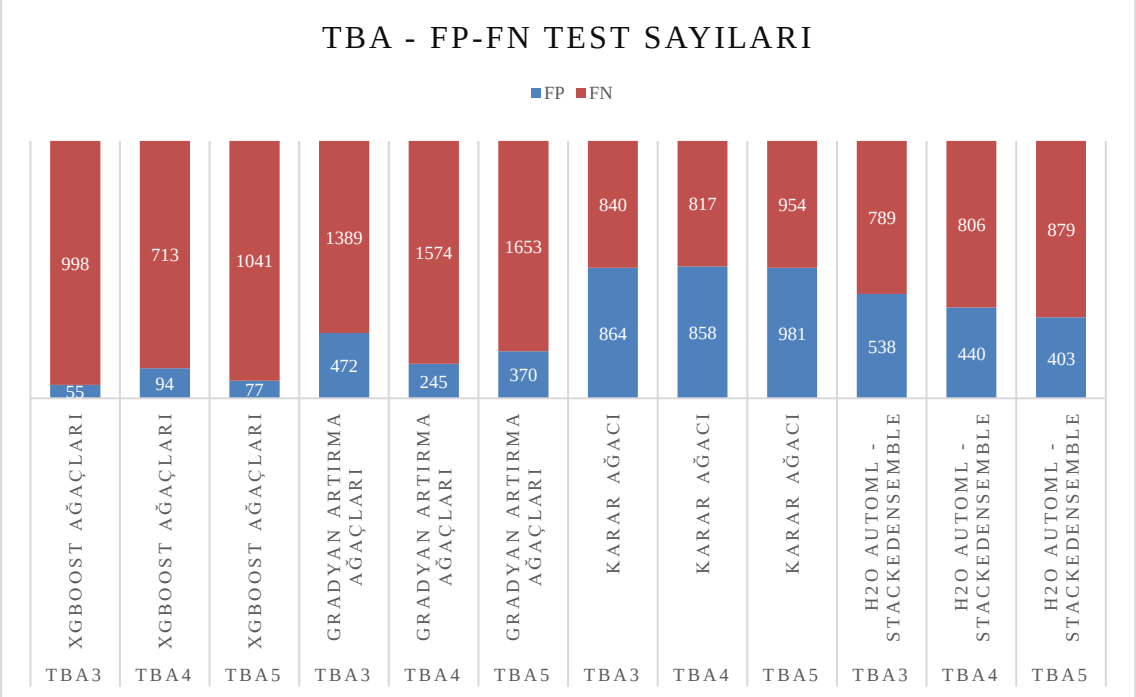


Şekil 3.9. TBA - 5 TB'ye ait Sonuçlar

TB olarak seçilen 3-4-5 içinden, 4 bileşen ile çalışan model, diğer ikisine göre biraz daha iyi performans göstermiştir.



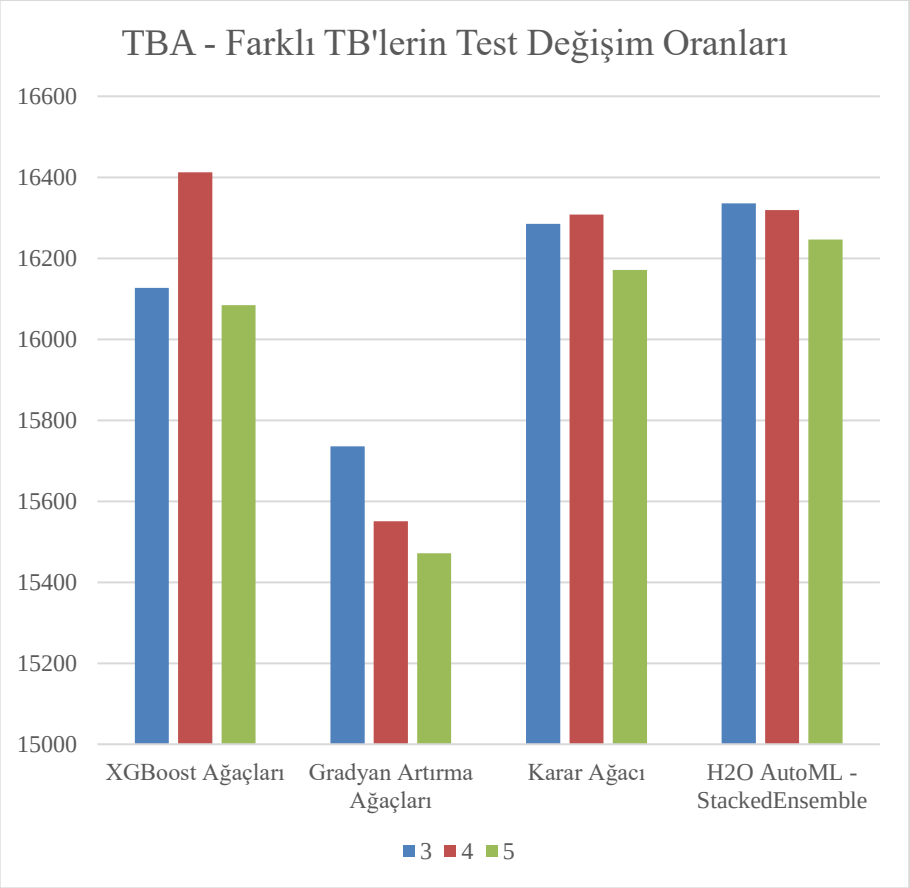
Şekil 3.10. TBA - TB 3-4-5'e ait FP ve FN Doğrulama Tablosu



Şekil 3.11. TBA - TB 3-4-5'e ait FP ve FN Test Tablosu

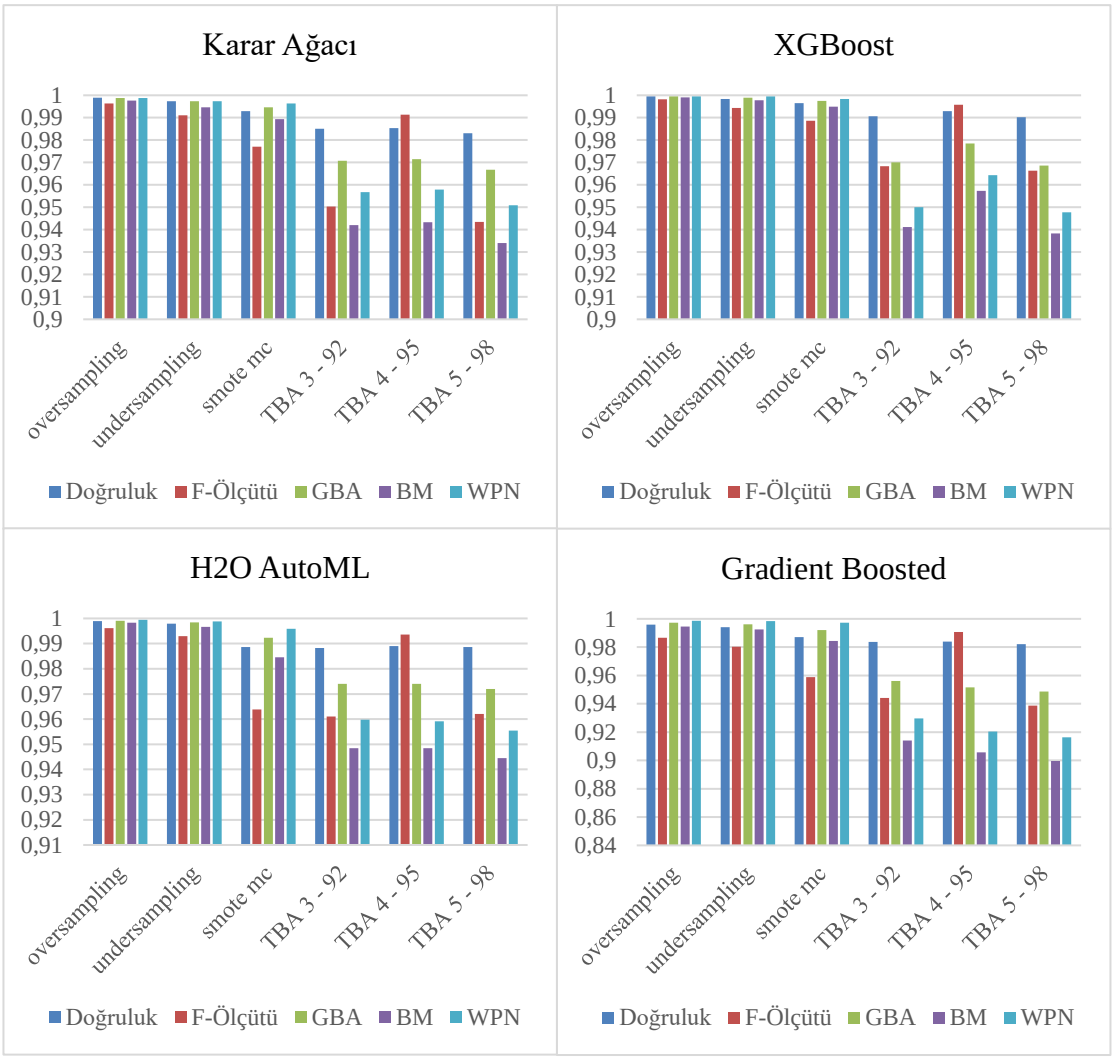
TBA ile geliştirilen modellerde hem doğrulama hem test aşamasında FP ve FN değerlerinde 0 yer almamaktadır. TBA-3 ve TBA-5'lik modelde FP oranları daha

düşüktür. Ayrıca TBA-3 ve TBA-5'lik modelde XGBoost doğrulama aşamasında 1 tane FP verirken; TBA-4 modelinde 70 tane vermiştir. Genel olarak TBA- 4 ile geliştirilen model daha iyi sonuçlar vermiştir.



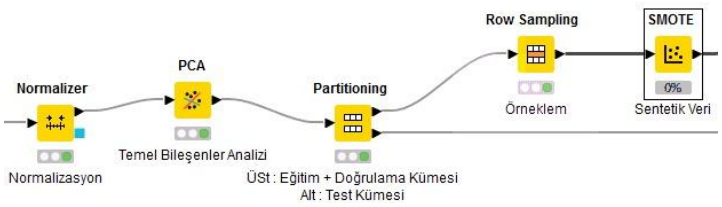
Şekil 3.12. TBA - TB 3-4-5'e ait Test Değişim Oranları

TB oranlarına bakıldığında TBA – 4 daha iyi sonuçlar verdiği söylenebilir. Aşağıdaki grafikte ise farklı teknikler ile yapılan modellere ait sonuçlar karşılaştırılmıştır.



Şekil 3.13. Farklı Tekniklere Ait Karşılaştırmalar

Undersampling ve Oversampling yöntemlerinde tüm modeller neredeyse aynı metrikler ile sonuçlanmıştır. SMOTE ile birlikte farklılıklar ortaya çıkmıştır. TBA ile birlikte SMOTE uygulandığında modelin performansında artış olmuştur. TBA-4 ile daha iyi sonuçlar alınmıştır.



Şekil 3.14. KNIME ile TBA + SMOTE

4. SONUÇLAR VE TARTIŞMA

4.1 Değerlendirme Ölçütleri

Çalışmada ele alınan veri büyük veridir. Tahmin etmeye çalışılan sınıflardan biri azınlık sınıfı, diğeri ise çoğunluk sınıfıdır. Dengesiz veri üzerinden ikili sınıflandırma problemini çözmek için veri çoklama ve yeni veri üretme teknikleri kullanılmıştır. Makine öğrenmesi algoritmaları kullanılarak müşteri memnuniyet durumu tahmin edilmeye çalışılmıştır. Bu bağlamda, şirketin yatırım yapacağı düşünülerek FP değeri üzerinden değerlendirme yapılmıştır.

Doğrulama ve test aşamasında doğru modeli seçmek için kullanılan ölçütlerden bir tanesi "TNR-TPR'a ait geometrik ortalama" (GBA) ölçütü olmuştur. Bundan farklı olarak ise "TNR+TPR-1" (BM) ve ağırlıklı TPR-TNR (WPN) ölçütleri de değerlendirmeye alınmıştır. Doğruluk metriğinin kullanılamayacağı birçok makale ve yazıda gösterilmiştir (Zhu, 2020). Bu nedenle dengesiz veriyi eğitecek modelin belirlenmesinde bir ölçüt olarak kullanılmamıştır.

Zhu'nun makalesinde yer alan test sonuçlarına göre; Veride dengesizlik arttıkça MCC ve CK metrik değerleri çarpıklığın artmasından dolayı doğrudan uygulanamayacağı görüşü referans alınmıştır. Bundan dolayı BM ve GBA ölçütleri yapılan test sonuçlarında değerlendirme ölçütü olarak kullanılmıştır. Bunlara ek olarak Jadhav tarafından bilimsel olarak test edilen WPN ölçütü de başka bir değerlendirme kriteri olarak kullanılmıştır. GBA, BM ve WPN ölçütleri tüm uygulanan yöntemlerde daha kararlı oldukları, birbirine yakın sonuçlar ürettikleri ve modeldeki azınlık sınıfına ait dengesizlikten etkilenmedikleri görülmektedir.

Test sonuçlarını değerlendirirken, şirketin harcama yapacağı zaman düşünülmüş, kaynak ve maliyet kriterleri dikkate alınarak FP üzerinden değerlendirme yapılmıştır. Buna göre FP sayısının az olması, daha az yanlış Memnuniyetsiz Müşteri tahmini anlamına gelmektedir.

4.2 Yöntemler

Doğrulama sonuçları değerlendirildiğinde TBA + SMOTE tekniğinin, oversampling ve undersampling tekniklerine göre daha iyi sonuçlar verdiği söylenebilir. Diğer bir farklılık ise veri setini çoğaltıldığında Derin Öğrenme'nin giderek daha iyi sonuçlar vermesidir.

Özet olarak GBA, BM ve WPN ölçütlerinin; yöntemler ve eğitim setine ait miktarlar değişse bile kararlı sonuçlar vermesidir. SMOTE ile yapılan tahminlerde FP ölçütüne ait değerlerde azalma görülmüş ve FP değerlerinin daha makul seviyelere geldiği görülmüştür.

TBA kullanıldığında model daha iyi sonuçlar vermiştir. TBA öncesinde veri normalleştirilmiş ve sonrasında rastgele örneklem ile veri küçültülmüştür. Sonrasında ise SMOTE – MC ile sentetik olarak çoğaltılan veri ile daha iyi modeller üretilmiştir.

4.3 Eğitim Hızı

Diğer bir kriter ise eğitim süresinin uzunluğudur. XGBoost, Sade Bayes, Gradyan Artırma, Karar Ağacı, H2O AutoML Çatısı, Lojistik Regresyon ve Rastgele Orman veri miktarı arttığında eğitim süresi belli aralıklarda kalmıştır. Derin Öğrenme ve Sinir Ağları gizli katmanlarının sayısına bağlı olarak ve veri miktarı ile orantılı olarak daha uzun sürelere çıkmaktadır.

Düşük Yüksek

Algoritma	min dk	max dk
Sade Bayes	4	8
GLM	6	9
H2O AutoML	8	8
Rastgele Orman	15	18
XGBoost Ağaçları	15	22
Lojistik Regresyon	30	36
Karar Ağacı	28	50
Gradyan Artırma Ağaçları	53	55
Sinir Ağları	65	118
Derin Öğrenme	55	125

Tablo 4.1 Algoritma Eğitim Süreleri Min-Max Değerleri

Genel olarak XGBoost hem sonuçları hem de hız olarak diğer algoritmalara göre öne çıkmaktadır. H2O AutoML çatısında hızın sabit kalmasının nedeni süre kısıtlaması eklenebilmesidir. Burada eğitim için konuşulan verinin büyüklüğü seçilirken rastgele örnekleme yönteminin seçilmesidir. Sonrasında örneklem algoritması uygulanmasına bağlı olarak veri 100.000 ile 300.000 arasında yer almıştır.

4.4 Gelecek Çalışmaları

Sahada yer alan cihazlara ait ek bilgilerin toplanması gelecekte geliştirilecek modeller açısından önem arz etmektedir. Bu bilgiler sahadaki dolap, ana dağıtım kutuları gibi internet hatlarını birbirine bağlayan ve sinyalleri gönderen cihazlardır. Cihazların sahip olduğu bilgileri alıp ve belli bir süre kaydetmek gerekmektedir. Bunun haricinde bu verinin yapısal olmamasından kaynaklı durumu dolayısıyla altyapı uzmanı ve elektronik uzmanı gibi kişilerle beraber çalışma gerekmektedir. Sahadaki verilerin işlenip hazırlanması sonrasında geliştirilecek modeller ile daha iyi sonuçlar vereceğini belirten makaleler mevcuttur.

Sahada dengesiz verinin daha iyi anlaşılmasını sağlamak adına farklı metrik ölçüleri de kullanılmaktadır. Cao ve ark. geliştirdikleri bir formül ile MCC ve F1 yöntemlerini birleştirerek MCC-F1 adını verdikleri bir metrik oluşturmuşlardır. Bu metrik ile dengesiz veriyi yorumlamada ROC ve Kesinlik metriklerine göre daha iyi sonuçlar aldıklarını söylemişlerdir. (Cao, 2020)

5. YORUM

Çalışılan veride yer alan internet ile alakalı özelliklerin normalleştirilip TBA'da 4 TB ile boyut indirgeme yapılması sonrasında veriyi rastgele örnekleme ve SMOTE yöntemi ile sentetik artırma yapılması ve çapraz doğrulama olarak k-5 uygulanması iyi sonuçlar vermiştir. Burada dikkat çeken konu TBA'da 4 TB çalıştığında modelin doğrulama sonuçlarında aşırı öğrenme göstermemiş olmasıdır. Bunu teyit eden ölçüt FP sayılarının TB 3'ten 4'e çıkıldığında yükselme göstermesidir. Tüm sonuçlara göre XGBoost algoritması tavsiye edilen yöntemdir. Aynı zamanda hız olarak da veri arttıkça çok fazla süre harcamadığı görülmüştür. FP değerlerinde de XGBoost diğer yöntemlere göre genel olarak düşük değerler verdiği görülmüştür. FP oranının düşürerek sahaya yönlendirilecek insanlar daha az zaman kaybına uğrayacaktır. Aynı zamanda şirketin bu abonelere yapacağı yatırımları daha ekonomik kullanmasını sağlayacaktır.

1. Sahadaki diğer cihazlara ait veriler modele eklendiğinde, modeli geliştirdiği ve daha iyi sonuçlar verdiği görülmektedir.

2. TBA ile SMOTE birlikte kullanıldığında ve öncesinde rastgele örneklem alındığında, dengesiz veri yetersiz özelliklere sahip olsa bile, doğru metrik seçimleriyle sonuca ulaşılabileceği gösterilmiştir.

3. Derin Öğrenme, yüksek maliyet ve hata alma riskine istinaden daha dikkatli kullanılması gereken bir yöntemdir. Tercih sebebi firma bazında değişebilir. Tezde kullanılan modelde Müşteri Memnuniyetsizliğini tahmin etmek, hızlı ve etkin bir çözüm gerektiğinden dolayı, derin öğrenme tercih sebebi değildir.

4. Genel olarak, Ağaç modelleri başarılı sonuçlar vermiştir.

5. TPR-TNR Geometrik ortalaması, TPR-TNR ağırlıklandırması ve BM ölçütleri dengesiz veride kullanıldığında daha kararlı sonuçlar vermiştir. Bu sonuçlar modem verileri üzerinden gösterilmiştir.

6. AutoML’de yer alan ölçütlendirme metriklerine dengesiz veri için de seçenekler eklenmelidir. GBA, WPN ve BM değerleri eklenmiş ve değerlendirmeler için ek geliştirme yapılmıştır.

7. Derin Öğrenme için Keras kütüphanesinin Python 3.6, Tensorflow 1 ve Keras 2 ile çalışması için çerçeve oluşturulması zorlu bir süreçtir. Çerçeve kurulduğunda AutoML çalıştırılsa bile, belirsiz nedenlerle "Keras Network Learner" hata verebilmektedir. Bu durumda ek geliştirme gerekmektedir.

8. AutoML yapısında ölçüt verilerinin değerlendirme için saklanması, modellerin hızlarının ölçülmesi ve test verilerinin değerlendirilmesi adına ek geliştirmeler gerekmektedir. Mevcut AutoML ideal senaryolar ve dengeli veriler için geliştirilmiştir ve doğrudan devreye alınmaya uygun değildir.

6. KAYNAKLAR

Ahsan, M., Gomes, R., Denton, A. (**2018**). Smote implementation on phishing data to enhance cybersecurity. In 2018 IEEE International Conference on Electro/Information Technology (EIT) (pp. 0531–0536). IEEE.

Aktepe, A. Ersöz, S. , Toklu, B. "Customer Satisfaction and Loyalty Analysis with Classification Algorithms and Structural Equation Modeling", Computers & Industrial Engineering, Volume 86, **2015**, Pages 95-106, ISSN 0360-8352, https://doi.org/10.1016/j.cie.2014.09.031.

Anonim - AutoML Düğümü KNIME Analitik Platformu, https://nodepit.com/workflow/com.knime.hub/Users/mumchakys05/Public/AutoML (Erişim tarihi: **22 Ağustos 2024**).

Anonim - Data Over Cable Service Interface Specification (DOCSIS) Proactive Network Maintenance (PNM), Best Practices Primer: HFC Networks (DOCSIS® 3.1), CM-GL-PNM-3.1-V01-200506, https://community.cablelabs.com/wiki/plugins/servlet/cablelabs/alfresco/download?id=4f3782eb-a1d7-4398-bc21-6d12bb70f915 (Erişim tarihi: **23 Aralık 2023**).

Anonim - DOCSIS® Best Practices and Guidelines, Proactive Network Maintenance Using Pre-equalization, CM-GL-PNMP-V02-110623, https://community.cablelabs.com/wiki/plugins/servlet/cablelabs/alfresco/download?id=fe685976-3dd1-44c6-91c3-a16d7031d79f (Erişim tarihi: **23 Aralık 2023**).

Arar, O.F., Ayan, K. (**2017**). "A feature dependent Naive Bayes approach and its application to the software defect prediction problem". Applied Soft Computing, 59, 197-209.

Bagheri, H., Shaltooki, A. (**2015**). Big Data: Challenges, Opportunities and Cloud Based Solutions. International Journal of Electrical and Computer Engineering (IJECE). 5. 340-343. 10.11591/ijece.v5i2.pp340-343.

Balaji, B., Raaj, A., Harsath, V., Pravin, R., Chinnappa Naidu, Rani, Aarthi, G. , Aggarwal, Geetika, Kumar, M. (**2024**). Taxi Revenue Optimization with Deep Q-Learning and Enhanced Data Visualization. 1-6. 10.1109/AIIoT58432.2024.10574699.

Basha, S. J., Madala, S. R., Vivek, K., Kumar, E. S., Ammannamma, T. "A Review on Imbalanced Data Classification Techniques," 2022 International Conference on Advanced Computing Technologies and Applications (ICACTA), Coimbatore, India, **2022**, pp. 1-6, doi: 10.1109/ICACTA54488.2022.9753392.

Bej, S., Davtyan, N., Wolfien, M., Nassar, M., Wolkenhauer, O. (**2021**). Loras: An oversampling approach for imbalanced datasets. Machine Learning, 110(2), 279–301.

Bolívar, D., Elreedy, D., Atiya, A. (**2022**). A theoretical distribution analysis of synthetic minority oversampling technique (SMOTE) for imbalanced learning. Machine Learning. doi:10.1007/s10994-021-06067-4.

Bunkhumpornpat, C., Sinapiromsaran, K., Lursinsap, C. (**2009**). Safe-level-smote: Safe-level-synthetic minority over-sampling technique for handling the class imbalanced problem. In Pacific-Asia Conference on Knowledge Discovery and Data Mining (pp. 475–482). Springer.

Burkov, A. (**2019**) The Hundred-Page Machine Learning Book.

Cao, C., Chicco, D., Hoffman, M. M. The MCC-F1 Curve: A Performance Evaluation Technique for Binary Classification, **2020**, 2006.11278, arXiv.

Chawla, N. V., Bowyer, K. W., Hall, L. O., Kegelmeyer, W. P., SMOTE: Synthetic Minority Over-sampling Technique, Journal of Artificial Intelligence Research, **2002**, Volume 16, P. 321-357, ISSN 1076-9757.

Chen, T., Guestrin, C. (**2016**), XGBoost: A Scalable Tree Boosting System. 785-794. 10.1145/2939672.2939785.

Christen, P., Hand, D. J., Kirielle, N. **2023**. A Review of the F-Measure: Its History, Properties, Criticism, and Alternatives. ACM Comput. Surv. 56, 3, Article 73 (March 2024), 24 pages. https://doi.org/10.1145/3606367.

Cubukcu, U., Erdogan, O., Pathak, S., Sannakkayala, S., Slot, M., Citus: Distributed PostgreSQL for Data-Intensive Applications, Virtual Event, June 20–25, China, **2021**, 2490-2502. 10.1145/3448016.3457551.

Cui, S., Sudjianto, A., Zhang, A., Li, R. (**2023**), Enhancing Robustness of Gradient-Boosted Decision Trees through One-Hot Encoding and Regularization, 2304.13761.

Cutler, A. (**2011**). Remembering Leo Breiman. Annals of Applied Statistics - ANN APPL STAT. 4. 10.1214/10-AOAS427.

Dablain, D., Krawczyk, B., Chawla, N. V. (**2022**). Deepsmote: Fusing deep learning and smote for imbalanced data. IEEE Transactions on Neural Networks and Learning Systems. https:// doi. org/ 10.1109/ TNNLS. 2021. 31365 03.

Elyan, E., Moreno-Garcia, C. F., Jayne, C. (**2021**). Cdsmote: Class decomposition and synthetic minority class oversampling technique for imbalanced-data classification. Neural Computing and Applications, 33(7), 2839–2851.

Fernandez, A., Garcia, S., Herrera, F., Chawla, N. V. (**2018**). Smote for learning from imbalanced data: Progress and challenges, marking the 15-year anniversary. Journal of Artificial Intelligence Research, 61, 863–905.

Guillen, M. D., Aparicio, J., Esteve, M., Gradient tree boosting and the estimation of production frontiers, Expert Systems with Applications, Volume 214, **2023**, 119134, ISSN 0957-4174.

Han, H., Wang, W. Y., Mao, B. H. (**2005**). Borderline-smote: A new over-sampling method in imbalanced data sets learning. In International Conference on Intelligent Computing (pp. 878–887). Springer.

Haykin, S. (**1999**). A comprehensive foundation. Neuralnet works. Pearson Education.

He, H., Bai, Y., Garcia, E. A., Li, S. A. (**2008**). Adasyn: Adaptive synthetic sampling approach for imbalanced learning. In IEEE International Joint Conference on Computational Intelligence, IJCNN (pp.1322–1328). IEEE.

He, H., Ma, Y., Imbalanced Learning, s.27 (**2013**).

He, Z., Lin, D., Lau, T., Wu, M. (**2019**). Gradient Boosting Machine: A Survey. arXiv preprint arXiv:1908.06951. Retrieved from https://arxiv.org/abs/1908.06951.

Jadhav, S. A., A Novel Weighted TPR-TNR Measure to Assess Performance of the Classifiers, Expert Systems with Applications, Volume 152, **2020**, 113391, ISSN 0957-4174.

Kalutarage, H. K., Nguyen, H. N., Shaikh, S. A. (**2017**). Naive Bayes: applications, variations and vulnerabilities. Soft Computing, 22(3), 855-870. doi:10.1007/s00500-016-2384-7.

Kishor, A., Chakraborty, C. (**2021**). Early and accurate prediction of diabetics based on FCBF feature selection and smote. International Journal of System Assurance Engineering and Management. https://doi. org/ 10. 1007/ s13198- 021- 01174-z.

Kreuzberger, D., Kühl, N., Hirschl, S. "Machine Learning Operations (MLOps): Overview, Definition, and Architecture," in IEEE Access, vol. 11, pp. 31866-31879, **2023**, doi: 10.1109/ACCESS.2023.3262138.L.TongHassani, Hossein & Huang, Xu & Silva, Emmanuel & Ghodsi, Mansi. (2016). A Review of Data Mining Applications in Crime. Statistical Analysis and Data Mining. 9. 10.1002/sam.11312.

Kumar, S., Zymbler, M. A Machine Learning Approach to Analyze Customer Satisfaction from Airline Tweets. J Big Data 6, 62 (**2019**). https://doi.org/10.1186/s40537-019-0224-1.

LeDell, E. "H2O AutoML: Scalable Automatic Machine Learning." (**2020**).

Lee, Y., Wang, Y., Lu, S., Hsieh, Y., Chien, C., Tsai, S., Dong, W. (**2016**). An Empirical Research on Customer Satisfaction Study: A Consideration of Different Levels of Performance. Springer Plus. 5. 10.1186/s40064-016-3208-z.

Liu, R. (**2023**). A novel synthetic minority oversampling technique based on relative and absolute densities for imbalanced classification. Applied Intelligence, 53(1), 786-803.

Madni, H., Umer, M., Ishaq, A., Abuzinadah, N., Saidani, O., Alsubai, S., Hamdi, M., Ashraf, I. (**2023**). Water-Quality Prediction Based on H2O AutoML and Explainable AI Techniques. Water. 15. 475. 10.3390/w15030475.

McAfee, A., Brynjolfsson, E. (**2012**). Big data: The Management Revolution. Harvard Business Review, 90(10), 60–68.

Meinzer, S., Thamm, A., Jensen, U., Hornegger, J., Eskofier, B. (**2017**). Can Machine Learning Techniques Predict Customer Dissatisfaction? A Feasibility Study for the Automotive Industry. Journal of Artificial Intelligence Research. 6. 80-90. 10.5430/air.v6n1p80.

Minaee, S., Kalchbrenner, N., Cambria, E., Nikzad, N., Chenaghlu, M., Gao, J. 2021. Deep Learning–based Text Classification: A comprehensive Review. Computing Surveys 54, 3 (**2021**), 1–40.

Mishra, D., Luo, Z., Jiang, S., Papadopoulos, T., Dubey, R. (**2017**). A Bibliographic Study on Big Data: Concepts, Trends and Challenges. Business Process Management Journal, 23(3), 555–573.

Mulla, G., Demir, Y., Hassan, M. (**2021**). Combination of PCA with SMOTE Oversampling for Classification of High-Dimensional Imbalanced Data. Bitlis Eren Üniversitesi Fen Bilimleri Dergisi. 10. 10.17798/bitlisfen.939733.

Nguyen, H. M., Cooper, E. W., Kamei, K. (**2011**). Borderline over-sampling for imbalanced data classification. International Journal of Knowledge Engineering and Soft Data Paradigms, 3(1), 4–21.

Nykodym, T. , Rao, A., Wang, A., Kraljevic, T., Lanford, J., Hussami, N. Generalized Linear Modeling with H2O, http://h2o.gitbooks.io/glm-with-h2o/ August, **2015**, ThirdEdition.

Oliveira, C., Guimarães, T., Portela, F., Santos, M. Benchmarking Business Analytics Techniques in Big Data, Procedia Computer Science, Volume 160, **2019**, Pages 690-695, ISSN 1877-0509, https://doi.org/10.1016/j.procs.2019.11.026.

Rosenbaum, P.R. (**2010**). Causal Inference in Randomized Experiments. In: Design of Observational Studies. Springer Series in Statistics. Springer, New York, NY. https://doi.org/10.1007/978-1-4419-1213-8_2.

Saito, T., Rehmsmeier, M. The Precision-Recall Plot is more Informative than the ROC Plot when Evaluating Binary Classifiers on Imbalanced Datasets. PLoS One. **2015** Mar 4;10(3):e0118432. doi: 10.1371/journal.pone.0118432. PMID: 25738806; PMCID: PMC4349800.

Salama, M. A., Hassanien, A. E., Fahmy, A. A. "Reducing the influence of normalization on data classification," 2010 International Conference on Computer Information Systems and Industrial Management Applications (CISIM), Krakow, Poland, **2010**, pp. 609-613, doi: 10.1109/CISIM.2010.5643523.

Settouti, N., El Louadi, M., Ares, S. (**2016**). "Comparative study of the use of feature selection and data mining techniques in detecting fraudulent credit card transactions". Procedia Computer Science, 83, 325-332.

Silveira, P., Morais, R., Petrella, S. (**2022**). A Communication Study of Young Adults and Online Dependency during the COVID-19 Pandemic. Societies. 12. 109. 10.3390/soc12040109.

Singh, V. K., Joshi, K. (**2022**) Automated Machine Learning (AutoML): An Overview of Opportunities for Application and Research, Journal of Information Technology Case and Application Research, 24:2, 75-85, DOI: 10.1080/15228053.2022.2074585.

Suleyman, M., Bhaskar, M. The Coming Wave, Dijital Baskı, Crown Publishing Group, **2023**.

Tong, L., Wang, Y., Wen, F., Li, X. "The Research of Customer Loyalty Improvement in Telecom Industry based on NPS Data Mining," in China Communications, vol. 14, no. 11, pp. 260-268, Nov. **2017**, doi: 10.1109/CC.2017.8233665.

Torres, H., Portela, F., Santos, M. F. (**2018**). An Overview of Big Data Architectures in Healthcare. https://doi.org/10.1007/978-3-319-77700-9.

Wei, Q., Shi, X., Li, Q., Chen, G. (**2020**). Enhancing Customer Satisfaction Analysis with a Machine Learning Approach: From a Perspective of Matching Customer Comment and Agent Note. 10.24251/HICSS.2020.178.

Wong, A., Marikannan, B. P. **2020** J. Phys.: Conf. Ser. 1712 012044.

Wu, X., Kumar, V., Quinlan, J. R., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G. J., Ng, A., Liu, B., Yu, P. S., & Steinberg, D. (**2008**). "Top 10 algorithms in data mining". Knowledge and Information Systems, 14(1), 1-37.

Yalovetzky, R., Kumar, N., Li, C., Pistoia, M. "QC-Forest: a Classical-Quantum Algorithm to Provably Speedup Retraining of Random Forest", **2024**, 2406.12008, arXiv, quant-ph, https://arxiv.org/abs/2406.12008.

Zaharia, M., Chowdhury, M., Franklin, M. J., Shenker, S., Stoica, I. (**2010**). Spark: Cluster computing with working sets. In 2nd USENIX workshop on hot topics in cloud computing (HotCloud 10).

Zhu, Q., On the Performance of Matthews Correlation Coefficient (MCC) for Imbalanced Dataset, Pattern Recognition Letters, Volume 136, **2020**, Pages 71-80, ISSN 0167-8655.

EKLER

EK 1 – SQL KODLARI

```
--Sorunlu Sinyal Değerlerinin saatlik bazda veri tabanından okunması
select id,hizmet_id,cmts_okuma_zamani::date,case
when extract(hour from cmts_okuma_zamani) between 0 and 3 then '00-04'
when extract(hour from cmts_okuma_zamani) between 4 and 7 then '04-08'
when extract(hour from cmts_okuma_zamani) between 8 and 11 then '08-12'
when extract(hour from cmts_okuma_zamani) between 12 and 15 then '12-16'
when extract(hour from cmts_okuma_zamani) between 16 and 19 then '16-20'
when extract(hour from cmts_okuma_zamani) between 20 and 23 then '20-24'
end as saat_araligi,
sum(case when sorunlu_mu_dn = 'false' then 0 else 1 end) sorunlu_port_sayisi_d,
sum(case when sorunlu_mu_up = 'false' then 0 else 1 end) sorunlu_port_sayisi_u
from (
select (sorunlu_mu_dn->'sorunlu')::text as sorunlu_mu_dn ,id,cmts_okuma_zamani
,(sorunlu_mu_up->'sorunlu')::text as sorunlu_mu_up,hizmet_id
from public.$${Schild}$$,
json_array_elements(sinyal_degerleri_dn::json) sorunlu_mu_dn,
json_array_elements(sinyal_degerleri_up ::json) sorunlu_mu_up
) dt
group by 1,2,3,4;
```