T.C.
KÜTAHYA DUMLUPINAR ÜNİVERSİTESİ
LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ
Bilgisayar Mühendisliği Anabilim Dalı

Yüksek Lisans Tezi

MAKİNE ÖĞRENMESİ YÖNTEMLERİYLE MÜŞTERİ KAYIP ANALİZİ

Danışman:
Dr. Öğr. Üyesi Muammer AKÇAY

Hazırlayan:
Zerrin ÇAKIR

Kütahya - 2021

KABUL VE ONAY

Lisansüstü Eğitim Enstitüsü Müdürlüğüne,

Bu çalışma, jürimiz tarafından

Bilgisayar Mühendisliği Anabilim Dalında YÜKSEK LİSANS TEZİ olarak kabul edilmiştir.

Tez Jürisi	**İmza**	
	Kabul	**Red**
Dr. Öğr. Üyesi Muammer AKÇAY (Danışman)		
Prof. Dr. Alpaslan DUYSAK		
Prof. Dr. Eyyüp GÜLBANDILAR		

Onay

İmza

Prof. Dr. Şahmurat ARIK

Enstitü Müdürü

BİLİMSEL ETİK BİLDİRİMİ

Yüksek Lisans tezi olarak hazırladığım "Makine Öğrenmesi Yöntemleriyle Müşteri Kayıp Analizi" adlı çalışmanın öneri aşamasından sonuçlandığı aşamaya kadar geçen süreçte bilimsel etiğe ve akademik kurallara özenle uyduğumu, tez içindeki tüm bilgileri bilimsel ahlak ve gelenek çerçevesinde elde ettiğimi, tez yazım kurallarına uygun olarak hazırladığımı, bu çalışmamda doğrudan veya dolaylı olarak yaptığım her alıntıya kaynak gösterdiğimi ve yararlandığım eserlerin kaynakçada gösterilenlerden oluştuğunu beyan ederim.

22/04/2021

Zerrin ÇAKIR

ÖZET

MAKİNE ÖĞRENMESİ YÖNTEMLERİYLE MÜŞTERİ KAYIP ANALİZİ

ÇAKIR, Zerrin
Yüksek Lisans Tezi, Bilgisayar Mühendisliği Anabilim Dalı
Tez Danışmanı: Dr. Öğr. Üyesi Muammer AKÇAY
Nisan, 2021, 75 Sayfa

Günümüzde müşteri kaybı analizi birçok şirket için önemli bir problem haline gelmiştir. Telekom şirketleri için bu analizlerin yapılması rekabet ve pazarlama açısından önemli faydalar sağlamaktadır. Bu tez çalışmasında bir telekomünikasyon şirketine ait veri seti kullanılmıştır. Veriler çok boyutlu olduğu için boyutsallık azaltma yöntemi olan Temel Bileşen Analizi tekniği ile boyutsallık azaltılmıştır. Problem dengesiz sınıf problemi olduğu için SMOTE tekniği ile dengeli veri dağılımı olan sınıf türü oluşturulmuş ve çözüm üretilerek daha iyi bir yaklaşım elde edilmiştir. Telekomünikasyon verileri veri madenciliği yöntemleri kullanılarak veri seti sınıflandırılıp veriler arasındaki ilişki matematiksel sonuçlar üretilerek modeller oluşturulmuştur. Bu çalışmada analiz ve tahmin işlemlerinde kullanılan sınıflandırma yöntemlerinden Destek Vektör Makineleri, Karar Ağaçları, Rastgele Orman ve Yapay Sinir Ağları alg1oritmaları kullanılmıştır. Modellerin tahmin başarı ve performans sonuçları literatürde kullanılan Doğruluk, Duyarlılık, Kesinlik, F1 Score ve ROC Eğrisi metrikleriyle kıyaslanmıştır. Yapılan eğitimler ve oluşturulan modeller sonucunda SMOTE işlemi uygulanmadan önce Rastgele Orman Algoritması %85, Destek Vektör Makineleri %80, Karar Ağaçları %78 ve Yapay Sinir Ağları %86 performans sonuçları elde edilmiştir. SMOTE işlemi uygulandıktan sonra ise Rastgele Orman Algoritması %97, Destek Vektör Makineleri %96, Karar Ağaçları %93 ve Yapay Sinir Ağları %95 performans sonuçları elde edilmiştir. Bu durumda SMOTE işlemi sonrası Rastgele Orman Algoritması %12, Destek Vektör Makineleri %16, Karar Ağaçları %15 ve Yapay Sinir Ağları %9 performans artışı sağlamıştır. Çalışmalar sonucunda en yüksek performansı gösteren algoritma Rastgele Orman Algoritması olmuştur.

Anahtar Kelimeler: Destek Vektör Makineleri, Karar Ağaçları, Makine Öğrenmesi, Müşteri Kaybı, Rastgele Orman Algoritması, SMOTE, Yapay Sinir Ağları, Yapay Zekâ

ABSTRACT

CUSTOMER LOSS ANALYSIS WITH MACHINE LEARNING METHODS

ÇAKIR, Zerrin
M.S. Thesis, Department of Computer Engineering, 2021 Thesis
Supervisor: Asst. Prof. Dr. Muammer AKÇAY
April, 2021, 75 pages

Customer churn analysis has become a major problem for many companies nowadays. Conducting these analyzes for telecom companies provides significant benefits in terms of competition and marketing. In this thesis, a data set belonging to a telecommunication company was used. Since the data is multidimensional, dimensionality has been reduced with the Principal Component Analysis technique which is the dimensionality reduction method. Since the problem is an unbalanced class problem, a class type with balanced data distribution was created with the SMOTE technique and a better approach was obtained by generating a solution. The data set was classified using the telecommunication data mining methods and the models were created by producing mathematical results for the relationship between the data. In this study, Support Vector Machines, Decision Trees, Random Forest Algorithms and Artificial Neural Networks Algorithms, the classification algorithms used in analysis and prediction processes, were used. The predictive success and performance results of the models were compared with the Accuracy, Recall, Precision, F1 Score and ROC Curve metrics. As a result of the trainings and models created, the Random Forest Algorithm 85%, Support Vector Machines 80%, Decision Trees 78% and Artificial Neural Networks 86% performance results were obtained before the SMOTE process was applied. After the SMOTE process was applied, the Random Forest Algorithm 97%, Support Vector Machines 96%, Decision Trees 93% and Artificial Neural Networks 95% performance results. In this case, the results were obtained, after the SMOTE process, Random Forest Algorithm 12%, Support Vector Machines 16%, Decision Trees 15% and Artificial Neural Networks 9% performance increase. As a result of the studies achieved, the algorithm with the highest performance was the Random Forest Algorithm.

Keywords: Artificial Intelligence, Artificial Neural Networks, Churn, Decision Tree, Machine Learning, Random Forest Algorithms, SMOTE, Support Vector Machines

ÖNSÖZ

Çalışmalarım esnasında verdiği destek için ve göstermiş olduğu anlayışından dolayı saygıdeğer danışmanım Dr. Öğr. Üyesi Muammer AKÇAY'a teşekkürlerimi sunarım.

Bu çalışmada maddi ve manevi olarak destek veren eşime, anneme, babama, kardeşlerime ve çalışma arkadaşlarıma sonsuz teşekkür ederim.

İÇİNDEKİLER

TABLOLAR LİSTESİ

Sayfa

ŞEKİLLER DİZİNİ

Sayfa

KISALTMALAR

Simge	Açıklama
SMOTE	Synthetic Minority Oversampling Technique
AUC	Area Under Curve
ROC	Receiver Operating Curve
ANN	Artificial Neural Network
YSA	Yapay Sinir Ağları
DVM	Destek Vektör Makinesi
SVM	Support Vector Machine
TP	True Positive (Doğru Pozitif Değerler)
FP	False Positive (Yanlış Pozitif Değerler)
TN	True Negative (Doğru Negatif Değerler)
FN	False Negative (Yanlış Negatif Değerler)
CRISP	Çapraz Endüstri Standart Süreç Modeli
NB	Naive Bayes

GİRİŞ

Verileri işlemek ve bunlardan mantıklı sonuçlar almak Veri Madenciliğinin en önemli amaçlarından biridir. Veri Madenciliği üç önemli işlem adımına sahiptir. İlki, veri ön işleme ile ilgilidir. Gerçek Dünyada, veriler gürültülü, eksik ve tutarsızdır. Bundan dolayı veriyi iyi sonuçlar için kullanılabilir hale getirmek önemlidir. Veri ön işleme; veri temizleme, veri tamamlama, veri dönüşümü, veri azaltma ve veri ayrıştırma işlemlerini içerir. Bu adımlardan sonra veri analizi gerçekleştirilebilir. İkincisi, problem için uygun algoritmayı belirlemektir. Kullanılacak algoritma seçimi probleme bağlıdır. Doğru algoritmanın kullanılması sonuçları olumlu yönde etkiler. Kayıp tahmini için bilinen dört algoritma vardır. Bu algoritmalar Yapay Sinir Ağları, Karar Ağaçları, Destek Vektör Makineleri ve Rastgele Orman Algoritmalarıdır. Kayıp tahmini probleminin performansını değerlendirmek için genellikle Karışıklık matrisi (confusion matrix) kullanılır. Matris üzerinde doğru ve yanlış tahmin değerleri gösterilir. Bu değerlere göre performans değerlendirmeleri yapılır (Tsai ve Lu, 2010).

Telekomünikasyon şirketleri çok fazla müşteriye sahip olduğu için müşterileri hakkında çok fazla işlenmemiş veri vardır. Müşteri verileri; yaş, cinsiyet, gelir ve telefon türü, abonelik türü gibi müşterinin kişisel bilgileri içermektedir. Bu büyük miktardaki veriyi yönetmek ve onları kullanılabilir hale getirmek zordur. Bu nedenle bu şirketler için veri madenciliği gereklidir. Şirketler bu verileri analiz ederek müşteri isteklerine göre hareket edebilir. Böylece müşteri ile daha uzun çalışma fırsatı elde etmiş olurlar. Bu da, şirketlere daha iyi ve güvenilir iş devamlılığı sağlar (Tsai ve Lu, 2010). Telefon hizmeti şirketleri, internet servis sağlayıcıları, ödemeli TV şirketleri ve sigorta şirketleri sık sık müşteri kaybı analizleri kullanır. Bu analizler onların temel ticari ölçümlerinden biridir (Joshi ve Gupta, 2019). Çünkü mevcut bir müşteriyi elde tutmanın maliyeti yeni bir müşteri elde etmekten daha azdır. Bu nedenle, bu şirketler için kayıp riski olan müşterileri tahmin etmek çok önemlidir. Şirketler için kayıp riski olan müşteriler tahmin edilerek müşteri kayıp riski azaltılabilir, yeni pazarlama yöntemleri geliştirilebilir ve şirket adına fayda sağlanabilir. Bu çalışmada, bir telekomünikasyon şirketi örneği verileri kullanılarak dengesiz sınıf problemlerine çözüm üretebilmek amaçlanmıştır. Başarılı sonuçlar elde edebilmek için veri indirgemesi ve performans iyileştirmeleri yapılarak dört farklı yeni model geliştirilmiştir. İleride bu konuda yapılacak olan çalışmalara yol göstermesi

hedeflenmiştir. Tahmin işlemi yapılırken sınıflandırma algoritmalarından DVM, Karar Ağaçları, Rastgele Orman ve YSA algoritmaları kullanılmıştır. Oluşturulan modellerin sonuçları SMOTE işlemi yapılmadan önce ve yapıldıktan sonra birbirleriyle karşılaştırılmıştır. Elde edilen modellerin başarısı Doğruluk, Duyarlılık, Kesinlik, F1 Score ve ROC Eğrisi performans değerlendirme ölçütlerine göre değerlendirilmiş ve en iyi tahmin yapan model bulunmuştur. Elde edilen sonuçlar ile literatürdeki sonuçlar karşılaştırılmıştır. Tezin içeriği aşağıdaki gibi organize edilmiştir:

Birinci bölümde, literatür çalışmalarına yer verilmiştir. İkinci bölümde veri madenciliği ve veri madenciliği yöntemlerine yer verilmiştir. Uygulamada kullanılan algoritmaların temelini oluşturan Karar Ağaçları, Rastgele Orman, DVM ve YSA algoritmaları anlatılmıştır. Üçüncü bölümde dengesiz sınıf problem setinin çözümü için kullanılan araçlar anlatılmıştır. Dördüncü bölümde problem seti, algoritmaların başarı performanslarını ölçmek için kullanılan metrikler anlatılmıştır. Beşinci bölümde, veriler üzerinde yapılan analizlere yer verilmiştir. Altıncı bölümde veri analizleri görselleştirme yapılmıştır. Yedinci bölümde oluşturulan dört farklı model için dengesiz sınıf problemi üzerine çözümler yapılarak deneysel çalışma sonuçlarına yer verilmiştir. Sekizinci bölümde eğitilen dört modelin dengesiz sınıf problem setindeki başarı performansları birbirleri ile ve yapılmış çalışmalardaki sonuçlar ile karşılaştırılması yer almaktadır.

BİRİNCİ BÖLÜM

LİTERATÜR TARAMASI

1. LİTERATÜR TARAMASI

Yapılan literatür taraması sonucu Müşteri Kayıp Analizi ile ilgili yapılmış tez çalışmaları Tablo 1.1'de gösterilmiştir.

Tablo 1.1: Literatür Taraması

Yazar	Amaç	Sonuç
(Hadden vd., 2006)	YSA, Regresyon ve Karar Ağaçları olmak üzere üç farklı teknik kullanıp, kayıp tahmini için en uygun modeli belirlemeyi amaçlamaktadır.	YSA algortiması %72, Karar Ağaçları %82, Regresyon algoritması %81 doğruluk oranı vermiştir.
(Kaynar vd., 2017)	Bu çalışmada telekomünikasyon sektöründe müşteri kaybını tahmin etmek için DVM, YSA ve Naive Bayes (NB) yöntemleriyle analiz yapılmıştır.	Başarı oranı, duyarlılık ve F ölçütü değerlerinde en iyi skorlar sırası ile %91,35, %91,34 ve %90,18 olup YSA en iyi performansı vermiştir.
(Kayaalp, 2017)	Gelen veriler test verisi olarak ele alınıp geriye kalan veriler eğitim verisi olarak kullanılarak C4.5 sınıflandırma yöntemi aracılığı ile müşteri kaybı bulunmaya çalışılmıştır.	Shiny'de kullanılan yöntemler karşılaştırıldığında en iyi performans sonucunu veren Karar Ağaçları algoritması olduğu bulunmuştur.
s(Kaptan, 2019)	Hava yolu üzerine faaliyet gösteren bir firmanın verileri kullanılarak müşteri kayıp analizi yapılmıştır.	C5.0 karar ağacı modelinin %88 TPR değeri, %94 doğruluk oranı ve %88 F puanı ile en yüksek sonucu vermiştir, YSA modeli %97 TNR değeri ve %90 hassasiyet oranı ile en yüksek değerlere sahip olmuştur.
(Başarslan, 2017)	Makine öğrenmesi süreci adımlarından olan veri madenciliği üzerine çapraz endüstri standart süreç modeli (CRISP) kullanılarak müşteri kayıp analizi yapılmıştır.	Yapılan performans değerlendirmelerinde ise k-kat çapraz geçerlemedeki gibi benzer sonuç veren C4.5 karar ağacı en iyi performansı göstermiştir.
(Çelik, 2019	Telekomünikasyon sektörüne ait veriler veri madenciliği yöntemleriyle analiz edilmiştir.	Lojistik regresyon doğruluk puanı %65, K en yakın komşu modelinde doğruluk puanı %72, Gauss Naive Bayes modelinde doğruluk puanı %57, Rastgele Orman modelinde doğruluk puanı %72, Karar ağacı modelinde doğruluk puanı %56'dır.
(Donat, 2019)	Telekomünikasyon sektöründe müşteri kayıp analizi yapılmıştır. Öznitelik seçiminin modelin performansına nasıl etki ettiği incelenmiştir.	Derin Öğrenme yöntemiyle 12. ve 14. değişkenler ile en iyi Doğruluk değeri %98,70 Hassasiyet değeri ise %84,94 elde edilmiştir.

Tablo 1.1'de İncelenen kaynaklardan direkt dengesiz veri setleri üzerinde modellemelerin yapıldığı görülmektedir. Bu çalışmada dengesiz veri seti dengeli veri setine dönüştürülerek tahmin işleminde hata payını en aza indirilerek yeni modeller

oluşturulmuştur. Amaç DVM, Karar ağaçları, Rastgele Orman ve YSA algoritmalarını kullanarak eğitimler sonucunda en iyi performans veren modeli bulmak ve gerekli iyileştirme çalışmalarını yapmaktır.

İKİNCİ BÖLÜM
VERİ MADENCİLİĞİ

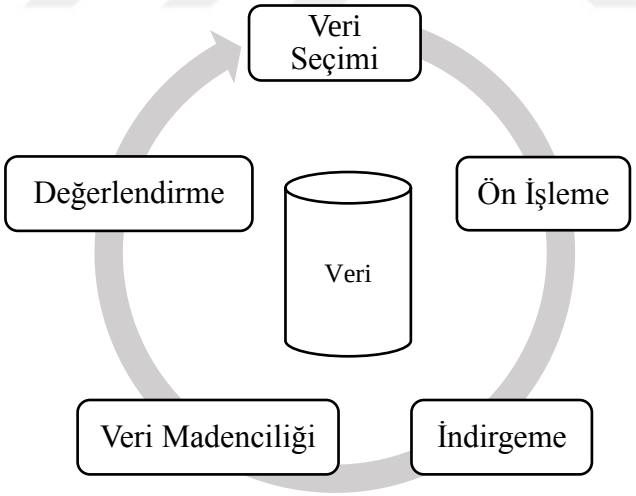
2.1.VERİ MADENCİLİĞİ TANIMI

Veri Madenciliği anlamında birçok tanımlamalar bulunmaktadır. Genel olarak işlenmemiş ham halde olan kayıtlar veri olarak tanımlanmaktadır. Veriler veri tabanı adı verilen depolama yerlerinde depolanırlar. Veri tabanı, verilerin birbirleriyle ilişkili şekilde depolayan alanlardır. Veri tabanına eklenen veriler daha da artarak büyük veriyi oluşturmaktadır. Elimizde var olan verileri ayrıştırmak, karşılaştırmak, sınıflandırıp analiz etmek için çeşitli tekniklerin kullanılması gerekmektedir (Akbulut, 2016).

Temel olarak veri madenciliği, büyük verilerin birbirleriyle olan ilişkilerini inceleyip analiz ederek, daha anlamlı veriler haline getirmek anlamına gelmektedir. Bu yöntemler, toplumun hemen hemen her alanda gelişmesine katkı sağlamaktadır.

2.2. VERİ MADENCİLİĞİ AŞAMALARI

Veri madenciliği çeşitli süreçlerden oluşmaktadır. Bu süreçler Şekil 2.1'de gösterilmiştir.



Şekil 2.1: Veri Madenciliği Aşamaları

Bu süreçler veri seçimi (data selection), ön işleme (preprocessing), indirgeme (data reduction), veri madenciliği (data mining), değerlendirme (evoluation) aşamalarından oluşmaktadır (Albayrak, 2008).

2.2.1. Veri Seçimi

Veri seçimi, veri madenciliği aşamalarından en önemli olan kısımlarından biridir. Öncelikle süreçlerin içerisinde uygulanacak olan model için verilerin iyi bir şekilde bilinmesi gerekmektedir. Bu yüzden süreç veriyi hazırlama safhasıyla başlamaktadır (Ali, 2017).

2.2.2. Ön İşleme

Veri Madenciliğinde kullanılacak modellerde daha başarılı sonuçlara ulaşabilmek için gerekli olan aşama ön işleme aşamasıdır. Verileri kullanabilmek için gerekli düzenlemeler bu aşamada yapılmaktadır (Koyuncugil, 2009). Bu aşamada veri azaltma, veri bir araya getirme, veri dönüştürme, ve veri temizleme işlemleri yapılarak, veriler analize uygun duruma getirilir (Gaber vd., 2005). Farklı algoritmalarda farklı işlemler yapmak gerekebilir.

2.2.3. İndirgeme

Veri seti her ne kadar ön işlemeden geçse de bir sonraki aşamalarda kullanmak için indirgeme işleminin yapılması gerekmektedir.

2.2.4. Veri Madenciliği

Veri seti bu aşamaya geldiğinde kullanıma hazırdır. Bu aşamada verinin durumuna göre en uygun olan metodun seçildiği aşamadır (Vadim, 2018). Öğrenme kümesi ve test kümesi bu aşamada seçilir. Bu amaçları gerçekleştirmek için de birden fazla yöntem denenerek analiz yapılıp, performanslar karşılaştırılabilir.

2.2.5. Değerlendirme

Veri madenciliğinin son aşamasıdır. Beklenilen çıktıları elde edebilmek amacıyla yeni verilerle çıkarım ve tahmin işlemleri yapılır. Belirlenen metotlar uygulandıktan sonra sonuçlar karşılaştırılarak en iyi sonuç çıkarımı yapılabilir.

2.3. VERİ MADENCİLİĞİ YÖNTEMLERİ

Bu çalışmada Veri Madenciliği yöntemlerinden 4 farklı algoritma kullanılarak yeni modeller oluşturulmuştur. Bu algoritmalar; Destek Vektör Makineleri, Rastgele Orman, Yapay Sinir Ağları ve Karar Ağaçları Algoritmalarıdır.

2.3.1. Destek Vektör Makineleri (DVM)

DVM beklenen genelleme hatasının üst sınırını azaltmak için ayırıcı hiper düzlem ile örnekler arasında mümkün olan en büyük mesafeyi oluşturarak, maksimize etmeye dayanan en güçlü denetimli makine öğrenme tekniklerinden biridir (Kim vd, 2005). Eğitim veri setindeki bazı örnekler, ayırıcı hiper düzlemine yakın olurlar ve sınıflandırma için en faydalı bilgileri sağlarlar. Bunlara destek vektörleri denir. DVM, Vladimir Vapnik ve Alexey Chervonenkis tarafından 1963 senesinde temelleri atılmıştır. DVM gözetimli öğrenme algoritmasıdır. İstatiksel öğrenme teorisine dayanır ve 1995 senesinde Vladimir Vapnik ve Alexey Chervonenkis ve Isabelle Guyon tarafından geliştirilmiştir (Akpınar, 2014). Amaç, sınıfları birbirinden ayıracak olan hiper düzlemin çıkarılmasıdır. Böylece destek vektörleri arasındaki uzaklık maksimize edilmiş olabilecektir.

Veri setinin geri kalan öznitelik vektörleri, karar sınırının konumunun belirlenmesinde herhangi bir etkiye sahip değildir. Ek olarak, DVM'de doğrusal ayrımcılık işlevlerini kullanmak amacıyla verileri yüksek bir boyuta dönüştürmek için uygun bir fonksiyon kullanılır (Ali, 2017). Başlıca Destek Vektör Makinesi kütüphaneleri şunlardır:

- libDVM
- DVM-Light
- DVMTorch

Önce veriler incelenip gerek olmayan ya da az etkileyen özniteliklerin çıkarılması gerekmektedir. Böylece veriler indirgenerek doğru sonucu elde etme imkanı daha da arttırılmaktadır. Bu işleme veri indirgeme adı verilmektedir. Sınıflandırma veya regresyon problemleri için kullanılmaktadır. Uygulama alanları şu şekildedir (Kim vd., 2005):

- Metin sınıflandırmalarında
- Görüntü sınıflandırmalarında
- Genetik ve diğer bilim dalları

Destek Vektör Makineleri Algoritması avantajları şu şekilde sıralanabilir (Guo-en, Wei-dong 2008):

- Yüksek karmaşık yapılarda etkilidir.
- Çok yönlüdür.
- Sınıflandırma problemini kareli en uygun (optimizasyon) türüne dönüştürüp çözebilir.
- Daha hızlı çözüm üretebilir
- Büyük verilerde uygulanabilmektedir.
- Uygulaması kolaydır.
- Aşırı yükleme sorunu yoktur.

Destek Vektör Makineleri Algoritması dezavantajları şu şekilde sıralanabilir (Ilgın vd., 2013):

- Sonuçları kesindir, belirsiz değerler üretemez.
- Fonksiyonlar negatif olmayan fonksiyonlar olmalıdır.

2.3.2. Rastgele Orman Algoritması

Rastgele Orman Algoritması, Brieman'ın 1996 da geliştirdiği Bagging tekniği ile Kimhon'un geliştirdiği tekniklerinin birleşimi olarak ortaya çıkmıştır. 2001 yılında Leo Breiman tarafından geliştirilmiştir (Gupta vd., 2018). Hem sınıflandırma hem de regresyon sorunlarını çözmek için tahmin modelleri oluşturmayı amaçlar. Topluluk (ensemble) yöntemleri daha iyi tahmin sonuçları elde etmek için birçok öğrenme modellerini kullanır. Bir Rastgele Orman modeli söz konusu olduğunda, model mümkün olan en iyi cevaba ulaşmak için rastgele ilişkisiz Karar Ağaçlarını içeren bir orman oluşturur (Xie, 2009).

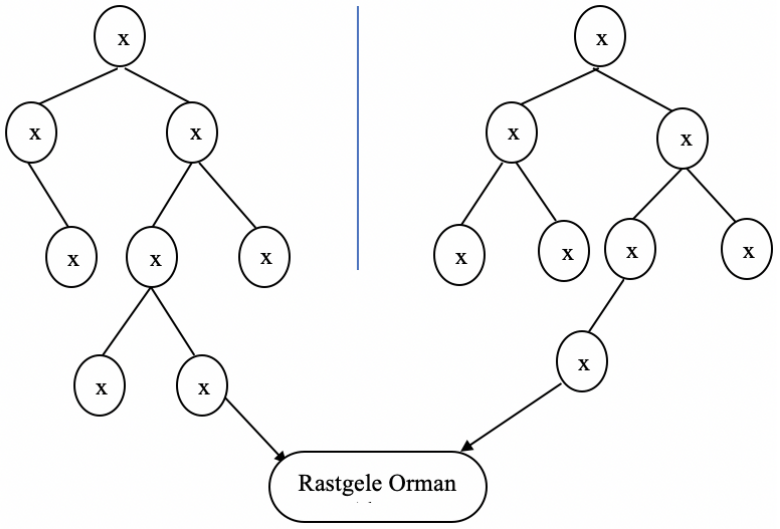
Amacı veri setinde bulunan bağımlı ve bağımsız özniteliklere bakarak kurallar oluşturup bu kuralları öğrenerek bir çıktı etiket değerini tahmin eden modeli meydana getirmektir. Hem regresyon hem sınıflandırma işlemlerinde kullanılmaktadır. Rastgele Orman Algoritması avantajları şu şekilde sıralanabilir (Wadikar, 2020):

- Kullanılan uygulamalarda ortaya çıkabilecek olan uyumsuzlukları önler.
- Hem sınıflandırma hem de regresyon konularında kullanılır.
- Ormandaki ağaç sayısı arttıkça hata payı azalır ve gerçeğe daha uygun sonuçlar verir.
- Veri setinde kullanılan öznitelikler arasından en önemli özniteliğin çıkarılmasını sağlar.
- Cevap verme süresi diğer algoritmalara bakılarak daha kısadır.

Rastgele Orman dezavantajları şu şekilde sıralanabilir:

- Hazırlamak için zaman kaybı yaratır.
- İyi bir korelasyon değeri göstermeyebilir.

Rastgele Orman algoritmasının çalışma mantığı Şekil 2.2'de aşağıda gösterilmiştir.



Şekil 2.2: Rastgele Orman Çalışma Mantığı (Akman, 2010)

- Veri setinde k tane özellik seçilir.
- Öğrenmek üzere kullanılacak olan verileri, gerçek veri seti içerisinden 2/3 olacak şekilde ayrılır. Bu işlem inBag olarak adlandırılmaktadır.
- Öğrenmek üzere ayrılan bu veri seti içinde bulunan özniteliklerden rastgele şekilde m tane özellik belirlenir.
- Belirlenen bu özelliklerden en iyi performans verecek özellikler bulunur.
- Sınıflandırma yapmak için:
- Rastgele seçilen öznitelik = $(\text{Toplam bağımsız özellikler})^{1/2}$
- Regresyon için:

- Rastgele seçilen öznitelik = (Toplam bağımsız özellikler/3)
- Kaç ağaç oluşturmak isteniyorsa bir önceki maddeden önceki adımlar tekrar gerçekleşir.
- Test verileri modele uygulanır.
- Ayrılan verilerden hata oranı ve Doğruluk oranları hesaplanır.

2.3.3. Yapay Sinir Ağları (YSA)

YSA, biyolojik sinir ağlarından ilham alan matematiksel bir modeldir. Birbirine bağlı yapay nöronlar grubundan oluşur ve bilgileri ilişkisel bir yaklaşım kullanılarak hesaplanır (Pendharkar, 2009). Nöronlar, YSA'ları oluşturan ağırlıklandırılmış şekilde birbirlerine bağlanmış sinir hücreleridir. Bu işlem birimindeki nöronlar sinyalleri alır, bunları birleştirir, dönüştürür ve sayısal bir sonuç ortaya çıkartır.

Çoğu durumda bir Yapay Sinir Ağı öğrenme aşamasında yapısını değiştiren uyarlanabilir bir sistemdir. Giriş ve çıkışlar arasındaki karmaşık ilişkileri modellemek ya da verilerdeki örüntüleri bulmak için kullanılır (Josji, 2019). YSA öngörü modellemelerinde, kontrol ve sistem tanımlama, görüntü işlemleri, tahmin yapma, tıp, haberleşme ve üretim yönetimi gibi pek çok dalda kullanılmaktadır (Pirim, 2006). YSA algoritmasının avantajları şu şekilde sıralanabilir (Faris, 2018):

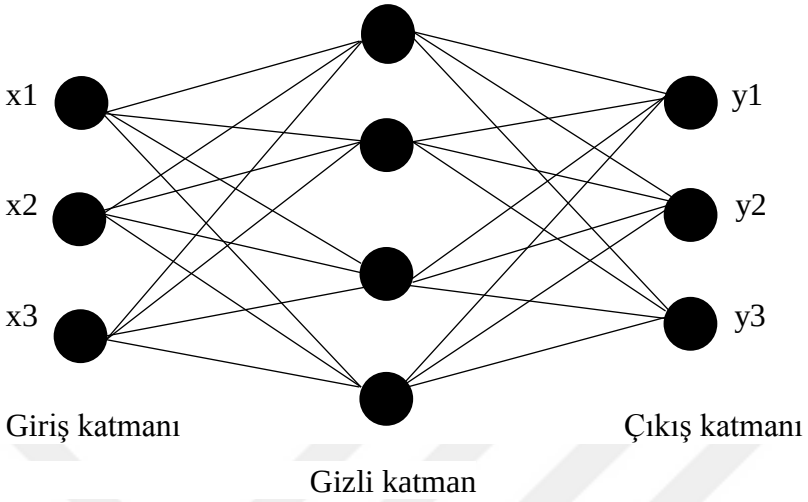
- YSA'lar eğitildikten sonra veri setinde eksik olsa dahi, sonuç üretebilirler.
- Öğrenerek yorum yaparak karar verebilmektedirler.
- Eş zamanlı işlemler yapabilirler.
- Dağıtık hafızaya sahiptirler.

YSA algoritmasının dezavantajları şu şekilde sıralanabilir (Faris, 2018):

- Hata toleransı vardır.
- Eğitim süresi bilinmez.
- Yavaş ve donanıma bağımlıdır.
- Ağa senaryo her haliyle gösterilemezse ağ doğru olmayan çıktılar verebilir.

Bir YSA hücre yapısı; girdiler, ağırlıklar, toplama fonksiyonu, aktivasyon fonksiyonu ve çıktılardan oluşmaktadır (Lu, 2009). Her hücre birbirine ağrılıklarla bağlıdır. Bu ağırlıklar iki hücre arasındaki bağlantı gücünü göstermektedir. YSA oluşturulurken, ağ içerisinde kaç adet katman olacağı, kaç işlemin gerçekleşeceği ve

katmanlar arası ağırlıkların verilmesi gerekmektedir. Bazı ağırlıklar 0 da olabilmektedir. YSA yapısı Şekil 2.3'te gösterilmiştir (Sharma ve Panigrahi, 2011).

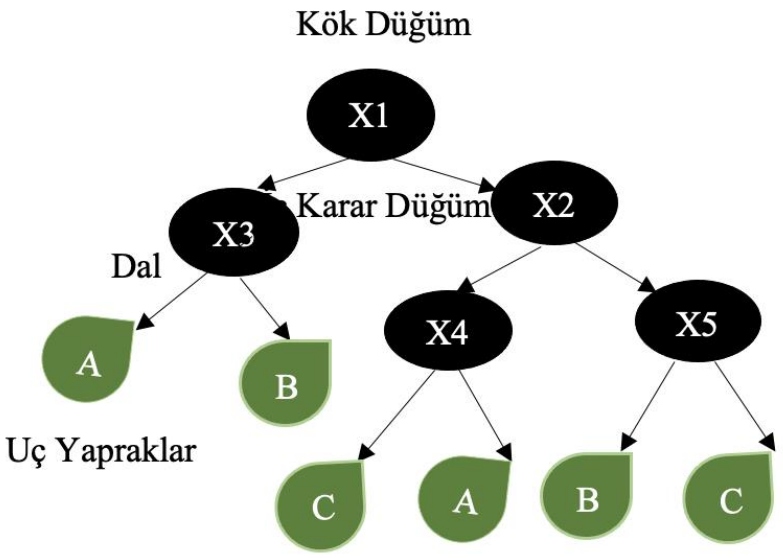


Şekil 2.3: YSA Yapısı

Şekil 2.3'te gösterilen modelin x1, x2, x3 ile gösterilen 3 tane girdi katmanı vardır. Bu giriş değerleri katmanlar arasındaki ağırlıklara bağlıdır. y1, y2, y3 ise çıkış katmanlarıdır.

2.3.5. Karar Ağaçları

Karar ağacı, ikili karar ağaçları oluşturan Sınıflandırma Ağaçları ve Regresyon Ağaçları metotları kullanılarak geliştirilmiştir. Ağaç tabanlı denetimli öğrenme yöntemlerinden biridir. Tekrar tekrar iki alt düğüm oluşturmak için tüm tahmin edici öznitelikleri kullanarak veri kümesinin alt kümelerini bölerek oluşturulur (Hadden vd., 2006). En üstteki karar düğümüne kök düğüm de denilmektedir. İki veya daha fazla kollara ayrılabilir. Amaç, hedef etikete göre olabildiğince homojen olan veri alt kümeleri üretmektir (Pedregosa vd., 2011). Karar Ağaçları veri madenciliği algoritmaları içerisinde en bilinen algoritmadır. Karar ağaç yapısı Şekil 2.4. te gösterilmiştir.



Şekil 2.4: Karar Ağaçları Yapısı

Yukarıdaki şekil 2.4'te gösterildiği gibi karar ağacı; kök, iç karar düğümleri, dal ve uç yapraklardan meydana gelmektedir. Verilerin test edildiği ve dalların hangi yöne eğileceklerini belirleyen iç karar düğümleridir. Eğilmesine yön verecek olan sorulara cevap verecek olan dallardır. Bulunduğu etiket ise uç yapraklar olarak tanımlanmaktadır.

Ağacın karmaşıklığı, bu ağacın doğruluğuyla da ilişkilidir. Ağaç karmaşıklığı şu etmenlere bağlıdır (Pedregosa vd., 2011):

- Toplam düğüm sayısı
- Toplam dal sayısı
- Ağaç derinliği
- Öznitelik sayısı

Veri madenciliği teknikleri aşırı uyum (overfitting) karşılama ihtimali olduğu için ağacın budanması gerekmektedir. Karışıklık parametresi hesaplanıp, fazla karmaşıklığı olan alt ağaçlar budanmaktadır (Çınar, 2019). En yaygın bilinen algoritmalar şu şekildedir:

- Cart
- Chaid
- C5.4
- IDC

Karar Ağaçları Algoritması avantajları şu şekilde sıralanabilir (Chen, 2017):

- Doğruluğu yüksektir.
- Yüksek kararlılık ve yorumlama anlamında kolaylık sağlar.
- Doğrusal olmayan ilişkileri iyidir.
- Hem kategorik hem sayısal verilerle çalışabilir
- Düşük maliyetlidir.

Karar Ağaçları Algoritması dezavantajları şu şekilde sıralanabilir (Chen, 2017):

- Aşırı uyum sorunu çıkabilir. Bu yüzden budama yapılması gerekmektedir.
- Karmaşıklığı yüksek olan büyük ağacın anlaşılması ve yorumlanması güçtür.

ÜÇÜNCÜ BÖLÜM

MATERYAL VE YÖNTEM

3.1.KULLANILAN ARAÇLAR

Uygulama, phyton proglama dilinde yazılmıştır. Uygulamayı yazmak için Anaconda Navigator (Anaconda Navigator, 2021) içerisinde bulunan Jupyter Notebook (Jupyter Notebook, 2021) kullanılmıştır. Veri işleme işlemleri için Scikit learn (Scikit-Learn, 2021) ve pandas (Pandas, 2021) kütüphaneleri kullanılmıştır. Çizim ve görüntüleme işlemleri için matplotlib (Matplotlib, 2021) kütüphanesi kullanılmıştır.

3.1.1.Python Programlama Dili

- Makine öğrenmesi, veri bilimi ve yapay zekâ için kullanılan açık kaynaklı bir programlama dilidir.
- Modüler ve nesne yönelimli olduğundan kullanımı kolaydır.
- Her türlü platformda çalışabilmektedir.
- Sistem programlama, kullanıcı arabirimi programlama, ağ programlama, web programlama, uygulama ve veri tabanı yazılımı programlama gibi birçok alanda kullanılmaktadır.

3.1.2. Scikit-learn Kütüphanesi

- Python programlama dili, bilimsel bilgi işleme alanındaki en popüler programlama dillerinden biridir.
- Scikit-learn, Python programlama dili için yazılmış ücretsiz bir makine öğrenmesi kütüphanesidir.
- Scikit-learn, biçok iyi bilinen makine öğrenmesi algoritmasının en gelişmiş uygulamalarını sunmaktadır.
- Veri analitiği için sunduğu seçenekler sayesinde veri işleme Scikit-learn kütüphanesi ile yapılabilmektedir (Blondel vd., 2011).

3.1.3. Pandas Kütüphanesi

- Pandas, yapılandırılmış verilerle çalışmayı hızlı, kolay ve etkileyici yapmak için tasarlanmış zengin veri yapıları ve işlevler sunmaktadır.
- Python'u güçlü ve üretken bir veri analizi ortamı yapan kritik bileşenlerden biridir.

- Numpy'ın yeterli kalmadığı alanlarda çözüm üretmektedir (Numpy, 2021).
- Yeniden şekillendirmeyi, paylaştırmayı, kümelemeyi kolaylaştıran ve veri alt kümelerini seçen gelişmiş bir indeksleme işlevi sunmaktadır (Barrett vd., 2014).

3.1.4. Matplotlib Kütüphanesi

- Matplotlib öncelikle bilimsel, mühendislik ve finansal verileri görselleştirmeyi amaçlayan taşınabilir bir 2D çizim ve görüntüleme paketidir.
- Python'ın çok boyutlu dizi kütüphanelerine bütünleşmiş (entegre) bir şekilde çalışabilmektedir.
- Programlamayı ve geliştirmeyi kolaylaştıran tamamen nesne odaklı bir tasarım sağlamaktadır.
- Tarihsel ve finansal çizimler yapılabilmektedir.
- W3C (World Wide Web Consortium) uyumlu yazı tipi yönetimi kullanılmaktadır (Bowyer vd., 2002).

3.1.5. Jupyter Notebook

- Eskiden IPython olarak bilinen, her şeyi birbirine bağlayan standart bilimsel Python araç setidir.
- Bilgi işlem süreci için sağlam ve üretken bir ortam sağlar.
- Python kodunun yazılmasını, test edilmesini ve hata ayıklanmasını hızlandırmak için geliştirilen bir Python kabuğudur.
- Veri ile etkileşimde bulunmak ve verileri Matplotlib ile görselleştirmek için kullanışlıdır (Barrett vd., 2014).

3.1.6. SMOTE

SMOTE algoritması, sentetik azınlık örnekler üretmeye dayanan bir aşırı örnekleme yöntemidir. Açılımı Synthetic Minority Over-Sampling Technique 'dir (Bowyer vd., 2002). Makine öğrenmesi akıllı sistem çalışmalarından biri olup, makinelerin kendi kendine karar verebilme fikrinden ortaya çıkmıştır. Makinelerin öğrenme aşamaları iki temel yöntemden oluşmaktadır. Bunlar danışmanlı öğrenme ve danışmansız öğrenmedir. Danışmalı öğrenmede makinenin öğrenmesi için bir danışmana ihtiyacı vardır (Janabi, 2020). Danışmansız öğrenmede makineye öğrenme

sırasında sadece örnek parametreler verilmektedir. Sistemin kendi kendine öğrenip gruplaması beklenmektedir. Danışmanlı öğrenme yöntemlerinin kullanılması kullanılan algoritma parametreleri ve veri seti içinde oldukça önem arz etmektedir. SMOTE yönteminin özellikleri şu şekilde özetlenebilir (Müslim, 2020):

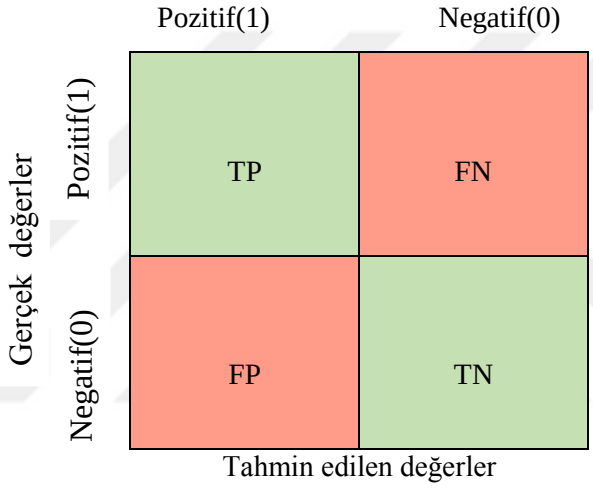
- SMOTE metodu dengesiz sınıf problemine çözüm olarak kullanılmaktadır.
- Örneklemeye dayanan bir yaklaşımdır.
- Rasgele aşırı örneklemeden farklı olarak sentetik örnekler oluşturulur.
- Rastgele aşırı örneklemede mevcut örneklerin kopyaları oluşturulur ve sınıflar bu şekilde dengelenmeye çalışılır.
- SMOTE yaklaşımında azınlık sınıfının her bir örneğini alınır ve her bir örnek için en yakın k komşusunun birleştirilmesiyle oluşturulan hat üzerinde yapay örnekler oluşturulur.
- Azınlık sınıfına ait örnek sayısı arttırılır ve sınıflar dengelenir.

DÖRDÜNCÜ BÖLÜM
PERFORMANS DEĞERLENDİRME METRİKLERİ

Performans değerlendirme metrikleri; Karışıklık matrisi, Duyarlılık, Kesinlik, Doğruluk ve ROC eğrisinden oluşmaktadır (Odabas, 2017).

4.1. KARIŞIKLIK MATRİSİ (CONFUSİON MATRIX)

Karışıklık matrisi, hata oranı olarak tanımlanmaktadır. Veri setinde bulunan sonuçların ve sınıflandırma modelinin doğru ve yanlış tahmin sayısının tablo olarak Şekil 4.1'de gösterilmiştir. İki veya daha fazla çıktının performansını ölçmek için kullanılmaktadır (Hamalainen vd., 2018).



Şekil 4.1: Karışıklık Matrisi

Karışıklık matris ve parametreleri Şekil 4.1'de gösterilmiştir. Parametre açıklamaları şu şekildedir (Yıldız, 2015);

TP: Gerçek değer ve tahmin edilen değer 1 olan değerlerdir.

TN: Gerçek değer ve tahmin edilen değer 0 ise olan değerlerdir.

FP: Gerçek değer 0 olup tahmin edilen değer 1 olan değerlerdir.

FN: Gerçek değer 1 olup tahmin edilen değer 0 olan değerlerdir.

Karışıklık matrisini bulmanın en önemli avantajı Duyarlılık, Kesinlik, Doğruluk, F1 Score ve AUC-ROC eğrisini ölçüp sonuç değerlerinin bulunmasını sağlamaktır.

4.1.1 Duyarlılık (Recall)

Pozitif durumların başarı derecesini göstermektedir. Formülü şu şekildedir:

$$\text{Duyarlılık} = TP / (TP + FN) \quad (4.1)$$

4.1.2 Kesinlik (Precision)

Doğru tahmin sonuç veren örneklerdeki başarı değerini gösteren durumdur. Formülü şu şekildedir:

$$\text{Kesinlik} = TP / (TP + FP) \quad (4.2)$$

4.1.3 Doğruluk (Accuracy)

Doğru sonuç bulunan tahmin değerlerinin hepsine oranı şeklinden tanımlanmaktadır. Formülü şu şekildedir:

$$\text{Doğruluk} = (TP + TN) / (FP + FN + TP + TN) \quad (4.3)$$

4.1.4 F1 Score

F1 Score, Precision ve Recall değerlerinin harmonik bir ortalamasıdır. F1 Score değerinin kullanılmasının en önemli nedeni eşit dağılmayan veri setlerinde doğru modelin seçilmesine yardımcı olmasıdır. F1 Score sadece FN ya da FP değerlerini değil tüm hata seçeneklerini de içeren bir ölçme metriğidir. Bu yüzden en önemli metriklerden biridir.

Formülü şu şekildedir:

$$\text{F1 Score} = 2 / [(1 / \text{Recall}) + (1 / \text{Precision})] \quad (4.4)$$

4.1.5 ROC Eğrisi

İlk 1971 yılında Lee Lusted tarafından tanı testlerinin geliştirilebileceğini düşündüğü için kullanılmıştır. Kullanılan modellerin performansını ölçmek için kullanılan grafiklerden biridir (Köksal, 2011). Özellikle verileri dengesiz olan veri setlerinde kullanılmaktadır. Modelin ne kadar başarılı olup olmadığı sonucunu göstermektedir. Farklı modelleri bir eğride karşılaştırıp yorumlanmasına yardımcı olmaktadır. Analiz sonucunda AUC değeri bulunur. AUC değeri eğri altında kalan alanı ifade eder. Modelin yanlış pozitif oranı x ekseninde, doğru pozitif oranını ise y eksenindedir (Bek vd., 2010). Eğri ne kadar 1 e yakınsa model o kadar başarılı sonuç vermiş demektedir. %100 lük bir başarı durumunda ise tahmin 1'e eşit olmaktadır.

BEŞİNCİ BÖLÜM
VERİ SETİ

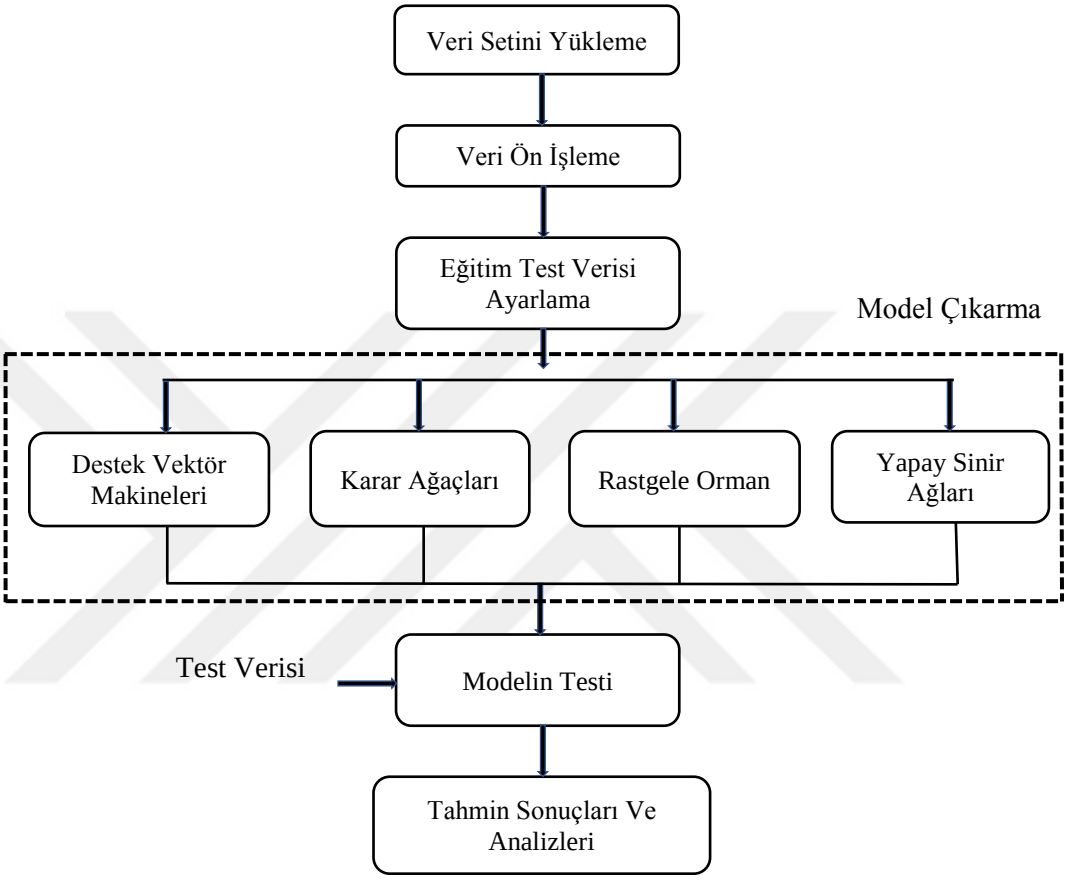
Veri seti, çeşitli değişkenlerden ve bu değişkenlerin değerlerinden oluşan ikili bir matristir. Kullanılacak olan veri seti öznitelik ve açıklamaları Tablo 5.1'de gösterilmiştir (PangKW, 2018).

Tablo 5.1: Veri Seti Öznitelikleri Ve Açıklamaları

Alan Adları	**Alan Açıklamaları**
customerID	Müşteri kimliği
Gender	Müşterinin cinsiyeti
SeniorCitizen	Yaşlı olup olmadığı
Partner	Müşterinin ortağı olup olmadığı
Dependents	Bakmakla yükümlü olduğu kişi sayısı
Tenure	Kullanım ayı süresi
PhoneService	Telefon hizmetinin olup olmadığı
MultipleLines	Birden fazla hattının olup olmadığı
InternetService	Müşterinin internet servis sağlayıcısı
OnlineSecurity	Çevrimiçi güvenliği olup olmadığı
OnlineBackup	Çevrimiçi yedeklemesi olup olmadığı
DeviceProtection	Cihaz korumasına sahip olup olmadığı
TechSupport	Müşterinin teknik desteği olup olmadığı
StreamingTV	Müşterinin TV yayını olup olmadığı
StreamingMovies	Müşterinin film akışı olup olmadığı
Contract	Müşterinin sözleşme süresi
PaperlessBilling	Kâğıtsız faturalandırması olup olmadığı
PaymentMethod	Müşterinin ödeme yöntemi
Churn	Müşterinin geri dönüp dönmediği
MaritalStatus	Medeni hal
InternationalPlan	Uluslararası plan
VoiceMailPlan	Sesli posta planı
NumbervMailMesaage	Mesaj numarası
TotalDayMinutes	Günlük toplam dakika
TotalDayCalls	Günlük toplam aramalar
TotalEveMinutes	Akşamları toplam dakika
TotalNightCalls	Gece aramalar toplamı
TotalintlMinutes	Toplam uluslararası dakika
TotalIntlCalls	Toplam uluslararası aramalar
CustomerServiceCalls	Müşteri hizmetleri çağrıları
TotalCall	Toplam çağrı
TotalRevenue	Toplam gelir

İlk olarak veri seti yüklenerek ve bir takım ön işlemlerden geçirilmesi gerekmektedir. Verinin ön işlemlerden geçmesi daha iyi sonuçlar alınmasını sağlamaktadır. Hazır hale getirilen veri seti eğitim ve test olmak üzere ayrılacaktır.

Eğitim verisi kullanılarak DVM, Karar Ağaçları, Rastgele Orman ve YSA kullanılarak eğitim yapılacak ve modeller oluşturulmuştur. Her bir model test verisi ile test edilecek ve Karışıklık matrisi, Doğruluk değeri, Kesinlik değeri, Duyarlılık değeri ve F1 Score ölçüm metrikleri ile değerlendirilecektir. Şekil 5.1 de çalışmanın akış şeması gösterilmektedir.



Şekil 5.1: Veri Akış Şeması

Veri kümesi 3332 müşteriye ait 33 öznitelikten (32 öznitelik + 1 label) oluşmaktadır. Veri kümesindeki öznitelikler belli gruplar altında incelenebilir. Veri setinde sınıf etiketi "Churn" dur. Korelasyonlara bakılarak veya görselleştirme yapılarak, öznitelikler arasındaki ilişki görülebilir. Veri setinde genel analiz yapılmıştır ve analizler sonucu bazı çıkarımlar elde edilmiştir.

5.1. KORELASYON ANALİZİ

Veri seti ilk öncelikle eğitime hazır hale getirilmiştir. Özniteliklerin bir kısmı kategorik verilere sahiptir. Ancak bu kategorik veriler metin şeklindedir. Metin değerlere sahip öznitelikler olduğu için şifreleme (encoding) işlemi uygulanmıştır ve veriler numerik değerlere sahip bir şekilde kategorik hale getirilmiştir. Bu işleme şifreleme (encoding) adı verilmektedir. Her kategori için yeni bir sütun oluşturulmuştur. Örneğin ait olduğu kategori değeri 1 olurken diğer kategoriler 0 değerini almaktadır. Bu şekilde toplam 40 sütun (39 öznitelik + 1 etiket) elde edilmiştir. Korelasyon analizi, öznitelikler arasındaki ilişkinin yönü, ilişki derecesi ve önem derecesini gösteren matematiksel yöntemdir. Bu ilişkinin derecesini ve yönünü gösteren katsayıya da korelasyon katsayısı adı verilmektedir. Sütun sayısı çok fazla olduğundan boyutsallık artacaktır. Bu yüzden korelasyon değerleri incelenerek hedef etiket ve öznitelikler arasındaki ilişki incelenmiştir. Korelasyon değerleri Tablo 5.2'de gösterilmektedir (Alsakran vd, 2014).

Tablo 5. 2: Veri Seti Korelasyon Değerleri

Öznitelik Adı	**Korelasyon değeri**
Churn	1,00
SeniorCitizen	0,110
tenure	-0,221
TotalDayCalls	0,018
TotalEveCalls	0,009
TotalNightMinutes	0,035
TotalNightCalls	0,005
TotalCall	0.023
Gender	-0,023
MaritalStatus	-0,122
Dependents	-0,107
OnlineSecurity_Yes	-0,101
MultipleLines	-0,131
InternetService_DSL	-0,046
InternetService_No	-0,211
OnlineSecurity_No	0,293
OnlineSecurity_No internet service	-0,211
OnlineBackup_No	0,253
OnlineBackup_No internet service	-0,211
OnlineBackup_Yes	-0,051
DeviceProtection_Yes	-0,047
TechSupport_No	0,294
DeviceProtection_No	0,248
DeviceProtection_No internet service	-0,211
TechSupport_No internet service	-0,211

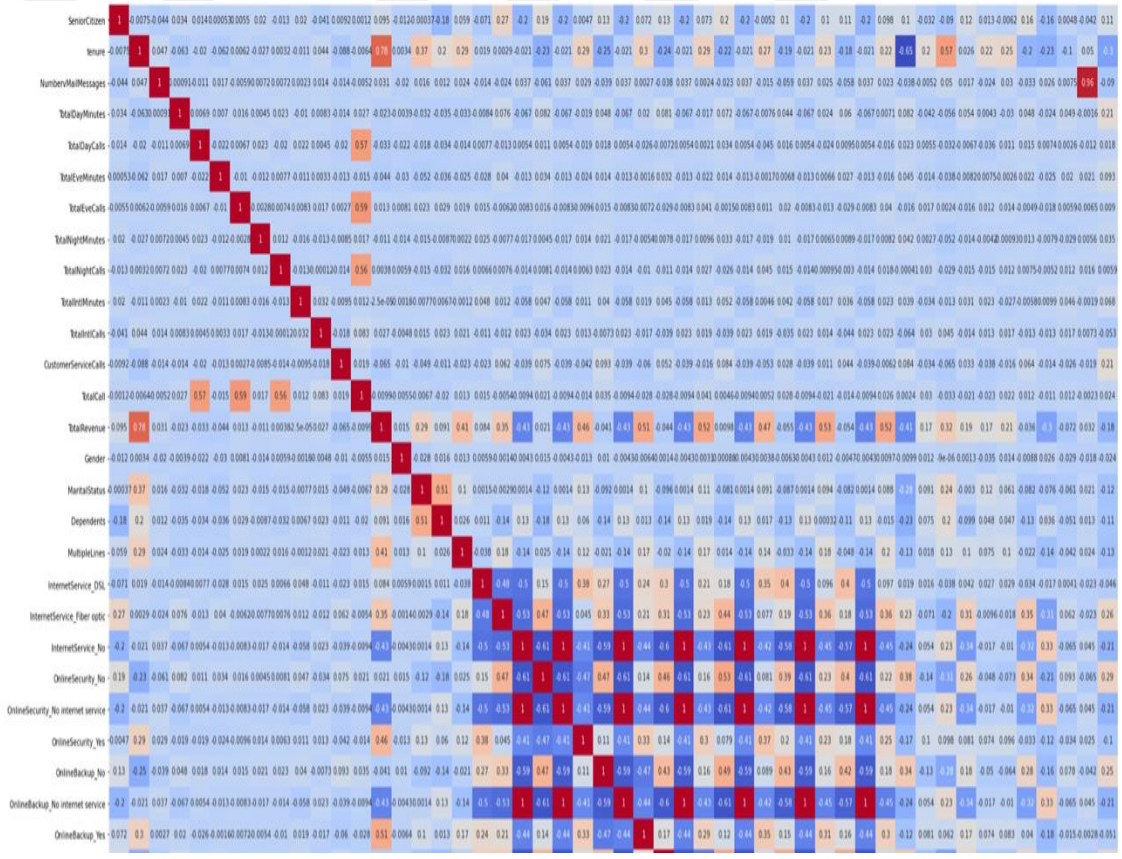
TechSupport_Yes	-0,102
StreamingTV_No	0,167
StreamingTV_No internet service	-0,211
StreamingMovies_No	0,174
StreamingMovies_No internet service	-0,211

Tablo 5. 2: Veri Seti Korelasyon Değerleri (Devam)

Öznitelik Adı	**Korelasyon değeri**
PaymentMethod_Credit card (automatic)	-0,113
PaymentMethod_Bank cart (automatic)	-0,075
Contract_One year	-0,161
Contract_Two year	-0,221
PaperlessBilling	0,124
StreamingTV_Yes	0,044
PaymentMethod_Electronic check	0,247
PaymentMethod_Mailed check	0,070
InternationalPlan	0,260
VoiceMailPlan	-0,102
PaymentMethod_Credit card (automatic)	-0,113
StreamingMovies_Yes	0,037
Contract_Month-to-month	0,323

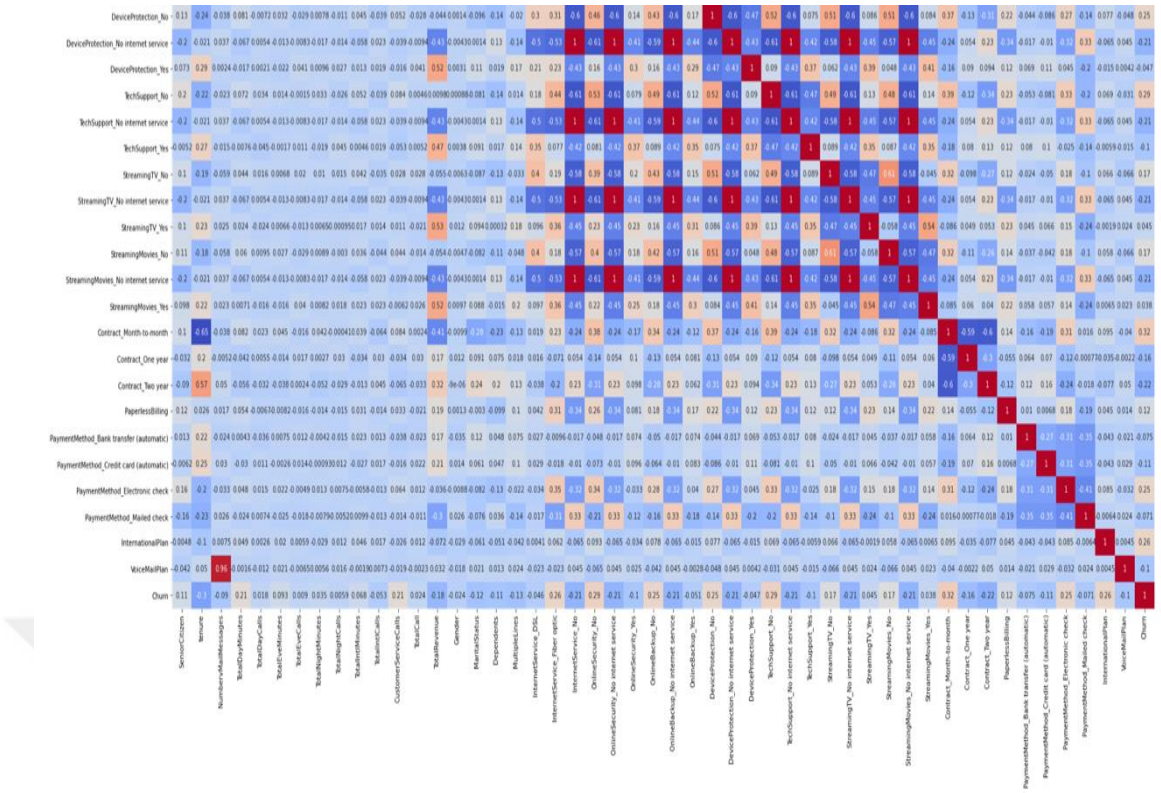
CustomerServiceCalls	0,209

Yapılan korelasyon analizi sonucunda, ilişkinin doğrusal (lineer) olup olmadığı hesaplanmıştır. Korelasyon katsayısı “r” ile ifade edilir ve değer aralığı -1 ile +1 arasındadır (Alsakran vd, 2014). Daha fazla veya daha az bir değer alamaz. Öznitelik değerleri beraber artıyor veya beraber azalıyorsa pozitif yönlü, özniteliklerden biri artarken diğeri azalıyorsa ise negatif yönlü bir ilişki vardır. OnlineBackup_No, TechSupport_No, DeviceProtection_No, StreamingTV_No, StreamingTV_Yes, StreamingMovies_No, StreamingMovies_Yes, Contract_Month-to-month, CustomerServiceCalls, PaperlessBilling, PaymentMethod_Electronic check, PaymentMethod_Mailed check ve InternationalPlan özniteliklerinin pozitif yönlü ilişki olduğu gözlenmiştir. Diğer öznitelikler ile negatif yönlü bir ilişki görülmüştür.



Şekil 5.2: Öznitelikler Arasındaki Korelasyon Grafiği-1

Şekil 5.2 ve Şekil 5.3’te, etiket sınıfı (Churn) ile diğer öznitelikler arasındaki korelasyon değerleri ve aralarındaki ilişki gösterilmektedir.



Şekil 5.3: Öznitelikler Arasındaki Korelasyon Grafiği-2

Korelasyonlara bakıldığında korelasyon değeri 0.1'in altında olan birçok kolon vardır. Bu kolonlar eğitim için bir fark oluşturmayacak aksine boyutsallığı artıracaktır. Şekil 5.2 ve Şekil 5.3 incelendiğinde korelasyon değerleri çok düşük olan kolonların müşteri kaybı olma durumunu etkilemediği görülmüştür.

5.2. TEMEL BİLEŞEN ANALİZİ (PCA)

Temel Bileşen Analizi (PCA), veri setindeki korelasyonların tanımlanmasına imkan sağlayan bir boyut indirgeme tekniğidir. Bu şekilde hiç bir gerekli bilgi kaybı olmadan daha az boyutlu bir veri setine dönüştürülebilir. PCA, sonucu etkileyen öznitelikler arasında ilişki olmama durumuna göre analiz sonuçlarını olumsuz yönde etkilemekte ve analizin yorumlanmasını zorlaştırmaktadır. PCA bu durumlar için kullanılan tekniktir. Çok fazla özniteliğe sahip bir veri seti, modeli eğitmek için daha fazla süre harcar ve veri işleme ile veri analizini daha karmaşık duruma getirir. PCA'nın avantajları şu şekildedir (Farag ve Elhabian, 2009):

- Eğitim süresini kısaltır.
- Verilerin boyutunu indirger.

- İlişkili öznitelikleri kaldırır.
- Gürültüyü azaltır.
- Kolay veri görselleştirme sağlar (maksimum 3D veri).

PCA'da indirgeme yöntemi çalışma mantığı şu şekildedir (Abdi ve Lynne, 2010):

- İndirgenecek olan k boyutuna karar verilir.
- Verilere standartlaştırma işlemi uygulanır.
- Öz değerler elde edilir.
- Seçilen k öz değerden kovaryans / korelasyon matris oluşturulur.
- Orijinal veri kümesinin kovaryans / korelasyon matrisi kullanılarak veriler dönüştürülür.
- k boyutlu yeni veri seti elde edilir.

Veri setine PCA tekniği uygulanarak değişken sayısı 11'e indirilerek boyutsallık indirgenmiştir.

5.3. STUDENT T HİPOTEZ TESTİ

Student t testi, iki bağımsız grup sayısal verinin arasında istatistiksel açıdan anlamlı bir fark olup olmadığını test etmek için kullanılır. Student t testi uygulama adımları şu şekildedir (Chakraborti vd., 2012):

- Bağımlı öznitelik sürekli olmalıdır.
- Bağımsız öznitelik kategorik olmalıdır.
- Her bir grupta bağımlı değişken normal dağılım göstermelidir.
- Varyanslar birbirine yakın (homojen) olmalıdır.
- Aykırı değerler barındırmamalıdır. Diğer değerlerden çok farklı olan veriler belirlenmeli, bunlar veri setinden çıkarılmalıdır.

Yapılan Student t test sonucunda eğer değerler > 0.05 ise örneklem gruplarından elde edilen ortalamaların eşit olduğu kabul edilir (Leys ve Schumann, 2010). Yani istatistiksel olarak gruplar arasında anlamlı bir fark yoktur sonucu elde edilir. P değeri küçüldükçe istatistiksel olarak anlamlı farklılığın kanıtı artar. P değeri Tablo 5.3'de gösterilmiştir (McLeod, 2019).

Tablo 5. 3: P Değeri Yorumlama Tablosu

P Değeri	Yorumu
0,01 <= p <0,05	İstatistiksel olarak anlamlı fark vardır.
0,001 <= p < 0,01	Yüksek anlamlı fark vardır.
p < 0,001	Çok yüksek seviyede istatistiksel olarak anlamlı fark vardır.
0,05 <= p < 0,10	Sınırda anlamlılık vardır.
P > 0.10	İstatiksel olarak anlamlı fark gözlenmemiştir.

Yapılan çalışmada istatistiksel çözümlemede Student t testi tekniği kullanılmıştır. Güvenilirlik analizi yapılmış ve ölçeğin iç tutarlılık katsayısı (Cronbach Alpha) 0.78 olarak hesaplanmıştır.

İki tip hata vardır. Testin sonunda iki karardan biri verilecek: 1. sıfır hipotezini reddetmek veya 2. sıfır hipotezini reddedememek. Sıfır hipotezi doğru olduğunda reddedilirse bir I. tip hata oluşur. Oluşabilecek hata tipleri Tablo 5.4'te gösterilmiştir (Goodman, 1993).

Tablo 5.4: Hipotez Hata Tipleri

Karar	H_0 doğru	H_1 yanlış
H_0 reddedilemez	Doğru Karar	II. Tip Hata
H_0 red	I. Tip Hata	Doğru Karar

SPSS programında Student t testi uygulanarak p değeri hesaplanmıştır. Özniteliklere yapılan test sonucunda müşteri kaybı olup olmama durumları arasında anlamlı bir fark olup olmadığı ve fark varsa hangi değişkenler arasında olduğu belirlenmiş ve bunun için bağımsız gruplarda Student t testi yapılmıştır. Elde edilen sonuçlar Tablo 5.5'te gösterilmiştir.

Tablo 5.5: Student T Testi Sonucu Bulunan P Değerleri

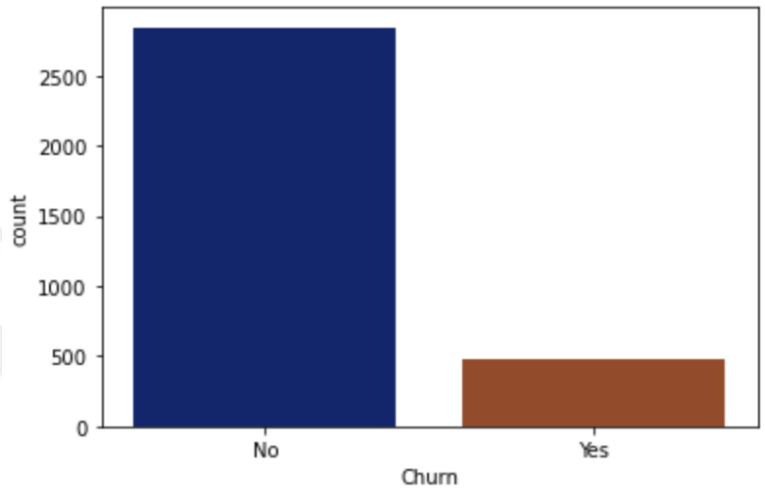
Öznitelik Adı	P Değeri
SeniorCitizen	0,937
tenure	0,120
TotalDayCalls	0,071
TotalEveCalls	0,902
TotalNightMinutes	0,067
TotalNightCalls	0,052
TotalCall	0.004
Gender	0,059
MaritalStatus	0,367
Dependents	0,014
MultipleLines	0,474
InternetService_DSL	0,710
InternetService_No	0,652
OnlineSecurity_No	0,426
OnlineSecurity_No internet service	0,277
OnlineSecurity_Yes	0,289
OnlineBackup_No	0,152
OnlineBackup_No internet service	0,832
OnlineBackup_Yes	0,458
DeviceProtection_Yes	1,000
TechSupport_No	0,353
TechSupport_No internet service	0,832
TechSupport_Yes	0,542
StreamingTV_No	0,560
StreamingTV_No internet service	0,340
StreamingTV_Yes	0,874
StreamingMovies_No	0,382
StreamingMovies_No internet service	0,711
StreamingMovies_Yes	0,260
Contract_Month-to-month	0,671
Contract_One year	0,367
Contract_Two year	0,491
PaperlessBilling	0,958
PaymentMethod_Credit card (automatic)	0,041
PaymentMethod_Bank cart (automatic)	0,032
PaymentMethod_Electronic check	0,011
PaymentMethod_Mailed check	0,382
InternationalPlan	0,630
VoiceMailPlan	0,442

H_0 hipotezinin doğruluğu altında 0.05 anlamlılık seviyesinde elde edilen p sonuç değerlerinin bulunduğu Tablo 5.4 incelendiğinde sadece 5 tane değerin anlamlılık düzeyi 0.05'den küçük ya da yakın olduğu gözlenmiştir. Bunlar PaymentMethod_Bank cart (automatic), PaymentMethod_Credit card (automatic), PaymentMethod_Electronic check, Dependents, TotalCall özniteliklerdir. Diğer kalan 34 tane özniteliğin p değerleri ise H_0 hipotezini kabul etmek için yeterlidir.

ALTINCI BÖLÜM
VERİ ANALİZİ

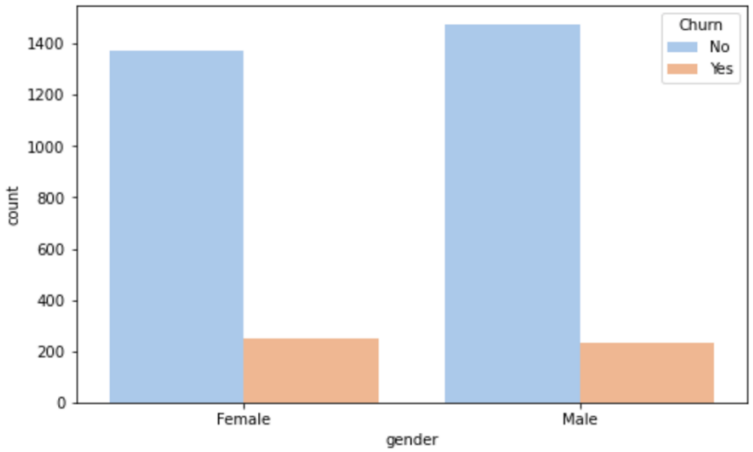
6.1. VERİ ANALİZİ

Veri analizi, model oluşturmada en önemli aşamalardan biridir. Veri setini doğru anlayıp analiz ederek doğru model oluşturulacaktır. Veriler Churn olan ve olmayan olarak karşılaştırıldığında, müşterilerin %85.49 u abone olmaya devam ederken %14.51'i abonelikten ayrıldığı Şekil 6.1'de gösterilmiştir.



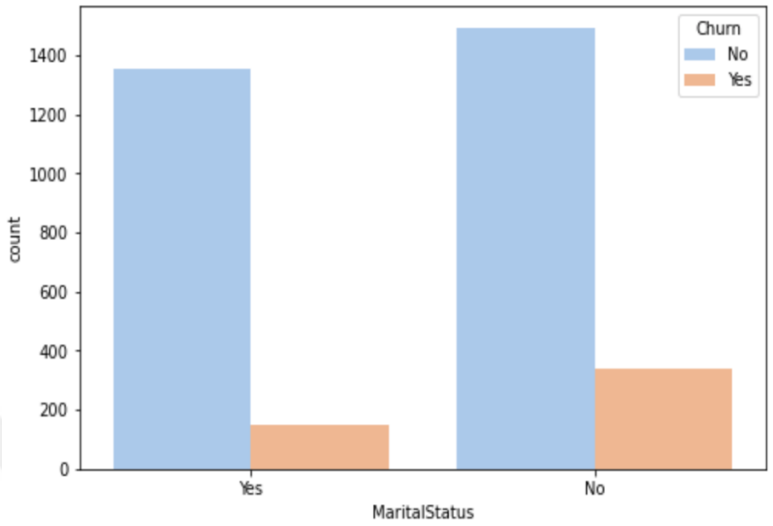
Şekil 6.1: Müşteri Kayıp Grafiği

Şekil 6.2 de cinsiyete göre müşteri kaybı olup olmama dağılımı gösterilmiştir. Grafiğe göre, erkeklerin kadınlara oranla daha uzun olarak müşteri oldukları görülmektedir. Ama müşteri kaybı olan kadınların sayısı erkeklerden daha fazla olduğu görülmektedir. Bu durumda kadınlara yönelik kampanya yapılarak kadın müşteri kayıp oranları azaltılabilir.



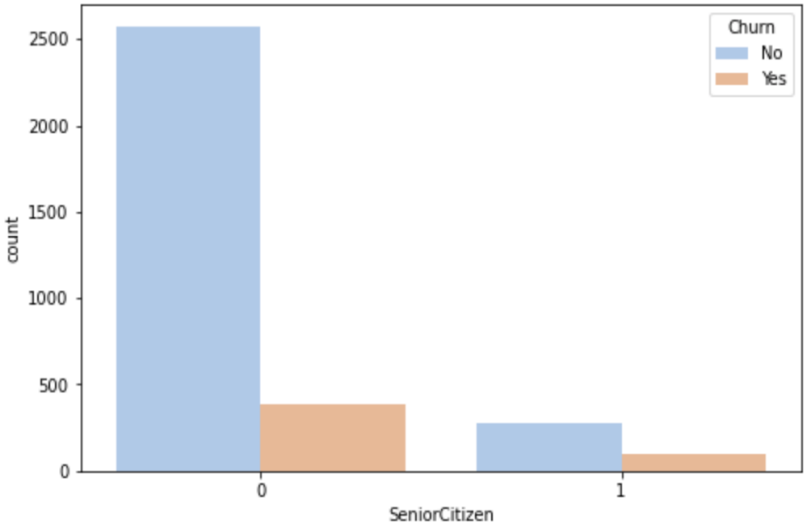
Şekil 6.2: Cinsiyete Göre Müşteri Kayıp Grafiği

Şekil 6.2 de kadın ve erkek müşterilerin ayrılma durumları karşılaştırmalı olarak gösterilmiştir. Grafikte kadın müşterilerin erkek müşterilere oranla daha çok ayrılma olduğu görülmektedir.



Şekil 6.3: Medeni Hale Göre Müşteri Kayıp Grafiği

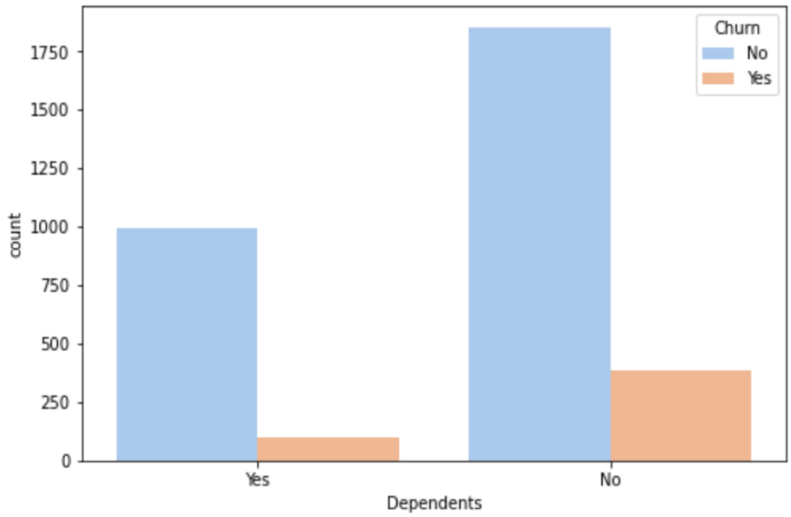
Şekil 6.3 grafiği incelendiğinde medeni durumu evli olmayanların evli olan müşterilere göre fazla ayrılma gösterdiği görülmektedir. Evli olan müşterilerin daha az müşteri kaybı olduğu sonucu çıkarılmıştır.



Şekil 6.4: Yaşlı Gruba Göre Müşteri Kayıp Grafiği

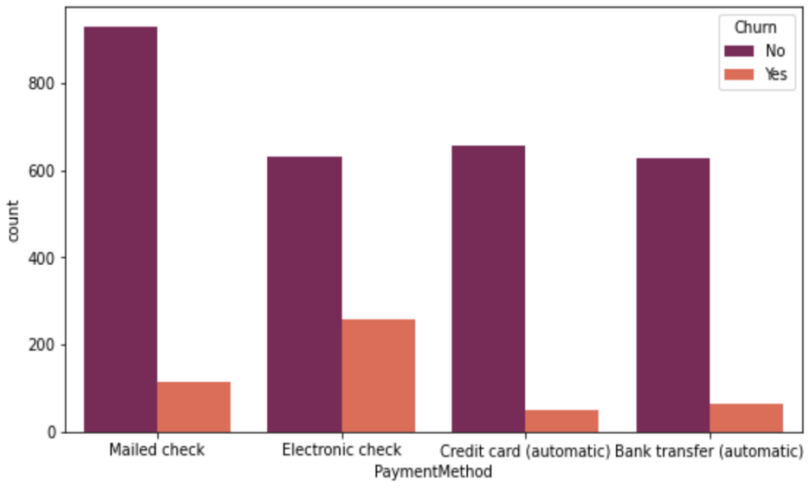
Şekil 6.4 teki grafik incelendiğinde (65 yaş üstü) yaşlı müşterilerin daha kısa süreli abone müşteri olduğu görülmektedir. Eğer yaş ortalamaları alınarak yaşlı

kullanıcıların daha fazla müşteri devamlılıklarını arttırmaya yönelik çalışmalara ağırlık verilirse yaşlı müşterilerin kayıp sayısında azalma gözlenebilir.



Şekil 6.5: Bakmakla Yükümlü Olduğu Kişilere Göre Müşteri Kayıp Grafiği

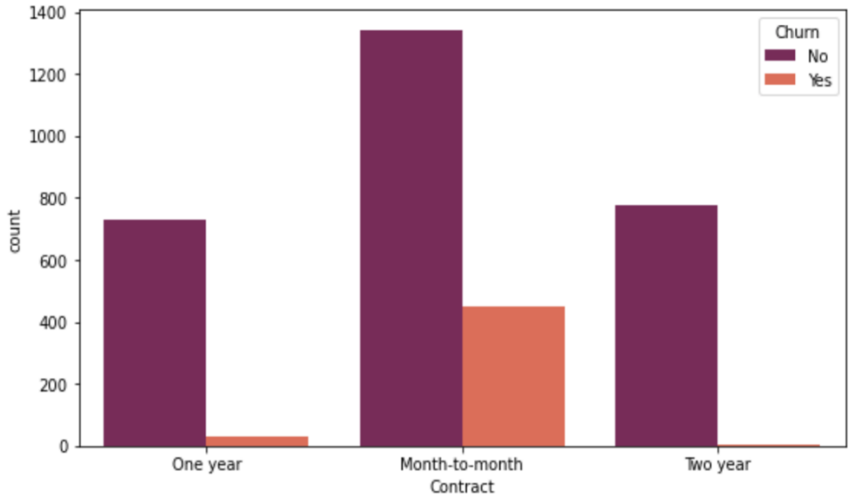
Şekil 6.5. grafiği incelendiğinde müşterilerin bakmakla yükümlü olduğu kişi sayıları arttıkça müşteriyi kaybetme ihtimallerinin arttığı görülmektedir. Bu durumda, müşterilerin bakmakla yükümlü olduğu kişi sayısı azaldıkça müşteri kaybının azalacağı yorumu yapılabilmektedir. Eğer, aynı aileden gelip müşteri olan kullanıcılara özel çeşitli kampanyalar ya da indirimler düzenlenerek bakmakla yükümlü olduğu kişi sayısı fazla olsa bile müşterileri elde tutma ihtimali artabilecektir.



Şekil 6.6: Ödeme Yöntemlerine Göre Müşteri Kayıp Grafiği

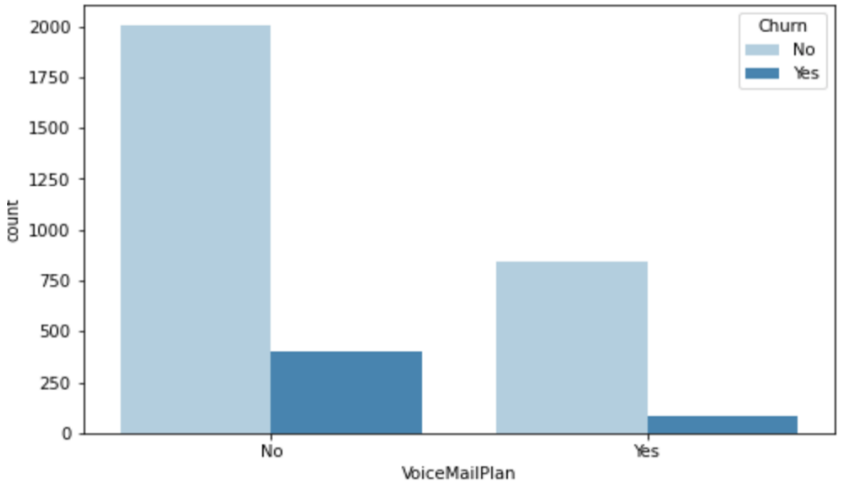
Şekil 6.6. grafiğinde ödeme yöntemlerine göre müşteri kaybı olup olmama durumları gösterilmiştir. En fazla müşteri kaybı olanların elektronik çek ile ödeme

yapan müşteriler olduğu görülmüştür. En az müşteri kaybı olanlar ise kredi kartı ile ödeme yapanlar olduğu görülmüştür. Dört ödeme yöntemi müşteri kaybı olma durumuna göre sıralandığında ise elektronik çekler, posta ile gönderilen çekler, banka transferleri ve en az kredi kartı şeklinde olduğu sonucuna varılmıştır.



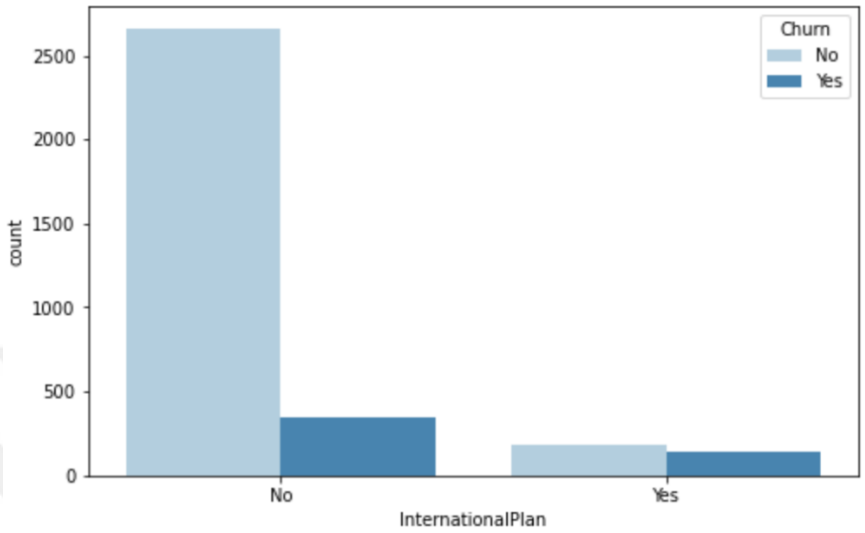
Şekil 6.7: Sözleşme Süresine Göre Müşteri Kayıp Grafiği

Şekil 6.7. incelendiğinde bir yıllık, iki yıllık ve aydan aya sözleşme yapan müşterin kayıp grafiği gösterilmiştir. Kaybedilen müşteriler, sözleşme sürelerine göre karşılaştırıldığında en fazla müşteri kaybı olanların aydan aya sözleşme yapan müşteriler olduğu görülmüştür. Daha sonra bir yıllık ve en az müşteri kaybı olanların ise iki yıllık sözleşme yapan müşteriler olduğu görülmüştür. Sözleşme süresi arttıkça müşteri ayrılma olasılığı azalma göstermiştir. Eğer müşterilere daha uzun süreli sözleşme yaptırmak için teşvik edici çalışmalar yapılırsa müşteri kayıp oranı azaltılmış olacaktır.



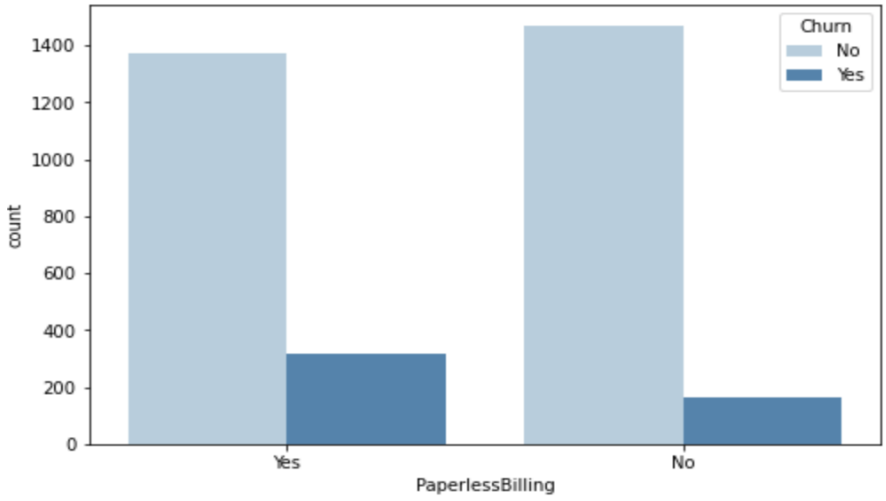
Şekil 6.8: Sesli Posta Alan Müşteri Müşteri Kayıp Grafiği

Şekil 6.8 teki grafik incelendiğinde kaybedilen müşterilerin çoğunun sesli mesaj aldığı görülmektedir. Sesli mesaj almayanların alanlara göre daha fazla daha müşteri kaybı olduğu sonucu çıkarılmıştır. Canlı destek gibi yöntemlerle müşterilerle iletişime geçilirse daha fazla müşteri memnuniyeti sağlanabilir.



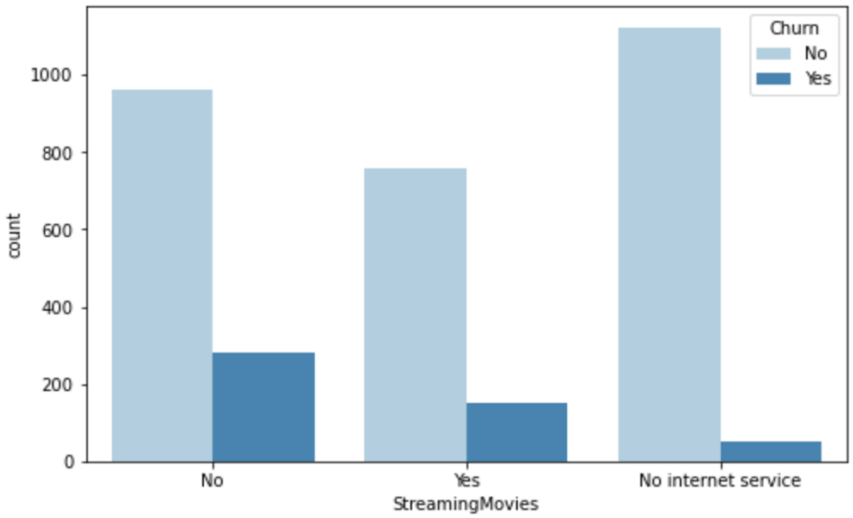
Şekil 6.9: Uluslararası Arama Müşteri Kayıp Grafiği

Şekil 6.9 grafiği incelendiğinde uluslararası müşteri portföyünde kaybedilen müşteri sayısının daha fazla olduğu görülmektedir. Uluslararası kampanyalar etkinlik düzenlenerek müşteri kaybı azaltılabilir.



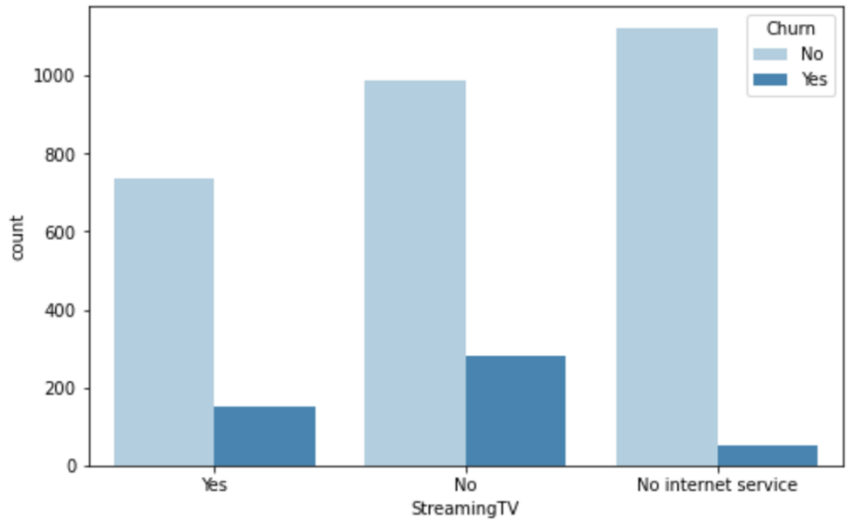
Şekil 6.10: Kâğıt Fatura İsteme Durumlarına Göre Müşteri Kayıp Grafiği

Şekil 6.10 grafiği incelendiğinde müşteri faturalarını isterse kâğıt olarak alabilmektedir. Kâğıt fatura isteyen müşterilerin kayıp oranı istemeyen müşterilerden daha fazla olduğu gözlenmiştir.



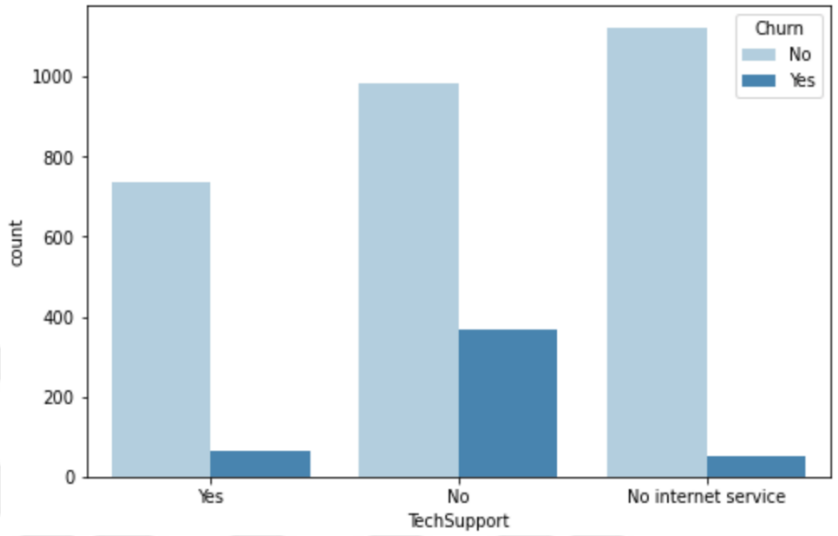
Şekil 6.11: Film Paketi Alma Durumlarına Göre Müşteri Kayıp Grafiği

Şekil 6.11 grafiği incelendiğinde müşteriler film paketi alma durumlarına göre sınıflandırıldığında film paketi alan müşterileri kaybetme oranının daha fazla olduğu görülmektedir. Eğer internet paketi ve film paketlerine yönelik ikili kampanyalar yapılarak müşteri kayıp riski daha da azaltılabilir.



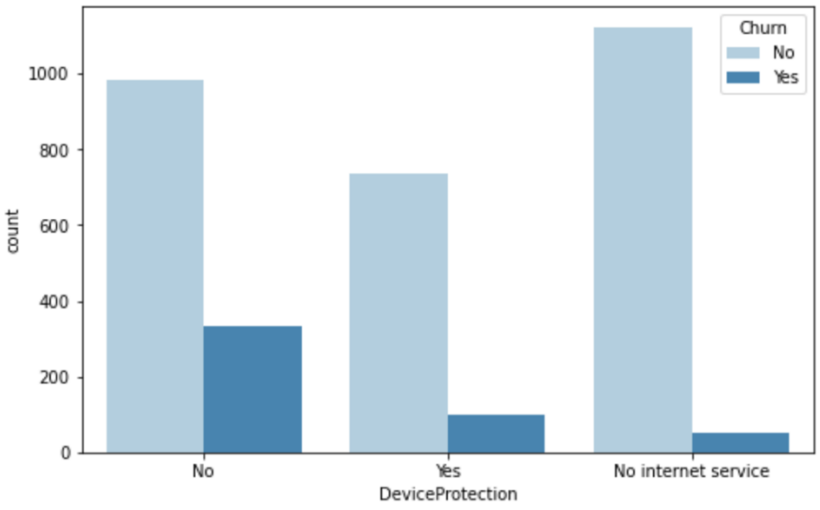
Şekil 6.12: TV Yayın Akışı Paketi Alan Müşterilerin Müşteri Kayıp Grafiği

Şekil 6.12 grafiği incelendiğinde müşteriler TV paketi alma durumlarına göre sınıflandırıldığında TV paketi alan müşterilerin almayan müşterilere göre daha az müşteri kaybı olduğu görülmektedir. Eğer internet paketi ve TV paketlerine yönelik ikili kampanyalar yapılırsa müşteri kayıp riskleri daha da azaltılabilir.



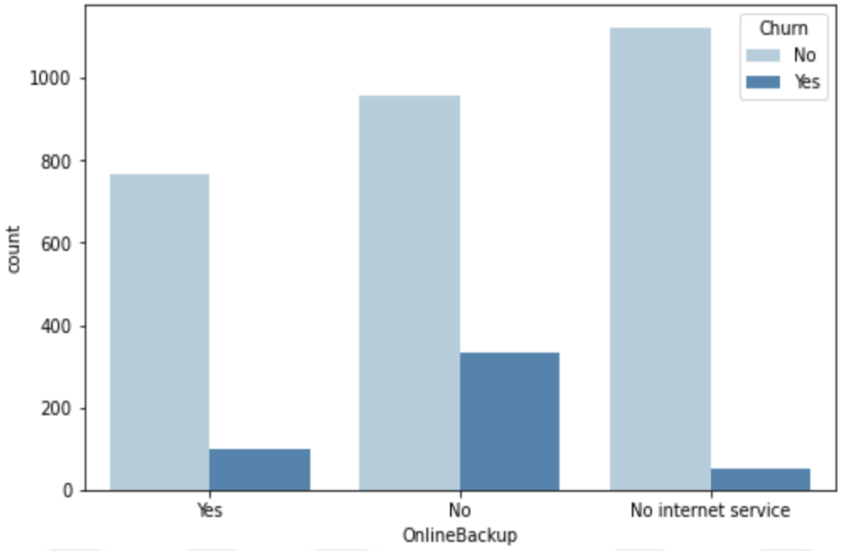
Şekil 6.13: Teknik Destek Durumlarına Göre Müşteri Kayıp Grafiği

Şekil 6.13 grafiği müşterilerin teknik destek alıp almadıklarını göstermektedir. Grafik incelendiğinde teknik destek aldıkça müşteri kayıp oranının azaldığı sonucuna varılmaktadır. O müşterileri tespit edip teknik destek verilirse müşterileri kaybetme riskleri azaltılabilir.



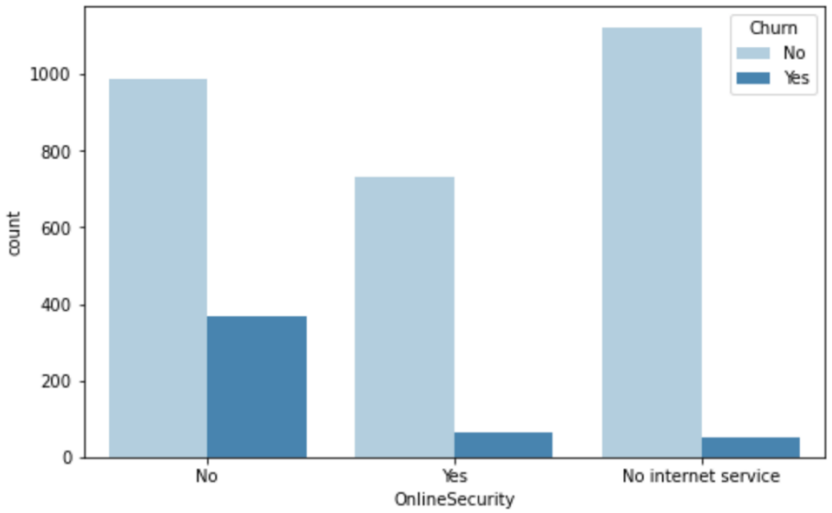
Şekil 6.14: Cihaz Korumasına Göre Müşteri Kayıp Grafiği

Şekil 6.14 grafiği kullandıkları cihazlarda güvenlik koruması yaptırıp yaptırmamalarını göstermektedir. Grafik incelendiğinde müşterilerin kullandıkları cihazlarının güvenlik koruması olmayanların sayısının, cihaz güvenlik koruması olanlara göre daha fazla olduğu görülmektedir. O müşterilerin çoğu cihaz güvenlik koruması yaptırmayıp sıkıntı yaşamış olabilecekleri anlaşılmıştır. Eğer müşterilerin kullandığı cihazlar daha güvenli hala getirilebilir veya zorunlu güvenlik koruması alınması teşvik edilirse müşteri kaybı azaltılabilir.



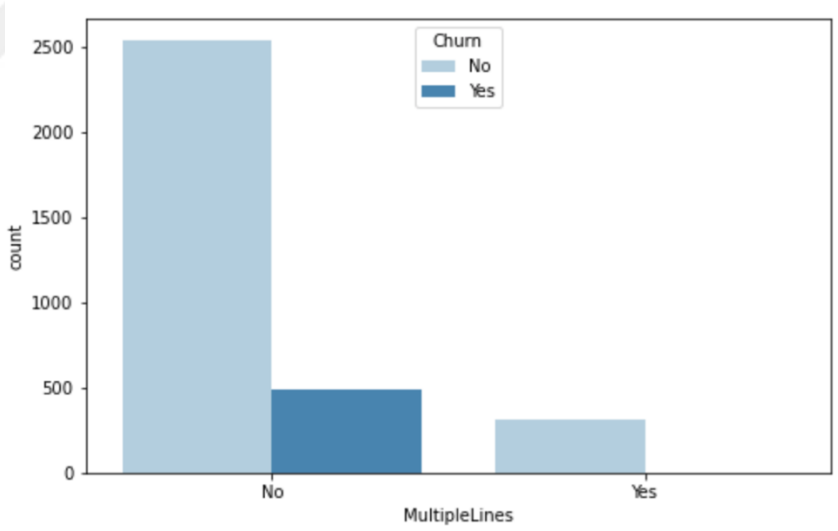
Şekil 6.15: Çevrimiçi Yedeklemesi Olmasına Göre Müşteri Kayıp Grafiği

Şekil 6.15 grafiği kullandıkları cihazlarda yedekleme yaptırıp yaptırmamalarını göstermektedir. Grafik incelendiğinde sistem yedeklemesi yaptıranların yaptırmayanlara göre az olduğu görülmüştür. Bu müşterilerin çoğu sistem yedeklemesi yaptırmamıştır. O müşterilere doğru yönlendirme yapılırsa müşteri kaybı azaltılabilir.



Şekil 6.16: Çevrimiçi Güvenlik Durumuna Göre Müşteri Kayıp Grafiği

Şekil 6.16 grafiği kullandıkları çevrimiçi güvenliğe sahip olup olmadıklarını göstermektedir. Grafik incelendiğinde Çevrimiçi güvenlik sistemi olmayanların sayısının güvenlik sistemi alan müşterilere göre daha çok oluğu görülmektedir. Daha çok çevrimiçi güvenlik sistemi olmayan müşterilerde kayıp yaşandığı gözlenmiştir.



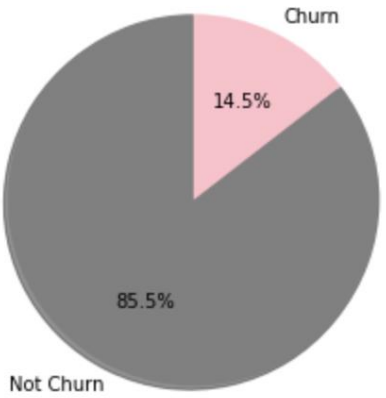
Şekil 6.17: Birden Fazla Hat Sahibi Olma Durumuna Göre Müşteri Kayıp Grafiği

Şekil 6.17 grafiği müşterilerin kullandıkları hat sayısının göstermektedir. Müşteriler genelde tek hat kullandığı gözlenmiştir. Birden fazla hat kullanan kullanıcılarda hiç müşteri kaybı yaşanmadığı görülmektedir. Eğer çoklu hat kullanımına yönelik kampanyalar yapılırsa kayıp oranı ciddi azalış gözlenebilir.

YEDİNCİ BÖLÜM

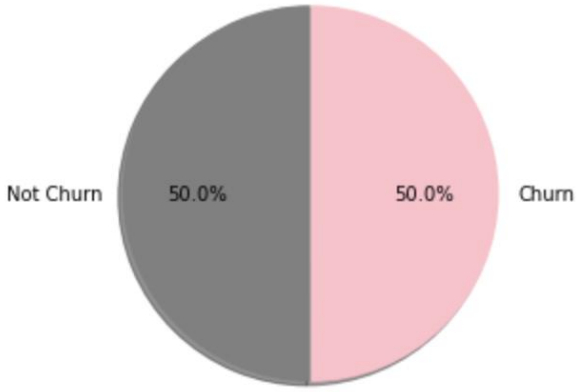
DENEYSEL ÇALIŞMA SONUCU

Dengesiz veri setleriyle karşılaşıldığında makine öğrenmesi algoritmaları, iki sınıftaki eşit olmayan dağılımı fark etmeyerek doğru sonuçlar vermeyebilir. Yani elde edilen sınıflandırıcılar genel hata oranını en aza indirmeye çalışırken ağırlıklı olan sınıfı kategorize ederek azınlık sınıfı göz ardı edebilir. Bu durumla karşılaşıldığında kullanılan yöntemlerden en bilineni SMOTE işlemidir (Müslim, 2020). SMOTE işlemi özellikle dengesiz veri setlerinde uygulanmaktadır. Tahmin yapılacak verilerin %85,49'unda müşteri kaybı olmazken %14,51'lik kısmında müşteri kaybı olmaktadır. Dağılım %50 - %50 şeklinde ayrılmadığı görülmektedir. Bu yüzden veri setine SMOTE işlemi uygulanıp dengesizliğin önüne geçilerek veri setini dengeli şekilde öğrenmesi sağlanmaktadır. Şekil 7.1'de görüldüğü gibi veri seti %85.5'lik gri kısım müşteri kaybı olmayan (Not-Churn) temsil etmektedir ve %14.5'lık pembe kısım olan (Churn) kısmını temsil etmektedir. Şekilde 7.1'de ki yüzdelik oranlardan anlaşılacağı gibi bir dengesizlik bulunmaktadır.



Şekil 7.1: SMOTE Öncesi Veri Seti Dağılım Grafiği

Bu sorunun üstesinden gelmek için SMOTE yöntemi uygulanmıştır. SMOTE uyguladıktan sonraki sınıf dağılımı Şekil 7.2'de verilmiştir. Şekil 7.2'de görüldüğü üzere müşteri kaybı olmayan (Not-Churn) ve olan (Churn) sınıfları eşit oranda dağılmıştır. SMOTE sonucu veri setinde toplam 5121 adet veri bulunmaktadır.



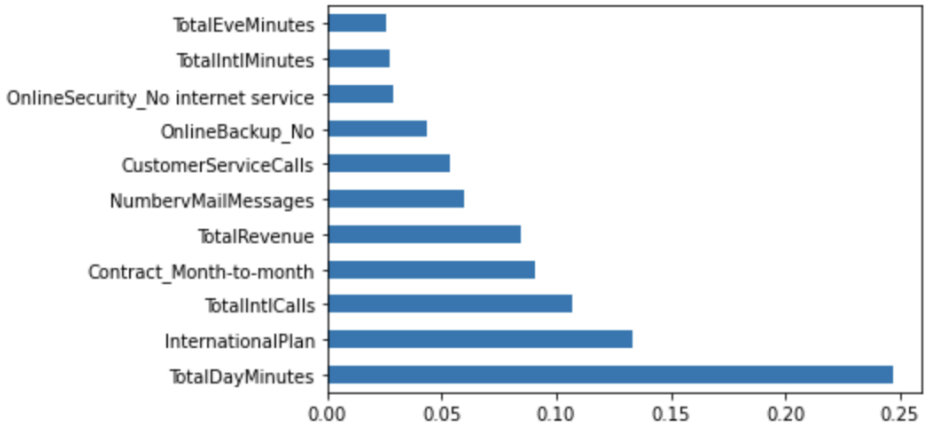
Şekil 7.2: SMOTE Sonrası Veri Seti Dağılım Grafiği

Verilerin değerlerinde aşırı ölçek farklılıkları olduğunda değerleri belirli aralığa sığdırma işlemi yapılır. Bu işleme ölçekleme (scaling) denir. Veri setine ölçekleme işlemi uygulanmamıştır. Çünkü çoğu öznitelik kategorik değerdir, nümerik değildir. Veri seti ön işleme kısmı burada tamamlanmış ve eğitim kısmına geçilmiştir. Veri setinin yüzde 90'ı eğitim için ayrılırken yüzde 10'luk kısmı test için ayrılmıştır. Eğitim seti 5121 örnekten oluşurken, test seti 569 örnekten oluşmaktadır. Eğitim için Rastgele Orman, Karar Ağaçları, DVM ve YSA algoritmaları kullanılmıştır.

Uygulama, Anaconda programında Jupyter Notebook ortamı kullanılarak kodlanmıştır. Sonuçlar ölçülürken Karışıklık matrisi, Doğruluk, Kesinlik, Duyarlılık ve ROC eğrisi metrikleri kullanılmıştır. Önce kullanılacak algoritmalara ait Karışıklık matrisi incelenmiştir. Her algoritma için verilere SMOTE uygulaması yapılmıştır. SMOTE işleminin etkisini görmek için SMOTE uygulanmadan önce ve sonra Karışıklık matrisleri ve analizleri eklenerek sonuçlar karşılaştırılmıştır.

7.1. KARAR AĞAÇLARI DENEYSEL SONUÇLARI

Model oluştururken tahmin yapmak için öncelikle kullanılacak özniteliklerin oluşturulacak model için ne kadar yararlı olup olmadığı incelenmelidir. Bu şekilde her özniteliğe bir puan atayarak, oluşturulacak olan modele ilişkin öngörü ve sorunla ilgili tahmine dayalı olan modelin verimliliğini ve etkinliğini arttıran boyutluluk azaltma ve özellik seçimi için temel oluşturulmaktadır. Bu şekilde önemli özniteliklere odaklanarak model iyileştirilmiştir. Şekil 7.3'te SMOTE öncesi özniteliklerin karşılaştırmalı grafiği gösterilmiştir.



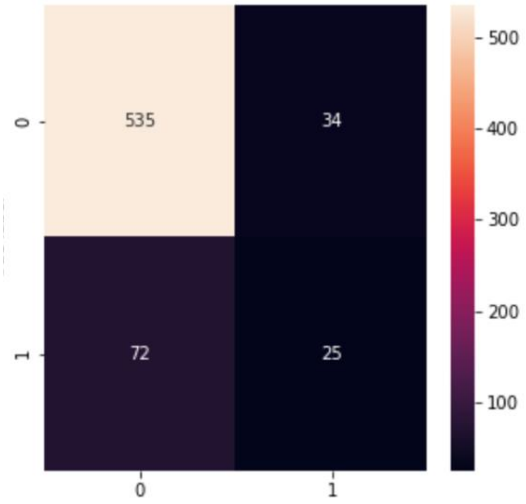
Şekil 7.3: SMOTE Öncesi Karar Ağaçları Öznitelik Dereceleri Karşılaştırma Grafiği

SMOTE öncesi Karar Ağaçları öznitelik önem dereceleri Şekil 7.3'de gösterilmiştir. Şekil incelendiğinde en yüksek önem değerine sahip başlıca öznitelikler sırasıyla TotalDayMinutes, InternationalPlan, TotalIntlCallss, Contract_Month-to-month, TotalRevenue, NumbervMailMessages, CustomerServiceCalls, OnlineBackup_No, OnlineSecurity_No internet service, TotalIntlMinutes, TotalEveMinutes özniteliklerdir. Özniteliklerin önem değerleri sırası ile Tablo 7.1'de detaylı gösterilmiştir.

Tablo 7.1: SMOTE Öncesi Karar Ağaçları Öznitelik Önem Değerleri

Öznitelik Adları	**Önem Değerleri**
TotalDayMinutes	0.246779
InternationalPlan	0.133198
TotalIntlCalls	0.106857
Contract_Month-to-month	0.090631
TotalRevenue	0.084224
NumbervMailMessages	0.059994
CustomerServiceCalls	0.053326
OnlineBackup_No	0.043114
OnlineSecurity_No internet service	0.028804
TotalIntlMinutes	0.027303
TotalEveMinutes	0.025379

Şekil 7.4'te Karar Ağaçları Algoritmasına ait Karışıklık matrisini göstermektedir. SMOTE uygulanmadan önceki Karışıklık matrisidir.

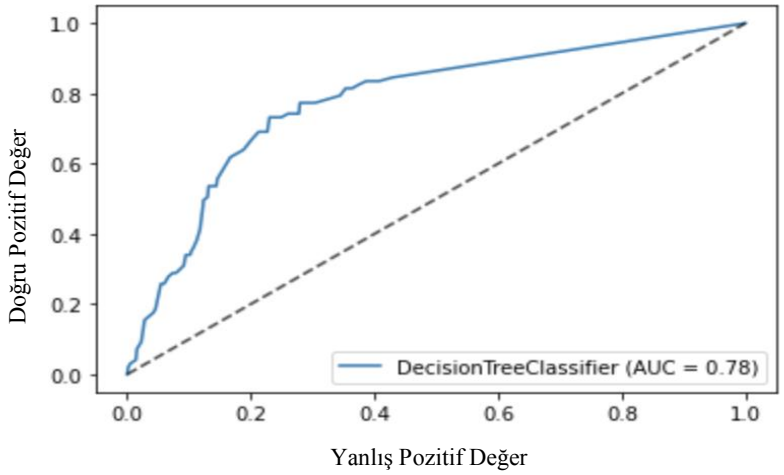


Şekil 7.4: SMOTE Öncesi Karar Ağaçları Karışıklık Matrisi

Tablo 7.2: SMOTE Öncesi Karar Ağaçları Model Performans Sonuçları

Duyarlılık	Kesinlik	Doğruluk	F1 Score
0.257732	0.423729	0.840841	0.320513

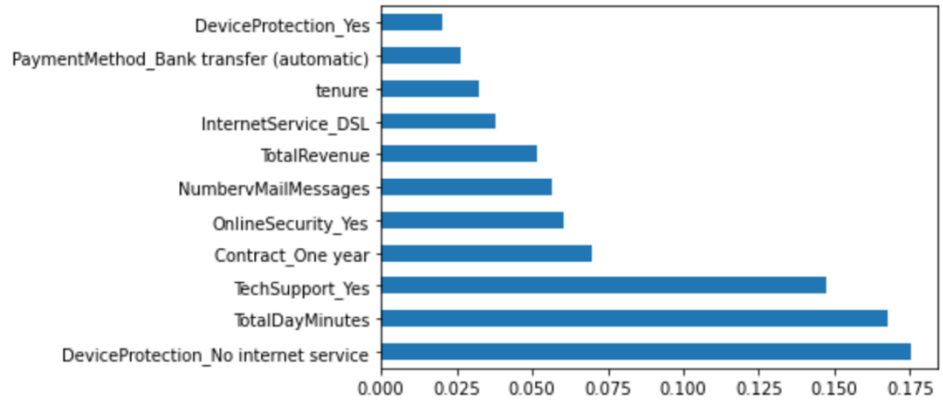
Tablo 7.2 incelendiğinde Duyarlılık 0.257732, Kesinlik 0.423729, Doğruluk 0.840841ve F1 Score 0.320513 sonuçları hesaplanmıştır.



Şekil 7.5: SMOTE Öncesi Karar Ağaçları ROC Eğrisi

Modelin SMOTE öncesinde ROC eğrisinden performansı değerlendirildiği zaman, AUC değeri 0,78 hesaplanmıştır. Grafik incelendiğinde yanlış pozitif ve doğru pozitif değerler 0,8 eşiğine kadar düzensiz artış gösterirken yanlış pozitif değerler 0,4 olduktan sonra doğru pozitif değerlerde hızlı bir artış ve 0,8 eşik değerinden sonra

doğrusal bir artış olduğu Şekil 7.5'te gözlenmiştir. Şekil 7.6'de SMOTE sonrası öznitelik önem dereceleri verilmiştir.



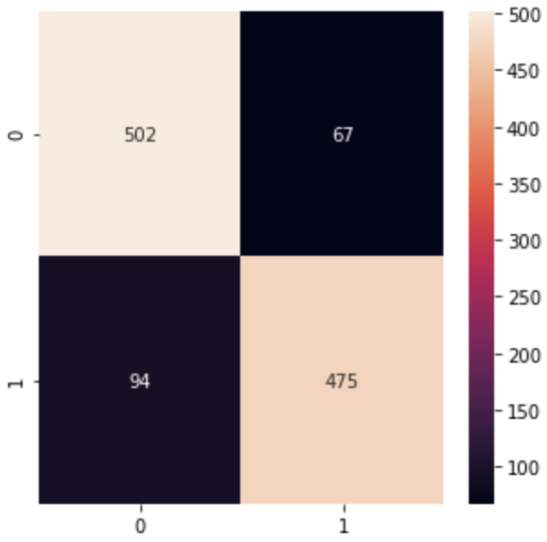
Şekil 7.6: SMOTE Sonrası Karar Ağaçları Öznitelik Dereceleri Karşılaştırma Grafiği

Şekil 7.6 incelendiğinde en yüksek önem değerine sahip başlıca öznitelikler sırasıyla DeviceProtection_No internet service, TotalDayMinutes, TechSupport_Yes, Contract_One year, OnlineSecurity_Yes, NumbervMailMessages, TotalRevenue, InternetService_DSL, tenure, PaymentMethod_Bank transfer(automatic), DeviceProtaction_Yes öznitelikleri olduğu gözlenmiştir. Özniteliklerin önem değerleri sırası ile Tablo 7.3'te detaylı gösterilmiştir.

Tablo 7.3: SMOTE Sonrası Karar Ağaçları Öznitelik Önem Değerleri

Öznitelik Adları	**Önem Değerleri**
DeviceProtection_No internet service	0.175405
TotalDayMinutes	0.167647
TechSupport_Yes	0.147217
Contract_One year	0.069880
OnlineSecurity_Yes	0.060162
NumbervMailMessages	0.056548
TotalRevenue	0.051714
InternetService_DSL	0.037689
Tenure	0.032298
PaymentMethod_Bank transfer (automatic)	0.026453
DeviceProtection_Yes	0.020475

Şekil 7.7'te Karar Ağaçları Algoritmasına ait Karışıklık matrisi gösterilmektedir. SMOTE sonrası Karışıklık matrisidir. Algoritma iki sınıfta da başarılı bir performans göstermiştir.

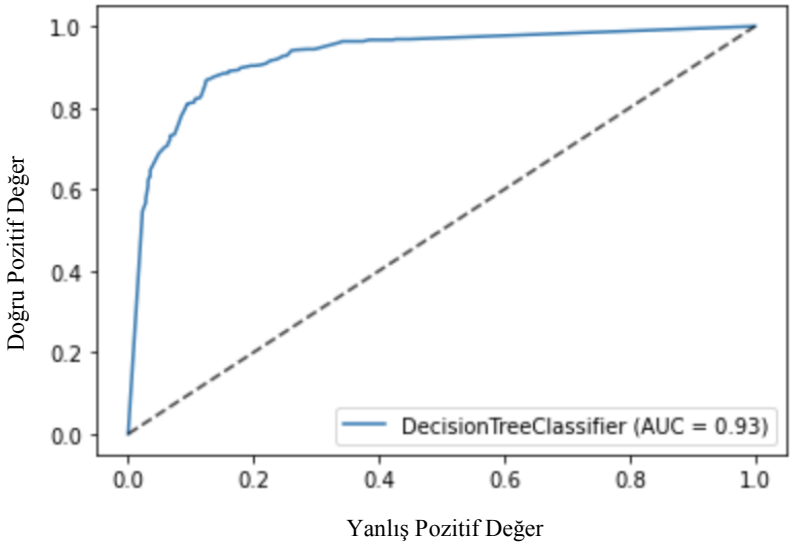


Şekil 7.7: SMOTE Sonrası Karar Ağaçları Karışıklık Matrisi

Tablo 7.4: SMOTE Sonrası Karar Ağaçları Model Performans Sonuçları

Duyarlılık	Kesinlik	Doğruluk	F1 Score
0.834798	0.876384	0.858524	0. 855086

Tablo 7.4. incelendiğinde Duyarlılık 0. 834798, Kesinlik 0. 876384, Doğruluk 0.858524 ve F1 Score 0.855086 sonuçları bölüm 5'teki formüller kullanılarak hesaplanmıştır. SMOTE öncesi ve sonrası performans değerleri karşılaştırıldığında SMOTE işlemi sonrasında oluşturulan modelde hesaplanan sonuçlar yüzdelik olarak karşılaştırıldığında Duyarlılık değerinde %58, Kesinlik değerinde % 238 ve Doğruluk değerinde %1 ve F1 Score %53 artış olduğu ve daha başarılı sonuçlar alınmıştır.

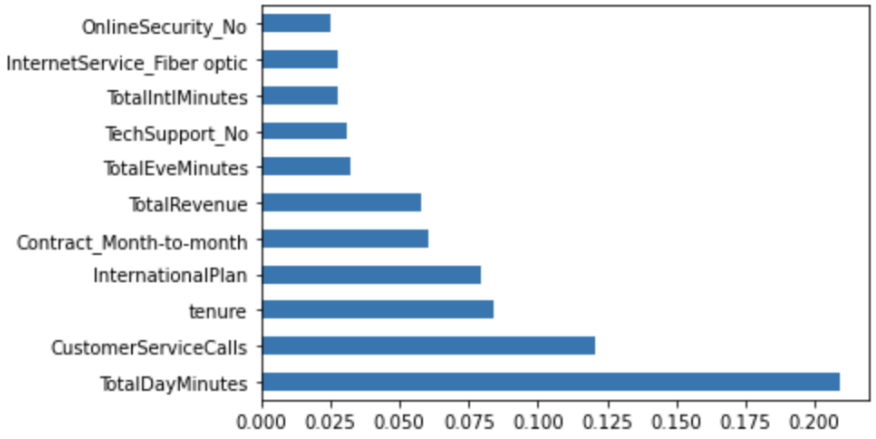


Şekil 7.8: SMOTE Sonrası Karar Ağaçları ROC Eğrisi

Modelin SMOTE sonrasında ROC eğrisinden performansı değerlendirildiği zaman, AUC değeri 0,93 hesaplanmıştır. Şekil 7.8 incelendiğinde yanlış pozitif değerler yaklaşık 0,1 eşiğine kadar çok az artış gösterirken doğru pozitif değerlerin 0,9 değerine kadar hızla arttığı ve yanlış pozitif değerler 0,1 değerine ulaştıktan sonra ise doğru pozitif değerlerde çok daha yavaş bir artış gözlemlenerek 0,9 eşik değerinde yanlış pozitif değerler artarken doğru pozitif değerler çok az artmıştır. Sonuçta doğru tahmin ettiği değerler yanlış tahmin ettiği değerlere göre azalmaktadır. SMOTE öncesine göre daha dengeli bir performans elde edilmiştir.

7.2. RASTGELE ORMAN ALGORİTMASI DENEYSEL SONUÇLARI

Model için kullanılacak olan özniteliklerin önem değerleri hesaplandı. Önemli parametrelere odaklanılarak modelde iyileştirilme yapılmıştır.



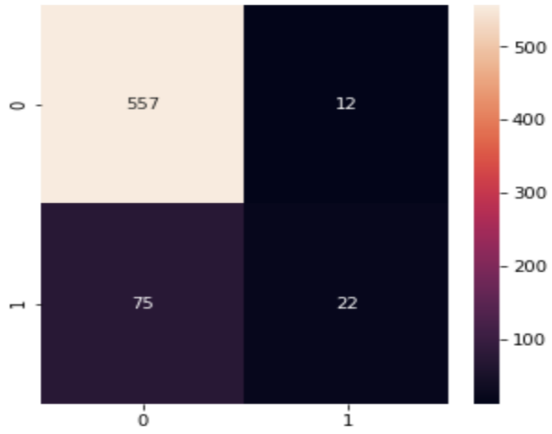
Şekil 7.9: SMOTE Öncesi Rastgele Orman Öznitelik Önem Dereceleri Grafiği

Şekil 7.9'da SMOTE öncesi öznitelik önem dereceleri gösterilmiştir. Şekil 7.9 incelendiğinde en yüksek önem değerine sahip başlıca öznitelikler sırasıyla TotalDayMinutes, CustomerServiceCalls, tenure, InternationalPlan, ContracMonth-to-month, TotalRevenue, TotalEveMinutes, TechSupport_No, TotalIntlMinutes, InternetService_Fiber optik, Online Security_No öznitelikleri olduğu gözlenmiştir. Özniteliklerin önem değerleri ayrıntılı olarak Tablo 7.5'te gösterilmiştir.

Tablo 7.5: SMOTE Öncesi Rastgele Orman Öznitelik Önem Değerleri

Öznitelik Adları	Önem Değerleri
TotalDayMinutes	0.208716
CustomerServiceCalls	0.120814
tenure	0.084065
InternationalPlan	0.079154
Contract_Month-to-month	0.060533
TotalRevenue	0.058089
TotalEveMinutes	0.032579
TechSupport_No	0.030650
TotalIntlMinutes	0.027654
InternetService_Fiber optic	0.027544
OnlineSecurity_No	0.025328

Şekil 7.10 Rastgele Orman Algoritmasına ait Karışıklık matrisini göstermektedir. SMOTE uygulanmadan önceki Karışıklık matrisidir.

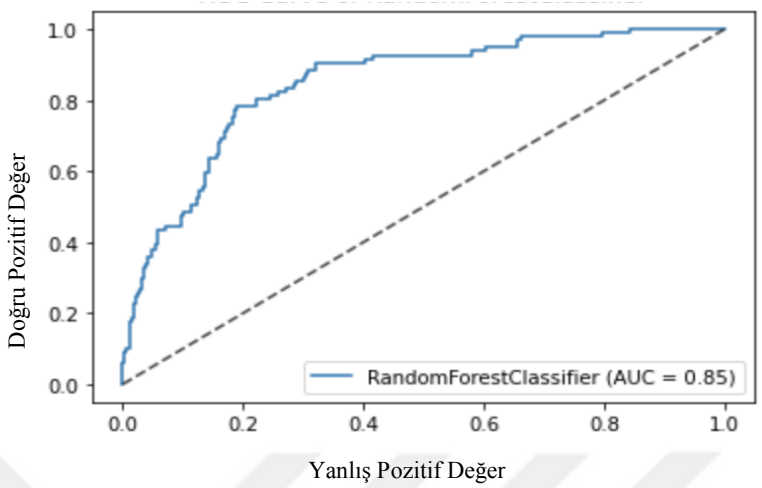


Şekil 7.10: SMOTE Öncesi Rastgele Orman Karışıklık Matrisi

Tablo 7.6: SMOTE Öncesi Rastgele Orman Model Performans Sonuçları.

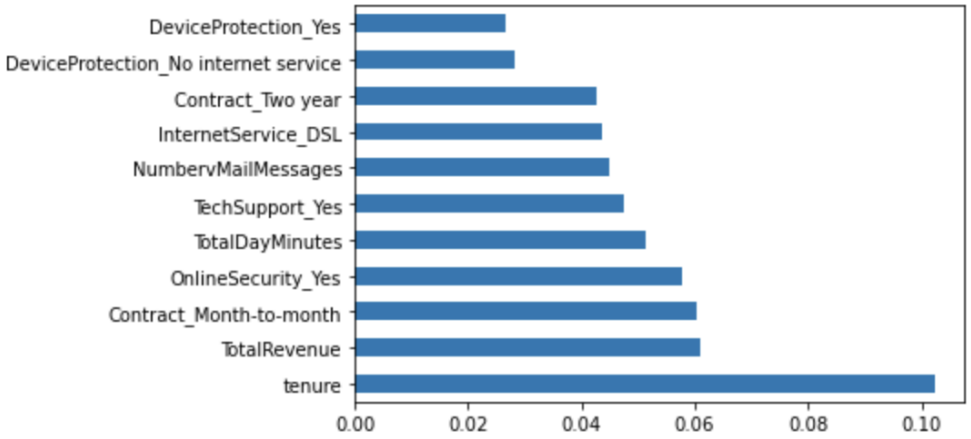
Duyarlılık	Kesinlik	Doğruluk	F1 Score
0.226804	0.647059	0.869369	0.335878

Tablo 7.6 incelendiğinde Duyarlılık 0.226804, Kesinlik 0.647059, Doğruluk 0.869369 ve F1 Score 0.335878 sonuçları hesaplanmıştır.



Şekil 7.11: SMOTE Öncesi Rastgele Orman ROC Eğrisi

Modelin SMOTE öncesinde ROC eğrisinden performansı değerlendirildiği zaman, AUC değeri 0,85'tir. Şekil 7.11 incelendiğinde dengesiz bir artış olduğu gözlenmektedir. Bu da veri setindeki verilerin dengesizliğinden kaynaklanmaktadır. Şekil 7.12 SMOTE sonrası öznitelik önem dereceleri gösterilmiştir.



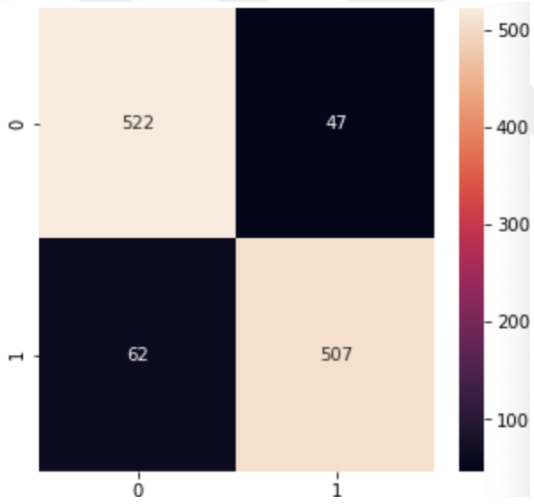
Şekil 7.12: SMOTE Sonrası Rastgele Orman Öznitelik Önem Dereceleri Grafiği

Şekil 7.12 incelendiğinde en yüksek önem değerine sahip başlıca öznitelikler sırasıyla tenure, TotalRevenue, Contract_Month-to-month, OnlineSecurity_Yes, TotalDayMinutes, TechSupport_Yes, NumbervMailMessages, InternetService_DSL, Contract_Two year, DeviceProtection_No internet service, DeviceProtection_Yes öznitelikleri olduğu gözlenmiştir. Özniteliklerin önem değerleri ayrıntılı şekilde Tablo 7.7'de gösterilmiştir.

Tablo 7.7: SMOTE Sonrası Rastgele Orman Öznitelik Önem Değerleri

Öznitelik Adları	Önem Değerleri
tenure	0.102325
TotalRevenue	0.060953
Contract_Month-to-month	0.060233
OnlineSecurity_Yes	0.057707
TotalDayMinutes	0.051338
TechSupport_Yes	0.047534
NumbervMailMessages	0.044979
InternetService_DSL	0.043580
Contract_Two year	0.042528
DeviceProtection_No internet service	0.028277
DeviceProtection_Yes	0.026651

Şekil 7.13 de Rastgele Orman Algoritmasına ait SMOTE uygulandıktan sonraki Karışıklık matrisidir.

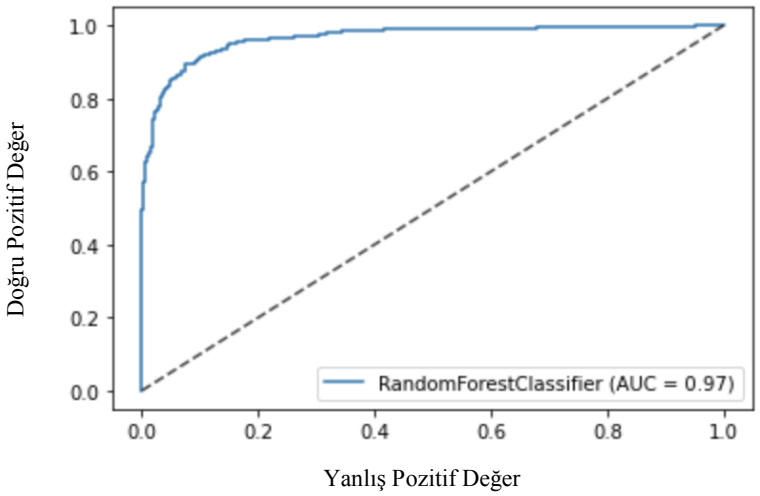


Şekil 7.13: SMOTE Sonrası Rastgele Orman Karışıklık Matrisi

Tablo 7.8: SMOTE Sonrası Rastgele Orman Model Performans Sonuçları

Duyarlılık	Kesinlik	Doğruluk	F1 Score
0.89103	0.915162	0.904218	0.902939

Tablo 7.8 incelendiğinde Duyarlılık 0.89103, Kesinlik 0.915162, Doğruluk 0.904218, F1 Score 0.902939 sonuçları hesaplanmıştır. SMOTE öncesi ve sonrası performans değerleri karşılaştırıldığında SMOTE işleminden sonra modelden daha başarılı sonuçlar alınmıştır.

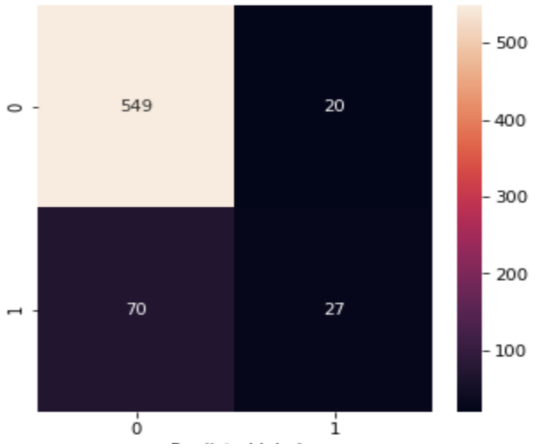


Şekil 7.14: SMOTE Sonrası Rastgele Orman ROC Eğrisi

Modelin SMOTE sonrasında ROC eğrisinden performansı değerlendirildiği zaman, AUC değeri 0,97 hesaplanmıştır. Rastgele Orman Algoritması AUC=0.97 değeriyle performansı en yüksek çıkan algoritma olmuştur. SMOTE öncesine göre daha dengeli bir performans gözlenmiştir.

7.3. YAPAY SİNİR AĞLARI (YSA) DENEYSEL SONUÇLARI

YSA algoritmasına ait SMOTE işlemi yapılmadan önceki Karışıklık matrisi Şekil 7.15'te gösterilmiştir.

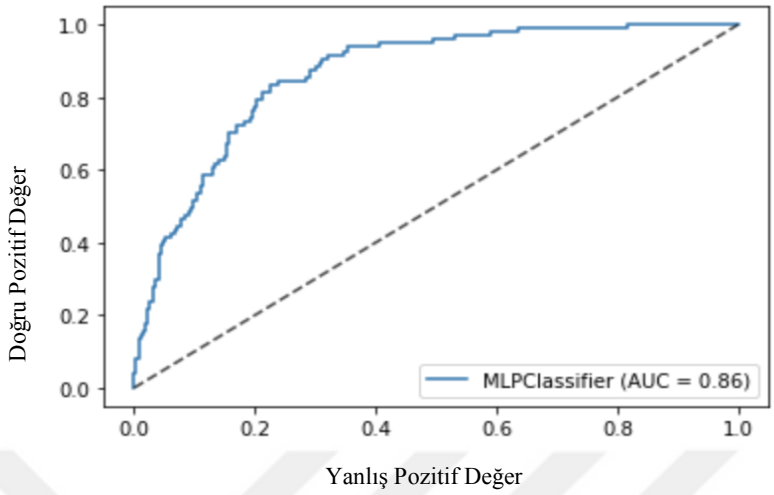


Şekil 7.15: SMOTE Öncesi YSA Matrisi

Tablo 7.9: SMOTE Öncesi YSA Model Performans Sonuçları

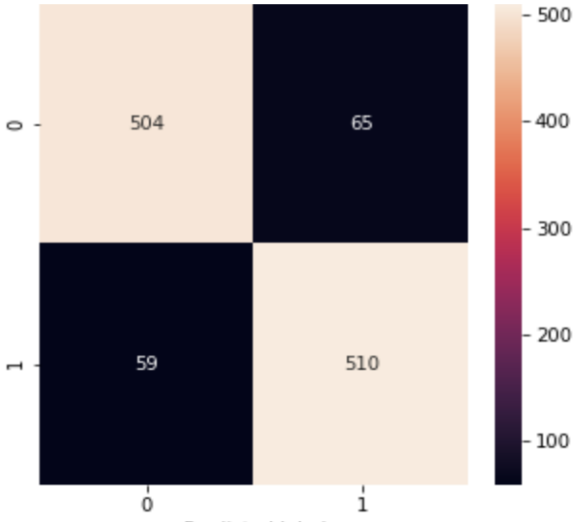
Duyarlılık	Kesinlik	Doğruluk	F1 Score
0.278351	0.574468	0.864865	0.375

Tablo 7.9 incelendiğinde Duyarlılık 0.278351, Kesinlik 0.574468, Doğruluk 0.864865, F1 Score 0.375 sonuçları hesaplanmıştır.



Şekil 7.16: SMOTE Öncesi YSA ROC Eğrisi

Modelin SMOTE öncesinde ROC eğrisinden performansı değerlendirildiği zaman, AUC değeri 0,86 çıkmıştır. Şekil 7.16'de dengesiz bir artış azalış olduğu gözlenmektedir. Şekil 9.15 YSA algoritmasına ait Karışıklık matrisini göstermektedir. SMOTE uygulandıktan sonraki Karışıklık matrisidir. Algoritma iki sınıfta da başarılı bir performans sonucu elde edilmiştir.

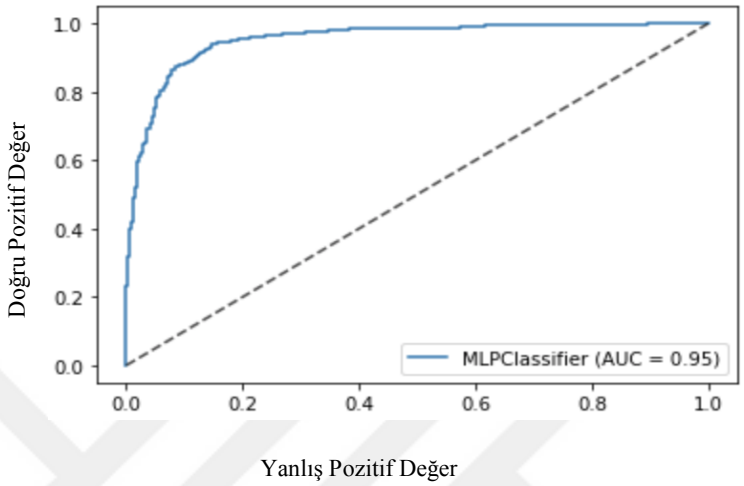


Şekil 7.17: SMOTE Sonrası YSA Karışıklık Matrisi

Tablo 7.10: SMOTE Sonrası YSA Model Performans Sonuçları

Duyarlılık	Kesinlik	Doğruluk	F1 Score
0.896309	0.886957	0.891037	0.891608

Tablo 7.10 incelendiğinde Duyarlılık 0.896309, Kesinlik 0.86957, Doğruluk 0.891037, F1 Score 0.891608 sonuçları hesaplanmıştır. SMOTE öncesi ve sonrası performans değerleri karşılaştırıldığında SMOTE işlemi sonrasında modeli ile daha başarılı sonuçlar alınmıştır.

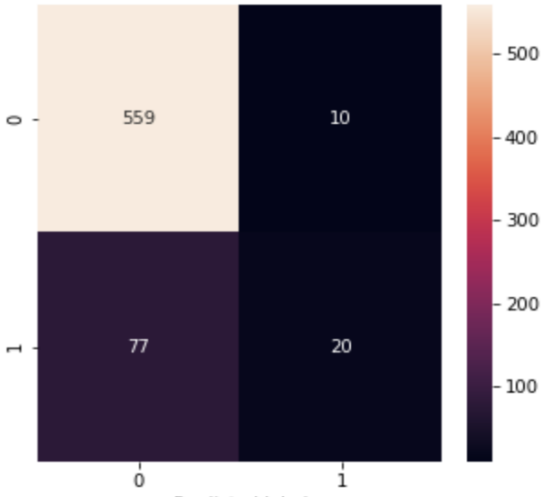


Şekil 7.18: SMOTE Sonrası YSA ROC Eğrisi

Modelin SMOTE sonrasında ROC eğrisinden performansı değerlendirildiği zaman, AUC değeri 0,95 çıkmıştır. Şekil 7.18 incelendiğinde yanlış pozitif değerler 0.1 değerine kadar dengesiz artış gösterirken, doğru pozitif değerler hızla artış göstermiştir. Yanlış pozitif değerler 0,1 değerinden sonra hızla artarken doğru pozitif değerler daha yavaş bir artış eğilimi göstermiştir. YSA modeli öğrendikten sonra 0,6 değerinden sonra bir değişiklik gözlenmemiştir. SMOTE öncesine göre daha dengeli bir performans sonucu vermiştir.

7.4. DESTEK VEKTÖR MAKİNELERİ (DVM) DENEYSEL SONUÇLARI

Şekil 7.19 DVM algoritmasına ait Karışıklık matrisini göstermektedir. SMOTE uygulanmadan önceki Karışıklık matrisidir.

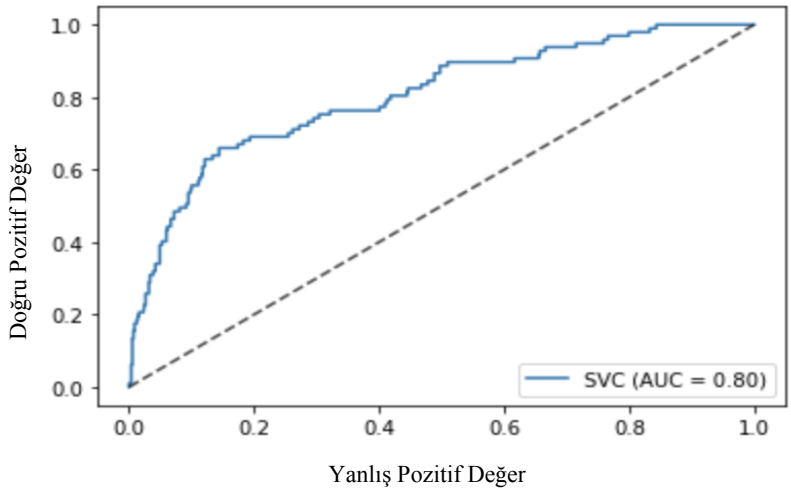


Şekil 7.19: SMOTE Öncesi DVM Karışıklık Matrisi

Tablo 7.11: SMOTE Öncesi DVM Model Performans Sonuçları

Duyarlılık	Kesinlik	Doğruluk	F1 Score
0.206186	0.666667	0.869369	0.314961

Tablo 7.11 incelendiğinde Duyarlılık 0.206186, Kesinlik 66666, F1 Score 0.314961 ve Doğruluk puanı 0.869369 bulunmuştur.

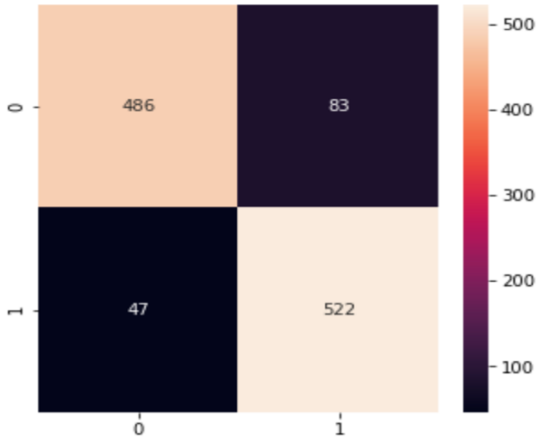


Şekil 7.20: SMOTE Öncesi DVM ROC Eğrisi

Modelin SMOTE öncesinde ROC eğrisinden performansı değerlendirildiği zaman, AUC değeri 0,80 çıkmıştır. Şekil 7.20 incelendiğinde dengesiz bir eğri olduğu

gözlenmektedir. Doğru pozitif değerler ve negatif pozitif değerler arasında sürekli dengesiz bir oran eğrisi olduğu gözlenmiştir.

Şekil 7.21 DVM algoritmasına ait Karışıklık matrisini göstermektedir. SMOTE uygulandıktan sonraki Karışıklık matrisidir. Algoritma iki sınıfta da başarılı bir performans sonucu vermiştir.

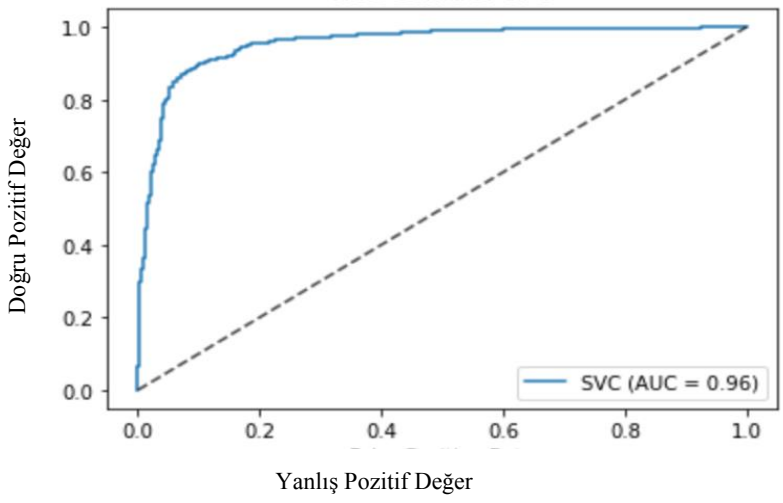


Şekil 7.21: SMOTE Sonrası DVM Karışıklık Matrisi

Tablo 7.12: SMOTE Sonrası DVM Model Performans Sonuçları

Duyarlılık	Kesinlik	Doğruluk	F1 Score
0.917399	0.86281	0.885764	0.889267

Tablo 7.12 incelendiğinde Duyarlılık 0. 917399, Kesinlik 0.86281, Doğruluk 0.885764, F1 Score 0.889267 değerleri hesaplanmıştır. SMOTE öncesi ve sonrası performans değerleri karşılaştırıldığında SMOTE işlemi sonrasında modelde daha başarılı sonuçlar alınmıştır.



Şekil 7.22: SMOTE Sonrası DVM ROC Eğrisi

Modelin SMOTE sonrasında ROC eğrisinden performansı değerlendirildiği zaman, AUC değeri 0,96 hesaplanmıştır. Şekil 7.22 incelendiğinde yanlış pozitif 0.25 değerine kadar dengesiz ve yavaş bir artış gösterirken, doğru pozitif değeri hızlı artış göstermiştir. Yanlış pozitif değerli 0,25 değerinden sonra hızla artarken doğru pozitif değeri daha yavaş bir artış eğilimi göstermiştir. SMOTE öncesine göre daha dengeli bir performans gözlenmiştir.

7.5. ALGORİTMA PARAMETRELERİNİN BELİRLEMESİ

Eğitilecek modellerde kullanılan DVM, YSA, Rastgele Orman ve Karar Ağaçları Algoritmaları için ayrı ayrı en iyi sonuç verecek olan değerler hesaplanmıştır (Kunt, 2019). Rastgele Orman Algoritması için en iyi performans çıktısı verecek olan parametre değerleri aşağıdaki gibi kullanılmıştır (Scikit- Learn, 2021c):

N_estimators: Tahmin işlemi için inşa edilmek istenilen ağaç sayısıdır. 1000 ağaç oluşturulmuştur. En iyi değer 822 bulunmuş ve uygulanmıştır.

Max_features: En iyi bölünmenin nerede olacağını bulmak için kullanılacak olan maksimum öznitelik sayısıdır. Auto ve sqrt fonksiyon değerleri denenmiştir. En iyi parametre değeri auto bulunmuş ve kullanılmıştır.

Max_depth: Ağaçlar için en çok (max) derinlik sayısıdır. En iyi parametre değeri 15 bulunmuş ve kullanılmıştır.

Min_samples_split: Bölünme için kullanılacak olan en çok (max) öznitelik sayıdır. Default değeri 2 dir. 1-10 arası değerler denenmiştir. En iyi değer 10 bulunmuş ve kullanılmıştır.

Min_samples_leaf: Son düğümün en az (min) boyutu belirlenir. Default değeri 1 dir. Ama yine de 1-10 arası değerler denenmiştir. En iyi değer 10 bulunmuş ve kullanılmıştır.

Karar Ağacı Algoritması için en iyi performans çıktısı verecek olan parametre değerleri aşağıdaki gibidir (Scikit- learn, 2021d):

Max_features: Varsayılan değeri auto dur. En iyi bölünmenin nerede olacağını bulmak için kullanılacak olan en çok (max) öznitelik sayıdır. Auto ve sqrt fonksiyon

değerleri denenmiştir. Parametre değeri auto kullanıldığında en iyi performansı vermiştir.

Max_depth: 1000'e kadar derinlik sayıları denenmiştir. En iyi değer 185 hesaplanmıştır. Algoritma parametre değeri olarak 185 uygulanmıştır.

Min_samples_split: Bölünme için kullanılacak olan en çok (max) öznitelik sayıdır. En iyi değer 17 bulunmuştur. Algoritma parametre değeri olarak 17 uygulanmıştır.

Min_samples_leaf: Son düğümün en az (min) boyutu belirlenir. En iyi değer 14 bulunmuştur. Algoritma parametre değeri olarak uygulanmıştır.

N_iter: 50 kez iterasyon yapılmıştır.

YSA algoritması için en iyi performans çıktısı verecek olan parametre değerleri aşağıdaki gibidir (Scikit- Learn, 2021b):

Solver: Ağırlık güncellemeleri yapılırken kullanılan en iyileme fonksiyonunu belirleyen parametredir. Fonksiyon olarak lbfgs, sgd ve adam fonksiyonları denenmiştir. Bu modelde gradyan tabanlı adam fonksiyonu en iyi sonucu verecek değer olarak bulunmuştur. Algoritma parametre değeri olarak uygulanmıştır.

Learning rate init: Oluşturulan modeldeki ağırlıklar için bulunan hataya karşılık, bir sonraki güncellemede ağırlıklar üzerinde ne kadar değişim ve güncelleme olması gerektiğini gösteren parametredir. 0.00001,0.0001,0.001,0.1,1 değerleri denenmiştir. En iyi değer 1 hesaplanmıştır. Algoritma parametre değeri olarak bulunmuş ve uygulanmıştır.

Alpha: Modelin ağırlıklarının boyutunu belirli bir sınırda tutmak için modelin aşırı öğrenme olmasını engelleyen parametredir. 0.00001,0.0001,0.001,0.1,1değerleri denenmiştir. En iyi değer 1 hesaplanıp kullanılmıştır.

Activasyon: Gizli katmanlar için aktivasyon fonksiyonudur. Identity, logistic sigmoid, tanh ve relu fonksiyonları uygulanmıştır. Identitiy, doğrusal durumlarda kullanılan fonksiyondur. Lojistic sigmoid, Y değerleri x teki değişikliklere çok az cevap verir ve öğrenme olayı minimum değerde gerçekleşir. Relu, [0, +∞) aralığında değer alır. Tanh, öğrenme hızı daha yüksektir ve sınıflama işlemi için daha geniş aralığa sahip olduğu için daha verimlidir. Tanh fonksiyonu kullanılıp en iyi değer hesaplanmıştır.

Hidden_layer_size: Nöron sayısıdır. 0, 50 arası nöron değerleri verilerek en iyi sonucu 50 değeri çıktığı için nöron sayısı 50 alınmıştır.

Max_iter: 560 kez iterasyon yapılmıştır.

DVM algoritması için en iyi performans çıktısı verecek olan parametre değerleri aşağıdaki gibidir (Scikit- Learn, 2021a):

Kernel: Modelde kullanılacak olan çekirdek çeşidini belirlemek için kullanılır. Linear, poly, rbf ve sigmoid fonksiyonları denenmiştir. Poly fonksiyonu kullanılmıştır.

C: Modelin aşırı öğrenme olmasını engellemek için kullanılan parametrelerden biridir. Kesinlikle olumlu değer olmalıdır. Varsayılan değeri 1'dir. 0.00001, 0.0001, 0.001,0.1 ve 1 değerleri denenmiş En iyi değer 1 bulunmuştur. Algoritmada bu değer kullanılmıştır.

Gamma: Seçilen çekirdek için kullanılan katsayı parametresidir. 0.0001, 0.001, 0.1, 1, scale ve auto değerleri denenmiştir. En iyi sonuç Scale fonksiyonuyla alınmış ve kullanılmıştır.

Max_iter: Yapılan iterasyon sayıdır. 1120 kez iterasyon yapılmıştır.

SEKİZİNCİ BÖLÜM
PERFORMANS ANALİZ SONUÇLARI

Tablo 8.1'de Karar Ağaçları, Rastgele Orman, YSA, DVM Algoritmalarının karşılaştırmalı performans sonuç değerleri verilmiştir.

	Doğruluk		**Kesinlik**		**Duyarlılık**		**F1 Score**	
	SMOTE Öncesi	**SMOTE Sonrası**	**SMOTE Öncesi**	**SMOTE Sonrası**	**SMOTE Öncesi**	**SMOTE Sonrası**	**SMOTE Öncesi**	**SMOTE Sonrası**
Karar Ağaçları	0.84	0.85	0.42	0.87	0.25	0.83	0.32	0.85
Rastgele Orman	0.86	0.90	0.64	0.91	0.22	0.89	0.33	0.90
YSA	0.86	0.89	0.57	0.88	0.27	0.89	0.37	0.89
DVM	0.86	0.88	0.66	0.86	0.20	0.91	0.31	0.88

Tablo 8.1: SMOTE Sonrası Karşılaştırmalı Model Performans Sonuçları

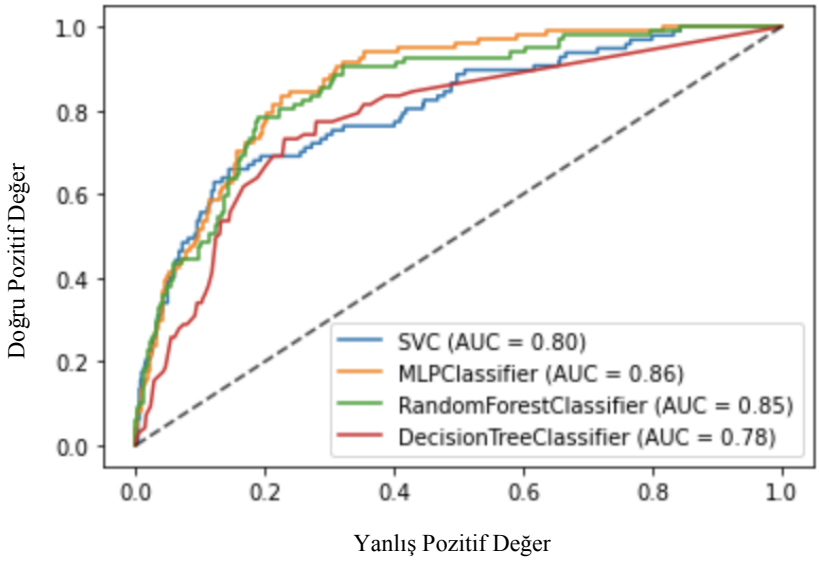
Tablo 8.1 incelendiğinde SMOTE öncesi ve sonrası değerler arasında fark olduğu görülmektedir. Rastgele Orman Algoritması SMOTE uygulanmadan önce de orta bir performans sergilemiştir ancak SMOTE ile performansında daha fazla artış sağlanmıştır ve problem çözümü için daha uygun algoritma olmuştur. Rastgele Orman Algoritmasından sonra YSA en iyi performansı göstermiştir. Karar Ağaçları ise en düşük performansı gösteren algoritmadır. Algoritmaların Doğruluk, Kesinlik, Duyarlılık ve F1 Score değerlerinde SMOTE öncesine göre artış görülmüştür.

Tablo 8.2'de eğitilen modellerin ROC eğrisi performans sonuçları karşılaştırmalı olarak verilmiştir.

Tablo 8.2: ROC Eğrisi AUC Performans Değerlendirme Sonuçları

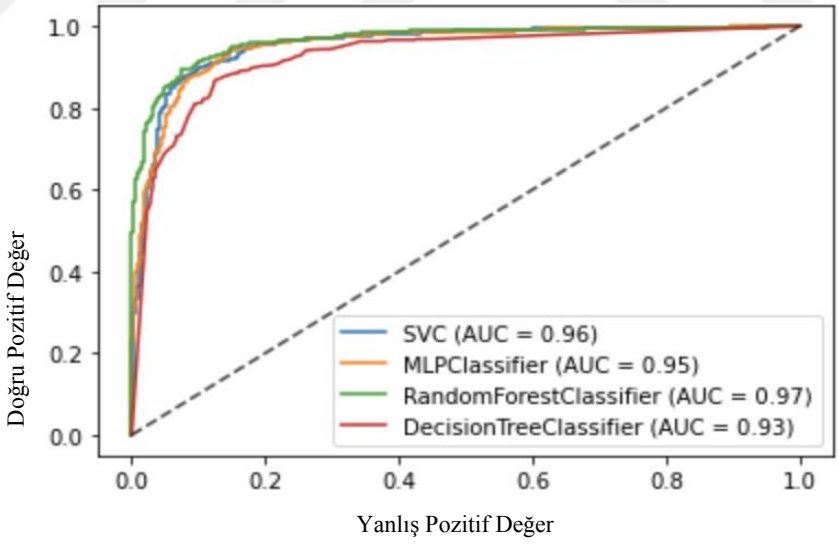
	Orijinal Veri	**SMOTE**
Karar Ağaçları	0.78	0.93
Rastgele Orman	0.85	0.97
YSA	0.86	0.95
DVM	0.80	0.96

ROC eğrisi modelin tahmininde ne kadar iyi olduğunu gösterir. AUC değeri ne kadar yüksekse iki sınıfın birbirinden hatasız ayrılması o kadar yüksek olduğu sonucu elde edilmektedir. Algoritmaların ne kadar doğru sonuç verdiğini ROC eğrisi ile analiz edilebilir.



Şekil 8.1: SMOTE Öncesi Karşılaştırmalı ROC Eğrisi

SMOTE öncesinde Şekil 10.1'de ROC eğrisinde görüldüğü gibi dengesiz bir eğri olduğu görülmektedir. SVM Algoritması %80, YSA Algoritması %86, Rastgele Orman Algoritması %85 ve Karar Ağaçları Algoritması %78 performans gerçekleşmiştir. Daha sonra aynı karşılaştırma SMOTE işlemi yapıldıktan sonra tekrar yapılmıştır.



Şekil 8.2: SMOTE Sonrası Karşılaştırmalı ROC Eğrisi

SMOTE işlemi yapıldıktan sonraki AUC sonuçları Şekil 8.2'de gösterilmiştir. İşlem sonrasında, SVM Algoritması %96, YSA Algoritması %95, Rastgele Orman Algoritması %97 ve Karar Ağaçları Algoritması %93 performans sonuçları elde edilmiştir. SMOTE sonrası yapılan bu analiz sonucunda karşılaştırma yapıldığında diğerlerinden daha iyi sonuç %97 ile Rastgele Orman Algoritması hesaplanmıştır.

Sonuç olarak sınıflandırma probleminde dengesizlik problemine çözüm önerilmiş ve performans değerleri iyileştirilmiştir.

8.1. ÇALIŞMANIN KATKILARI

Yapılan çalışmalardaki uygulanan algoritmalara ilave olarak performans iyileştirme ve SMOTE çalışması yapılmıştır. SMOTE işlemi öncesi ve sonrasındaki performans başarıları kıyaslanmıştır. Bunun sonucunda performans sonucunun SMOTE uygulandıktan sonra daha başarılı olduğu görülmüştür. Özniteliklerin algoritmaları ne kadar etkiledikleri izlenerek PCA tekniği uygulanmış ve öznitelik indirgeme işlemi yapılmıştır. Bu şekilde hem daha hızlı hem de daha başarılı performans sonuçları elde edilmiştir. Dengeli dağılmayan veri setlerinde model başarısını sadece Doğruluk metriği ile ölçmek yeterli değerlidir. Duyarlılık, Kesinlik ve F1 metriklerini mutlaka kontrol edilmelidir. F1 metriği sınıflandırmada modelin gerçek başarısını göstermektedir. Değerlendirme yapılırken telekomünikasyon sektörü örneği için Duyarlılık, Doğruluk, Kesinlik ve F1 Score değerlerinin beraber değerlendirilmesi gerektiği sonucuna varıldı.

Eğitimler bir telekomünikasyon şirketi örneği için uygulanmıştır. Literatürdeki çalışmalara ek SMOTE işlemi uygulanarak bu yöntemle daha iyi tahmin ve analiz sonuçlarına ulaşıldığı gözlenmiştir. Dengesiz sınıf problemlerine yeni bir yaklaşım getirilmiş ve yapılacak olan çalışmalara yeni bir bakış açısı kazandırmak hedeflenmiştir.

SONUÇ

Bu çalışmada, bir telekomünikasyon şirketi veri setini kullanarak müşteri kaybını öngören bir sistem geliştirilmiştir. Daha önceki makaleler, tezler ve ilgili çalışmalar incelenerek dengesiz sınıf problemlerine çözüm önerilmiştir. Bu tür öngörüleri oluşturmak için Rastgele Orman, Karar Ağaçları, YSA ve DVM algoritmaları kullanılarak modeller tasarlanmıştır. Veri seti incelenmiş öznitelik önem değerleri bulunarak öznitelik indirgeme işlemleri yapılmıştır. Makine öğrenmesi algoritmaları doğrudan kategorik verilerle çalışamadığı için verilere şifreleme (encoding) işlemi uygulanarak kategorik sayılara dönüştürülmüştür. Öznitelikler arasındaki ilişkiler gözlenmiş, veri setinden genel çıkarımlar yapılmıştır. İnceleme sonucunda dengesiz sınıf problemiyle karşılaşılmıştır. Bunu çözmek için SMOTE işlemi kullanılmıştır. SMOTE tekniğinin Doğruluk değeri üzerindeki etkisini görmek için iki farklı yöntem uygulanmıştır. İlk olarak modeller dengesiz sınıf problemini içeren orijinal verilerle eğitildi. Daha sonra SMOTE tekniği orijinal verilere uygulandı ve modeller bu yeni verilerle tekrar eğitildi. Bu iki veri setinde dört farklı algoritma ile eğitilmiş dört farklı modelin sonuçları gösterilmektedir.

Yapılan deneysel sonuçlara göre SMOTE uygulanan veriler için elde edilen Doğruluk değerleri oluşturulan modeller için daha başarılı performans sergilediği görüldü. Yapılan eğitimler ve oluşturulan modeller sonucunda SMOTE işlemi uygulanmadan önce Rastgele Orman Algoritması %85, Destek Vektör Makineleri %80, Karar Ağaçları %78 ve Yapay Sinir Ağları %86 performans sonuçları elde edilmiştir. SMOTE işlemi uygulandıktan sonra ise Rastgele Orman Algoritması %97, Destek Vektör Makineleri %96, Karar Ağaçları %93 ve Yapay Sinir Ağları %95 performans sonuçları elde edilmiştir. Bu durumda SMOTE işlemi sonrası Rastgele Orman Algoritması %12, Destek Vektör Makineleri %16, Karar Ağaçları %15 ve Yapay Sinir Ağları %9 performans artışı sağlamıştır. Bu çalışmada diğerlerinden daha iyi performansı veren algoritma Rastgele Orman algoritmasıdır. Daha az öznitelik ile daha fazla performans sergilediği ve modelin daha kısa zaman içerisinde çalışması da sağlanmıştır.

KAYNAKÇA

Abdi, H., Lynne, J. (2010). *Principal Component Analysis. Wiley Interdisciplinary Reviews: Computational Statistics*, 2(4), 433-459.

Akbulut, S. (2006). *Veri Madenciliği Teknikleri İle Bir Kozmetik Markanın Ayrılan Müşteri Analizi Ve Müşteri Segmentasyonu* (Yüksek Lisans Tezi). Gazi Üniversitesi Fen Bilimleri Enstitüsü, Ankara, 103.

Akpınar, H.(2014), Data Veri Madenciliği Veri Analizi, *Papatya Bilim*, 1-448.

Albayrak, M. (2008). *EEG Sinyallerindeki Epileptiform Aktivitenin Veri Madenciliği Süreci İle Tespiti* (*Doktora Tezi*). *Sakarya Üniversitesi Fen Bilimleri Enstitüsü*, Sakarya, 130.

Ali, W. (2017). Phishing Website Detection Based On Supervised Machine Learning With Wrapper Features Selection. *International Journal Of Advanced Computer Science And Applications (IJACSA)*, 8(9), 72-78.

Alsakran, J., Al-Kadi, O., Faris, H., Fayyoumi, A., Rodan, A. (2014). Negative Correlation Learning For Customer Churn Precision: A Comparison Study. *Hindawi Publishing Corporation The Scientific World Journel*, 1(1), 1-6.

Anaconda Navigator, Erişim Adresi: https://www.anaconda.com/products/individual, Erişim: 07.04.2021.

Barrett, P., Greenfield, P., Hunter, J., Hsu, J. C., Miller, J. T. (2004). *A Portable Python Plotting Package, Astronomical Data Analysis Software And Systems XIV ASP Conference Series.* Proceedings Of The Conference, In Pasadena, California, USA, 347(1), 91-95.

Başarslan, M. S. (2017). *Telekomünikasyon Sektöründe Müşteri Kayıp Analizi* (Yüksek Lisans Tezi). Düzce Üniversitesi Fen Bilimleri Enstitüsü, Düzce, *126.*

Bek, Y., Tomak, L. (2010). İşlem Karakteristik Eğrisi Analizi Ve Eğri Altında Kalan Alanların Karşılaştırılması. *Journal Of experimental and Clinical Medicine*, 27(2), 58-65.

Blondel, M., Cournapeau, D., Dubourg, V., Gramfort, A., Grisel ,O., Michel, V., Passos,

A., Pedregosa, F., Prettenhofer, P., Thirion, B., Vanderplas, J., Varoquaux, G. Weiss, R. (2011). Scikit-learn: Machine Learning In Python. *Journal Of Machine Learning Research*, 12(2011), 2826-2830.

Bowyer, K.W., Chawla, N. V., Hall, L. O., Kegelmeyer, W. P. (2002). SMOTE: Synthetic

Minority Over-sampling Technique. *Journal Of Artificial Intelligence Research*, 16(1), 321–357.

Chakraborti, S., Michaelson, G. And Mccracken, A.K. (2012). *Shortest Expected Length*

Confidence Interval For The Power Of The T-Test. Communications In Statistics

Simulation And Computation, 41(1): 1336– 1345.

Çelik, E. (2019). *Customer Churn Analysis Based On Machine Learning By Using Data Mining Techniques In Telecommunication Sector* (Yüksek Lisans Tezi). Yeditepe Üniversitesi Yönetim Bilişim Sistemleri, İstanbul, 121.

Chen, W., Finley, T., Ke, G., Ma, W., Ye, Q., Wang, T. (2017). *A Highly Efficient Gradient Boosting Decision Tree.* Advances Neural Information Processing System, 31(1) ,3149–3157.

Çınar, A. (2019). Veri Madenciliğinde Sınıflandırma Algoritmalarının Performans Değerlendirmesi Ve R Dili İle Bir Uygulama. *Öneri Dergisi*, 14(51), 90-111.

Donat, H. (2019). *Öznitelik Seçimi İle Telekomünikasyon Sektöründe Kayıp Müşteri Analizi* (Yüksek Lisans Tezi). İstanbul Üniversitesi Cerrah Paşa Lisansüstü Eğitim Enstitüsü, İstanbul, 104.

Farag, A. A., & Elhabian, S. (2009). A Tutorial On Principal Component Analysis. Erişim

Adresi: http://dai.fmph.uniba.sk/courses/ml/sl/PCA.pdf, Erişim: 07.04.2021.

Faris, H. (2018). *A Hybrid Swarm Intelligent Neural Network Model For Customer*

Churn Precision And Identifying The Influencing Factors. King Abdullah II School For Information Technology, The University Of Jordan, Jordan, 9(1),1-288.

Gaber, M. M., Krishnaswamy, S., Zaslavsky, A. (2005). *Mining Data Streams: A Review*,
ACM Sigmod Record, 34(2),1-22.

Goodman, S. N. (1993). P Values, Hypothesis Tests And Likelihood: Impli-Cations For Epidemiology Of A Neglected Historical Debate. *American Journal Of Epidemiology*, 137(5), 485–496.

Guo-en, X., Wei-dong, J. (2008). *Model of Customer Churn Precision On Support Vector*
Machine. System Engineering – Theory & Practice, 28(1), 71-77.

Gupta, A., Jain, S., Jangid, Y., Sawant, R., Tiwari, T. (2018). *Comprehensive Analysis Of Housing Price Precision In Pune Using Multi-Featured Random Forest Approach*. 2018 Fourth International Conference On Computing Communication Control And Automation (ICCUBEA), India, 1-5.

Hadden, J., Roy, R., Ruta, D., Tiwari, A. (2006). Churn Precision using Complaints Data.
World Academy Of Science Engineering and Technology, 19(1), 158-163.

Hamalainen, W., Matilainen, P., Mononen, J., Mughal M., Ruuska, S. (2018). Evaluation
Of The Confusion Matrix Method In The Validation Of An Automated System For Measuring Feeding Behaviour Of Cattle. *Behavioural Processes*, 148(1), 56-62.

Ilgın, H. A., Kozal, A. Ö., Teke, M. (2013). *Comparative Analysis Of Hyperspectral Dimension Reduction Methods*. 21 St Signal Processing And Communications Applications Conference, 1(1), 1-10.

Joshi, P., Gupta, S. (2019). Predicting Customers Churn In Telecom Industry Using Centroid Oversampling Method And KNN Classifier. *International Research Journal Of Engineering And Technology (IRJET)*, 6(4), 3708-3712.

Jupiter Notebook, Erişim Adresi: https://jupyter.org/, Erişim: 07.04.2021.

Kaptan, F. (2019). *Müşteri Kayıp Analizi: Hava Yolu Sektöründe Bir Uygulama* (Yüksek Lisans Tezi). İstanbul Teknik Üniversitesi Fen Bilimleri Enstitüsü, İstanbul, 129.

Kayaalp, F. (2017). Telekomünikasyon Sektöründe Müşteri Ayrılma Tahmin Analizi Çalışmaları Derlemesi. *Karaelmas Fen ve Mühendislik Dergisi*, 7(2), 696-705.

Kim, S., Shin, K.S., Park, K. (2005). *An Application Of Support Vector Machines For Customer Churn Analysis: Credit Card Case*. International Conference On Natural Computation, Advances In Natural Computation, 3611(1), 636-647.

Koyuncugil, A. S., Özgülbaş, N. (2009). Tıp ve Sağlık Hizmetlerinde Kullanımı Ve Uygulamaları. *Bilişim Teknolojileri Dergisi*, 2(2), 21-30.

Köksal, B. (2011). *Regresyon Analizinde ROC Eğrisi Kestirimi İle Model Seçimi* (Yüksek Lisans Tezi). Marmara Üniversitesi Sosyal Bilimler Enstitüsü, İstanbul, 157.

Kunt, M. S. (2019). Telekomünikasyon Sektöründe Müşteri Kayıp Analizi (Yüksek Lisans Tezi). Ankara Üniversitesi Fen Bilimleri Enstitüsü, Ankara, 67.

Leys, C. And Schumann, S. (2010). A Nonparametric Method To Analyze Ineraction. *Journal Of Experimental Social Psychology*, 46(1), 684–688.

Li, B., X, Jiuzuo. (2020). Study on the Precision of Imbalanced Bank Customer Churn Based On Generative Adversarial Network. *Journal Of Physics: Conferance Series, Applied Mathematics, Data Analysis and Data Mining*,1624(1), 1-7.

Lu, Y. H., Tsai, C. F. (2010). Data Mining Techniques In Customer Churn Precision. *Recent Patents On Computer Science*, 3(1), 28–32.

Lu, Y.- H., Tsai, C. -F. (2009). Customer Churn Precision By Hybrid Neural Networks. *Expert Systems With Applications*, 36(10), 12547-12553.

Matplotlib, Erişim Adresi: https://matplotlib.org/, Erişim: 07.04.2021.

McLeod, S. A. (2019). *What A P-Value Tells You About Statistical Significance.* Simply Psychology. Erişim Adresi: https://www.simplypsychology.org/p-value.html, Erişim: 07.04.2021.

Müslim, M. A., Safitri, A. R. (2020). Improved Accuracy Of Naive Bayes Classifier For

Determination Of Customer. *Journal Of Soft Computing, Exploration*, 1(1), 70-75.

Numpy, Erişim Adresi: https://numpy.org/, Erişim: 07.04.2021.

Odabaş, Ö. (2017). *Veri Madenciliği Teknikleri İle Telekom Sektöründe Ayrılan Müşteri Analizi* (Yüksek Lisans Tezi). İstanbul Ticaret Üniversitesi Fen Bilimleri Enstitüsü, İstanbul, 60.

PangKW, (2018), Erişim Adresi: https://www.kaggle.com/pangkw/telco-Churn, Telco Churn, Erişim: 07.04.2021.

Pandas, Erişim Adresi: https://pandas.pydata.org/, Erişim: 07.04.2021.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E. (2011), Scikit-Learn: Machine Learning in Python, *Journal of Machine Learning Research*, 12(1), 2825-2830.

Pendharkar, P. C. (2009). Genetic Algorithm Based Neural Network Approaches For Predicting Churn In Cellular Wireless Network Services. *Expert Systems With Applications*, 36(3), 6714-6720.

Pirim, H. (2006). Yapay Zeka, *Journal Of Yaşar University*, 1(1), 81-93.

Sharma, A., Panigrahi, P. K. (2011). A Neural Network Based Approach For Predicting Customer Churn In Cellular Network Services. *International Journal Of Computer Applications*, 27(11), 1-30.

Scikit-Learn, Erişim Adresi: https://scikit-learn.org/, Erişim: 07.04.2021.

Scikit-Learn, (2021a). Scikit-Learn Machine Learning in Phyton, Erişim Adresi: https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html/, Erişim: 07.04.2021.

Scikit-Learn, (2021b). Scikit-Learn Machine Learning in Phyton, Erişim Adresi: https://scikit-learn.org/stable/modules/generated/sklearn.neural_network.MLPClassifier.html/, Erişim: 07.04.2021.

Scikit-Learn, (2021c). Scikit-Learn Machine Learning in Phyton, Erişim Adresi:

https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html/, Erişim: 07.04.2021.

Scikit-Learn, (2021d). Scikit- learn Machine Learning in Phyton, Erişim Adresi: https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html/, Erişim: 07.04.2021.

Wadikar, D., (2020). *Customer Churn Precision*. Technological University Dublin, 80.

Xie, Y., Li, X., Ngai, E. W. T.,Ying, W. (2019). Customer Churn Precision Using Improved Balanced Random Forests Elsevier. *Expert System With Application*, 36(1), 5445-5449.

Vadim, K. (2018). Overview Of Different Approaches To Solving Problems Of Data Mining. *Procedia Computer Science*, 123(2018), 234-239.

Yıldız, M. (2015). Telekomünikasyon Sektöründe Müşteri Ayrılma Tahmini. *Karaelmas Fen ve Müh. Dergisi* 7(2), 696-705.

DİZİN

LEE
Logo

Kütahya Dumlupınar Üniversitesi
Lisansüstü Eğitim Enstitüsü Bilgisayar Mühendisliği
Anabilim Dalı

MAKİNE ÖĞRENMESİ YÖNTEMLERİYLE MÜŞTERİ KAYIP ANALİZİ

Zerrin ÇAKIR

2021