T.C.
YILDIZ TEKNİK ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

TELEKOMÜNİKASYON SEKTÖRÜNDE MÜŞTERİ SADAKATİNİN VERİ MADENCİLİĞİ TEKNİKLERİYLE MODELLENMESİ

MÜMİN YILDIZ

YÜKSEK LİSANS TEZİ
BİLGİSAYAR MÜHENDİSLİĞİ BÖLÜMÜ
BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI

DANIŞMAN
DOÇ. DR. SONGÜL ALBAYRAK

İSTANBUL, 2017

T.C.

YILDIZ TEKNİK ÜNİVERSİTESİ

FEN BİLİMLERİ ENSTİTÜSÜ

TELEKOMÜNİKASYON SEKTÖRÜNDE MÜŞTERİ SADAKATİNİN VERİ MADENCİLİĞİ TEKNİKLERİYLE MODELLENMESİ

Mümin YILDIZ tarafından hazırlanan tez çalışması 03.04.2017 tarihinde aşağıdaki jüri tarafından Yıldız Teknik Üniversitesi Fen Bilimleri Enstitüsü Bilgisayar Mühendisliği Bölümü, Bilgisayar Mühendisliği Anabilim Dalı'nda **YÜKSEK LİSANS TEZİ** olarak kabul edilmiştir.

Tez Danışmanı

Doç. Dr. Songül ALBAYRAK

Yıldız Teknik Üniversitesi

Jüri Üyeleri

Doç. Dr. Songül ALBAYRAK

Yıldız Teknik Üniversitesi _____

Prof. Dr. Oya KALIPSIZ

İstanbul Üniversitesi _____

Prof. Dr. Hasan DAĞ

Kadir Has Üniversitesi _____

ÖNSÖZ

Öğrenim hayatım boyunca bana emeği geçen tüm hocalarıma, tez sürecinde beni en iyi şekilde yönlendiren değerli danışmanım Doç. Dr. Songül ALBAYRAK Hocama ve her türlü desteğini hiçbir zaman benden esirgemeyen aileme ve dostlarıma sonsuz teşekkür ederim.

Tez çalışmamda kullandığım veri kümelerini benim ile paylaşan Chih-Fong TSAI ve Wouter VERBEKE'ye teşekkürü borç bilirim.

Nisan, 2017

Mümin YILDIZ

İÇİNDEKİLER

Sayfa

SİMGE LİSTESİ vi

KISALTMA LİSTESİ vii

ŞEKİL LİSTESİ viii

ÇİZELGE LİSTESİ ix

ÖZET x

ABSTRACT xii

BÖLÜM 1

GİRİŞ 1

1.1 Literatür Özeti 1
1.2 Tezin Amacı 8
1.3 Hipotez 10

BÖLÜM 2

VERİ MADENCİLİĞİ YÖNTEMLERİ 11

2.1 Aykırı Değer Analizi Yöntemleri 11
2.2 Öznitelik Seçme Yöntemleri 13
2.3 Sınıflandırma Yöntemleri 15
2.4 Melez Yöntemler 19
2.4.1 Birinci Melez Model (M1) 19
2.4.2 İkinci Melez Model (M2) 19
2.4.3 Üçüncü Melez Model (M3) 19

BÖLÜM 3

VERİ KÜMELERİ 21

3. Veri Kümleri: 21

3.1 Tsai Veri Kümesi 21

3.1.1 Tsai Veri Kümesinde Kayıp Değer Doldurulması 22
3.1.2 Tsai Veri Kümesi Öznitelik Seçme İşlemleri 22
3.1.3 Tsai Veri Kümesi Aykırı Değer Tespiti 23
3.1.3.1 Kutu Grafiği Analizi 23
3.1.3.2 Ortalama – Standart Sapma Aykırı Değer Analizi 26
3.2 Larose Veri Kümesi 28
3.2.1 Larose Veri Kümesi Öznitelik Seçme İşlemi 28
3.2.2 Larose Veri Kümesi Aykırı Değer Tesbiti 29
3.2.2.1 Kutu Grafiği Analizi 29
3.2.2.2 Ortalama – Standart Sapma Aykırı Değer Analizi 31
3.2.2.3 Uzaklık Tabanlı Aykırı Değer Analizi 32
3.2.3 Larose Veri Kümesi Veri Alt Örnekleme 32

BÖLÜM 4

DEĞERLENDİRME KRİTERLERİ 34

BÖLÜM 5

SINIFLANDIRMA SONUÇLARI 37

5.1. Tsai Veri Kümesi 37
5.1.1 Kayıp Değer Doldurma İşlemi Sınıflandırma Sonuçları 37
5.1.2. Öznitelik Azaltımı Sınıflandırma Sonuçları 37
5.1.3 Aykırı Değer Analizi Sonucu Kalan Veri İle Yapılan Sınıflandırmalar . 38
5.1.4 Sınıflayıcıların Parametreleri Değiştirilerek Yapılan Çalışmalar 41
5.1.5 Melez Yöntemler 42
5.2 Larose Veri Kümesi 43
5.2.1 Öznitelik Azaltımı Sınıflandırma Sonuçları 43
5.2.2 Alt Örnekleme İşlemi Sınıflandırma Sonuçları 44
5.2.2.1 Tekil Veri Kümesinde Sınıflandırma 44
5.2.2.2 Çoklu Veri Kümesinde Sınıflandırma 46
5.2.3 Melez Yöntemler 50
5.2.3.1 Birinci Melez Model 50
5.2.3.2 İkinci Melez Model 51
5.2.3.3 Üçüncü Melez Model 52

BÖLÜM 6

SONUÇ VE ÖNERİLER 54

KAYNAKLAR 56

EK-A

Tsai Veri Kümesi Öznitelikleri ve Açıklamaları 59

EK-B

Larose Veri Kümesi Öznitelikleri ve Açıklamaları 64

ÖZGEÇMİŞ 65

SİMGE LİSTESİ

d	Uzaklık denklemi
H	Entropi
p	Bir olayın gerçekleşme olasılığı
x^2	Ki-kare testi
A_{ij}	Gözlenen frekans değeri
E_{ij}	Beklenen frekans değeri
P	İncelenen olayın gözlenme olasılığı
β_0	Bağımsız değişkenler sıfır değerini aldığında bağımlı değişkenin değeri(sabiti)
β_k	Bağımsız değişkenin regresyon katsayısı
X_k	Bağımsız değişkenler
K	Bağımsız değişken katsayısı
e	Euler Sayısı(2,71)
OP	Sınıflandırıcı sonuçlarının sayısal ölçümleri

KISALTMA LİSTESİ

AD1 Kümesi	Toplam örnek sayısının %75'ini kapsayan 1 ve üstü aykırı özniteliği olan örnekler
AD2 Kümesi	Toplam örnek sayısının %10'unu kapsayan 26 ve üstü aykırı özniteliği olan örnekler
ALBA	Active Learning Based Approach for SVM Rule Extraction
CART	Classification and Regression Tree
ÇKA	Çok Katmanlı Algılayıcılar
DVM	Destek Vektör Makinaları
FN	False Negative
FP	False Positive
GA	Genetik Programlama Algoritmaları
ICA	Independent Component Analaysis(Bağımsız Bileşen Analizi)
KD1	Kayıp değerlerin sınıflarına bakılmasızın doldurulmasıyla oluşan veri kümesi
KD2	Kayıp değerlerin sınıflarına bakılarak doldurulmasıyla oluşan veri kümesi
LOLIMOT	Locally Linear Model Tree
M1	Birinci Melez Model
M2	İkinci Melez Model
M3	Üçüncü Melez Model
PCA	Principal Component Analaysis (Temel Bileşen Analizi)
Q1	Birinci Çeyrek
Q2	İkinci Çeyrek
Q3	Üçüncü Çeyrek
RO	Rastgele Orman
ROC	Receive Operating Curves
RotO	Rotasyon Ormanı
SRP	Sparse Random Projection (Seyrek Rastgele Yansıma)
TN	True Negative
TP	True Positive

ŞEKİL LİSTESİ

Sayfa

Şekil 2.1 Çok Katmanlı Ağ Mimarisi 16
Şekil 2.2 Melez Yöntemler 20
Şekil 3.1 Tekil veri kümesi 33
Şekil 3.2 Çoklu veri kümesi 33
Şekil 5.1 Tsai veri kümesi 1. melez model sınıflandırma sonuçları 43
Şekil 5.2 Larose veri kümesi 1. melez model sınıflandırma sonuçları 50
Şekil 5.3 Larose veri kümesi 2. melez model sınıflandırma sonuçları 51
Şekil 5.4 Larose veri kümesi 3. melez model sınıflandırma sonuçları 52

ÇİZELGE LİSTESİ

Çizelge 3.1 Tsai veri kümesi en iyi 10 özniteliği 23
Çizelge 3.2 Tsai veri kümesinde kutu grafiği yöntemiyle bulunan özniteliklerine göre aykırı abone sayıları..... 24
Çizelge 3.3 Tsai veri kümesinde kutu grafiği yöntemi ile bulunan özniteliklerin aykırı değer sayıları..... 25
Çizelge 3.4 Tsai veri kümesinde ortalama standart sapma yöntemiyle bulunan özniteliklerine göre aykırı abone sayıları..... 26
Çizelge 3.5 Tsai veri kümesinde ortalama standart sapma yöntemi ile bulunan özniteliklerin aykırı değer sayıları..... 27
Çizelge 3.6 Larose veri kümesi en iyi 10 özniteliği..... 29
Çizelge 3.7 Larose veri kümesinde kutu grafiği yöntemiyle bulunan özniteliklerine göre aykırı abone sayıları..... 30
Çizelge 3.8 Larose veri kümesinde kutu grafiği yöntemi ile bulunan özniteliklerin aykırı değer sayıları 30
Çizelge 3.9 Larose veri kümesinde ortalama standart sapma yöntemiyle bulunan özniteliklerine göre aykırı abone sayıları..... 31
Çizelge 3.10 Larose veri kümesinde ortalama standart sapma yöntemi ile bulunan özniteliklerin aykırı değer sayıları 31
Çizelge 3.11 Larose veri kümesinde uzaklık tabanlı aykırı değer analizi sonuçları..... 32
Çizelge 4.1 Sınıf karışıklık matrisi..... 34
Çizelge 5.1 Tsai veri kümesinde kayıp değer doldurulmasıyla yapılan sınıflandırma sonuçları 39
Çizelge 5.2 Tsai veri kümesi öznitelik azaltımı sınıflandırma sonuçları 39
Çizelge 5.3 Tsai aykırı değer analizi sınıflandırma sonuçları..... 40
Çizelge 5.4 Tsai Veri Kümesi 77 Öznitelik ve 1'den Fazla Aykırı Değeri Olan Abonelerin Atılmasıyla Yapılan Sınıflandırma Sonuçları 40
Çizelge 5.5 Tsai veri kümesinde ÇKA parametre değişimleri sınıflandırma sonuçları..... 41
Çizelge 5.6 Tsai veri kümesinde KNN parametre değişimi sınıflandırma sonuçları 42
Çizelge 5.7 Tsai veri kümesi 1. melez yöntem sınıflandırma sonuçları 42
Çizelge 5.8 Larose veri kümesi öznitelik azaltımı sınıflandırma sonuçları..... 45
Çizelge 5.9 Larose tekil veri kümesi sınıflandırma sonuçları 45
Çizelge 5.10 Larose çoklu alt veri kümesi geri çağırma oranı sonuçları 47
Çizelge 5.11 Larose çoklu alt veri kümesi kesinlik oranı sonuçları 48
Çizelge 5.12 Larose çoklu alt veri kümesi f-ölçütü sonuçları..... 49
Çizelge 5.13 Larose veri kümesi birinci melez model sınıflandırma sonuçları 50
Çizelge 5.14 Larose veri kümesi ikinci melez model sınıflandırma sonuçları..... 51
Çizelge 5.15 Larose veri kümesi üçüncü melez model sınıflandırma sonuçları 52

ÖZET

TELEKOMÜNİKASYON SEKTÖRÜNDE MÜŞTERİ SADAKATİNİN VERİ MADENCİLİĞİ TEKNİKLERİYLE MODELLENMESİ

Mümin YILDIZ

Bilgisayar Mühendisliği Bölümü

Yüksek Lisans Tezi

Tez Danışmanı: Doç. Dr. Songül ALBAYRAK

Son yıllarda önemli ölçüde artan rekabet ortamında, müşterilerin kolayca alternatif hizmetlere yönelebilmelerinden dolayı şirketlerin stratejilerinde ayrılacak müşteriyi önceden tahmin etmek önem kazanmıştır. Ayrılacağı önceden tahmin edilen müşterilere promosyon, indirim, hediye ya da avantajlar sağlanması müşterinin ayrılmasını engelleyebilecek ve böylece şirketler uzun vadede daha fazla kâr elde edebileceklerdir. Ancak yapılacak yanlış tahminler, ayrılmayacak müşteriye ayrılacakmış gibi davranmamıza neden olabilir. Bu da müşteriye gereksiz promosyon ya da hediye uygulanmasını gerektirir. Yani şirket için bunun anlamı, gereksiz para kaybıdır. Bu nedenle ayrılacak müşteriyi doğru tahmin etmekte şirketler için önemlidir. Ayrılacak müşteri tahmin modelini oluşturmak için de eskiden ayrılmış müşterilerin verileri kullanılır.

Müşteri sadakati tahmin modelinin kalitesi iki önemli faktöre bağlıdır. Bunlar, müşteri sadakatini elde etmek amacıyla kullanılan eldeki veriler ve veri madenciliği yöntemleridir. Bu çalışmada da telekomünikasyon sektöründe hizmet veren şirketler için ayrılacak müşterileri önceden tahmin edebilmek amacıyla çeşitli veri madenciliği teknikleri kullanılmış ve başarıyı artırmak için melez modeller denenmiştir. Yapılan veri ön işlem adımları ve melez sınıflandırma teknikleri, klasik sınıflandırma teknikleriyle karşılaştırılmıştır. Sınıflandırma tekniklerinin çalışma süresini düşürmek ve başarımını artırmak amacıyla öznitelik sayısı azaltılarak da denemeler yapılmış ve performansları

ölçülmüştür. Bu işlemler 2 farklı Amerikan telekomünikasyon şirketinden alınan veriler üzerinde denenmiş ve performans ölçütü olarak Geri Çağırma Oranı, Kesinlik Oranı ve F-Ölçütü kullanılmıştır.

Anahtar Kelimeler: Müşteri Sadakati, Müşteri Ayrılma Tahmini, Telekomünikasyon, Veri Madenciliği, Melez Model

ABSTRACT

MODELLING OF CUSTOMER CHURN PREDICTION IN TELECOMMUNICATION SECTOR WITH DATA MINING TECHNIQUES

Mümin YILDIZ

Hardware Department of Computer Engineering

MSc. Thesis

Adviser: Assoc. Prof. Dr. Songül ALBAYRAK

Customer churn is gained importance for companies because of increasing competition become more important of marketing strategies, more conscious act of the customer in recent years. Giving promotions, discounts or gifts can prevent the customer who is predicted as churn. Thus companies will be able to get more profits in the long run. However, predictions which will be incorrect can cause to behave nonchurn customer as if churn. This mean is giving unnecessarly promotions or gifts to customer. In other words it is unnecessary waste of money for companies. For this reason, correct prediction of churn customer is very important for companies. Information of formerly churn customers is used to create customer churn prediction model.

Quality of customer churn prediction model depends on two important factors. These are the available data and data mining techniques. In these research, various data mining techniques are used and hybrid models are tried to increase performance in order to correct prediction of churn customer for companies which service in telecommunication sector. Performed pre-processing steps and hybrid classification techniques are compared with classic classification models. In order to decrease run time of classification techniques and increase success, feature reduction process was applied and performance was measured. These processes are tried in two dataset that were taken two different American Telecommunication companies. Recall, precision and F-measure are used as performance criteria.

Keywords: Customer Loyalty, Customer Churn Prediction, Telecommunication, Data Mining, Hybrid Model

BÖLÜM 1

GİRİŞ

1.1 Literatür Özeti

Özellikle son yıllarda rekabetin artması, pazarlama stratejilerinin önem kazanması, müşterilerin daha bilinçli hareket etmesi gibi nedenlerden dolayı müşteri sadakati şirketler için önem kazanmaktadır. Bu gibi sebeplerden dolayı, müşteriler çeşitli nedenlerle kolay bir şekilde alternatif hizmetlere yönelebilmektedirler. Bunun engellenebilmesi için şirketler verdikleri hizmete bağlı olarak çeşitli stratejiler geliştirmelidir. Bu stratejiler iki farklı şekilde müşteriye uygulanabilirler. Bunlar reaktif (tepkisel) ve proaktif (önleyici) yaklaşımlardır [1]. Reaktif yaklaşım; şirket, müşteri üyeliğini ya da aboneliğini iptal edene kadar bekler. İptal durumda şirket, müşterinin kalması için teşvik verir. Proaktif yaklaşımda ise müşteri ayrılmadan, ayrılacağı tahmin edilmeye çalışılır ve onlara teşvik ya da özel programlar verilir. Bu yaklaşımla ayrılacak müşteriyi elde tutmak daha az maliyetlidir. Ancak bu sistem, eğer tahmin modeli hatalı çalışıyorsa şirket için zararlıdır. Çünkü şirket ayrılmayan müşterilerine de promosyonlar sunacaktır. Bu da şirket için para kaybıdır.

Müşteri sadakati hizmet alınan alanda başka alternatiflerinde bulunduğu durumlarda, müşterinin bir şirkete duymuş olduğu bağlılık olarak tanımlanabilir. Müşterinin bağlılığı, şirkete, hizmete, satıcıya ya da ürüne karşı olabilir.

Müşteri kaybı çeşidi iki farklı şekilde incelenebilir: Gönüllü ve gönülsüz [2]. Gönülsüz müşteri kaybında, şirket kendisi müşteriye hizmet vermeyi bırakır. Bunun nedenleri hizmete kötüye kullanma ya da faturanın ödenmemesi gibi nedenler olabilir. Gönüllü

ayrılan müşteriler ise iki farklı alt gruba ayrılabilir. Birinci gruptaki müşterilerin ayrılma nedenleri, ev taşınması ya da meslek değişimi gibi nedenlerdir. İkinci gruptaki müşteriler ise daha iyi hizmet yada daha ucuz ürün veren şirketlere geçtikleri için ayrılanlardır. Birinci gruptakilerin sayısı ise ikinci gruptakilere göre çok düşüktür. Müşteri yönetiminde ikinci gruptaki gibi müşteri kayıplarını belirlemeye çalışır ve çözümler sağlar [3], [4].

Müşterilerin bir şirketle çalışmaktan vazgeçmelerinin nedenleri ise çalışanların ilgisizliği, memnuniyetsizlik, çevre etkisi ve daha ucuz ya da kaliteli hizmet veren başka bir şirkete geçmeleri olarak sıralanabilir. Şirketler yıllık gelirlerinin %10'unu müşteri tatmini sağlayamama yüzünden kaybetmektedirler [5].

Araştırmalara göre memnun olmayan müşterilerin %4'ü şirkete şikayet nedenlerini bildirip gitmekte geri kalan %96'sı ise herhangi bir geri bildirim yapmadan ayrılmaktadır. Ayrılan müşterilerin %91'ide şirketle asla tekrar çalışmamaktadır. Hizmetten memnun kalmayan bir müşteri ise memnuniyetsizliğini ortalama 8-10 kişiye söylemektedir. Ayrılan bir müşterinin tekrar şirkete dönme maliyeti yıllık getirisinin 5 katıdır. Şirketle çalışmayı bırakan müşterilerin %68'i kendileriyle ilgilenilmedikleri ve değersiz olduklarını hissettikleri nedeniyle ayrıldıklarını belirtmişlerdir. Hizmetten memnun olmayan müşterilerin %95'i sorunlarının hemen çözülmesi durumunda şirketle çalışmaya devam edeceklerini söylemişlerdir. Ayrıca şirketle ilişkisi güçlenen müşterilerin şirkette kalma ve şirkete harcanan para miktarı her geçen yıl daha da artmaktadır. [5]

Ayrılacak müşteriyi elde tutmak için şirketler var olan yöntemleri kullanabilir ya da yeni stratejiler üretebilirler. Müşteriyi elde tutmak için;

- Müşteriyle birebir iletişime geçilir ve ayrılma nedenleri öğrenilir.
- Ayrılacağı öngörülen müşterilere yeni promosyonlar, indirimler ve hediyeler sunulur.
- Müşteriyle sürekli iletişim halinde olarak sorunları dinlenir.
- Kişiye özel jestler yapılır;
    - Teşekkür Notları,
    - Doğum Günleri, Yıl Dönümleri Kutlamaları gibi yöntemler izlenebilir.

Müşteri sadakati tahmininde eski müşterilerin verileri kullanılarak gelecek dönemde ayrılma eğiliminde olan müşteriler tahmin edilmeye çalışılır. Verimli bir ayrılma modelinin birkaç yolla şirkete yararlı etkileri vardır. Ayrılacak müşterilerin belirlenmesi ile pazarlama stratejilerinde uygun maliyetli yolların seçilmesi sağlanır. Ayrıca müşteriyi elde tutma kampanyaları, seçilen müşterilerle sınırlandırılabilir ancak müşterilerin çoğunu kapsamalıdır. Yapılacak yanlış tahminler ise, zaten ayrılmayacak müşteriye indirim yapılması nedeniyle şirkete para kaybına neden olur. Ancak, ayrılacak müşteriye ayrılmayacakmış gibi davranıp kaybetmek, ayrılmayacak müşteriyi ayrılacakmış gibi düşünmekten çok daha önemlidir. (Bunun anlamı False Negative'i düşük tutmak, false positive'i düşük tutmaktan daha önemlidir.) Müşteri sadakati tahmin modelinin kalitesi iki önemli faktöre bağlıdır. Bunlar, müşteri sadakatini elde etmek amacıyla kullanılan eldeki veriler ve veri madenciliği yöntemleridir.

Literatürde, telekomünikasyon sektöründe müşteri ayrılma tahmini başarısının artırılması için sınırlı sayıda çalışma vardır. Bu çalışmalarda yeni model üretme, var olan modeli geliştirme, var olan modelleri birleştirme ve çeşitli veri önişleme gibi adımları uygulanarak başarı artırılmaya çalışılmıştır.

Veri madenciliğinin önemli adımlarından olan veri önişleme adımında normalizasyon, öznitelik azaltma, öznitelik birleştirme, sürekli özniteliklerin ayrıklaştırılması, sınıflar arası dengesizliği gidermek amacıyla örnek üretme, kayıp değer doldurma gibi yöntemler kullanılmıştır.

Huang ve arkadaşları [6] veri kümesinden yeni öznitelikler türeterek ve öznitelikleri bazı özelliklerine göre gruplayarak doğruluğu artırmaya çalışmışlardır. Öznitelik türetme işleminde veri kümesinin büyüklüğünden yararlanılmıştır. Oluşturulan gruplar; değişmeyen sabit bilgiler (cinsiyet, doğum yeri, yaşadığı bölge, …), bağış bilgileri, müşteri hesap bilgileri (geniş bant göstergeleri, veri erişimi, …), servis istekleri, telefon hattı bilgileri, şikayet bilgileri, fatura bilgileri, ödeme bilgileri ve arama detaylarıdır. Öznitelik grupları kendi aralarında üç farklı şekilde birleştirilerek doğruluk ve performans açısından en verimli model oluşturulmaya çalışılmıştır. Sınıflandırma işleminde lojistik regresyon, lineer sınıflandırma, naif bayes, karar ağacı, çok katmanlı algılayıcılar, destek vektör makinaları ve deneysel veri işleme algortimaları kullanılmıştır. Sonuçlara göre

ROC (Receive Operating Curves) eğrisi kullanılarak yapılan değerlendirmelerde tüm özniteliklerin kullanıldığı veri kümesinin destek vektör makinalarıyla yapılan sınıflandırma işlemi en iyi sonucu vermiştir.

Tsai ve Lu'nun [7] yüz bin kişilik veri kümesinde uyguladıkları iki farklı melez yöntem doğruluk oranını önemli ölçüde artırmıştır. Bu melez yöntemlerin ilkinde yapay sinir ağları iki kez kullanılmış ikincisinde ise kendinden organize haritalar (Self – Organized Map - SOM) ile yapay sinir ağları birlikte kullanılmıştır. Kullanılan ilk hibrit yöntemin ilk algorimasında veri indirgemesi yapılmış ve bu uygulamanın sonunda ise doğru tahmin edilen örnekler asıl modeli oluşturmak için kullanılmıştır. Farklı parametre değerleri ile denenen bu hibrit model %94,32 başarıya ulaşmış ve tek başına kullanılan yapay sinir ağlarından %1,52 daha iyi sonuç vermiştir. İkinci melez modelde ise kendinden organize haritalar kümeleme algoritmasıyla kümeler oluşturulmuştur. Sonrasında ayrılan ve ayrılmayan müşteri gruplarının en yüksek oranını içeren iki küme, kümeleme sonuçları olarak seçilmiştir. Seçilen bu grupların birleştirilmesiyle oluşan veri kümesi kullanılarak yapay sinir ağları algoritması ile asıl model oluşturulmuştur. Bu modelde %93,06 başarı sağlayarak yapay sinir ağlarının tek başına kullanılmasından daha iyi performs göstermiştir ancak ilk melez model bu yönteme göre daha iyidir.

Kişioğlu ve Topçu [8] öznitelikler arası ilişkileri belirlemeye çalışmış ve tahmin modelini bayes belief network algoritmasını kullanarak inşa etmiştir. Değişkenlerin birbiriyle etkileşimini görmek amacıyla korelasyon analizi ve çoklu bağlantı (multicollinearity) testi kullanılmış ve herbir değişkenin bağımsızlığı kontrol edilmiştir. Değişkenler arası ilişkiler oluşturulurken test sonuçlarının yanı sıra uzman görüşlerinede başvurulmuştur. Daha sonra bu ilişkilere göre bayes inanç ağları kullanılarak tahmin modeli oluşturulmuştur. Ayrıca bu araştırmada "fatura miktarı eğilimi" adında farklı bir öznitelik denenmiştir. Bu öznitelik tüm müşterilerin faturalarının ortalamasıdır ve her birinin genele göre fatura miktarının artması, azalması veya sabit kalmasıdır. Algoritmanın performansını ölçmek amacıyla üç farklı senaryo kurgulanmış ve bu senaryo sonuçlarına göre müşteriyi elde tutmak amacıyla öneriler sunulmuştur. Bu önerilerin sonuçlarının işe yarayıp yaramadığı ise test edilmiş ve başarıya ulaşılmıştır. Öneriler ise;

- Ortalama dakika kullanımı artarken ortalama olarak 9-15TL ya da 15-30TL arası fatura ödeyenlere 10 dk hediye verilirse ayrılma oranı düşmektedir.
- Ortalama fatura miktarı 0-15TL arasındaki müşteriler eğer en az bir giden arama yapıyorlarsa düşük ayrılma oranına sahiptirler. Bu bölümdeki müşteriler için 5 dakika bedava verilirse ayrılma oranı daha da düşürecektir.
- Bu öneri tarife çeşidiyle alakalıdır. İki çeşit tarife vardır. Bu tarifeler birinci çeşit ve ikinci çeşit tarife olarak adlandırılmıştır. Birinci tarife, çok fazla giden arama yapan müşteriler içindir. İkinci tarifenin sabit ücreti daha azdır ancak dakika başına ücret birinciye göre çok daha maliyetlidir. Ortalama fatura miktarı 30 TL'den küçükken birinci tarifeyi kullanan müşteriler, ikinci tarifeye göre ayrılma yüzdesi daha yüksektir. Ayrılma miktarını azaltmak için birinci tarifeyi kullanan müşterilere, tarifelerini ikinci tarifeye geçirmeleri önerilebilir.

Yeshwarth ve arkadaşları [9] yeni melez model geliştirmiş ve farklı bir öznitelik denemiştir. Geliştirilen melez modelde C4.5 karar ağacı ve genetik programlama algoritmaları (GA) kullanılmıştır ve oylama esasına dayanmaktadır. İki algoritmanın sonucu farklı olduğunda hangisinin sonucuna dikkat edileceğine karar vermek amacıyla bir mekanizma inşa edilmiştir. Bu mekanizma orantısal olarak sonuç verdiği için eşik değeri isteğe göre değiştirilebilmektedir. Bu sayede risk daha iyi yönetilebilmektedir. Yeshwarth müşteri ayrılma oranını aşağıdaki gibi hesaplamıştır;

$$Müşteri\ Ayrılma\ Oranı = (OP_{C4.5} + OP_{GA})/2 \quad (1.1)$$

$$OP_{C4.5} = çıktı(C4.5) * C4.5'in\ doğruluğu \quad (1.2)$$

$$OP_{GA} = çıktı(GA) * GA'nın\ doğruluğu \quad (1.3)$$

Bu formülde OP, kişisel sınıflandırıcıların sonuçlarının sayısal ölçümlerini gösterir.

Müşteri kaybı için kullanılan yeni öznitelik, toplum etkisidir. Bir kişinin ayrılmaya etkisi, sadece kendi hayatındaki etmenler değildir. Bir kişi aynı zamanda tanıdıklarının da etkisinde kalabilir. Komşularının etkisini hesaplamak amacıyla oyun kuramı kullanılmıştır. Müşteri kaybına toplum etkisinin hesaplanması için Monte Carlo görüşü benimsenmiş ve %72,18 başarıya ulaşılmıştır.

Bu yöntemde eşik değerleriyle oynandığında doğruluk ve yanlış pozitif değerleri ayarlanabilmektedir. Bu şekilde, operatör isterse müşteri kaybetmemek için eşik değerini aşağıya çekip daha fazla muhtemel ayrılacak müşteri bulabilecektir. Ancak ayrılmayacak müşteriler ayrılacakmış gibi sınıflandırılacağı için, ayrılmayacak müşterilere de promosyon uygulanacaktır. Buda şirket için para kaybı demektir. Bu şirket için bir tercih meselesidir.

Verbeke ve arkadaşları [10] iki yeni yöntem önermişlerdir. Bu yöntemler karıncakolonisi+ ve ALBA(Active Learning Based Approach for SVM Rule Extraction)'dır. Karınca kolonisi algoritması kural çıkarımı için karınca kolonisi optimizasyonu tekniğini kullanır. Bu teknik karıncaların buldukları yiyecek ile yuvaları arasındaki en iyi yolu bulma kabiliyetleri gözlemlenerek gerçek hayattan oluşturulmuştur. Karıncalar geçtikleri yollarda feromen adı verilen özel bir koku bırakmaktadırlar. Birden fazla yolun olduğu durumlarda başlangıçta eşit olarak dağılan karıncalar zamanla obtimum yolu kullanmaktadır. Bunu kısa olan yolda feromen miktarının daha yoğun olmasından anlamaktadırlar. ALBA yöntemi destek vektör makinalarının bazı kısımlarını kullanan bir kural çıkarma algoritmasıdır. Aktif öğrenme, veri kümesinde gürültünün en yüksek olduğu alanlardaki problemlere odaklanır. Bu bölgeler iki sınıf arasındaki DVM karar sınırlarının yakınında bulunur. İlk olarak bu bölgedeki gürültüler kaldırılır. Sonrasında aktif öğrenme yaklaşımı ek veri örneklerini kapsamak için, örnekler karar sınırlarına yakın oluşturulur. KarıncaKolonisi ve ALBA yöntemleri C4.5, DVM, Lojistik regresyon gibi yöntemlerle karşılaştırılmıştır. Özgünlük (Specificity) oranına bakıldığında karınca kolonisi ve ALBA klasik yöntemlere göre çok daha iyidir. Doğruluk oranında ve duyarlılık (sensitivity) oranında ise klasik yöntemlerin oldukça gerisine düşülmüştür. Karınca kolonisin özgünlük, hassasiyet ve doğruluk başarıları sırasıyla %99,71, %37,09 ve %90,85'dir. C4.5 algoritmasının başarısı ise aynı ölçüt sıralamasıyla %98,34, %64,93 ve %93,59'dur.

Bock ve Poel'un çalışmasında [11] iki farklı yöntem gerçek hayatta 4 farklı sektörden alınan veriler üzerinde denemiştir. Kullanılan yöntemler rotasyon ormanı ve rotasyon ormanıyla adaboostun birleşmesiyle oluşan rotboostdur. Rotasyon tabanlı topluluk algoritması için 3 farklı öznitelik çıkartma algoritması denenmiştir. Bunlar temel bileşen analizi (PCA), bağımsız bileşen analizi (ICA) ve seyrek rastgele yansıma (SRP)'dır.

Rotboost algoritmasının rotasyon ormanından farkı, temel sınıflandırıcı olarak adaboost yönteminin kullanılmasıdır. Veri kümeleri ise kendin yap uygulamaları, banka, telekominikasyon ve postayla giysi siparişidir. Rotasyon ormanı ve rotboost algoritmaları yaygın olarak kullanılan 3 farklı sınıflandırcıyla karşılaştırılmıştır. Bu sınıfladırıcılar rastgele altuzaylar (RSM), rastgele orman ve karar ağaçlarıdır (C4.5 ve CART). Sonuçlara göre RSM algoritması banka (%79,45), telekomünikasyon (%69,71) ve postayla giysi siparişi (%78,85) veri kümelerinde en yüksek doğruluk oranına ulaşmıştır. Kendin yap uygulamalarında ise C4.5 (%70,01) daha iyidir.

Zhao ve arkadaşları [12] tek-sınıf destek vektör makinaları (DVM) kullanarak müşteri sadakati tahmin modeli oluşturmuşlardır. Tek-sınıf DVM aykırılık bulmak için kullanılmıştır [1]. Temel fikir, ilk olarak öznitelik uzayında çalıştırmaktır ve sadece başlangıç sınıfı olduğu varsayılır. Ancak başlangıç noktası için "yeterli yakın" bütün veri noktaları, aykırı değer ya da anormal veri noktası olarak düşünülebilir. Eğer girdi verisiyle seçilen örnekler eşleşirse, sonrasında onlar anormal veri olarak hesaba katılabilir. Geliştirilen tek sınıf DVM'de farklı çekirdek fonksiyonlar kullanılmıştır. Bunlar lineer, polinamik ve gaus fonksiyonlarıdır. Tek sınıf DVM karar ağacı, naif bayes ve yapay sinir ağlarıyla karşılaştırılmıştır. Sonuçlara göre en iyi doğruluk oranı gaus çekirdeğinin kullanıldığı tek sınıf destek vektör makinalarında %87,15 ile elde edilmiş ve en yakın algoritma olan naif bayese %3,91 fark atmıştır.

Gharbani ve arkadaşları [13] yapay sinir ağları, ağaç modelleri ve fuzzy (bulanıklık) modelleri birleştirilerek yeni bir model oluşturulmuştur. Bu melez model yerel lineer model ağacı (Locally Linear Model Tree - LOLIMOT) olarak adlandırılmıştır ve oluşumunda kullanılan yöntemlerin avantajlarını alırken dezavantajlarını da ortadan kaldırmaktadır. Bu model öğrenme prosedürünü optimum yapıya göre otomatik olarak seçer.

LOLIMOT dört iteratif adımı temel alır. Başlangıçta iki model, tüm girdi verilerini alır ve sonrasında tahmin için doğrusal en küçük kareler kullanılarak, yerel doğrusal nöronlarla başlanır. Sonra en kötü nöron bulunur. Bu yerel kayıp fonksiyonlarıyla hesaplanarak yapılır. Üçüncü adımda tüm bölümler kontrol edilir. En kötü nöron, hiper küplerin ekseni

dik kesecek şekilde ikiye bölünmesiyle geliştirilir. Tüm boyutlardaki bölümler denenir ve p bölümleri için aşağıdaki adımlar uygulanır [10].

1. Çok boyutlu üye fonksiyonları, oluşturulan hiber dikdörtgenlerin ikisi içinde inşa edilir.
2. Tüm fuzzy geçerlilik fonksiyonları oluşturulur.
3. Kural sonuç parametreleri yeni üretilmiş yerel lineer modeller için tahmin edilir. Onların merkezleri yeni bölümlerin merkezleridir ve standart sapma genellikle 0,7'dir.
4. Kayıp fonksiyonu mevcut model için hesaplanır.

Sonraki adımda, önceki adımdaki en iyi p seçilir. Eğer eğitim ve doğrulama veri kümelerinde, p kayıplarını azaltıyorsa geçerli fonksiyonlar ve nöronlar güncellenir. Nöron sayısı bir artırılır ve algoritma ikinci adıma döner. Aksi halde öğrenme algoritması sonlandırılır. Bu otomatik öğrenme algoritması en iyi doğrusal ve doğrusal olmayan modeli oluşturur. Bu model sinir ağları, lojistik regresyon ve karar ağacı algoritmalarıyla karşılaştırılmış ve sonuçlara göre LOLIMOT kesinlik oranında %82,1 başarı sağlayarak kendine en yakın olan yöntem olan karar ağacından %17,4 daha iyi sonuç vermiştir.

Gharbani ayrıca Geppart'ın savunduğu müşteri ayrılma oranının hesaplandığı aşağıdaki formülü benimsemiştir[14];

$$Aylık\ Ayrılma = \frac{C_0 + A_1 - C_1}{C_0} \tag{1.4}$$

Burada:

C_0: Ay başındaki müşteri sayısı

C_1: Ay sonundaki müşteri sayısı

A_1: Ay boyunca yeni gelen müşteri sayısı

1.2 Tezin Amacı

Müşteriyi elde tutmanın şirketlere maddi açıdan ve zaman açısından yararları vardır. Bunları kısaca dört maddede özetlemek gerekirse;

- Var olan müşteriyi tanırsınız ve bu kişilerin yaptığı hizmetleri ve satışları tahmin edebilir ve sizde ona göre üretimi satın almayı ve stoklamayı daha verimli yapabilirsiniz.
- Var olan müşteri de sizi tanımaktadır. Bu yüzden müşteriye, kendinizi ya da ürününüzü tanıtmakla vakit kaybetmezsiniz.
- Yeni müşteriler bulmak ve onları ürününüzü almaya ikna etmek, size bağlı olan müşteriyi elde tutmaktan her zaman daha maliyetlidir.
- Var olan müşterilerin size bağlılığı ve bu bağlılıkla ilgili olarak sizi başka müşterilere de tavsiye etmesi, hem yeni müşteri bulmak için hem de yeni müşteriye harcanacak maliyeti düşürmek açısından önemlidir.

Bu çalışmanın amacı, telekomünikasyon sektöründe verimli bir müşteri sadakati tahmin modeli oluşturmaktır. Telekomünikasyon sektörü Dünya'da çok önemli bir noktaya gelmiş ve uygulanan müşteri sadakati tahmin modelleri ve bu modellere göre oluşturulan stratejiler oldukça önem kazanmıştır. Türkiye'de Bilgi Teknolojileri İletişim Kurumu'nun(BTK) açıkladığı verilere göre, kasım 2008 ile mayıs 2012 arasında mobilde taşınan numara sayısı 43 milyonu geçmiştir.

Dünyada müşteri sadakati önemini rakamlarla ifade edecek olursak;

- Müşteriyi tutma oranlarındaki %1'lik artış, firma değerini ortalama olarak %5 artırmaktadır [15].
- Kablosuz ağ telekomünikasyon endüstrisinde aylık müşteri ayrılma oranı %2.2 ve yıllık %27'dir [16].
- Avrupa ve Amerika'da müşteri ayrılma maliyeti yıllık 4 milyar, Dünya'da ise 10 milyar dolardır [17].
- Aylık doğru tahmin oranının %1 geliştirilmesiyle, 1.5 milyon abone şirkette kalacağını varsayalım. Bunun da yıllık kazanç oranı 54 milyon dolar civarındadır [12].
- ABD'de 2005 yılında açıklanan tüketici raporunda; 2004 yılında mobil telefon kullanıcılarının %35'i, hizmet aldıkları firmayı değiştirmeyi düşünmektedirler [18].

- Toplam abone sayısının 580 milyon olduğu AB'de taşınan numara yüzdesi 10,3'dür [19].
- Avrupa'da numara taşımacılığının en yüksek olduğu ülke olan Finlandiya'da bu oran %68,7'dir [19].
- Türkiye'de Bilgi Teknoloji Kurumunun(BTK) verilerine göre kasım 2008 ile mart 2017 tarihleri arasında sabit ve mobil numaralarda toplam numara taşıma sayısı 104.137.096'dir.
- Türkiye'de yıllık ayrılma oranı %18'dir [19].

1.3 Hipotez

Telekomünikasyon sektöründe ayrılacak müşteri tahmini üzerine yapılan çalışmalar tez süresince incelenmiştir. Çalışmalarda kullanılan yöntemler yeni model üretme, var olan modeli geliştirme ve var olan modelleri birleştirme olarak özetlenebilir. Kullanılan ön işleme adımlarında genel olarak öznitelik indirgeme, özniteliklerin birleştirilmesi, yeni öznitelik üretme ve sınıfların dengeye getirilmesi gibi çeşitli yöntemler denenmiştir. Ayrıca bu alanda yapılan çalışmalarda veri kümesinin doğruluğu çok büyük önem arz etmektedir.

Bu tez çalışmasında telekomünikasyon sektöründe ayrılacak müşteriyi var olan yöntemlerden daha doğru şekilde tahmin etmek amacıyla araştırma yapılmıştır. Bunun için var olan veri ön işleme yöntemleriyle veri kümesindeki eksik veriler giderilecek sonrasında öznitelik indirgeme işlemleri uygulanmıştır. Çeşitli öznitelik sayılarıyla oluşturlan alt veri kümelerinde klasik yöntemler kullanılarak sınıflandırmalar yapılacak ve optimum veri kümesi elde edilmeye çalışılmıştır. Sınıflar arası dengesizliği gidermek amacıyla 2 farklı şekilde alt örnekleme işlemi yapılmış ve tahmin esnasında çoğunluk sınıfa olan kayma engellenmiştir. Aykırı değer analizi de yapılarak gürültüler giderilmeye çalışılmıştır. En sonunda ise ayrılacak müşteriyi en doğru şekilde bulmak amacıyla klasik yöntemler ve melez modeller denenmiş ve yeni melez modeller ile karşılaştırılmıştır.

Ayrıca bazı makalelerde geçen yöntemler birbirleriyle karşılaştırılarak elimizdeki veri kümesi için en uygun yöntem belirlenmiştir.

BÖLÜM 2

VERİ MADENCİLİĞİ YÖNTEMLERİ

Veri madenciliği büyük ölçekli veriler arasından bilgiye ulaşılması ya da büyük veri yığınları içerisinden gelecekle ilgili tahminde bulunabilmemizi sağlayabilecek bağıntıların aranmasıdır. Bu büyük verilerin işlenmesinde bilgisayar programlarından yararlanılır. Veri madenciliğinde ham bilginin kullanılabilmesi için bazı adımlardan geçmesi gerekebilir. Bu adımları veri temizleme, veri bütünleştirme, veri seçme, veri dönüşümü ve örüntü değerlendirme olarak sıralayabiliriz. Bu çalışmada da bazı ön işleme adımları kullanılmıştır. Kullanılan veri madenciliği teknikleri aykırı değer analizi, öznitelik seçme yöntemleri, veri alt örnekleme ve sınıflandırma teknikleridir. Bu bölümde kullanılan tekniklerden kısaca bahsedilmiştir.

2.1 Aykırı Değer Analizi Yöntemleri

Kutu Grafiği: Bu yöntem bir değişkenin sıklık dağılımını göstermek için kullanılır. Verinin dağılım şeklini, merkezi eğilimini, değişkenlerin yayılım düzeyini ve uç değerlerini göstermesi açısından kullanışlıdır. Kutu grafiğinin gösterimi için en küçük değer, birinci çeyrek (Q1), ikinci çeyrek (medyan- Q2), üçüncü çeyrek (Q3) ve en büyük değer bulunur. Kutu grafiğinin çiziminde ise;

- Kutunun uç noktaları Q1 ve Q3 işaretlenir,
- Kutunun uzunluğu Q3 – Q1'dir. Bu fark, verilerin ortadaki yarısının yayılma ölçüsüdür.
- Q2, kutunun içinde çizgi ile işaretlenir.
- Kutu dışındaki iki çizgi, alt uç değer ve üst uç değere kadar uzatılır.

Bu şekilde dağılımın merkezi, verilerin yayılma genişliği ve uç değerleri kolaylıkla görülür. Ayrıca eğer bir değer 1,5*(Q3-Q1)'den büyük ya da küçükse aykırı değer olarak kabul edilir. 1,5 kat sayısı veriye göre değiştirilebilir. Bu çalışmada 1,5 katsayısı benimsenmiştir.

Ortalama-Standart Sapma: Bu yöntemde isminden de anlaşılabileceği gibi verideki özniteliklerin değerlerinin ortalaması ve standart sapması hesaplanarak aykırı değerler bulunmaya çalışılır. Bunun için her bir özniteliğin değerleri ayrı ayrı hesaplanır. Sonrasında hesaplanan ortalama ve standart sapma değerleri kullanılarak eşik değerleri bulunur. Bu eşik değerlerinin altında ve üstünde kalan değerler aykırı değer olarak kabul edilir. Eşik değeri ise denklem (2.1)'deki formül ile hesaplanır.

$$Eşik\ Değeri = Ortalama\ -/+\ 2 * Standart\ Sapma \quad (2.1)$$

Burada standart sapma ile çarpılan 2 katsayısı veriye ve duruma göre değiştirilebilir. Bu çalışmada 2 katsayısı kabul edilmiştir.

Uzaklık Tabanlı Hesaplama: Bu yöntemde her bir örneğin diğer komşularına olan uzaklığı bulunur. Bu uzaklık denklem (2.2)'deki formül ile hesaplanır.

$$d = \sqrt{\sum_{i=1}^{n}(X_i - Y_i)^2} \quad (2.2)$$

Bu formüldeki n değeri veri kümesindeki her bir örneğin özellik sayısını göstermektedir. Bu işlem her bir örnek için yapılır ve diğerleriyle olan uzaklığı bulunmuş olur. Sonrasında ise;

- Bir uzaklık eşik değeri ve uzak komşu sayısı (p) belirlenir.
- Her bir örnek için uzaklık eşik değerinin üstünde kaç tane komşu olduğu hesaplanır.
- Belirlenen p değerinden daha fazla uzak komşuya sahip olan değerler aykırı değer olarak kabul edilir.

Bu çalışmada eşik değeri ve uzak komşu sayısı sabitleri 3 olarak kabul edilmiştir.

2.2 Öznitelik Seçme Yöntemleri

Veri kümesindeki bazı özellikler sınıf etiketini etkilemeyebilir ya da etkisi başka bir özellikle aynı olabilir. Bu özniteliklerin sonuca yararlı etkisi olmayacağı gibi performansı da ciddi ölçüde düşürürler. Bunun engellenebilmesi için gereksiz özniteliklerin bulunup çıkarılması işlem karmaşıklığını azaltacak ve tahmin modelinin oluşturulma zamanını ciddi ölçüde düşürecektir. Bu bağlamda çeşitli öznitelik çıkarma yöntemleri geliştirilmiştir. Bu tez çalışmasında kullanılan öznitelik yöntemleri entropi, bilgi kazancı, kazanç oranı, simetrik belirsizlik, ki-kare ve one-r algoritmalarıdır.

Entropi (Bilgi Yitimi): bilgiyi ölçmek için kullanılan kavramdır. Entropi bir veri kümesindeki rastgeleliği ve belirsizlik oranını ölçmeye yarar. Mantık olarak bir veri kümesindeki tüm örnekler tek bir sınıfa ait ise belirsizlik yoktur. Bu durumda entropi değeri sıfır değerine sahiptir. Eğer bir olayın gerçekleşme olasılığı p = 1 değerine sahip ise bu olayın gerçekleşmesinde bir belirsizlik olmadığı söylenir. Bu olasılık sıfır değerine yaklaştıkça belirsizlik artar.

Entropi tek başına kullanılabilse de genelde entropi kullanılarak geliştirilen yöntemler daha çok kullanılır. Bilgi kazancı, kazanç oranı ve karar ağaçları bu yöntemlere örnek olarak verilebilir.

Bilgi Kazancı (Information Gain): veri kümesinde en ayırt edici özelliği bulmak için kullanılır. Temelinde entropi metodu vardır. Özniteliklerin ayrık değerler olması gerekmektedir. Sürekli değişkenlere uygulanabilmesi için değerlerin ayrıklaştırılması gerekmektedir. Y'nin sınıf etiketi X'inde öznitelik olduğunu varsayarsak X'in Y'ye etkisini (Entropisini) eşitlik (2.3) ve (2.4) deki gibi bulabiliriz.

$$H(y) = \sum_y p(y) \log_2\big(p(y)\big) \tag{2.3}$$

$$H(Y|X) = \sum_x p(x) \sum_y p(y|x) \log_2\big(p(y|x)\big) \tag{2.4}$$

Bir eğitim veri seti için entropi bu veri setinin ilişkisinin değerlendirilmesinde önemli bir ölçüttür. X bilgisi sağlandığında Y'nin entropisinin azalma oranı dikkate alınarak Y bilgisi ile ilgili daha fazla bilgi sağlayacak bir ölçüt tanımlanabilmektedir. Bu ölçüte bilgi kazancı adı verilmektedir [20].

$$Bilgi\ Kazancı = H(Y) - H(Y|X) = H(X) - H(X|Y) \tag{2.5}$$

Bilgi kazancı simetrik bir ölçüttür (Eşitlik 2.5). X gözlemlendikten sonra Y ile ilgili kazanılan bilgi Y gözlemlendikten sonra X ile ilgili kazanılan bilgiye eşittir. Bilgi kazancının önemli bir zayıf noktası çok sayıda değere sahip özniteliklere daha fazla bilgi değerleri olmasa bile daha fazla eğilimli olmasıdır [20].

Kazanç Oranı: Bu yöntem bilgi kazancı yönteminin çok sayıda değere sahip olan özniteliklere kaymasından kaynaklanan dezavantajını gidermek amacıyla geliştirilmiştir (Eşitlik 2.6) [20].

$$Kazanç\ Oranı = \frac{Bilgi\ Kazancı}{H(X)} \tag{2.6}$$

Bu formülde gösterildiği gibi bilgi kazancının entropiye bölünmesiyle normalizasyon yapılır. Bu oran sonucunda tüm değerler 0-1 arasında elde edilir. Kazanç oranı 1 olduğunda X ile Y arasındaki ilişkinin kesin olarak tahmin edilebildiği, 0 olduğundaysa aralarında hiçbir ilişkinin olmadığı anlamına gelmektedir.

Simetrik Belirsizlik Katsayısı: bu yöntemde kazanç oranı gibi bilgi kazancı yönteminin dezavantajını gidermek amacıyla geliştirilmiştir. Simetrik belirsizlik katsayısında bilgi kazancı, özniteliğin ve sınıfın entropilerinin toplamlarına bölünmektedir. Bu bölümün 2 ile çarpım nedeni ise değerleri 0-1 arasına normalize etmektir. (Eşitlik 2.7) [20].

$$Simetrik\ Belirsizlik\ Katsayısı = 2\frac{Bilgi\ Kazancı}{H(Y)+H(X)} \tag{2.7}$$

Simetrik belirsizlik katsayısı ve kazanç oranı bilgi kazancının aksine az sayıda değere sahip özniteliklerden yana eğilimlidir [20].

Ki-Kare Testi (x^2): İki değişken arasındaki ilişkinin bağımlı veya bağımsız olduğunu tespit etmeye yarayan ve değerlerin frekans dağılımlarını temel alan bir analiz yöntemidir. Bu analiz iki temel adımla gerçekleştirilmektedir. İlk adımda özelliklerin sınıflara göre frekansları hesaplanır. İkinci adımda önceden belirlenen serbestlik derecesi ve önem seviyesine göre ilk adımdaki değerleri ki-kaynaşımı (chi-merge) yöntemi ile kullanarak veri kümesindeki alakasız öznitelikler bulunur[21]. Her bir öznitelik için bulunan bu ki-kare değerleri, o özniteliğin sınıf içindeki ilişki derecesini göstermektedir. 0 değerine sahip olan bir öznitelik sınıf değişkeniyle alakasız olduğu anlamına gelir. Ki-kare değeri arttıkça özniteliğin sınıf değişkeniyle ilişkisi o kadar artmaktadır. Ki-kare testinin formülü aşağıdaki gibidir;

$$x^2 = \sum_{i=1}^{2}\sum_{j=1}^{k}\frac{(A_{ij}-E_{ij})^2}{E_{ij}} \quad (2.8)$$

Eşitlik 2.8'deki k değeri sınıf sayısını, A_{ij} gözlenen frekans değerini ve E_{ij} ise beklenen frekans değerini göstermektedir.

One-R: Holte tarafından önerilmiş basit bir algoritmadır. Eğitim veri setindeki her veri için bir kural oluşturur ve en az hata veren kuralı seçer. Tüm numerik değerli özniteliklerin sürekli değerler olduğunu varsayar ve düz mantık bir metot ile değerleri birçok ayrık aralığa böler [20].

2.3 Sınıflandırma Yöntemleri

Naif Bayes: Matematikçi Thomas Bayes tarafından geliştirilen bu yöntem, kısaca bir olay gerçekleştiğinde başka bir ihtimalin gerçekleşme olasılığını hesaplar. Naif bayes yönteminin uygulanabilmesi için özniteliklerin birbirlerini etkilemediği kabul edilir.

Naif Bayes yönteminin formülü eşitlik (2.9)'daki gibidir.

$$P(C_j|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|C_j)P(C_j)}{p(\boldsymbol{x})} = \frac{p(\boldsymbol{x}|C_j)P(C_j)}{\sum_k p(\boldsymbol{x}|C_k)P(C_k)} \quad (2.9)$$

Bu formülde $P(x|C_j)$, C_j olayı gerçekleştiğinde x olayının olma olasılığı, $P(C_j)$, C_j sınıfının olma olasılığı, $p(\mathbf{x})$ x olayının olma olasılığı ve $P(C_j|x)$ de x olayı gerçekleştiğinde C_j'nin olma olasılığıdır.

Lojistik Regresyon: Bu yöntem adını, bağımlı değişkene uygulanan logit dönüştürmeden (logit transformation) almaktadır [22]. Bu analizin temel amacı, örneklerin hangi grubun üyesi olduğunu tahmin etmek amacıyla regresyon denklemi oluşturmaktır. Değişkenlerin birbirinden bağımsız olması şartıyla kategorik ya da numerik verilerden oluşabilirler. Ayrıca bu analizde değişkenler arasındaki ilişkinin doğrusal olma şartı bulunmamaktadır. Üstel ya da polinom ilişkisi bulunabilir. Bu nedenle lojistik regresyon doğrusal olmayan denklemler oluşturabilir. Bu analizin formülü eşitlik 2.10'daki gibidir.

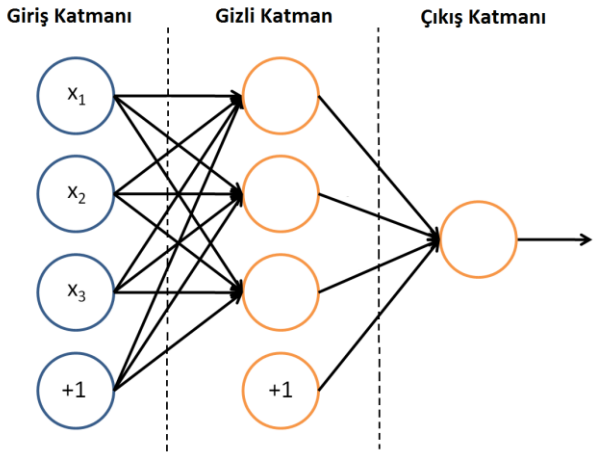
$$P = \frac{e^{\beta_0+\beta_1X_1+\cdots+\beta_kX_k}}{1+e^{\beta_0+\beta_1X_1+\cdots+\beta_kX_k}} \quad (2.10)$$

Bu formülde P: İncelenen olayın gözlenme olasılığını, β_0: Bağımsız değişkenler sıfır değerini aldığında bağımlı değişkenin değerini (başka bir ifadeyle sabiti), β_1 β_2 … β_k:

Bağımsız değişkenlerin regresyon katsayılarını, $X_1 X_2 \dots X_k$:Bağımsız değişkenleri, k: Bağımsız değişken sayısını, e: 2.71 sayısını göstermektedir.

Çok Katmanlı Algılayıcılar (ÇKA): İnsanların beyinleri örnek alınarak oluşturulan bu yöntem beyindeki sinir hücrelerinin (nöron) çeşitli şekillerde birbirlerine bağlanmasıyla oluşan sinir ağlarını (sinapsis) temel alır. Çok katmanlı algılayıcılarda insan beyni taklit edilmeye çalışılmış ve öğrenme, bilgileri kayıt etme, yorumlama gibi işlemler veri madenciliğinde kullanılmaya başlanmıştır.

ÇKA ağları giriş katmanı, gizli katman ve çıkış katmanı olmak üzere en az 3 katmandan oluşmaktadırlar. Gizli katman kendi içinde tek nöron katmanından oluşabileceği gibi birden fazla katmandan da oluşabilmektedir. Gizli katman sayısı genellikle deneme yanılma yöntemi kullanılarak belirlenir [23], [24]. Şekil 2.1'de çok katmanlı ağları gösteren bir mimari gösterilmiştir. Bu mimaride sinir hücreleri çemberlerle, sinir ağları ise oklar ile ifade edilmiştir. Sinir hücreleri kendilerine gelen verileri kendi denklemi ile işleyerek sonraki katmana sinir ağlarındaki katsayı ile çarparak gönderirler.



Şekil 2.1 Çok Katmanlı ağ mimarisi

AdaBoost (Adaptive Boosting): Adaboost algoritması boosting algoritmasının özelleştirilmiş halidir. Bu yöntem sınıflandırmada kullanıldığı gibi öznitelik seçme yöntemi olarak da uygulanabilmektedir. Çalışma prensibi ise temel sınıflandırıcıların kombinasyonundan daha güçlü bir model oluşturmaktır. Bu algoritmada temel sınıflandırıcılar birer özellikmiş gibi düşünülür ve sınıfları en doğru şekilde tahmin eden öznitelikler belirlenir. Bu özniteliklerin birleştirilmesiyle asıl model oluşturulur [25].

Torbalama (Bagging): Bu yöntem birden fazla sınıflandırma algoritmasının birlikte çalıştırılmasıyla elde edilen sonuçlara bakarak tahmin yapar. Tahmin sonucunu belirlemek için ortalama ya da oylama yöntemi kullanılır. Bu yöntemde sınıflandırıcılar için yeni alt veri kümeleri oluşturulur. Bu alt veri kümeleri orijinal veri kümesindeki örneklerden rasgele seçilerek oluşturulur. Bu nedenle bazı örnekler alt veri kümesinde hiç olmazken bazıları da bir ya da daha fazla kez aynı veri kümesinde olabilirler [26].

Rastgele Alt Uzaylar: Bu yöntem torbalama algoritması gibi birden fazla sınıflandırma yönteminin birlikte çalıştırılmasıyla tahmin üretir. Torbalama yönteminden farkı ise alt veri kümelerinin oluşturulmasında örneklerin değil özniteliklerin rastgele olarak orijinal veri kümesinden seçilmesidir. Ho tarafından yapılan çalışmalarda alt veri kümelerinin büyüklüğünün orijinal öznitelik sayısının yarısı kadar olması gerektiği belirtilmektedir [27]. Bu yöntemin başarısı sınıflandırıcıların kararlarının birbirinden farklı olmasıyla sağlanır. Bunun içinde her bir sınıflandırıcı için oluşturulan alt veri kümelerindeki özniteliklerinin mümkün olduğunca birbirinden farklı olmasıyla sağlanır [28].

Rotasyon Ormanı(RotO): Bu model son yıllarda kullanılmaya başlanan sınıflandırma performansını önemli ölçüde artıran yeni nesil sınıflandırma algoritmasıdır. Diğer kolektif öğrenme yöntemleri gibi birden fazla sınıflandırıcı kullanılır. Bu algoritmanın çalışma zamanı diğer kolektif öğrenme yöntemlerine göre daha uzundur. Bunun nedeni alt veri kümelerinin torbalama yöntemindeki gibi oluşturulmasından sonra her bir alt veri kümesi için öznitelik çıkarma yöntemi olan ana bileşen analizinin kullanılmasıdır. Ana bileşen analiziyle dönüştürülen alt veri kümelerinde ayırt ediciliği en yüksek olan öznitelikler seçilerek veri kümeleri son haline getirilir [29].

Karar Tablosu: Bu yöntemin çalışma prensibi farklı öznitelikleri rasgele seçerek yeni alt veri kümeleri oluşturmaktır. Oluşturulan bu alt veri kümeleri çapraz doğrulama yöntemi ile performansları ölçülür ve en iyi öznitelikleri içeren küme belirlenir. Daha sonra bu veri kümesi kullanılarak sınıflandırıcı model oluşturulur. Karar tablolarının incelenmesi, sistemin mantığını adım adım akışının takibi kolaydır. Bu avantaj sayesinde çapraz doğrulama sonuçlarına göre tablodaki ölçütler değiştirilebilir ve etkileri gözlemlenebilir.

Karar tabloları öznitelik seçiminde sarma yaklaşımı kullanmaktadır. Ayrıca test esnasında kullanılan örnek etiketlenemiyorsa çoğunluk olan sınıfa ya da kendisine en yakın örneğin etiketiyle sınıflandırılır [30].

OneR Algoritması: Veri kümesindeki tüm özniteliklerden sadece bir tanesini seçerek karar ağacı oluşturur. Bu nedenle algoritması basit ve çalışma zamanı azdır. Öznitelik seçme aşamasında veri kümesindeki tüm öznitelikler ayrı ayrı kullanılarak modeller oluşturulur ve bu modellerin performansları eğitim kümesindeki tüm örnekler kullanılarak ölçülür. Çıkan sonuçlara göre en iyi sonucu veren modelin kullanıldığı öznitelik seçilir [30].

Rastgele Orman (RO): Torbalama yönteminden geliştirilmiş olup düğüm noktalarını rasgele olarak seçmektedir. Torbalama algoritması tüm değişkenler arasından en iyi dalı kullanarak her bir düğümü dallara ayırır. Rastgele orman ise her bir düğümde rastgele olarak seçilen değişkenler arasından en iyisini kullanarak her bir düğümü dallara ayırır [31].

CART (Classification and Regression Tree): Bu yöntem sınıflandırma ağacının ve regresyon ağaçlarının birleşmesidir [32]. Algoritma özniteliğin sürekli ya da kategorik veri olmasına göre hangi yöntemi kullanacağını seçer. Eğer öznitelik sürekli verilerden oluşuyorsa regresyon ağacı, kategorik verilerden oluşuyorsa sınıflandırma ağacını kullanır. CART yöntemi aynı zamanda veri kümesindeki özniteliklerin birbiriyle olan ilişkilerini diyagram olarak gösterebilmektedir. Bu özellikle araştırmacılar için çok kullanışlıdır [33].

J48 Karar Ağacı: C4.5 karar ağacı algoritmasının geliştirilmesiyle ortaya çıkmıştır. Her bir öznitelik için sadece özniteliğin aldığı değerleri ve sınıf etiketlerini içeren yeni tablolar oluşturulur ve her özniteliğin entropisi bulunur. En yüksek entropi değerine sahip olan öznitelik düğüm olarak alınır. Ağaçtaki ilk düğüme kök düğüm adı verilir. Kök düğüm özniteliğin alabileceği değerlere göre alt düğümleri ya da sınıf etiketini gösterir. Seçilen özniteliğin alt düğüme geçme şartını sağlayan örneklerin kapsandığı yeni veri kümesi oluşturulur. Yeni oluşan bu veri kümesinde tekrar entropi hesabı yapılır ve en iyi çıkan öznitelik düğüm olarak kullanılır. Bu tüm örnekler sınıflanana kadar devam eder. Ancak

bu yöntem her verinin kendine özgü değeri olduğu gibi durumlarda aşırı kural oluşturma yapabilir.

2.4 Melez Yöntemler

Bu çalışmada 3 farklı melez metot denenmiştir. Bu metotlardan bir tanesi Tsai'nin kullandığı 2 klasik yöntemin birleştirilmesiyle oluşan yöntemdir. Bunun yanında iki farklı melez yöntem daha denenmiştir. Bu yöntemler her iki veri kümesi içinde klasik yöntemlerin kombinasyonları ile denenmiş ve en verimli model bulunmaya çalışılmıştır. Melez modellerin hepsinde Şekil 2.2'de gösterildiği gibi birinci adımda 10 kat çapraz doğrulama kullanılmıştır. Kullanılan melez modeller ayrıntılı olarak aşağıda anlatılmıştır.

2.4.1 Birinci Melez Model (M1)

Bu model Tsai'nin kullandığı yöntemdir. Bu modelde veri kümesi ilk olarak 10 parçaya bölünür. Bu parçaların birisi test diğer dokuz tanesi eğitim kümesi olarak ayrılır. Bu dokuz parçalık veri kümesi hem eğitim hem de test kümesi olarak kendi içinde test edilir. Bu adım 10 kez tekrarlanır. Her tekrarlanışta yanlış sınıflanan örnekler atılır. Bu 10 iterasyon sonunda doğru kalan örnekler ile asıl model oluşturulur ve ilk ayrılan test kümesiyle test edilir.

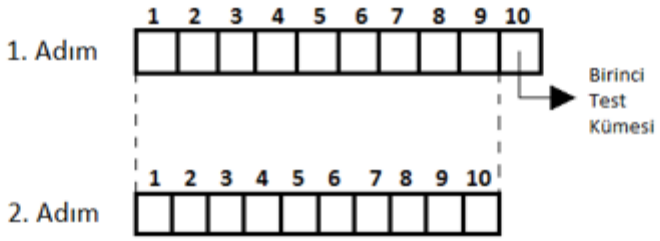
2.4.2 İkinci Melez Model (M2)

Bu modelde veri kümesi ilk olarak 10 parçaya bölünür. Bu parçaların birisi test diğer dokuz tanesi eğitim kümesi olarak ayrılır. Ayrılan dokuz parçada kendi içinde 10 alt kümeye bölünür. Bu alt kümelerden her bir tanesi sırayla alt test kümesi olarak, geri kalanda alt eğitim kümesi olarak kullanılır. Sonrasında kullanılan sınıflama yöntemiyle en iyi sonucu veren alt küme seçilir ve asıl model bu alt veri kümesiyle eğitilerek ilk ayrılan test kümesiyle test edilir ve performansı ölçülür.

2.4.3 Üçüncü Melez Model (M3)

Bu melez modelde de ilk olarak veri kümesi 10 parçaya ayrılır ve 1 parçası test 9 parçası eğitim olarak alınır. Bu dokuz parça kendi içinde tekrar 10 parçaya ayrılır ve 1 parçası alt test ve 9 parçası alt eğitim kümesi olarak kullanılır. Alt eğitim kümesiyle eğitilen modelde

test edilen her alt test kümesinin doğru sınıflanan örnekleri alınır ve yeni bir eğitim kümesi oluşturulur. Bu yeni eğitim kümesi kullanılarak asıl model oluşturulur ve ilk ayrılan test kümesiyle test edilir.



Birinci melez model (M1): 2. adımda tüm veri kümesi kendi içinde eğitilir ve test edilir. Yanlış sınıflanan örnekler atıldıktan kalan örnekler ile tekrar model oluşturulur. Bu işlem sonunda birinci test kümesiyle test edilir.
İkinci melez model (M2): 2. adımda 10 kat çapraz doğrulama yapıldıktan sonra en iyi sonucu veren alt küme alınarak asıl model oluşturulur ve birinci test kümesiyle test edilir.
Üçüncü melez model (M3): 2. adımda 10 kat çapraz doğrulama işlemi uygulanır ve her katmanda doğru sınıflanan örnekler alınarak yeni veri kümesi oluşturulur ve bu veri kümesi ile asıl model oluşturulup birinci test kümesi ile test edilir.

Şekil 2.2 Melez Yöntemler

BÖLÜM 3

VERİ KÜMELERİ

3. Veri Kümeleri

Bu araştırmada iki farklı veri kümesi kullanılmıştır. Bunlar: Larose (2005) ve Tsai (2009) veri kümeleridir. Veri kümesi için yapılan tüm işlemlerde WEKA ve kendi yazdığımız programlar kullanılmıştır.

3.1 Tsai Veri Kümesi

Tsai'nin veri kümesi, Amerikan Telekom Şirketinden Temmuz 2001 ile Ocak 2002 arasında (6 ay) alınmıştır. Veri kümesinde, 51.306 abonenin bilgisi vardır. Bu abonelerin 34.761'i şirketten ayrılmış, 16.545'ide şirketten hala hizmet almaktadır. Veri kümesinde müşterinin kayıp bilgisi dahil 172 öznitelik vardır. Veri kümesinde sınıflar arası dengesizlik olduğu için veriler Tsai'nin kullandığı yöntemlerle örnek artırımı işlemi uygulanmış ve sonucunda 49.652 ayrılan müşteri ve 50.438 ayrılmayan müşteri olmak üzere toplam 100.000 örnek olmuştur. Ayrıca %50'den fazla kayıp veriye sahip olan öznitelikler silinmiş ve toplam 148 öznitelik kalmıştır. Bu özniteliklerin ise 111 tanesi numerik değer, 37 tanesi ise kategorik değerdir. Veri kümesinde kayıp veri vardır. Bu veri kümesinin öznitelik kısaltmaları ve açıklamaları Ek-A'da mevcuttur.

3.1.1 Tsai Veri Kümesinde Kayıp Değer Doldurulması

Tsai'nin veri kümesi için, ilk olarak kayıp değerler doldurulmuştur. Kayıp değerler iki farklı yöntemle tamamlanmıştır. Birinci yöntemde sürekli değerler tamamlanırken her bir özniteliğin ortalaması alınmış ve sonrasında eksik değerler yerine özniteliğin ortalaması atanmıştır. Kategorik değerler için ise, her bir öznitelik için, en çok hangi değerin geçtiği hesaplanmış ve kayıp veriler için bu değer atanmıştır. Bu yöntemden ileride KD1 olarak bahsedilecektir. İkinci yöntemde sınıflara göre eksik değerler tamamlanmıştır. Yani sürekli değere sahip her öznitelik için ayrılan ve ayrılmayan müşteri sınıflarına göre örneklerin ortalamaları bulunmuş ve eksik değerler bu ortalamalar ile doldurulуştur. Kategorik öznitelikler için yine aynı yöntemle değeri eksik olan abonenin sınıfına göre, aynı sınıftan olan abonelerin en çok hangi değeri varsa o değerle doldurulmuştur. Bu yöntemden de ileride KD2 olarak bahsedilecektir.

3.1.2 Tsai Veri Kümesi Öznitelik Seçme İşlemleri

Öznitelik seçimi için WEKA'dan simetrik belirsizlik, oneR, bilgi kazanımı, kazanç oranı, ki-kare ve filtrelenmiş öznitelik metotları kullanılmıştır. Bu yöntemlerin sonuçlarını ortak bir şekilde değerlendirmek için başarı puanları toplanmıştır. Ancak bazı yöntemlerin başarı puanı -1 ile 1 arasında, bazıları -5 ile 5 arasında olabilmektedir. Bunun anlamı başarı puanlarının toplamı tam doğru bir sıralama vermeyecektir. Bu sorunu çözmek için tüm öznitelik seçme yöntemlerinin başarı puanları aralığını aynı aralığa normalize edilmiştir. Öznitelik seçme işleminde belirlenen ilk 10 öznitelik Çizelge 3.1'deki gibidir.

Bu veri kümesi için en önemli özniteliğin telefonun kullanım süresi olduğu görülmektedir. İkinci sırada hizmet alınan süre ve üçüncü sırada kullanılan cihazın fiyatı gelmektedir.

Bu veri kümesi için farklı sayıda öznitelikler ile alt veri kümeleri oluşturularak sınıflandırmalar yapılmış ve sonuçlar üzerindeki etkileri gözlemlenmiştir. Oluşturulan alt kümelerin öznitelik sayıları ise en seçici 36, 57, 77, 102, 126 ve 148 (tüm öznitelikler)'dir.

Çizelge 3.1 Tsai veri kümesi en iyi 10 özniteliği

Sıra No	Öznitelik	Açıklaması	Birimi
1	Eqpdays	Mevcut cihazın kullanıldığı süre	Ay
2	Months	Toplam hizmet alınan ay süresi	Ay
3	Hnd_price	Kullanılan cihazın fiyatı	$
4	Hnd_webcap	Cihazın internet destekleyebilmesi	Evet/Hayır
5	Totmrc_mean	Aylık toplam kontör yükleme ortalaması	$
6	Csa	Bölgesel iletişim miktarı	Dakika
7	Totmrc_range	Toplam aylık kontör yükleme sayısı	
8	Change_mou	Önceki üç aylık ortalama ile aylık kullanım dakikalarının yüzde değişimi	
9	Mou_mean	Aylık ortalama dakika kullanımı	Dakika
10	Asl_flag	Hesap harcama limiti	$

3.1.3 Tsai Veri Kümesi Aykırı Değer Tespiti

Veri kümesinde aykırı değerleri bulmak için 2 farklı yöntem denenmiştir. Bu yöntemler tek boyutlu veri analizi yöntemleri olan kutu grafiği (box-plot) ve ortalama standart sapma aykırı değer hesaplama yöntemidir. Bu yöntemlerin sonuçları her bir özniteliğin sahip olduğu aykırı değer sayısı olarak ve belirli bir sayıdan fazla özniteliğinde aykırı değeri olan örnek sayıları olarak iki farklı şekilde gösterilmiştir.

3.1.3.1 Kutu Grafiği Analizi

Tsai veri kümesi için bu yöntemle yapılan aykırı değer analizi sonuçları Çizelge 3.2 ve Çizelge 3.3'deki gibidir.

Çizelge 3.2 Tsai veri kümesinde kutu grafiği yöntemiyle bulunan özniteliklerine göre aykırı abone sayıları

Bir Abonenin #'dan Fazla Aykırı Değer Öznitelik Sayısı	**Abone Sayısı**
>0	74.416
>1	64.208
>5	40.205
>15	21.187
>25	11.017
>35	6.026
>45	3.240
>55	1.623
>65	709
>75	251
>85	57
>95	4
>98	1
>99	0

Çizelge 3.2'de belirli sayıdaki aykırı öznitelik değerine ve fazlasına kaç abonenin sahip olduğu gösterilmektedir. Yani örnek verecek olursak bir ve daha fazla özniteliğinde aykırı değeri olan abone sayısı 74.416 kişidir. Yirmi altı ve daha fazla sayıda özniteliğinde aykırı değeri olan abone sayısı 11.017'dir.

Çizelge 3.3 Tsai veri kümesinde kutu grafiği yöntemi ile bulunan özniteliklerin aykırı değer sayıları

	Öznitelik Adı	Aykırı Örnek Sayısı		Öznitelik Adı	Aykırı Örnek Sayısı		Öznitelik Adı	Aykırı Örnek Sayısı
1	change_rev	26321	38	ovrrev_Range	9294	75	mou_peav_Range	6504
2	totmrc_Range	22558	39	peak_dat_Mean	8942	76	avg3rev	6160
3	roam_Mean	18736	40	mou_pead_Mean	8940	77	rev_Mean	6030
4	roam_Range	18401	41	mou_pead_Range	8935	78	totmou	5987
5	plcd_dat_Mean	14980	42	mouowylisv_Range	8884	79	adjmou	5979
6	plcd_dat_Range	14750	43	mou_opkv_Range	8821	80	mou_peav_Mean	5869
7	cc_mou_Range	14399	44	peak_dat_Range	8806	81	mou_cvce_Mean	5821
8	callwait_Mean	14305	45	inonemin_Range	8706	82	avg6rev	5776
9	cc_mou_Mean	14265	46	threeway_Mean	8660	83	totrev	5684
10	datovr_Mean	14030	47	callwait_Range	8500	84	adjrev	5646
11	datovr_Range	14015	48	mouowylisv_Mean	8384	85	avg6qty	5606
12	change_mou	13768	49	recv_vce_Range	8357	86	peak_vce_Mean	5471
13	ccrndmou_Mean	13456	50	mou_opkv_Mean	8305	87	avg3qty	5385
14	ccrndmou_Range	13435	51	opk_vce_Range	8212	88	avgrev	5308
15	comp_dat_Mean	13393	52	inonemin_Mean	8151	89	avgqty	5252
16	mou_cdat_Mean	13393	53	rev_Range	8065	90	avg6mou	5224
17	mou_cdat_Range	13378	54	mou_rvce_Range	7929	91	mou_Mean	5201
18	comp_dat_Range	13189	55	drop_blk_Range	7805	92	avg3mou	5200
19	custcare_Mean	12710	56	unan_vce_Range	7755	93	attempt_Mean	5081
20	da_Mean	11994	57	drop_blk_Mean	7613	94	plcd_vce_Mean	5062
21	custcare_Range	11621	58	plcd_vce_Range	7214	95	complete_Mean	5002
22	vceovr_Mean	11601	59	mou_rvce_Mean	7202	96	comp_vce_Mean	4975
23	ovrmou_Mean	11527	60	complete_Range	7192	97	avgmou	4714
24	ovrrev_Mean	11386	61	drop_vce_Mean	7143	98	threeway_Range	4522
25	mouiwylisv_Mean	11316	62	opk_vce_Mean	7142	99	unan_dat_Mean	3120
26	blck_vce_Mean	10701	63	attempt_Range	7140	100	unan_dat_Range	3058
27	blck_vce_Range	10623	64	totcalls	7114	101	drop_dat_Mean	2600
28	da_Range	10373	65	adjqty	7114	102	eqpdays	2585
29	iwylis_vce_Mean	10363	66	mou_cvce_Range	7068	103	drop_dat_Range	2561
30	mouiwylisv_Range	10100	67	comp_vce_Range	7067	104	months	2294
31	mou_opkd_Mean	9618	68	recv_vce_Mean	6997	105	totmrc_Mean	1800
32	mou_opkd_Range	9612	69	unan_vce_Mean	6857	106	blck_dat_Mean	1228
33	opk_dat_Mean	9606	70	owylis_vce_Range	6853	107	blck_dat_Range	1223
34	ovrmou_Range	9532	71	peak_vce_Range	6815	108	recv_sms_Mean	872
35	iwylis_vce_Range	9529	72	mou_Range	6700	109	recv_sms_Range	860
36	opk_dat_Range	9467	73	owylis_vce_Mean	6660	110	callfwdv_Mean	433
37	vceovr_Range	9410	74	drop_vce_Range	6509	111	callfwdv_Range	431

Çizelge 3.3'de kutu grafiği analizinin tsai veri kümesi üzerinde kullanılmasıyla bulunan her bir öznitelik için kaç abone değerinin aykırı olduğu gösterilmektedir. Sonuçlara göre en çok aykırı değeri olan öznitelik change_rev (Önceki üç aylık ortalamayla aylık gelirin yüzde değişimi)'dir.

3.1.3.2 Ortalama – Standart Sapma Aykırı Değer Analizi

Tsai veri kümesinde ortalama-standart sapma aykırı değer analizinin kullanılmasıyla elde edilen sonuçlar Çizelge 3.4 ve Çizelge 3.5'deki gibidir.

Çizelge 3.4 Tsai veri kümesinde ortalama standart sapma yöntemiyle bulunan özniteliklerine göre aykırı abone sayıları

Bir Abonenin #'dan Fazla Aykırı Değer Öznitelik Sayısı	**Abone sayısı**
>0	43121
>5	16297
>15	6865
>25	3318
>35	1523
>45	630
>55	205
>65	46
>75	5
>85	1
>98	0

Çizelge 3.4'de kutu grafiği yönteminde olduğu gibi belirli sayıdaki aykırı öznitelik değerine ve fazlasına kaç abonenin sahip olduğu gösterilmektedir. Bu çizelgeye göre bir ve daha fazla özniteliğinde aykırı değeri olan abone sayısı 43.121'dir.

Çizelge 3.5 Tsai veri kümesinde ortalama standart sapma yöntemi ile bulunan özniteliklerin aykırı değer sayıları

	Öznitelik Adı	Aykırı Örnek Sayısı		Öznitelik Adı	Aykırı Örnek Sayısı		Öznitelik Adı	Aykırı Örnek Sayısı
1	totmrc_Range	5181	38	complete_Range	3997	75	inonemin_Range	2896
2	mou_opkv_Mean	5162	39	plcd_vce_Range	3951	76	rev_Range	2824
3	avg6mou	4882	40	avg3rev	3948	77	blck_vce_Mean	2785
4	avgmou	4880	41	attempt_Range	3945	78	callwait_Mean	2742
5	avg3mou	4863	42	owylis_vce_Range	3934	79	callwait_Range	2680
6	mou_Mean	4850	43	totrev	3930	80	change_rev	2674
7	mou_rvce_Mean	4840	44	adjrev	3906	81	threeway_Range	2634
8	mou_cvce_Mean	4798	45	iwylis_vce_Mean	3900	82	blck_vce_Range	2493
9	change_mou	4732	46	vceovr_Mean	3899	83	threeway_Mean	2129
10	eqpdays	4575	47	mouowylisv_Mean	3893	84	peak_dat_Range	980
11	mou_peav_Mean	4538	48	ovrrev_Mean	3867	85	plcd_dat_Range	954
12	attempt_Mean	4455	49	totmrc_Mean	3765	86	plcd_dat_Mean	944
13	opk_vce_Mean	4453	50	recv_vce_Mean	3753	87	comp_dat_Mean	929
14	complete_Mean	4440	51	opk_vce_Range	3749	88	mou_pead_Range	913
15	plcd_vce_Mean	4437	52	rev_Mean	3730	89	comp_dat_Range	911
16	owylis_vce_Mean	4411	53	totmou	3723	90	mou_pead_Mean	895
17	comp_vce_Mean	4397	54	adjmou	3714	91	datovr_Range	880
18	mou_opkv_Range	4345	55	mouiwylisv_Mean	3712	92	opk_dat_Mean	868
19	avgqty	4313	56	drop_blk_Mean	3648	93	peak_dat_Mean	851
20	avg6qty	4309	57	ovrmou_Mean	3602	94	opk_dat_Range	848
21	avg3qty	4278	58	drop_vce_Range	3602	95	mou_cdat_Range	809
22	da_Range	4253	59	cc_mou_Range	3577	96	datovr_Mean	798
23	mou_peav_Range	4241	60	iwylis_vce_Range	3552	97	mou_cdat_Mean	787
24	mou_cvce_Range	4203	61	ccrndmou_Range	3541	98	drop_dat_Range	702
25	months	4195	62	mouowylisv_Range	3509	99	mou_opkd_Range	701
26	peak_vce_Mean	4186	63	recv_vce_Range	3414	100	roam_Mean	661
27	mou_rvce_Range	4132	64	ccrndmou_Mean	3409	101	mou_opkd_Mean	598
28	unan_vce_Mean	4128	65	mouiwylisv_Range	3383	102	drop_dat_Mean	481
29	drop_vce_Mean	4113	66	cc_mou_Mean	3352	103	unan_dat_Mean	477
30	mou_Range	4110	67	adjqty	3333	104	unan_dat_Range	429
31	avgrev	4087	68	totcalls	3322	105	roam_Range	411
32	vceovr_Range	4079	69	da_Mean	3296	106	recv_sms_Range	296
33	ovrmou_Range	4061	70	custcare_Range	3202	107	callfwdv_Range	216
34	peak_vce_Range	4054	71	unan_vce_Range	3155	108	recv_sms_Mean	199
35	comp_vce_Range	4040	72	inonemin_Mean	3144	109	blck_dat_Range	161
36	ovrrev_Range	4022	73	drop_blk_Range	3032	110	callfwdv_Mean	125
37	avg6rev	4007	74	custcare_Mean	2946	111	blck_dat_Mean	114

Çizelge 3.5'de her bir öznitelik için kaç abone değerinin aykırı olduğu gösterilmektedir. Ortalama-standart sapma yöntemine göre en çok totmrc_range (aylık tekrar kontör yükleme aralığı) özniteliğinin aykırı değeri vardır.

Aykırı değer analizinde daha yaygın kullanıldığı için kutu grafiği yöntemi benimsenmiştir. Bunun için Çizelge 3.2'ye göre müşterilerin aykırı nitelik sayısına bakılarak sınıflandırmalar yapılmıştır. Bu işlemin sonuçları 5.1.1. kısımda anlatılmaktadır.

Bu veri kümesinde aykırı değer analizi iki farklı şekilde yapılmıştır. Birincisinde toplam örnek sayısının %75'ini kapsayan 1 ve üstü aykırı özniteliği olan örnekler (AD1 Kümesi), ikincisinde toplam örnek sayısının %10'unu kapsayan 26 ve üstü aykırı özniteliği olan örnekler (AD2 Kümesi) silinerek çalışılmıştır.

3.2 Larose Veri Kümesi

Larose'un veri kümesi, Kablosuz Ağ Telekom Şirketinden alınmıştır ve 5.000 abonenin bilgileri vardır. Her bir abonenin ayrılma bilgisi dahil 20 tane özniteliği vardır ve eksik veri yoktur. Geri kalanların ise 16 tanesi numerik veri 3 tanesi de kategorik veridir. Müşterilerin %14,3'ü (707 abone) sonraki 3 ayda şirketten ayrılmıştır. Veri kümesinde kayıp veri yoktur. Veri kümesi ve açıklamaları Ek B'de verilmiştir.

3.2.1 Larose Veri Kümesi Öznitelik Seçme İşlemi

Larose veri kümesinde de 10 farklı öznitelik seçme metodu denenmiştir. Bunlar; anlamsal belirsizlik, ki-kare, filtrelenmiş öznitelik, bilgi kazancı, filtrelenmiş öznitelik, kazanç oranı ve OneR, yöntemleridir. Bu veri kümesinde de Tsai veri kümesinde uygulanan başarı puanı metodu benimsenmiştir.

3.1.2 de anlatıldığı gibi öznitelikler başarı puanlarına göre sıralandığında ise en önemli öznitelik birinci sırada olacak şekilde aşağıdaki gibi sıralanmıştır (Çizelge 3.6).

Çizelge 3.6 Larose veri kümesi en iyi 10 özniteliği

Sıra	Öznitelik Adı	Açıklama	Birim
1	international_plan	Uluslararası arama kullanımı	Evet/Hayır
2	total_day_minutes	Günlük toplam konuşma süresi	Dakika
3	number_customer_service_calls	Müşteri hizmetlerini arama sayısı	
4	voice_mail_plan	Sesli posta kullanımı	Evet/Hayır
5	total_eve_minutes	Geceleri konuşma süresi	Dakika
6	state	Yaşadığı yer	
7	total_day_charge	Günlük toplam harcanan kontör	
8	number_vmail_messages	Sesli mesajlaşma sayısı	
9	total_intl_calls	Toplam uluslararası arama sayısı	
10	total_intl_charge	Uluslararası aramalarda harcanan toplam kontör sayısı	

Öznitelik seçme işlemi sonuçlarına göre en iyi özellik uluslararası arama yapıp yapmadığıdır. İkinci ve üçüncü sırada ise günlük konuşma süresi ve müşteri hizmetleri arama sayısı gelmektedir.

3.2.2 Larose Veri Kümesi Aykırı Değer Tesbiti

Larose veri kümesinde 3 farklı aykırı değer analizi yapılmıştır. Bunlar; tek boyutlu veri analiz yöntemi olan kutu grafiği, uzaklık tabanlı aykırı değer hesaplama ve çok boyutlu veri analizi yöntemi olan ortalama – standart sapma analizleridir. Tek boyutlu veri analizi yöntemlerinin sonuçları her bir özniteliğin sahip olduğu aykırı değer sayısı olarak ve belirli bir sayıdan fazla özniteliğinde aykırı değeri olan örnek sayıları olarak iki farklı şekilde gösterilmiştir. Çok boyutlu veri analizi yönteminde ise uzak komşu sayısı belirli bir sayıdan fazla olan abone sayıları gösterilmiştir. Bu veri kümesinde aykırı değer çok fazla olmadığı için aykırı değerler çıkarılarak analiz yapılmamıştır.

3.2.2.1 Kutu Grafiği Analizi

Larose veri kümesine uygulanan kutu grafiği analizinin sonuçları aşağıdaki gibidir. Çizelge 3.7'de her bir öznitelik için kaç tane abonenin aykırı değer olduğu hesaplanmıştır.

Çizelge 3.7 Larose veri kümesinde kutu grafiği yöntemiyle bulunan özniteliklerine göre aykırı abone sayıları

Bir Abonenin #'dan Fazla Aykırı Değer Öznitelik Sayısı	**Abone Sayısı**
>0	1889
>1	360
>2	67
>3	9
>4	1
>5	0

Çizelge 3.7'ye bakıldığında bir ve daha fazla özniteliğinde aykırı değeri olan kişi sayısı 1889'dur. Beş ve daha fazla özniteliğinde aykırı değeri olan hiç kimse yoktur.

Çizelge 3.8 Larose veri kümesinde kutu grafiği yöntemi ile bulunan özniteliklerin aykırı değer sayıları

Öznitelik Adı	**Aykırı Değer Sayısı**
account_length	24
area_code	1246
number_vmail_messages	60
total_day_minutes	34
total_day_calls	35
total_day_charge	34
total_eve_minutes	42
total_eve_calls	27
total_eve_charge	42
total_night_minutes	39
total_night_calls	43
total_night_charge	39
total_intl_minutes	72
total_intl_calls	118
total_intl_charge	72
number_customer_service_calls	399

Çizelge 3.8'de belirli sayıdaki aykırı öznitelik değerine ve fazlasına kaç abonenin sahip olduğu gösterilmektedir.

3.2.2.2 Ortalama – Standart Sapma Aykırı Değer Analizi

Çizelge 3.9 ve Çizelge 3.10 Ortalama – Standart Sapma Aykırı Değer Analizini göstermektedir. Çizelge 3.9'da her bir öznitelik için kaç tane abonenin aykırı değer olduğu hesaplanmıştır.

Çizelge 3.9 Larose veri kümesinde ortalama standart sapma yöntemiyle bulunan özniteliklerine göre aykırı abone sayıları

Bir Abonenin #'dan Fazla Aykırı Değer Öznitelik Sayısı	**Abone Sayısı**
>0	2067
>1	1017
>2	291
>3	84
>4	19
>5	2
>6	0

Bu analize göre bir ve daha fazla özniteliğinde aykırı değeri olan kişi sayısı 2067'dir (Çizelge 3.9). Altı ve daha fazla özniteliğinde aykırı değeri olan abone ise hiç yoktur.

Çizelge 3.10 Larose veri kümesinde ortalama standart sapma yöntemi ile bulunan özniteliklerin aykırı değer sayıları

Öznitelik Adı	**Aykırı Değer Sayısı**
account_length	232
area_code	0
number_vmail_messages	341
total_day_minutes	236
total_day_calls	249
total_day_charge	236
total_eve_minutes	236
total_eve_calls	242
total_eve_charge	236
total_night_minutes	223
total_night_calls	229
total_night_charge	223
total_intl_minutes	228
total_intl_calls	194
total_intl_charge	228
number_customer_service_calls	147

Çizelge 3.10'da belirli sayıdaki aykırı öznitelik değerine ve fazlasına kaç abonenin sahip olduğu gösterilmektedir.

3.2.2.3 Uzaklık Tabanlı Aykırı Değer Analizi

Uzaklık tabanlı Aykırı değer analizinin sonuçları Çizelge 3.11'de görülmektedir. Burada Uzak komşu sayısı belirli bir p miktarından fazla olan abone sayıları görülmektedir. Yani uzak komşusu 100 den fazla olan 911 kişi vardır.

Çizelge 3.11 Larose veri kümesinde uzaklık tabanlı aykırı değer analizi sonuçları

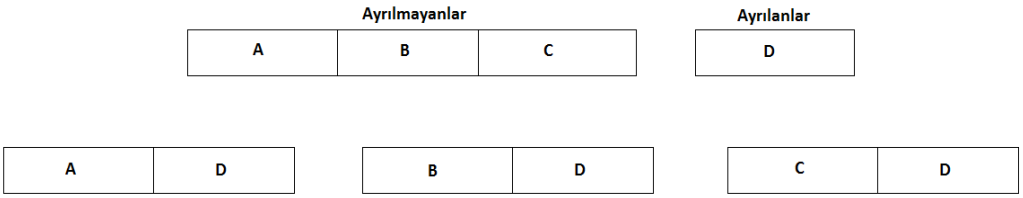
Uzak Komşu Sayısı # dan fazla Abonelerin Sayısı	Aykırı Değer Abone Sayısı
p>100	911
p>200	402
p>300	228
p>400	150
p>450	131
p>500	101
p>600	70
p>700	50
p>800	34
p>900	31
p>1000	25
p>1100	19
p>1400	8
p>1600	8
p>2000	6
p>2100	5
p>2200	3
p>2400	2
p>2450	1
p>2500	0

Tüm aykırı değer analizlerine bakıldığında bu veri kümesinde ciddi bir aykırı değer olmadığı gözlemlenmiştir. Veri kümesinin az olmasından dolayı da aykırı değerlerin atılmamasına karar verilmiştir.

3.2.3 Larose Veri Kümesi Veri Alt Örnekleme

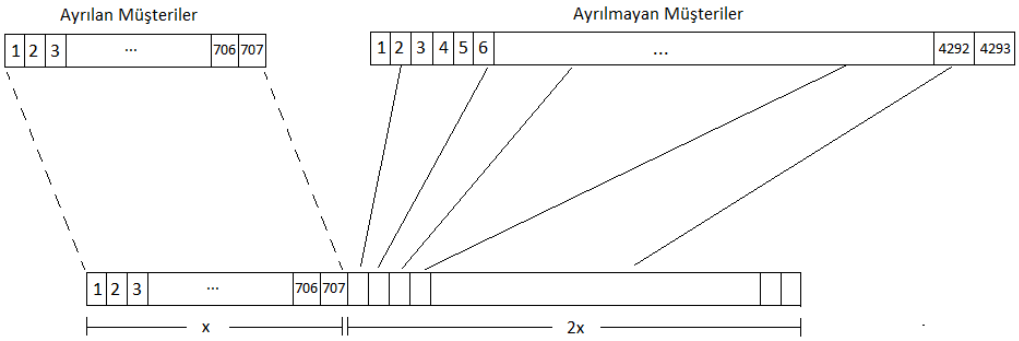
Larose veri kümesinde müşterilerin %14,3'ü ayrılmış, %86,7 sadık müşteri olduğundan çalışmanın başında veri alt örnekleme işlemi (down sampling) yapılmıştır. Bu alt

örnekleme işleminde x adet ayrılan müşteri varken 2x adet ayrılmayan müşteri olacak şekilde alt kümeler oluşturulmuştur. Bu alt kümeler iki farklı şekilde oluşturulmuştur. Birincisi, üç alt küme oluşturulmuş ve tüm ayrılan müşteriler bu üç alt kümenin hepsine de konmuştur. Ayrılmayan müşteriler ise bir müşteri bir kez seçilebilecek şekilde üç kümeye eşit sayıda rastgele olarak dağıtılmıştır. Bu yöntemden ileride "Tekil Veri Kümesi" olarak bahsedilecektir. İkinci yöntemde ise yirmi tane alt küme rastgele oluşturulmuş ve bu yirmi alt kümeye yine tüm ayrılan müşteriler eklenmiştir. Bu yöntemde ayrılmayan müşteriler için, bir müşteri birden fazla kez herhangi bir veri kümesinde olabilecek şekilde rastgele olarak seçilmiş ve veri kümeleri tamamlanmıştır. Bu yöntemden ileride "Çoklu Veri Kümesi" olarak bahsedilecektir.



Şekil 3.1 Tekil veri kümesi

Şekil 3.1'de görüldüğü gibi tekil veri kümesinde ayrılmayan müşteriler kendi aralarında rastgele olarak karıştırıldıktan sonra 3 eşit parçaya bölünmüştür ve her parça ayrılan müşterilerle birleştirilip tekil alt veri kümeleri oluşturulmuştur.



Şekil 3.2 Çoklu veri kümesi

Şekil 3.2'de görüldüğü gibi çoklu veri kümesinde ayrılan müşterilerin hepsi tüm alt kümelere eklenir. Geri kalanı ise ayrılmayanlar müşterilerden rastgele olarak rasgele olarak doldurulur.

BÖLÜM 4

DEĞERLENDİRME KRİTERLERİ

Sonuçların değerlendirilmesinde sınıf karışıklık matrisi kullanılmıştır. Bu matris veri madenciliği sınıflandırma yöntemlerini değerlendirmek için kullanılan ve sonuçların nasıl sınıflandırıldığını gösteren bir yöntemdir. Bu yöntem tahmin edilen sınıflara karşılık gerçek sınıflarını göstermektedir. Çizelge 4.1'de iki sınıflı bir sınıflayıcı için örnek gösterilmiştir;

Çizelge 4.1 Sınıf karışıklık matrisi

		TAHMİN EDİLEN		TOPLAM
		Pozitif	Negatif	
Gerçek	Pozitif	**TP** True Pozitif	**FN** False Negatif	Gerçek Pozitif Sayısı
	Negatif	**FP** False Pozitif	**TN** True Negatif	Gerçek Negatif Sayısı
TOPLAM		Tahmin Pozitif Sayısı	Tahmin Negatif Sayısı	Toplam Örnek Sayısı

Bu matris test edilen bir örneğin alabileceği bütün olasılıkları göstermektedir. İhtiyaca göre bu matrisi değerlendirmek için farklı kriterlere bakılabilir. Kullanılan yaygın değerlendirme kriterleri doğruluk (Accurancy), kesinlik (Precision), geri çağırma (Recall-

doğru pozitif oranı), yanlış pozitif oranı, özgünlük (specificity), hassaslık (sensitivity) ve F-ölçütü'dür.

- Doğruluk, bir sınıflandırıcının doğru sınıflandırdığı örnek sayısının toplam örnek sayısına oranıdır.

$$Doğruluk = \frac{TP+TN}{TP+TN+FP+FN} \quad (4.1)$$

- Hata ise doğruluk oranının tam tersidir.

$$Hata = \frac{FP+FN}{TP+TN+FP+FN} \quad (4.2)$$

- Kesinlik, gerçek değeri pozitif olup pozitif değere sınıflandırılan sayısının, pozitif değerlere sınıflandırılanların toplamına oranıdır.

$$Kesinlik = \frac{TP}{TP+FP} \quad (4.3)$$

- Geri çağırma, gerçek değeri pozitif olup pozitif değere sınıflandırılan sayısının, gerçek değeri pozitif olanların tümüne oranıdır.

$$Geri\ Çağırma = \frac{TP}{TP+FN} \quad (4.4)$$

- Özgünlük, gerçek değeri negatif olup negatif değere sınıflandırılan sayısının, gerçek değeri negatif olanların tümüne oranıdır.

$$Özgüllük = \frac{TN}{TN+FP} \quad (4.5)$$

- Hassaslık, gerçek değeri pozitif olup pozitif değere sınıflandırılan sayısının, gerçek değeri pozitif olanların tümüne oranıdır.

$$Hassasiyet = \frac{TP}{TP+FN} \quad (4.6)$$

- F-Ölçütü, iki sınıflandırıcının tek ölçüt ile değerlendirilmesi için kullanılır.

$$F-Ölçütü = \frac{2*Kesinlik*\ Geri\ Çağırma}{Kesinlik+\ Geri\ Çağırma} \quad (4.7)$$

Bu çalışmada, veri kümesinde dengesizlik olduğu için kesinlik, geri çağırma ve f-ölçütü oranı değerlerine bakılmıştır. Veri kümesinde dengesizlik olduğu durumlarda, sınıf etiketi fazla olan tarafa kayma olmaktadır. Eğer fazla olan sınıfı pozitif olarak düşünecek olursak, bu TP oranını artırmakla beraber FP oranını da artırmaktadır. Burada doğruluk

değeri yüksek olabilir ancak bu gerçeği yansıtmamaktadır. Bu nedenle doğruluk değerine bakılmamıştır.

Bu çalışmada pozitif sınıf olarak ayrılan müşteriler alınmıştır. Çünkü hedef bu tür müşterileri tespit etmektir.

BÖLÜM 5

SINIFLANDIRMA SONUÇLARI

5.1. Tsai Veri Kümesi

Tsai veri kümesi 3. bölümde anlatıldığı gibi kayıp değerlerin doldurulması, aykırı değer analizi, öznitelik azaltma işlemleri ve klasik sınıflama yöntemleri ve bu yöntemlerin bazılarının parametrelerinin değiştirilmesiyle yapılan sınıflandırma sonuçları ile melez modellerin sınıflandırma sonuçları gösterilmiştir. Veri kümesinin büyüklüğü ve tutarsızlığı nedeniyle sadece birinci melez modelin performansı ölçülmüş, diğer melez modeller bu veri kümesinde denenmemiştir.

5.1.1 Kayıp Değer Doldurma İşlemi Sınıflandırma Sonuçları

Üçüncü bölümde açıklandığı gibi kayıp değerler iki farklı şekilde tamamlanmıştır. Kayıp değer doldurma işlemi sonrası sınıflandırmaların sonuçları Çizelge 5.1'deki gibidir.

Sonuçlara bakıldığında eksik değerlerin sınıflara bakılarak ya da bakılmaksızın doldurulmasının sınıflandırma performansı neredeyse hiç değiştirmediği görülmektedir. Bundan sonraki işlemlere eksik değerlerin sınıflarına göre doldurulmasıyla oluşturulan veri kümesiyle devam edilmiştir.

5.1.1. Öznitelik Azaltımı Sınıflandırma Sonuçları

Farklı sayılarda özniteliklerden oluşan kümelerin (tüm veri kümesi, en seçici 126, 102, 77, 57 ve 36 öznitelik) 12 farklı klasik sınıflama yöntemi ile yapılan sınıflandırma sonuçları

Çizelge 5.2'de gösterilmiştir. Her veri kümesinin iki değerlendirme kriteri içinde en iyi sonuçlar kalın yazıyla gösterilmiştir.

Tsai öznitelik azaltımı sınıflandırma sonuçlarına göre genel olarak en iyi sonuçları geri çağırma oranı için naif bayes algoritması, kesinlik oranı için lojistik algoritması vermiştir. Öznitelik azaltımı işleminde geri çağırma için en iyi sonuç en seçici 77 öznitelik ile oluşturulan veri kümesinde görülmüş ve naif bayes sınıflandırıcısı için %81,5 ile tüm veri kümesine göre yaklaşık 8 puan fark atmıştır. Ancak lojistik sınıflandırıcısı için kesinlik ölçütü de 0.6 puan düşmüştür. Bunun dışındaki sonuçlarda önemli bir fark yoktur. Sonuç olarak hem çalışma zamanı hem de performans açısından 77 öznitelikten oluşan veri kümesinin kullanılabileceğine karar verilmiştir.

5.1.3 Aykırı Değer Analizi Sonucu Kalan Veri İle Yapılan Sınıflandırmalar

Tsai veri kümesinde aykırı değer analizi 3.1.3. bölümde anlatıldığı gibi iki farklı şekilde yapılmıştır. Birincisinde toplam örnek sayısının %75'ini kapsayan 1 ve üstü aykırı özniteliği olan örnekler (AD1 Kümesi), ikincisinde toplam örnek sayısının %10'unu kapsayan 26 ve üstü özniteliği olan örnekler (AD2 Kümesi) silinerek çalışılmıştır. Sonuçlar ise Çizelge 5.3'de gösterilmektedir.

Aykırı değer analizine (Çizelge 5.3) 1 ve üstü aykırı özniteliği olan örnekler atıldığında, 26 ve üstü aykırı özniteliği olanların atılmasından, kendi içlerinde en iyi sonuçları veren sınıflandırıcılar için daha iyi sonuç vermiştir. Aykırı değer analizi ile yapılan sınıflandırmalar öznitelik azaltımı ile yapılan sınıflandırma sonuçlarıyla karşılaştırıldığında, yine kendi içlerinde en iyi sonuçları veren sınıflandırıcılar için 1 ve üstü aykırı özniteliği olan örneklerin atıldığı veri kümesi hem başarı açısından hem de performans açısından daha iyi sonuç vermiştir.

Öznitelik azaltımı işleminde ve aykırı değer analizi yönteminde en iyi sonuçları veren veri kümelerinin özellikleri birleştirilmiştir. Yani 1 ve üstü aykırı değeri olan abonelerin veri kümesinden çıkarılmasıyla oluşan alt veri kümesinin en seçici 77 özniteliğiyle yapılan sınıflandırma sonuçları da Çizelge 5.4'deki gibidir.

Çizelge 5.1 Tsai veri kümesinde kayıp değer doldurulmasıyla yapılan sınıflandırma sonuçları

		Naif Bayes (%)	Lojistik (%)	Çok Katmanlı Algılayıcılar (%)	Adaboosts (%)	Torbalama (%)	Rastgele Alt Uzaylar (%)	Karar Tablosu (%)	OneR (%)	Rastgele Orman (%)	Rastgele Ağaç (%)	CART (%)	J48 (%)	Ortalama (%)
KD1	Geri Çağırma	**73,60**	59,2	60,00	65,8	53,20	58,60	70,30	51,30	56,70	53,20	68,80	58,40	60,76
	Kesinlik	52,20	**59,20**	49,60	56,50	53,60	57,60	57,60	50,40	55,20	52,50	57,00	56,70	54,84
KD2	Geri Çağırma	**73,60**	59,30	50,00	65,80	53,30	59,20	71,00	51,30	57,60	52,70	64,00	58,40	59,68
	Kesinlik	52,20	**59,20**	49,60	56,50	53,80	58,10	58,00	50,40	55,90	52,40	58,00	57,60	55,14

Çizelge 5.2 Tsai veri kümesi öznitelik azaltımı sınıflandırma sonuçları

		Naif Bayes (%)	Lojistik (%)	Çok Katmanlı Algılayıcılar (%)	Adaboosts (%)	Torbalama (%)	Rastgele Alt Uzaylar (%)	Karar Tablosu (%)	OneR (%)	Rastgele Orman (%)	Rastgele Ağaç (%)	CART (%)	J48 (%)	Ortalama (%)
Tüm Veri Kümesi	Geri Çağırma	**73,60**	59,20	60,00	65,80	53,20	58,60	70,30	51,30	56,70	53,20	68,8	58,40	60,76
	Kesinlik	52,20	**59,20**	49,60	56,50	53,60	57,60	57,60	50,40	55,20	52,50	57,00	56,70	54,84
En İyi 126 Öznitelik	Geri Çağırma	**77,30**	59,20	60,00	65,80	53,30	58,30	70,30	51,30	56,80	52,70	68,90	58,90	61,07
	Kesinlik	51,60	**59,10**	49,60	56,50	53,50	42,10	57,60	50,40	55,10	52,70	57,00	56,90	53,51
En İyi 102 Öznitelik	Geri Çağırma	**80,30**	59,20	60,00	65,80	53,20	57,90	70,30	51,30	57,10	52,70	68,90	58,70	61,28
	Kesinlik	51,20	**58,90**	49,60	56,50	53,60	57,60	57,60	50,40	55,30	52,20	57,00	57,30	54,77
En İyi 77 Öznitelik	Geri Çağırma	**81,50**	59,40	50,00	66,60	53,40	58,70	68,90	51,30	57,30	52,50	68,20	59,60	60,62
	Kesinlik	51,00	**58,60**	49,60	56,40	53,70	58,00	57,70	50,40	55,50	52,50	57,00	57,50	54,825
En İyi 57 Öznitelik	Geri Çağırma	**79,30**	58,30	40,00	66,60	53,70	59,10	68,70	51,30	57,10	51,30	69,00	57,00	59,28
	Kesinlik	51,30	**58,30**	49,60	56,40	54,30	**58,30**	57,70	50,40	55,70	50,40	57,10	57,00	54,71
En İyi 36 Öznitelik	Geri Çağırma	**77,30**	57,60	50,00	66,60	54,30	59,70	69,40	51,30	57,80	52,90	69,50	57,10	60,29
	Kesinlik	51,90	57,50	49,60	56,40	54,60	58,90	**57,70**	50,40	55,80	53,40	57,00	57,20	55,03
Ortalama	Geri Çağırma	**78,22**	58,82	53,33	66,20	53,52	58,72	69,65	51,30	57,13	52,55	68,88	58,28	60,55
	Kesinlik	51,53	**58,60**	49,60	56,45	53,88	55,42	57,65	50,40	55,43	52,28	57,02	57,10	54,61

Çizelge 5.3 Tsai aykırı değer analizi sınıflandırma sonuçları

		Naif Bayes (%)	Lojistik (%)	Çok Katmanlı Algılayıcılar (%)	Adaboosts (%)	Torbalama (%)	Rastgele Alt Uzaylar (%)	Karar Tablosu (%)	OneR (%)	Rastgele Orman (%)	Rastgele Ağaç (%)	CART (%)	J48 (%)	Ortalama (%)
AD1 Kümesi	Geri Çağırma	53,70	60,60	50,00	61,70	54,70	58,60	65,50	52,50	58,00	55,00	**83,20**	60,50	59,50
	Kesinlik	59,00	**60,40**	50,60	59,00	52,30	56,00	59,20	51,40	56,20	53,90	55,40	58,00	55,95
AD2 Kümesi	Geri Çağırma	**74,90**	60,60	30,00	71,00	54,60	60,80	72,20	52,40	57,60	53,60	73,40	57,40	59,88
	Kesinlik	53,10	**59,90**	49,90	56,60	54,20	58,10	57,90	50,80	55,50	53,10	57,10	57,60	55,30

Çizelge 5.4 Tsai Veri Kümesi 77 Öznitelik ve 1'den Fazla Aykırı Değeri Olan Abonelerin Atılmasıyla Yapılan Sınıflandırma Sonuçları

	Naif Bayes (%)	Lojistik (%)	Çok Katmanlı Algılayıcılar (%)	Adaboosts (%)	Torbalama (%)	Rastgele Alt Uzaylar (%)	Karar Tablosu (%)	OneR (%)	Rastgele Orman (%)	Rastgele Ağaç (%)	CART (%)	J48 (%)	Ortalama (%)
Geri Çağırma	67,50	60,40	60,00	59,70	53,90	60,20	65,10	52,50	60,50	53,80	**81,00**	61,60	61,35
Kesinlik	56,60	**60,00**	50,60	59,70	51,90	57,90	59,30	51,40	59,60	53,70	55,50	58,20	56,20

Bu sonuçlara göre sınıflandırma yöntemlerinin ortalaması küçük miktar artmıştır. Ancak en iyi sonuçları veren yöntemlere baktığımızda geri çağırma oranı için CART algoritması tüm veri kümesinde yüzde 83,2 başarı sağlarken en seçici 77 özniteliğin kullanıldığı veri kümesi yüzde 81 başarı elde etmiştir. Yine kesinlik oranı için lojistik algoritması tüm veri kümesinde yüzde 60,4 başarı sağlamışken en seçici 77 özniteliğin kullanıldığı veri kümesinde yüzde 60 başarı elde etmiştir. Bu sonuçlara göre başarı bir miktar azalmıştır. Ayrıca CART algoritmasının geri çağırma oranında diğer yöntemlere göre daha iyi sonuç vermesinin nedeni sınıflandırmaların ayrılan müşteri yönüne kaymasıdır.

5.1.4 Sınıflayıcıların Parametreleri Değiştirilerek Yapılan Çalışmalar

Hem sınıflandırma yöntemlerinin parametre değişiminin sonuçlarını görmek hem de veriyi daha iyi tanımak amacıyla bazı klasik sınıflandırma algoritmalarının parametreleri değiştirilerek performansları ölçülmüştür. Bunun için çok katmanlı algılayıcılar ve k en yakın komşu algoritmaları weka platformunu kullanarak orijinal veri kümesinde performansları ölçülmüştür.

Çizelge 5.5 Tsai veri kümesinde ÇKA parametre değişimleri sınıflandırma sonuçları

İterasyon Sayısı	50			100			200			500		
Katman Sayısı	8	12	16	8	12	16	8	12	16	8	12	16
Doğruluk	50,09	49,91	50,00	50,09	50,00	50,00	50,09	50,09	50,00	50,09	50,00	50,00
Geri Çağırma	40,00	60,00	50,00	40,00	50,00	50,00	40,00	40,00	50,00	40,00	50,00	50,00
Kesinlik	49,00	49,6	49,60	49,60	49,60	49,60	49,60	49,60	49,60	49,60	49,60	49,60

Çizelge 5.5'de yapay sinir ağlarının parametrelerinin değiştirilmesiyle elde edilen sonuçlar gösterilmektedir. Bu yöntemde değiştirilen parametreler eğitim sayısı ve katman sayısıdır. Elde edilen sonuçlara göre doğruluk oranı ve kesinlik ölçütü neredeyse hiç değişmemiştir. En iyi geri çağırma oranına katman sayısının 12 eğitim sayısının 50 olduğu parametrelerle %60 olarak elde edilmiştir. Bu sonuçlara bakıldığında Tsai veri kümesi için yapay sinir ağlarında parametre değiştirilmesi performansı etkilemediği gözlemlenmiştir.

Çizelge 5.6 Tsai veri kümesinde KNN parametre değişimi sınıflandırma sonuçları

K Değeri	**1**	**3**	**5**	**7**	**9**	**19**	**39**	**59**
Geri Çağırma	51,20	50,90	50,70	51,20	51,40	51,60	**51,70**	**51,70**
Kesinlik	51,90	52,40	53,00	53,50	53,90	54,70	55,50	**55,90**

Çizelge 5.6'da K en yakın komşu algoritması için komşu sayısına göre geri çağırma ve kesinlik oranları gösterilmektedir. Sonuçlara göre bakılan komşu sayısı arttığında performans bir miktar artmaktadır. Ancak tahmin değerleri zaten çok düşük olduğu için bu artış önemsizdir. Bu çizelgeye göre K en yakın komşu algoritması Tsai veri kümesi için uygun bir yöntem değildir.

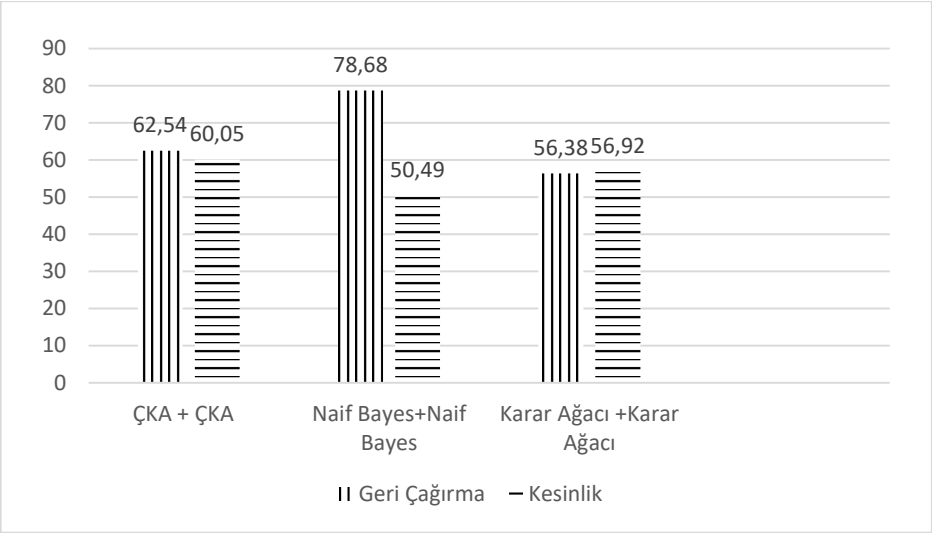
5.1.5 Melez Yöntemler

Tsai veri kümesine 3. bölümde anlatılan 1. melez model uygulanmıştır. Bu melez model ÇKA + ÇKA, naif bayes + naif bayes ve karar ağacı + karar ağacı yapısından oluşmaktadır. Bu melez model adımlarının her iki sınıflandırıcısının inşasında da 10 kat çapraz doğrulama kullanılmıştır. Burada ki sınıflandırmalar için matlab programlama dili kullanılmıştır. Sonuçlar Çizelge 5.7'de ve Şekil 5.1'de gösterilmektedir.

Çizelge 5.7 Tsai veri kümesi 1. melez yöntem sınıflandırma sonuçları

	ÇKA + ÇKA	Naif Bayes + Naif Bayes	Karar Ağacı + Karar Ağacı
Geri Çağırma	62,54	**78,68**	56,38
Kesinlik	**60,05**	50,49	56,92

Çizelge 5.7 ve Şekil 5.1'de gösterilen sonuçlara göre, geri çağırma oranında en iyi sonucu naif bayes algoritmasının kullanıldığı melez model vermiştir. Kesinlik oranı için ise ÇKA'nın kullanıldığı melez model en iyi performansı göstermiştir. Bu algoritmaları tek başına kullanıldığı (Çizelge 5.2) yöntemlerle karşılaştıracak olursak ÇKA'nın geri çağırma oranında küçük bir miktar artış gözlemlenmiştir. Kesinlik oranında ise %10,45 artış olmuştur. Naif bayes yöntemi için geri çağırmada melez model %5 daha iyi sonuç vermiştir. Ancak kesinlik oranında %1,71 geri kalmıştır. Karar ağacı algoritmasında melez model geri çağırma oranında %2,02 daha kötü sonuç vermiştir. Kesinlik oranı için ise aynı sonucu verdiği söylenebilir. Bu sonuçlar değerlendirildiğinde herhangi bir melez model kesin bir fark ortaya koyamamıştır.



Şekil 5.1 Tsai veri kümesi 1. melez model sınıflandırma sonuçları

Tsai veri kümesi üzerinde yapılan tüm denemeler düşünüldüğünde en seçici 77 özniteliğin kullanıldığı veri kümesi üzerinde naif bayes tekniği önerilmektedir. Ancak uygulanan tüm yöntemlere rağmen neredeyse hiç performans artışı gözlemlenememiştir. Bunun nedeninin veri kümesinin tutarsızlığı olduğu düşünülmektedir.

5.2 Larose Veri Kümesi

Larose veri kümesinde 3. bölümde anlatıldığı gibi öznitelik azaltma işlemleri, alt örnekleme işlemleri, klasik sınıflama yöntemleri ve bu yöntemlerin bazılarının parametrelerinin değiştirilmesiyle yapılan sınıflandırma sonuçları ile melez modellerin sınıflandırma sonuçları gösterilmiştir.

5.2.1 Öznitelik Azaltımı Sınıflandırma Sonuçları

Larose veri kümesinde daha önce bahsedildiği gibi tüm veri kümesi, en seçici 12, 8 ve 5 özniteliğin oluşturduğu alt kümeler kullanılarak sınıflandırmalar yapılmıştır. Bu sınıflandırmaların sonuçları ise Çizelge 5.8'de gösterilmektedir.

Öznitelik indirgeme sonuçlarına göre genel olarak en iyi sınıflandırma sonuçları kesinlik ölçütü için rastgele alt uzaylar ve geri çağırma ölçütü için tüm veri kümesinde rotasyon ormanı, en seçici 12 özniteliğin kullanıldığı veri kümesinde CART (Sınıflandırma ve

Regresyon Ağaçları) algoritması, en seçici 8 ve 5 özniteliğin kullanıldığı veri kümelerinde rasgele ağaç algoritması vermiştir. Bu işlemde en seçici 8 ve 5 özniteliğin kullanıldığı veri kümelerinde başarı düşmüştür. Tüm veri kümesinin kullanımıyla, en seçici 12 öznitelik kullanımı arasında ciddi bir fark yoktur. Geri çağırma ölçütü için en seçici 12 özniteliğin kullanıldığı veri kümesi tüm veri kümesinden %2,4 daha iyidir. Kesinlik ölçütü için ise tüm veri kümesinin kullanılması en seçici 12 özniteliğin kullanılmasından %1,5 daha iyidir.

Çizelge 5.8'deki sonuçlar değerlendirildiğinde en seçici 12 özniteliğin oluşturduğu veri kümesinin yeterli olduğu belirlenmiştir.

5.2.2 Alt Örnekleme İşlemi Sınıflandırma Sonuçları

Larose veri kümesinde sınıflar arası dengesizlik olduğu için alt örnekleme işlemiyle alt veri kümeleri oluşturulmuştur. Bu üçüncü bölümde bahsedildiği gibi iki farklı şekilde yapılmıştır. Bunlar tekil ve çoklu alt veri kümeleridir.

5.2.2.1 Tekil Veri Kümesinde Sınıflandırma

Larose veri kümesi tekil sınıflandırma sonuçları Çizelge 5.9'daki gibidir. Bu sınıflandırma sonuçlarına bakıldığında en iyi sonuçları CART ve J48 algoritmaları vermiştir. Tekil veri kümesi ile Çizelge 5.8'deki tüm verilerin kullanılmasıyla elde edilen sonuçları karşılaştıracak olursak genel ortalamalara göre tüm örneklerin kullanıldığı veri kümesi geri çağırmada %42,45 başarı elde ederken tekil veri kümesi %63,96 başarı elde etmiştir. Kesinlik ölçütüne bakacak olursak tüm örneklerin kullanıldığı veri kümesi %72,37 iken tekil veri kümesi %75,63'dür. Yapılan sınıflandırmalarda kendi içlerinde en iyi sonuçları veren sınıflama algoritmaları karşılaştıracak olursak kesinlik ölçütü için tüm veri kümesinde rastgele alt uzaylar yöntemi %92,9 başarı elde etmiştir. Tekil veri kümesinde rotasyon ormanı algoritması ise %89,33 puan almıştır. Geri çağırma oranına bakacak olursak tekil veri kümesinde rotasyon ormanı algoritmasıyla yapılan sınıflandırma tüm veri kümesinde yine rotasyon ormanı yöntemiyle yapılan sınıflandırmadan %12,13 daha iyidir. Bu sonuçlar değerlendirildiğinde tekil veri kümesi oluşturulmasının yararlı olduğu görülmektedir.

Çizelge 5.8 Larose veri kümesi öznitelik azaltımı sınıflandırma sonuçları

		Naif Bayes (%)	Lojistik (%)	Çok Katmanlı Algılayıcılar (%)	Adaboosts (%)	Torbalama (%)	Rastgele Alt Uzaylar (%)	Rotasyon Ormanı (%)	Karar Tablosu (%)	OneR (%)	Rastgele Orman (%)	Rastgele Ağaç (%)	CART (%)	J48 (%)	Ortalama (%)
Tüm Veri Kümesi (19 öznitelik)	Geri Çağırma	43,30	23,60	61,80	11,50	53,70	35,40	**67,20**	35,20	16,80	43,10	52,20	66,90	65,90	42,45
	Kesinlik	65,40	56,40	75,50	56,60	73,40	**92,90**	92,40	73,20	61,70	85,70	53,40	84,60	89,60	72,37
En Seçici 12 Öznitelik	Geri Çağırma	43,00	22,30	51,60	11,50	53,70	34,50	69,40	36,40	16,80	45,80	56,90	**69,60**	68,00	42,51
	Kesinlik	64,80	55,40	70,60	56,60	74,40	**91,40**	88,50	73,90	61,70	79,40	56,20	83,70	86,20	71,19
En Seçici 8 Öznitelik	Geri Çağırma	41,00	19,10	50,60	11,50	40,60	26,20	54,20	33,90	16,80	46,00	**52,10**	42,90	50,90	35,97
	Kesinlik	61,10	51,90	65,20	56,20	58,30	**89,80**	79,80	72,90	61,70	66,10	52,90	79,70	80,70	66,38
En Seçici 5 Öznitelik	Geri Çağırma	24,30	18,40	34,10	11,70	40,00	14,40	41,90	31,10	16,80	39,50	**46,70**	43,10	36,90	29,92
	Kesinlik	63,00	50,00	55,30	58,50	57,90	**81,00**	68,40	66,70	61,70	55,50	46,50	78,40	75,70	62,52

Çizelge 5.9 Larose tekil veri kümesi sınıflandırma sonuçları

		Naif Bayes (%)	Lojistik (%)	Çok Katmanlı Algılayıcılar (%)	Adaboosts (%)	Torbalama (%)	Rastgele Alt Uzaylar (%)	Rotasyon Ormanı (%)	Karar Tablosu (%)	OneR (%)	Rastgele Orman (%)	Rastgele Ağaç (%)	CART (%)	J48 (%)	Ortalama (%)
1.2F+T	Geri Çağırma	66,90	71,90	56,90	59,30	69,20	65,60	**79,90**	68,70	34,20	65,50	58,40	74,00	74,70	63,78
	Kesinlik	73,60	76,90	69,40	75,80	72,20	88,81	88,60	80,50	63,00	82,80	61,10	86,00	**90,10**	76,62
2.2F+T	Geri Çağırma	65,30	75,00	55,90	46,40	68,00	64,50	76,80	68,30	35,60	66,10	62,90	**77,40**	75,10	63,38
	Kesinlik	73,00	71,30	66,90	72,20	73,10	84,10	89,70	80,50	63,80	80,50	61,30	86,10	**92,00**	75,40
3.2F+T	Geri Çağırma	68,70	71,30	57,40	60,50	68,70	61,20	**81,30**	76,40	34,10	63,40	59,80	75,50	79,80	64,73
	Kesinlik	73,60	75,70	66,90	76,40	71,50	85,40	**89,70**	76,20	62,30	78,60	60,50	86,80	84,60	74,88
Ortalamalar	Geri Çağırma	66,97	72,73	56,73	55,40	68,63	63,77	**79,33**	71,13	34,63	65,00	60,37	75,63	76,53	63,96
	Kesinlik	73,40	74,63	67,73	74,80	72,27	85,83	**89,33**	79,07	63,03	80,63	60,97	86,3	88,9	75,63

5.2.2.2 Çoklu Veri Kümesinde Sınıflandırma

Larose veri kümesi çoklu sınıflandırmasının geri çağırma, kesinlik ve f-ölçütü oranları Çizelge 5.10, Çizelge 5.11 ve Çizelge 5.12'deki gibidir.

Larose çoklu veri kümesi sonuçlarının ortalamalarına bakıldığında(Çizelge 5.10, Çizelge 5.11 ve Çizelge 5.12) en iyi sonuçları rotasyon ormanı algoritması vermiştir. Kesinlik ölçütü için %90,46, geri çağırma oranı için ise %81,47 başarı elde edilmiştir. Çoklu veri kümelerinin sonuçlarını birbirleriyle karşılaştıracak olursak kesinlik oranında en iyi sonucu 12. alt küme j48 algoritmasıyla %94,63 olarak elde etmiştir. Geri çağırma oranı için 14. alt küme rotasyon ormanı algoritması ile %82,89 başarı kazanmıştır.

Çoklu veri kümeleri arasından en iyi alt kümeyi bulmak amacı ile kesinlik ve geri çağırma oranlarıyla elde edilen Çizelge 5.12'de gösterilen f-ölçütü oranına bakılmıştır. Bu kritere göre en iyi sonucu 16. alt kümede rotasyon ormanı algoritmasıyla %87,37 elde edilmiştir. 16. alt kümenin aynı sınıflandırma algoritması için kesinlik oranı ve geri çağırma oranı başarıları ise sırayla %92,56 ve %82,74'dür.

Çoklu alt veri kümelerinin ortalamaları orijinal veri kümesi ve tekil veri kümelerinin ortalamaları ile karşılaştırılmıştır. Sonuçlara göre orijinal veri kümesi kesinlik ölçütü için %1,94 daha iyidir ancak geri çağırmada %14,27 geri düşmüştür. Tekil veri kümesine göre ise kesinlik oranı için %1,13, geri çağırma oranı için %2,14 çoklu alt veri kümelerinin ortalamaları daha iyidir.

Bu sonuçlar değerlendirildiğinde çoklu alt örnekleme işleminin sonuçlar üzerindeki etkisi çok fazladır.

Çizelge 5.10 Larose çoklu alt veri kümesi geri çağırma oranı sonuçları

Alt Küme	Naif Bayes (%)	Lojistik (%)	Çok Katmanlı Algılayıcılar (%)	Adaboosts (%)	Torbalama (%)	Rastgele Alt Uzaylar (%)	Rotasyon Ormanı	Karar Tablosu (%)	OneR (%)	Rastgele Orman (%)	Rastgele Ağaç (%)	CART (%)	J48 (%)	Ortalama
1	65,21	54,88	72,56	65,35	68,60	66,48	82,32	76,52	39,04	73,55	59,69	76,52	77,37	66,42
2	67,33	56,01	74,82	57,99	68,46	60,25	82,04	75,25	36,07	73,83	62,52	75,67	77,37	66,00
3	65,77	57,85	73,13	61,67	69,73	63,51	80,62	75,67	39,89	73,13	63,65	76,66	77,51	66,86
4	65,77	58,27	73,41	59,41	66,20	59,55	81,61	71,99	39,18	73,13	65,77	76,10	75,81	65,97
5	65,49	57,99	73,83	58,27	68,88	58,42	81,61	71,43	37,06	73,55	67,47	75,25	74,12	65,72
6	65,49	56,15	71,00	53,04	68,32	57,28	81,47	74,40	40,31	72,56	67,47	75,39	74,54	65,23
7	65,49	57,71	73,83	61,39	68,46	61,53	82,32	72,28	32,96	74,26	64,21	77,37	80,62	66,42
8	68,74	57,99	73,13	55,87	66,76	57,28	80,34	67,19	35,93	72,84	63,37	76,24	79,63	65,32
9	65,91	55,30	74,12	46,96	69,31	60,11	80,06	67,75	35,79	71,29	58,13	77,51	73,41	63,75
10	66,34	57,57	76,24	68,74	68,32	58,42	80,62	70,01	37,48	73,41	59,97	75,39	76,38	66,36
11	66,76	55,16	72,42	58,56	67,61	56,72	81,90	72,28	38,05	72,84	60,68	76,52	78,36	65,15
12	66,34	55,59	72,84	66,05	68,74	60,25	81,19	71,29	39,60	74,68	56,86	76,80	77,23	66,03
13	64,92	55,45	70,72	68,03	69,17	58,27	81,61	69,59	37,20	71,99	66,62	77,65	78,22	66,13
14	67,75	59,26	73,41	63,65	68,46	62,09	82,89	73,97	39,18	74,96	64,07	76,94	77,79	67,39
15	65,91	55,73	73,55	64,50	67,04	56,72	80,20	72,14	38,76	70,44	67,89	75,95	81,05	66,10
16	66,05	54,88	75,25	60,68	68,60	59,26	82,74	71,29	35,22	72,70	64,07	75,53	74,12	65,43
17	67,89	57,71	75,81	62,09	69,73	57,57	81,61	72,42	38,05	70,86	61,53	74,68	77,51	66,06
18	67,47	56,15	75,25	65,77	68,74	61,10	81,61	75,81	36,63	75,53	66,76	76,10	74,96	67,05
19	66,2	59,12	72,98	62,94	70,44	60,54	81,33	74,54	38,47	73,55	57,99	77,79	78,93	66,68
20	65,06	57,57	70,01	57,71	70,16	58,98	81,33	73,69	41,16	72,28	64,07	76,80	76,24	65,82
Ortalama	66,29	56,81	73,41	60,93	68,58	59,710	**81,47**	72,47	37,80	73,06	63,13	76,34	77,05	66,00

Çizelge 5.11 Larose çoklu alt veri kümesi kesinlik oranı sonuçları

Alt Küme	Naif Bayes (%)	Lojistik (%)	Çok Katmanlı Algılayıcılar (%)	Adaboosts (%)	Torbalama (%)	Rastgele Alt Uzaylar (%)	Rotasyon Ormanı	Karar Tablosu (%)	OneR (%)	Rastgele Orman (%)	Rastgele Ağaç (%)	CART (%)	J48 (%)	Ortalama
1	72,37	68,79	79,66	76,62	73,82	88,35	89,81	76,30	64,64	88,89	63,36	86,70	88,80	77,67
2	72,34	68,28	82,40	76,21	69,34	85,71	89,23	79,28	60,71	87,58	67,79	85,74	87,38	77,16
3	73,93	71,38	78,21	78,99	74,81	84,40	89,48	78,33	65,13	87,93	66,96	87,42	93,36	78,65
4	72,54	70,07	79,00	77,49	72,22	85,92	91,01	79,91	60,09	88,98	67,78	88,49	89,04	78,02
5	71,67	71,55	79,09	74,64	75,74	91,78	90,02	79,40	62,98	89,81	70,04	89,11	93,07	79,37
6	72,80	70,02	80,45	73,82	73,52	85,26	91,57	76,56	61,16	87,24	70,15	87,95	91,65	78,04
7	74,32	70,47	83,92	75,87	72,56	88,24	91,80	81,89	54,69	88,38	67,86	87,66	89,34	78,29
8	72,75	69,61	79,17	78,37	71,95	87,10	88,20	81,34	58,80	85,98	64,37	86,38	87,56	77,21
9	73,27	68,60	78,92	76,85	75,50	87,99	91,44	84,78	63,89	88,58	65,13	89,11	92,18	78,93
10	74,33	69,93	79,97	75,94	73,29	86,76	90,19	82,09	62,21	90,10	63,00	88,10	93,59	78,57
11	73,07	68,54	78,53	76,38	73,31	84,78	92,20	84,46	61,98	88,49	65,70	92,01	89,64	78,39
12	71,71	68,47	83,74	76,56	73,30	85,54	91,84	83,31	61,95	90,57	60,82	88,87	94,63	78,59
13	72,86	67,82	79,74	74,11	73,87	87,66	88,91	81,73	62,92	87,16	70,3	86,87	89,48	77,95
14	74,61	70,90	82,91	75,00	73,22	87,10	89,88	80,21	62,67	88,33	66,81	89,18	87,03	78,54
15	72,81	67,81	79,27	74,03	71,60	88,91	91,45	81,60	63,57	87,68	70,28	88,32	86,04	78,02
16	72,63	69,29	81,97	74,87	72,93	87,66	92,56	81,16	62,41	89,39	67,71	87,54	91,61	78,75
17	73,73	70,10	81,34	77,98	75,38	84,97	90,30	81,53	64,05	86,38	64,54	87,13	90,28	78,49
18	74,18	69,16	83,00	76,73	74,77	86,06	90,16	79,64	60,51	88,70	70,87	88,20	89,23	78,63
19	75,12	70,49	79,26	79,46	72,81	87,70	89,29	78,77	61,68	87,54	61,01	87,44	86,38	77,70
20	73,02	70,66	81,15	76,40	74,81	86,51	89,84	81,15	63,26	90,44	67,11	87,72	92,29	78,94
Ortalama	73,20	69,6	80,59	76,32	73,44	86,92	**90,46**	80,67	61,97	88,41	66,58	87,99	90,13	78,30

Çizelge 5.12 Larose çoklu alt veri kümesi f-ölçütü sonuçları

Alt Küme	Naif Bayes (%)	Lojistik (%)	Çok Katmanlı Algılayıcılar (%)	Adaboosts (%)	Torbalama (%)	Rastgele Alt Uzaylar (%)	Rotasyon Ormanı (%)	Karar Tablosu (%)	OneR (%)	Rastgele Orman (%)	Rastgele Ağaç (%)	CART (%)	J48 (%)	Ortalama
1	68,60	61,05	75,94	70,54	71,11	75,87	85,90	76,41	48,68	80,50	61,47	81,29	82,69	71,38
2	69,75	61,54	78,43	65,86	68,9	70,76	85,48	77,21	45,25	80,12	65,05	80,39	82,07	70,90
3	69,61	63,91	75,58	69,26	72,18	72,48	84,82	76,98	49,48	79,85	65,26	81,69	84,70	72,08
4	68,99	63,63	76,10	67,26	69,08	70,34	86,05	75,74	47,43	80,28	66,76	81,83	81,89	71,31
5	68,44	64,06	76,37	65,45	72,15	71,40	85,61	75,20	46,66	80,87	68,73	81,60	82,52	71,62
6	68,95	62,32	75,43	61,73	70,82	68,52	86,23	75,46	48,59	79,23	68,78	81,19	82,21	70,83
7	69,63	63,45	78,55	67,87	70,45	72,50	86,8	76,79	41,13	80,71	65,98	82,19	84,76	71,68
8	70,69	63,27	76,03	65,23	69,26	69,11	84,09	73,59	44,60	78,87	63,87	80,99	83,41	70,49
9	69,40	61,24	76,44	58,30	72,27	71,43	85,37	75,31	45,88	79,00	61,43	82,91	81,73	70,21
10	70,11	63,15	78,06	72,16	70,72	69,82	85,14	75,57	46,78	80,90	61,45	81,25	84,11	71,72
11	69,77	61,13	75,35	66,29	70,34	67,97	86,75	77,90	47,15	79,91	63,09	83,55	83,62	70,95
12	68,92	61,36	77,91	70,92	70,95	70,70	86,19	76,83	48,32	81,86	58,77	82,4	85,05	71,62
13	68,66	61,01	74,96	70,94	71,44	70,01	85,10	75,17	46,76	78,85	68,41	82,00	83,47	71,32
14	71,01	64,56	77,87	68,86	70,76	72,50	86,24	76,96	48,22	81,10	65,41	82,61	82,15	72,36
15	69,19	61,18	76,30	68,94	69,25	69,26	85,46	76,58	48,16	78,12	69,06	81,67	83,47	71,30
16	69,18	61,25	78,47	67,03	70,70	70,72	87,37	75,91	45,03	80,19	65,84	81,09	81,94	71,21
17	70,69	63,30	78,48	69,13	72,45	68,64	85,74	76,71	47,74	77,85	63,00	80,43	83,41	71,51
18	70,67	61,98	78,94	70,83	71,63	71,46	85,67	77,68	45,63	81,59	68,75	81,7	81,47	72,19
19	70,38	64,31	75,99	70,24	71,61	71,63	85,12	76,60	47,39	79,94	59,46	82,33	82,49	71,55
20	68,81	63,45	75,17	65,75	72,41	70,14	85,37	77,24	49,87	80,35	65,55	81,90	83,50	71,59
Ortalama	69,57	62,56	76,82	67,63	70,92	70,76	**85,73**	76,29	46,93	80,00	64,80	81,75	83,03	71,39

5.2.3 Melez Yöntemler

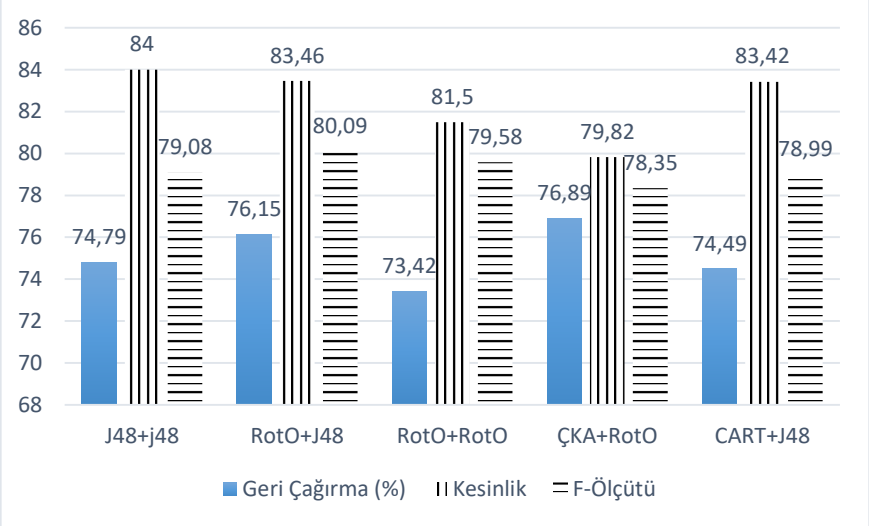
Üçüncü bölümde anlatılan melez modellerin hepsi larose veri kümesinde çoklu alt örnekleme yöntemiyle oluşturulan 20 alt kümede uygulanmıştır. Bu yöntemde uygulanan sınıflandırmalar java programlama dilinde weka kütüphanesi kullanılarak yapılmıştır.

Bu melez yöntemde yaygın olarak kullanılan sınıflandırma yöntemlerinin kendi aralarında kombinasyonları kullanılmıştır. Ancak çizelgelerde 20 alt kümenin sonuçlarının ortalamaları en iyi beş sonucu veren modelle birlikte gösterilmiştir.

5.2.3.1 Birinci Melez Model

Çizelge 5.13 Larose veri kümesi birinci melez model sınıflandırma sonuçları

	Geri Çağırma (%)	Kesinlik (%)	F-Ölçütü (%)
J48+j48	74,79	**84,00**	79,08
RotO+J48	76,15	83,46	**80,09**
RotO+RotO	73,42	81,5	79,58
ÇKA+RotO	**76,89**	79,82	78,35
CART+J48	74,49	83,42	78,99



Şekil 5.2 Larose veri kümesi 1. melez model sınıflandırma sonuçları

Çizelge 5.13 ve Şekil 5.2'deki sonuçlara göre geri çağırma oranı için en iyi sonucu çok katmanlı algılayıcılar + rotasyon ormanı melezi (%76,89), kesinlik oranı için ise karar ağacı

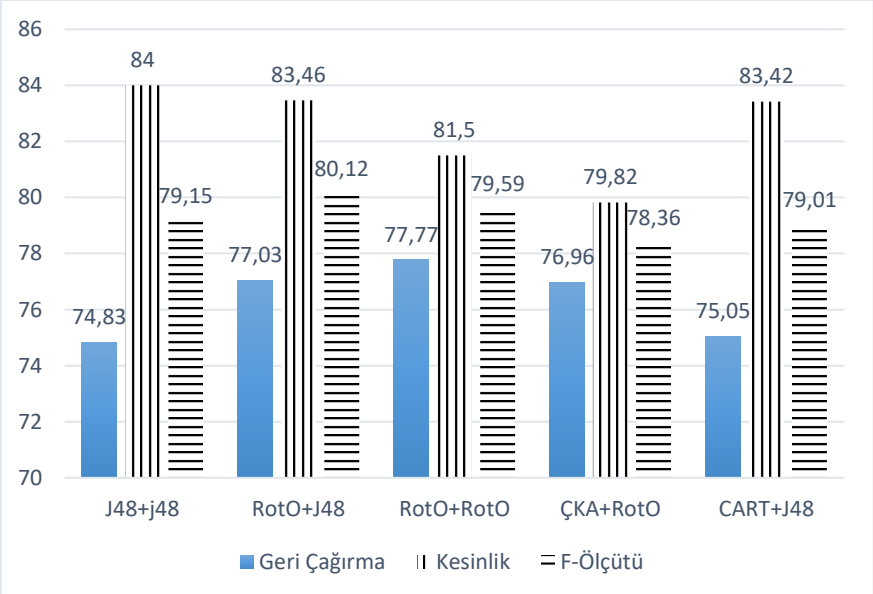
+ karar ağacı melezi (%84) vermiştir. F-ölçütünde ise en iyi performans rotasyon ormanıyla karar ağacının birleşiminde (%80,09) elde edilmiştir.

Bu sonuçları aynı veri kümelerin kullanıldığı klasik yöntemlerden en iyi sonucu veren rotasyon ormanı algoritması ile karşılaştıracak olursak bu melez model kesinlik oranı için %6,46, geri çağırma için %4,58 geri kalmıştır. Sonuçlar değerlendirildiğinde bu model hem başarı acısından hem de çalışma zamanı açısından bu veri kümesinde iyi bir performans elde edememiştir.

5.2.3.2 İkinci Melez Model

Çizelge 5.14 Larose veri kümesi ikinci melez model sınıflandırma sonuçları

	Geri Çağırma (%)	Kesinlik (%)	F-Ölçütü (%)
J48+j48	74,83	**84,00**	79,15
RotO+J48	77,03	83,46	**80,12**
RotO+RotO	**77,77**	81,50	79,59
ÇKA+RotO	76,96	79,82	78,36
CART+J48	75,05	83,42	79,01



Şekil 5.3 Larose veri kümesi 2. melez model sınıflandırma sonuçları

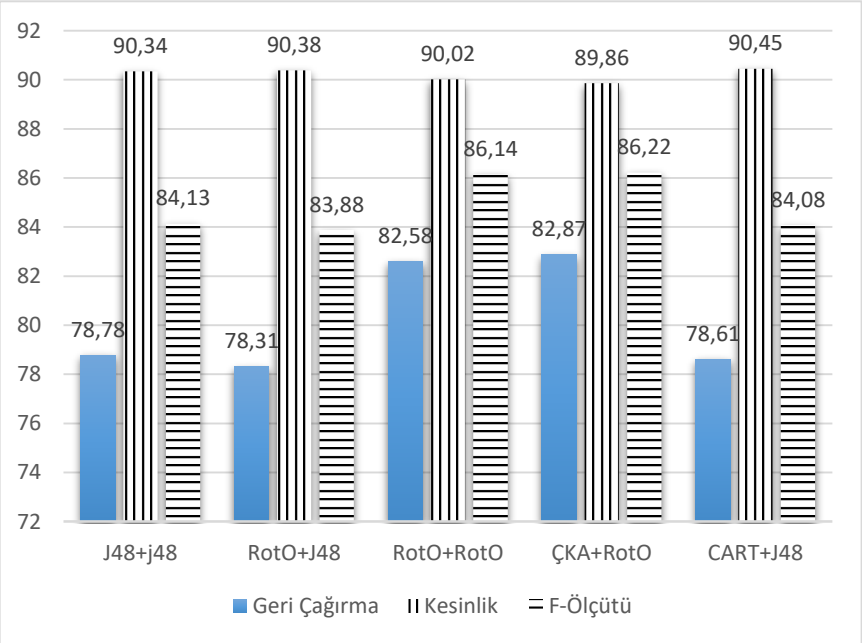
Çizelge 5.14 ve Şekil 5.3'e göre kullanılan melez model için en iyi sonuçlar geri çağırma oranı için rotasyon ormanı + rotasyon ormanı melezi (%77,77 başarı), kesinlik oranı için j48+j48 melezi (%84 başarı) ve f-ölçütü için ise rotasyon ormanı + j48 melezi (%80,52 başarı) algoritmalarıdır.

Bu melez modelin sonuçlarını çoklu alt veri kümesinde klasik yöntemlerin tek başlarına kullanıldığı modellerle karşılaştıracak olursak (Çizelge 5.10 ve Çizelge 5.11), klasik yöntem rotasyon ormanı algoritması ile kesinlik oranı için %6,46, geri çağırma oranı için %3,7 daha iyidir. Bu sonuçlar değerlendirildiğinde kullanılan melez model hem başarı acısından hem de çalışma zamanı açısından bu veri kümesinde iyi bir performans elde edememiştir.

5.2.3.3 Üçüncü Melez Model

Çizelge 5.15 Larose veri kümesi üçüncü melez model sınıflandırma sonuçları

	Geri Çağırma (%)	Kesinlik (%)	F-Ölçütü (%)
J48+j48	78,78	90,34	84,13
RotO+J48	78,31	90,38	83,88
RotO+RotO	82,58	90,02	86,14
ÇKA+RotO	**82,87**	89,86	**86,22**
CART+J48	78,61	**90,45**	84,08



Şekil 5.4 Larose veri kümesi 3. melez model sınıflandırma sonuçları

Üçüncü melez yöntem ile yapılan sınıflandırmalara göre(Çizelge 5.15 ve Şekil 5.4) en iyi sonuçları geri çağırma oranı için çok katmanlı algılayıcılar + rotasyon ormanı melezi (%82,87), kesinlik oranı için CART + karar ağacı melezi (%90,45) ve f-ölçütü için çok katmanlı algılayıcılar + rotasyon ormanı melezi (%86,22) vermiştir.

Bu melez model klasik yöntemlerin en iyi sonucunu veren rotasyon ormanı tek başına kullanımıyla karşılaştırıldığında çok düşük miktar olsa da daha iyidir. Geri çağırma oranı için %1,4 ve f-ölçütü için %0,49 daha iyi, kesinlik oranı için %0,01 daha kötüdür. Çalışma zamanı düşünüldüğünde bu melez modelin çalışma süresi tek başına rotasyon ormanından çok daha fazladır. Ancak süre sınırlaması olmadığından ve doğru tahmin yapmanın daha önemli olmasından dolayı bu veri kümesinde üçüncü melez yöntem önerilmektedir.

Bu veri kümesi için kullanılan üç melez model kendi aralarında karşılaştırıldığında en iyi performans üçüncü önerilen melez modelde elde edilmiştir. Kendi içlerinde en iyi sonuçları veren algoritmaların sonuçlarına göre üçüncü melez yöntem geri çağırma oranında Tsai'nin önerdiği birinci melez yönteme %5,98, ikinci melez yönteme %5,1 fark atmıştır. Kesinlik ölçütü için üçüncü yöntem birinci ve ikinci yöntemden %6,45 daha iyidir. F-ölçütünde ise üçüncü yöntem ikinci yönteme %6,13 birinci yönteme %6,1 fark atmıştır. Bu sonuçlar düşünüldüğünde en iyi melez yöntem üçüncü önerilen modeldir.

BÖLÜM 6

SONUÇ VE ÖNERİLER

Bu çalışmada telekomünikasyon sektöründen alınmış iki adet veri kümesi üzerinde ayrılan müşteriler tahmin edilmeye çalışılmıştır. Bunun için veri kümeleri üzerinde eksik değerler doldurulmuş, aykırı değer analizi yapılmış, var olan sınıflandırma yöntemlerinin parametreleri değiştirilmiş, alt örnekleme işlemi uygulanmış ve yeni melez modeller oluşturularak tahmin gücü artırılmaya çalışılmıştır.

Tsai veri kümesi 100.000 örnekten ve 172 öznitelikten oluşan büyük bir kümedir. Bu kümedeki kayıp değerler iki farklı şekilde doldurularak sonuçları gözlemlenmiştir. Daha sonra öznitelik azaltımı işlemi yapılarak tahmin gücü ve performans artırılmaya çalışılmıştır. En seçici 77 özniteliğin kullanıldığı veri kümesi orijinal veri kümesinden hem performans hem de tahmin gücü açısından daha iyi olduğu anlaşılmıştır. Ayrıca aykırı değer analizi de uygulanmış ve belirli sayılarda aykırı değeri olan örnekler veri kümesinden atılmıştır. Bu işlemde bazı yöntemler belirli değerlendirme oranları için daha iyi sonuç vermiş ama iki kriter için tek bir yöntem başarıyı yakalayamamıştır. Daha sonra orijinal veri kümesi üzerinde klasik yöntemlerin parametreleri değiştirilerek denemeler yapılmıştır. Kullanılan yöntemler yapay sinir ağları(çok katmanlı algılayıcılar) ve k en yakın komşu algoritmasıdır. Ancak yine sonuç neredeyse değişmemiş ve çok düşük kalmıştır. En son Tsai'nin uygulamış olduğu model orijinal veri kümesiyle denenmiş ancak onun ulaştığı başarıya ulaşılamamıştır. Sonuç olarak bu veri kümesi üzerinde denenen tüm işlemler başarısız olmuş ve performans artırılamamıştır. Bunun nedeninin veri kümesinin tutarsız olduğu düşünülmektedir. Aykırı değer analizinde çıkan aykırı değer oranları da bu ihtimali doğrulamaktadır.

Larose veri kümesi 5000 aboneden oluşmaktadır ve 20 tane öznitelik içerir. Bu veri kümesinde kayıp değer ve önemli sayıda aykırı değeri yoktur. Ancak veri kümesinin sınıf dağılımı %14,6'ya %85,4 gibi bir oranla çok dengesizdir. Bu veri kümesinde ilk olarak öznitelik indirgeme işlemi yapılmıştır. Bu işlemin sonucunda tüm veri kümesi kullanımıyla en seçici 12 özniteliğin oluşturduğu veri kümesi arasında neredeyse fark olmadığı gözlemlenmiştir. Bu nedenle çalışma zamanı açısından daha avantajlı olan en seçici 12 özniteliğin kullanıldığı veri kümesi önerilmektedir. İkinci olarak veri kümesinin dengesiz olmasından dolayı alt örnekleme işlemi tüm veri kümesine iki farklı şekilde uygulanmıştır. Uygulanan yöntemler tekil alt örnekleme ve çoklu alt örnekleme işlemleridir. Kullanılan bu iki yöntemde başarıyı artırmıştır ancak çoklu alt örnekleme işleminin başarısı çok daha fazladır. Sonraki denemelerde çoklu alt örnekleme işleminin uygulanması ile oluşturulan alt veri kümeleri kullanılarak başarı artırılmaya devam edilmiştir. Üçüncü olarak üç farklı melez model kullanılarak başarı artırılmaya çalışılmıştır. Kullanılan birinci ve ikinci melez model klasik yöntemlerden en iyi sonucu veren rotasyon ormanı algoritmasından geri kalmıştır. Ancak üçüncü melez model düşük miktar da olsa daha iyi sonuç vermiştir. Sonuç olarak alt örnekleme işlemi (down sampling) işlemi bu veri kümesinde başarıyı artırmıştır. Özellikle çoklu alt örnekleme yöntemi önerilir. Ayrıca geliştirilen üçüncü melez model tüm yöntemlerden daha başarılıdır.

KAYNAKLAR

[1] Burez, J. ve Van den Poel, D., (2007). "Crm at a Pay-TV Company: Using Analytic Models to Reduce Customer Attrition by Targeting Marketing for Subscription Services", Expert Systems with Applications, 32: 277-288.

[2] Jamal, Z. ve Bucklin, R., (2006). "Improving the Diagnosis and Predicition of Customer Churn: A Heterogeneous Hazard Modeling Approach", Journal of Interaktif Marketing, 20.

[3] Hand, D. ve Crowder, M. (2005). "Measuring Customer Quality in Retail Banking", Statistical Modeling, 5(2): 2902-2917.

[4] Kim, H. ve Yoon, C., (2004). "Determinants of Subscriber Churn and Customer Loyalty in the Korean Mobile Tlephony Market", Telecommunications Policy, 2004. 28(9-10): 751-765.

[5] Ürkmez, İ., (2015). Müşteri Sadakati Kârlılığın Anahtarıdır, http://ilhanurkmez.com/2015/06/27/musteri-sadakati-karliligin-anahtaridir, 16 Ocak 2017.

[6] Huang, B., Kechadi, M. T. ve Buckley, B., (2012). "Customer Churn Prediction in Telecommunications", Expert Systems with Applications, 39(1): 1414-1425.

[7] Tsai, C. ve Lu, Y., (2009). "Customer Churn Prediction by Hybrid Neural Networks", Expert Systems Application, 36(10): 12547–12553.

[8] Kisioglu, P. ve Topcu, Y.I., (2010). "Applying Bayesian Belief Network Approach to Customer Churn Analysis: a Case Study on The Telecom Industry of Turkey", Expert Systems with Applications, 38: 7151–7157.

[9] Yeshwanth, V., Vimal Raj, V. ve Saravanam, M., (2011). "Evolutionary Churn Prediction in Mobile Networks Using Hybrid Learning", Proc. of XXIV Florida Artificial Intelligence Research Society Conference, 471- 476.

[10] Verbeke, W., Martens, D., Mues, C. ve Baesens, B., (2011). "Building Comprehensible Customer Churn Prediction Models with Advanced Rule Induction Techniques", Expert Systems with Applications, 38: 2354–2364.

[11] De Bock, K. W. ve Van den Poel, D., (2011). "An Empirical Evaluation of Rotation-Based Ensemble Classifiers for Customer Churn Prediction", Expert Systems with Applications, 38(10): 12293-12301.

[12] Zhao, Y., Li, B. ve Li, X., (2005). "Customer Churn Prediction Using Improved One-class Support Vector Machine", Lecture Notes in Artificial Intelligence, 3584: 300–306.

[13] Ghorbani, A., Taghiyareh, F. ve Lucas, C., (2009). "The Application of The Locally Linear Model Tree on Customer Churn Prediction", Proceedings of The International Conference of Soft Computing and Pattern Recognition, 472–477.

[14] Geppert, C., (2002). "Customer Churn Management: Retaining High-Margin Customers with Customer Relationship Management Techniques", Eine Studie der KPMG LLP, USA.

[15] Gupta, S., Lehmann, D. ve Stuart, J., (2004). "Valuing Customers", Journal of Marketing Research, 41(1): 7-18.

[16] Wei, C., ve Chiu, I., (2002). "Turning Telecommunications Call Details to Churn Prediction: A Data Mining Approach", Expert Systems with Applications, 23: 103–112.

[17] SAS International, (2001). Predicting Churn, SAS Institute, Cary, NC.

[18] Kim, H., Widdows, R. ve Park, J., (2006). "E-Loyalty: Winning Strategies for Mobile Carriers", Journal of Consumer Marketing, 23(4): 208-218.

[19] Güngör, M., Evren, G., (2010). "Mobil Numara Taşınabilirliği", Bilgi Teknolojileri ve İletişim Kurumu, Ankara.

[20] NOVAKOVIĆ, J., STRBAC, P. ve BULATOVIĆ, D., (2011). "Toward Optimal Feature Selection Using Ranking Methods and Classification Algorithms", Yugoslav Journal of Operations Research, 21: 119-135.

[21] Liu, H. ve Setiono, R., (1995). "Chi2: Feature Selection And Discretization of Numeric Attributes", Seventh International Conference on Tools with Artificial Intelligence, 388-391.

[22] Lemeshow, S. ve Hosmer, D., (2000). "Applied Logistic Regression (Wiley series in Probability and Statistics", Wiley-Interscience, 2. Sub edition, 2-4.

[23] Haykin, S., (1999). "Neural Networks, A Comprehensive Foundation", Upper Saddle River, Prentice Hall.

[24] Chaudhuri, B. B. ve Bhattacharya, U., (2000). "Efficient Training and Improved Performance of Multilayer Perceptron in Pattern Classification", Neurocomputing, 34: 11-27.

[25] Çelik, V., Adaboosts Algoritması, http://vakkascelik.com/?p=22, 19 Eylül 2015.

[26] Yılmaz, H., (2014). Random Forests Yönteminde Kayıp Veri Probleminin

İncelenmesi ve Sağlık Alanında Bir Uygulama, Yüksek Lisans Tezi, Osmangazi Üniversitesi Sağlık Bilimleri Enstitüsü Biyoistatistik Anabilim Dalı, Eskişehir.

[27] Ho , T. K., (1998). "The Random Subspace Method for Constructing Decision Forests", IEEE Transactions on Pattern Analysis and Machine Intelligence, 20(8).

[28] Amasyalı, M. F., (2013). "Sınıflandırıcı Toplulukları için Yarı Rastgele Altuzaylar", SİU, KKTC.

[29] Rodriguez, J. J., Kuncheva, L. I., ve Carlos J. A., (2006). "Rotation Forest: A New Classifier Ensemble Methods", IEEE Transactions on Pattern Analysis and Machine Intelligence, 28(10): 1619-1630.

[30] Çakır, M., (2005). Firma Başarısızlığının Dinamiklerinin Belirlenmesinde Makina Öğrenmesi Teknikleri: Amprik Uygulamalar ve Karşılaştırmalı Analiz, Uzmanlık Yeterlilik Tezi, TC Merkez Bankası İstatistik Genel Müdürlüğü, Ankara.

[31] Breiman, L., (2001). Random Forests, Machine Learning, 2001 Kluwer Academic Publishers, 45(1): 5-32.

[32] Chang, L. Y. ve Wang, H. W., (2006). "Analysis of Traffic Injury: An Application of Non-Parametric Classification Tree Techniques", Accident Analysis Prevention, 38: 1019-1027.

[33] Hébert, M., Collin-Vézina, D ., Daigneault, I., Nathalie Parent, N. ve Tremblay, C., (2006). "Factors Linked to Outcomes in Sexually Abused Girls: A Regression Tree Analysis", Coprehensive Psychiatry, 47: 443-455.

EK-A

TSAİ VERİ KÜMESİ ÖZNİTELİKLERİ VE AÇIKLAMALARI

Öznitelik Adı	Açıklama
ACTVSUBS	Evdeki aktif abone sayısı
ADULTS	Evdeki yetişkin sayısı
AGE1	İlk ev üyesinin yaşı
AGE2	İkinci ev üyesinin yaşı
AREA	Coğrafi bölge
ASL_FLAG	Hesap harcama limiti
CAR_BUY	Yeni ya da kullanılmış araç alıcısı
CARTYPE	Baskın araç yaşam tarzı
CHILDREN	Evde bulunan çocuk sayısı
CHURN	Gözlem tarihinden sonra 31-60 gün aralığında abonenin ayrılıp ayrılmadığı
CRCLSCOD	Kredi sınıfı kodu
CREDITCD	Kredi kartı göstergesi
CRTCOUNT	Bireysel kredi oranına yapılan düzeltmeler
CSA	Yerel hizmet alanı iletişimi
CUSTOMER_ID	Müşteri kimlik numarası
DIV_TYPE	Bölüm tipi kodu
DUALBAND	Çift Bant göstergesi
DWLLSIZE	Konut büyüklüğü
DWLLTYPE	Konut birim çeşidi
EDUC1	İlk ev üyesi eğitimi
ETHNIC	Etnik toplanma kodu(Hangi milletten olduğu)
FORGNTVL	Yurtdışı seyahat durumu
HND_PRICE	Kullanılan cihazın fiyatı
HHSTATIN	Ev halkının durum göstergesi
HND_WEBCAP	Cihazın webi destekleyebilmesi
INCOME	Tahmin edilen gelir
INFOBASE	İsim ve adres bilgilerinin veri tabanında olup olmaması
KID0_2	Evde 0-2 yaş arası çocuklar

KID3_5	Evde 3-5 yaş arası çocuklar
KID6_10	Evde 6-10 yaş arası çocuklar
KID11_15	Evde 11-15 yaş arası çocuklar
KID16_17	Evde 16-17 yaş arası çocuklar
LAST_SWAP	Son telefon değiştirme zamanı
LOR	Evdeki yaşam süresi
MAILFLAG	Mail kullanma durumu
MAILORDR	Posta ile yapılan alımlar
MAILRESP	Mail cevaplayıcısı
MARITAL	Evlilik durumu
MODELS	Satılan modellerin sayısı
MTRCYCLE	Motorsiklet göstergesi
NEW_CELL	Yeni telefon kullanıcısı
NUMBCARS	Bilinen araç sayısı
OCCU1	Birinci ev üyesinin mesleği
OWNRENT	Ev sahibi ya da kiracı olma durumu
PCOWNER	Bilgisayar sahibi olma durumu
PHONES	Değiştirdiği cihaz sayısı
PRE_HND_PRICE	Önceki cihaz ücreti
PRIZM_SOCIAL_ONE	Sosyal grup mesajlaşması
PROPTYPE	Mülkiyet türü detayı
REF_QTY	Toplam yönlendirme sayısı
REFURB_NEW	Cihaz: yenilenmiş ya da yeni
RV	RV(recreational vehicle) göstergesi
SOLFLAG	Telefon isteme göstergesi
TOT_ACPT	Desteklenen takımdan kabul edilen teklifler
TOT_RET	Desteklenen takıma toplam arama sayısı
TRUCK	Kamyon göstergesi
UNIQSUBS	Evdeki abone sayısı
WRKWOMAN	Evdeki çalışan kadın sayısı
ADJMOU	Müşterinin hizmet aldığı süre boyunca toplam dakika kullanımı
ADJQTY	Müşterinin hizmet aldığı süre boyunca toplam arama sayısı
ADJREV	Müşterinin hizmet aldığı süre boyunca toplam geliri
ATTEMPT_MEAN	Ortalama arama girişimi sayısı
ATTEMPT_RANGE	Arama girişimi sayısı aralığı
AVG3MOU	Önceki üç ayın ortalama aylık dakika kullanımı
AVG3QTY	Önceki üç ayın ortalama aylık arama sayısı
AVG3REV	Önceki üç ayın ortalama aylık geliri
AVG6MOU	Önceki altı ayın ortalama aylık dakika kullanımı
AVG6QTY	Önceki altı ayın ortalama aylık arama sayısı
AVG6REV	Önceki altı ayın ortalama aylık geliri
AVGMOU	Müşterinin hizmet aldığı süre boyunca ortalama aylık dakika kullanımı
AVGQTY	Müşterinin hizmet aldığı süre boyunca ortalama aylık arama sayısı

AVGREV	Müşterinin hizmet aldığı süre boyunca ortalama aylık geliri
BLCK_DAT_MEAN	Ortalama başarısız veri aramaları sayısı
BLCK_DAT_RANGE	Başarısız veri aramaları sayısı aralığı
BLCK_VCE_MEAN	Ortalama başarısız sesli aramaların sayısı
BLCK_VCE_RANGE	Başarısız sesli aramaların sayı aralığı
CALLFWDV_MEAN	Ortalama yönlendirme çağrıları sayısı
CALLFWDV_RANGE	Ortalama yönlendirme çağrıları sayı aralığı
CALLWAIT_MEAN	Ortalama bekletilen çağrı aramaları sayısı
CALLWAIT_RANGE	Ortalama beklenen çağrı aramaları sayı aralığı
CC_MOU_MEAN	Müşteri hizmetleri aramalarının ortalama yuvarlanmamış ortalama dakika kullanımı
CC_MOU_RANGE	Müşteri hizmetleri aramalarının yuvarlanmamış dakika kullanım aralığı
CCRNDMOU_MEAN	Müşteri hizmetleri aramalarının ortalama yuvarlanmış dakika kullanımı
CCRNDMOU_RANGE	Müşteri hizmetleri aramalarının yuvarlanmış dakika kullanım aralığı
CHANGE_MOU	Önceki üç aylık ortalamayla aylık kullanım dakikalarının yüzde değişimi
CHANGE_REV	Önceki üç aylık ortalamayla aylık gelirin yüzde değişimi
COMP_DAT_MEAN	Ortalama tamamlanan veri aramaları sayısı
COMP_DAT_RANGE	Tamamlanan veri aramaları sayı aralığı
COMP_VCE_MEAN	Ortalama tamamlanan sesli aramaların sayısı
COMP_VCE_RANGE	Tamamlanan sesli aramaların sayı aralığı
COMPLETE_MEAN	Ortalama tamamlanan aramaların sayısı
COMPLETE_RANGE	Tamamlanan aramaların sayısının aralığı
CUSTCARE_MEAN	Ortalama müşteri hizmetleri aramalarının sayısı
CUSTCARE_RANGE	Müşteri hizmetleri arama sayısı aralığı
DA_MEAN	Ortalama bilinmeyen numaralar servisi aramaları sayısı
DA_RANGE	Bilinmeyen numaralar servisi aramaları sayısı aralığı
DATOVR_MEAN	Ortalama veri kullanımı fazlalığı geliri
DATOVR_RANGE	Veri kullanımı fazlalığı geliri aralığı
DROP_BLK_MEAN	Ortalama bağlantı kesilmesi ya da engellenen aramaların sayısı
DROP_BLK_RANGE	Bağlantı kesilmesi ya da engellenen aramaların sayı aralığı
DROP_DAT_MEAN	Ortalama bağlantı kesilmeli veri aramaları sayısı
DROP_DAT_RANGE	Bağlantı kesilmeli veri araması sayı aralığı
DROP_VCE_MEAN	Ortalama bağlantı kesilmiş sesli aramaların sayısı
DROP_VCE_RANGE	Bağlantı kesilmeli sesli aramaların sayı aralığı
EQPDAYS	Kullanılan cihazın kullanıldığı gün sayısı
INONEMIN_MEAN	Ortalama bir dakikadan az gelen aramaların sayısı
INONEMIN_RANGE	Bir dakikadan az gelen aramaların sayı aralığı
IWYLIS_VCE_MEAN	Ortalama kablosuzdan kablosuza gelen sesli aramaların sayısı
IWYLIS_VCE_RANGE	Kablosuzdan kablosuza gelen sesli aramaların sayısının aralığı
MONTHS	Şirketten toplam hizmet alınan ay sayısı
MOU_CDAT_MEAN	Ortalama tamamlanmış veri aramalarının yuvarlanmamış dakika kullanımı
MOU_CDAT_RANGE	Tamamlanmış veri aramalarının yuvarlanmamış dakika kullanımı aralığı

MOU_CVCE_MEAN	Ortalama tamamlanmış sesli aramaların yuvarlanmamış dakika kullanımı
MOU_CVCE_RANGE	Tamamlanmış sesli aramaların yuvarlanmamış dakika kullanımı aralığı
MOU_MEAN	Aylık kullanılan dakikaların ortalama sayısı
MOU_OPKD_MEAN	İndirimli veri aramalarının yuvarlanmamış ortalama dakika kullanımı
MOU_OPKD_RANGE	İndirimli veri aramalarının yuvarlanmamış dakika kullanım aralığı
MOU_OPKV_MEAN	İndirimli sesli aramaların yuvarlanmamış ortalama dakika kullanımı
MOU_OPKV_RANGE	İndirimli sesli aramalarının yuvarlanmamış dakika kullanım aralığı
MOU_PEAD_MEAN	Ortalama yuvarlanmamış en fazla veri araması dakikası
MOU_PEAD_RANGE	Yuvarlanmamış en fazla veri araması dakika aralığı
MOU_PEAV_MEAN	Ortalama yuvarlanmamış en fazla sesli arama dakikası
MOU_PEAV_RANGE	Yuvarlanmamış en fazla veri araması dakika aralığı
MOU_RANGE	Dakika kullanımının sayı aralığı
MOU_RVCE_MEAN	Ortalama gelen sesli aramaların yuvarlanmasız dakika kullanımı
MOU_RVCE_RANGE	Gelen sesli aramaların yuvarlamasız dakika kullanım aralığı
MOUIWYLISV_MEAN	Kablosuz ağdan kablosuz ağa gelen sesli aramaların yuvarlamasız ortalama dakika kullanımı
MOUIWYLISV_RANGE	Kablosuz ağdan kablosuz ağa gelen sesli aramaların yuvarlamasız dakika kullanım aralığı
MOUOWYLISV_MEAN	Kablosuz ağdan kablosuz ağa giden sesli aramaların yuvarlamasız ortalama dakika kullanımı
MOUOWYLISV_RANGE	Kablosuz ağdan kablosuz ağa giden sesli aramaların yuvarlamasız ortalama dakika kullanım aralığı
OWYLIS_VCE_MEAN	Ortalama kablosuz ağdan kablosuz ağa giden sesli aramaların sayısı
OWYLIS_VCE_RANGE	Kablosuz ağdan kablosuz ağa giden sesli aramaların sayısının aralığı
OPK_DAT_MEAN	Ortalama indirimli veri aramalarının sayısı
OPK_DAT_RANGE	İndirimli veri aramalarının sayı aralığı
OPK_VCE_MEAN	Ortalama indirimli sesli aramaların sayısı
OPK_VCE_RANGE	İndirimli sesli aramaların sayı aralığı
OVRMOU_MEAN	Ortalama fazla kullanılan dakika
OVRMOU_RANGE	Fazla kullanılan dakika aralığı
OVRREV_MEAN	Ortalama fazla gelir
OVRREV_RANGE	Fazla gelir aralığı
PEAK_DAT_MEAN	Ortalama en fazla veri araması sayısı
PEAK_DAT_RANGE	En fazla veri arama sayısı aralığı
PEAK_VCE_MEAN	Ortalama giden ve gelen en fazla sesli arama sayısı
PEAK_VCE_RANGE	Giden ve gelen en fazla sesli arama sayısı aralığı
PLCD_DAT_MEAN	Belirli bir yerden ortalama veri araması girişimi sayısı
PLCD_DAT_RANGE	Belirli bir yerden veri araması girişimi sayısı aralığı
PLCD_VCE_MEAN	Belirli bir yerden ortalama sesli arama girişimi sayısı
PLCD_VCE_RANGE	Belirli bir yerden sesli arama girişimi sayısı aralığı
RECV_SMS_MEAN	Ortalama gelen SMS aramalarının sayısı
RECV_SMS_RANGE	Gelen SMS aramalarının sayı aralığı
RECV_VCE_MEAN	Ortalama gelen sesli aramaların sayısı
RECV_VCE_RANGE	Gelen sesli aramaların sayı aralığı

RETDAYS	Son aramadan beri geçen gün sayısı
REV_MEAN	Ortalama aylık gelir(kontör miktarı)
REV_RANGE	Gelir aralığı (kontör miktarı)
RMCALLS	Toplam gezici arama sayısı
RMMOU	Toplam gezici aramalarda kullanılan dakika
RMREV	Toplam gezici aramaların geliri
ROAM_MEAN	Gezici arama sayısı ortalaması
ROAM_RANGE	Gezici arama sayısı aralığı
THREEWAY_MEAN	Üç yönlü arama sayısı ortalaması
THREEWAY_RANGE	Üç yönlü arama sayısı aralığı
TOTCALLS	Müşterinin hizmet aldığı süre boyunca toplam arama sayısı
TOTMOU	Müşterinin hizmet aldığı süre boyunca toplam dakika kullanımı
TOTMRC_MEAN	Aylık toplam kontör yükleme ortalaması
TOTMRC_RANGE	Toplam aylık tekrar kontör yükleme aralığı
TOTREV	Toplam gelir
UNAN_DAT_MEAN	Ortalama cevapsız veri aramalarının sayısı
UNAN_DAT_RANGE	Cevapsız veri aramalarının sayı aralığı
UNAN_VCE_MEAN	Ortalama cevapsız sesli aramaların sayısı
UNAN_VCE_RANGE	Cevapsız veri aramaların sayı aralığı
VCEOVR_MEAN	Ortalama sesli arama geliri fazlalığı
VCEOVR_RANGE	Sesli arama geliri aralığı

EK-B

LAROSE VERİ KÜMESİ ÖZNİTELİKLERİ VE AÇIKLAMALARI

Öznitelik	**Açıklaması**
Churn	Müşterinin ayrılıp ayrılmadığı
state	Abonenin yaşadığı eyalet
account_length	Hesap uzunluğu
area code	Bölge kodu
phone number	Telefon numarası (ID)
international plan	Uluslar arası arama kullanıp kullanmadığı
voice mail plan	Sesli mail kullanıp kullanmadığı
number vmail messages	Sesli mail mesaj sayısı
Total_day_minutes	Günlük toplam konuşma süresi
total_day_calls	Günlük toplam arama sayısı
total_day_charge	Günlük toplam harcanan kontör
total_eve_minutes	Akşam toplam konuşma süresi
total_eve_calls	Akşam toplam konuşma sayısı
total_eve_charge	Akşam toplam harcanan kontör
total_night_minutes	Gece toplam konuşma süresi
total_night_calls	Gece toplam konuşma sayısı
total_night_charge	Gece toplam harcanan kontör
total_intl_minutes	Toplam uluslararası arama süresi
total_intl_calls	Toplam uluslararası arama sayısı
total_intl_charge	Toplam uluslararası aramalarda harcanan kontör
number_customer_service_calls	Müşteri hizmetleri arama sayısı

ÖZGEÇMİŞ

KİŞİSEL BİLGİLER

Adı Soyadı : Mümin YILDIZ

Doğum Tarihi ve Yeri : 11.10.1990 BOLU

Yabancı Dili : İngilizce

E-posta : muminyildiz@outlook.com

ÖĞRENİM DURUMU

Derece	**Alan**	**Okul/Üniversite**	**Mezuniyet Yılı**
Y. Lisans	Bilgisayar Mühendisliği	Yıldız Teknik Üniversitesi	2017
Lisans	Bilgisayar Mühendisliği	Fatih Üniversitesi	2013
Lisans	Elektrik-Elektronik Mühendisliği	Fatih Üniversitesi	2013
Lise	Fen Bilimleri	Bolu Atatürk Lisesi	2007

YAYINLARI

Bildiri

1.Yıldız, M. ve Albayrak, S., (2015). ”Telekomünikasyon Sektöründe Müşteri Ayrılma Tahmini”, 23. Sinyal İşleme ve İletişim Uygulamaları Kurultayı, 256-259.