



AKDENİZ ÜNİVERSİTESİ
SOSYAL BİLİMLER ENSTİTÜSÜ



Erokan CANBAZOĞLU

VERİ MADENCİLİĞİ TEKNİKLERİ İLE SAĞLIK TURİZMİNDE
REZERVASYON İPTALLERİNİN TAHMİNİ

Ekonometri Ana Bilim Dalı
Doktora Tezi

Antalya, 2024



AKDENİZ ÜNİVERSİTESİ
SOSYAL BİLİMLER ENSTİTÜSÜ



Erokan CANBAZOĞLU

VERİ MADENCİLİĞİ TEKNİKLERİ İLE SAĞLIK TURİZMİNDE
REZERVASYON İPTALLERİNİN TAHMİNİ

Danışman

Prof. Dr. Emre İPEKÇİ ÇETİN

Ekonometri Ana Bilim Dalı

Doktora Tezi

Antalya, 2024

T.C.
Akdeniz Üniversitesi
Sosyal Bilimler Enstitüsü Müdürlüğüne,

Erokan Canbazoğlu'nun bu çalışması, jürimiz tarafından Ekonometri Anabilim Dalı Doktora Programı tezi olarak kabul edilmiştir.

Başkan : Prof. Dr. Kenan Oğuzhan ORUÇ (İmza)

Üye (Danışmanı) : Prof. Dr. Emre İPEKÇİ ÇETİN (İmza)

Üye : Prof. Dr. Sezgin IRMAK (İmza)

Üye : Prof. Dr. Mehmet MERT (İmza)

Üye : Dr. Öğr. Üyesi Erdal ERDAL (İmza)

Tez Başlığı: Veri Madenciliği Teknikleri ile Sağlık Turizminde Rezervasyon İptallerinin Tahmini

Onay: Yukarıdaki imzaların, adı geçen öğretim üyelerine ait olduğunu onaylarım.

Tez Savunma Tarihi : 15/02/2024

Mezuniyet Tarihi : 14/03/2024

(imza)
Prof. Dr. Seda BAYRAKTAR
Müdür

AKADEMİK BEYAN

Doktora Tezi olarak sunduğum “Veri Madenciliği Teknikleri ile Sağlık Turizminde Rezervasyon İptallerinin Tahmini” adlı bu çalışmanın, akademik kural ve etik değerlere uygun bir biçimde tarafımdan yazıldığını, yararlandığım bütün eserlerin kaynakçada gösterildiğini ve çalışma içerisinde bu eserlere atıf yapıldığını belirtir, bunu şerefimle doğrularım.

Erokan CANBAZOĞLU





T.C.
AKDENİZ ÜNİVERSİTESİ
SOSYAL BİLİMLER ENSTİTÜSÜ



TEZ ÇALIŞMASI ORJİNALLİK RAPORU BEYAN BELGESİ

ÖĞRENCİ BİLGİLERİ	
Adı-Soyadı	Erokan CANBAZOĞLU
Öğrenci Numarası	20145245001
Ana Bilim Dalı	Ekonometri
Programı	Doktora
Danışman Unvanı, Adı-Soyadı	Prof. Dr. Emre İPEKÇİ ÇETİN
Tez Başlığı	Veri Madenciliği Teknikleri ile Sağlık Turizminde Rezervasyon İptallerinin Tahmini
Turnitin Ödev Numarası	2271826021
Rapor Tarihi	11/03/2024
Benzerlik Oranı	Alıntılar hariç: %3 Alıntılar dahil: %3

SOSYAL BİLİMLER ENSTİTÜSÜ MÜDÜRLÜĞÜNE,

Yukarıda bilgileri bulunan öğrenciye ait tez çalışmasının a) Kapak sayfası, b) Giriş, c) Ana Bölümler ve d) Sonuç kısımlarından oluşan toplam 117 sayfalık kısmına ilişkin olarak Turnitin adlı intihal tespit programından Sosyal Bilimler Enstitüsü Tez Çalışması Orijinallik Raporu Alınması ve Kullanılması Uygulama Esaslarında belirlenen filtrelemeler uygulanarak yukarıdaki detayları verilen ve ekte sunulan rapor alınmıştır.

Danışman tarafından uygun olan seçenek işaretlenmelidir:

Benzerlik oranları belirlenen limitleri aşmıyor ise:

(X) Yukarıda yer alan beyanın ve ekte sunulan Tez Çalışması Orijinallik Raporunun doğruluğunu onaylarım.

Benzerlik oranları belirlenen limitleri aşıyor, ancak tez danışmanı intihal yapılmadığı kanısında ise:

() Yukarıda yer alan beyanın ve ekte sunulan Tez Çalışması Orijinallik Raporunun doğruluğunu onaylar ve Uygulama Esaslarında öngörülen yüzdelik sınırlarının aşılmasına karşın, aşağıda belirtilen gerekçe ile intihal yapılmadığı kanısında olduğumu beyan ederim.

Gerekçe:

Benzerlik taraması yukarıda verilen ölçütlere uygun olarak tarafımca yapılmıştır. İlgili tezin orijinallik raporunun uygun olduğunu beyan ederim.

Danışman Öğretim Üyesi
Unvanı, Adı-Soyadı
Prof. Dr. Emre İPEKÇİ ÇETİN
(imza)

İÇİNDEKİLER

ŞEKİLLER LİSTESİ	iv
TABLOLAR LİSTESİ	vi
KISALTMALAR LİSTESİ	vii
ÖZET	viii
ABSTRACT	ix
TEŞEKKÜR.....	x
ÖNSÖZ	xi
GİRİŞ.....	1

BİRİNCİ BÖLÜM

SAĞLIK TURİZMİ REZERVASYON İPTAL TAHMİNİ

1.1. Sağlık Turizmi	4
1.2. Sistematik Literatür Taraması.....	7
1.2.1. Veri Toplama ve Veri Ön İşleme	10
1.2.2. Konu Daraltma	12
1.2.3. Otomatik Analizler	16
1.2.4. Manuel Analiz	18

İKİNCİ BÖLÜM

VERİ MADENCİLİĞİ

2.1. Veri Madenciliği Terimleri	25
2.2. Veri Madenciliği Tanımı	28
2.3. Veri Madenciliği Yazılımları	29
2.4. CRISP-DM Veri Madenciliği Süreci	31
2.4.1. İşi Anlama	32
2.4.2. Veriyi Anlama	35
2.4.3. Veriyi Hazırlama	37
2.4.4. Modelleme.....	40
2.4.5. Değerlendirme	42
2.4.6. Uygulama	43

ÜÇÜNCÜ BÖLÜM

VERİ MADENCİLİĞİ SINIFLANDIRMA YÖNTEMLERİ

3.1. Veri Madenciliği Sınıflandırma Yöntemleri.....	45
3.1.1. ZeroR.....	45
3.1.2. Naive Bayes	46
3.1.3. K-En Yakın Komşu.....	48
3.1.4. Destek Vektör Makineleri	51
3.1.5. Lojistik Regresyon	53
3.1.6. Karar Ağaçları	55
3.1.7. Yapay Sinir Ağları	58
3.1.7.1. Biyolojik Sinir Hücresi.....	59
3.1.7.2. Yapay Sinir Hücresi Yapısı	60
3.1.7.3. Yapay Sinir Ağı Genel Yapısı	61
3.1.7.3.1. Tek Katmanlı Yapay Sinir Ağları	61
3.1.7.3.2. Çok Katmanlı Yapay Sinir Ağı Genel Yapısı.....	62
3.1.7.4. Nöronlar Arası Bağlantı Yapılarına Göre YSA.....	63
3.1.7.4.1. İleri Beslemeli YSA.....	63
3.1.7.4.2. Geri Beslemeli YSA	63
3.1.8. Topluluk Öğrenmesi Teknikleri	65
3.1.8.1. Rastgele Orman	67
3.1.8.2. AdaBoost	67
3.1.8.3. XGBoost	67
3.1.8.4. CatBoost	68
3.1.8.5. LightGBM	68
3.2. Veri Madenciliği Yöntemlerinin Karşılaştırılması	68
3.3. Model Değerlendirme Kriterleri	70

DÖRDÜNCÜ BÖLÜM

VERİ MADENCİLİĞİ TEKNİKLERİ İLE ÇEVİRİM İÇİ KARAR DESTEK SİSTEMİ TASARIMI: SAĞLIK TURİZMİ UYGULAMASI

4.1. Araştırma Problemi.....	76
4.2. Araştırma Amacı.....	76
4.3. Araştırmanın Önemi	77
4.4. Araştırma Kısıtları	77
4.5. Araştırmanın Katkısı.....	78

4.6. Araştırma Süreci	78
4.6.1. İşi Anlama	78
4.6.2. Veriyi Anlama	78
4.6.3. Verinin Hazırlanması	86
4.6.4. Modelleme.....	89
4.6.4.1. ZeroR	90
4.6.4.2. Naive Bayes.....	90
4.6.4.3. K-En Yakın Komşu	91
4.6.4.4. Destek Vektör Makineleri	93
4.6.4.5. Lojistik Regresyon.....	94
4.6.4.6. Karar Ağaçları	96
4.6.4.7. Yapay Sinir Ağları.....	97
4.6.4.8. Rastgele Orman	98
4.6.4.9. AdaBoost	100
4.6.4.10. XGBoost	101
4.6.4.11. CatBoost	103
4.6.4.12. LightGBM	104
4.6.5. Değerlendirme.....	105
4.6.6. Uygulama	110
TARTIŞMA VE SONUÇ	113
KAYNAKÇA.....	117
ÖZGEÇMİŞ	132

ŞEKİLLER LİSTESİ

Şekil 1.1 Literatür Tarama Metodolojisi	8
Şekil 1.2 Scopus Arama Dizisi ve Çalışma Sayısı	10
Şekil 1.3 Web of Science Arama Dizisi ve Çalışma Sayısı.....	11
Şekil 1.4 Veri Seti İndirgeme Adımları.....	11
Şekil 1.5 Unigram ve Bigram Kelime Bulutları	13
Şekil 1.6 Idatuning Paketi ile LDA Konu Sayısına Karar Verme	14
Şekil 1.7 Tüm Dokümanlara ait Konu-Kelime Dağılımı	15
Şekil 1.8 Tüm Araştırmaların Konu Başlıklarına Göre Dağılımları	15
Şekil 1.9 İndirgenmiş Son Veri Seti için LDA Analizi	16
Şekil 1.10 İndirgenmiş Son Dokümanlara ait Konu- ve Kelime Dağılımları	17
Şekil 1.11 İndirgenmiş Dokümanların Konu Başlıklarına Göre Dağılımları.....	18
Şekil 2.1 VEBB Piramidi.....	25
Şekil 2.2 Veriden Bilgelik Yolculuğu	27
Şekil 2.3 Veri Biliminde Tercih Edilen Teknolojiler	31
Şekil 2.4 CRISP-DM Veri Madenciliği Süreci	32
Şekil 2.5 CRISP-DM Birinci Basamak: İş Anlama.....	33
Şekil 2.6 CRISP-DM İkinci Basamak: Veriyi Anlama	36
Şekil 2.7 CRISP-DM Üçüncü Basamak: Veriyi Hazırlama	38
Şekil 2.8 CRISP-DM Dördüncü Basamak: Modelleme Görev ve Çıktıları	40
Şekil 2.9 CRISP-DM Beşinci Basamak: Değerlendirme Görev ve Çıktıları	42
Şekil 2.10 CRISP-DM Altıncı Basamak: Uygulama Görev ve Çıktıları	43
Şekil 3.1 Temel Karar Ağacı Yapısı.....	56
Şekil 3.2 Sinir Hücresi Yapısı	59
Şekil 3.3 Yapay Sinir Hücresi Yapısı	60
Şekil 3.4 Tek Katmanlı YSA	61
Şekil 3.5 Çok Katmanlı YSA	62
Şekil 3.6 Geri Beslemeli YSA Kaynak: (Can, 2022: 54)	64
Şekil 3.7 Torbalama.....	66
Şekil 3.8 Güçlendirme	67
Şekil 3.9 Alıcı Çalışma Karakteristik (ROC) Örnek Eğrileri	74
Şekil 4.1 Sağlık Turistlerinin Geldikleri Ülkelere Göre Dağılımları	81
Şekil 4.2 Özel İsteklere Göre İptal Durumu	82

Şekil 4.3 Tekrarlı Sağlık Turistinin Rezervasyon İptal Etme Durumu	82
Şekil 4.4 Depozito Durumuna Göre İptal Durumu.....	83
Şekil 4.5 Otel Türlerine Göre Aylık Sağlık Turisti Sayıları.....	83
Şekil 4.6 Tedavi Tarihi ve Rezervasyon Tarihi Arasında Geçen Süre	84
Şekil 4.7 Tedavi ve Varış Süresinin İptallere Etkisi.....	84
Şekil 4.8 Varış Süresinin Yoğunluk Dağılımı	85
Şekil 4.9 İptal Durumunun Veri Setindeki Dağılımı	86
Şekil 4.10 Değişkenlere ait Korelasyon Matrisi	87
Şekil 4.11 Naive Bayes ROC ve En Çok Katkı Sağlayan Değişkenler.....	91
Şekil 4.12 KNN için ROC ve En Çok Katkı Sağlayan Değişkenler	92
Şekil 4.13 DVM için ROC ve En Çok Katkı Sağlayan Değişkenler.....	94
Şekil 4.14 LR için ROC ve En Çok Katkı Sağlayan Değişkenler	95
Şekil 4.15 Karar Ağaçları için ROC ve En Çok Katkı Sağlayan Değişkenler	97
Şekil 4.16 YSA için ROC ve En Çok Katkı Sağlayan Değişkenler	98
Şekil 4.17 RO için ROC ve En Çok Katkı Sağlayan Değişkenler	99
Şekil 4.18 AdaBoost için ROC ve En Çok Katkı Sağlayan Değişkenler	101
Şekil 4.19 XGBoost için ROC ve En Çok Katkı Sağlayan Değişkenler	102
Şekil 4.20 CatBoost için ROC ve En Çok Katkı Sağlayan Değişkenler	104
Şekil 4.21 LightGBM için ROC ve En Çok Katkı Sağlayan Değişkenler	105
Şekil 4.22 Google Colab Çalışma Ortamı	111
Şekil 4.23 Modelin Pickle ile Dışarı Aktarılması.....	111
Şekil 4.24 Sağlık Turizmi Rezervasyon Formu Örneği	112
Şekil 4.25 Tahmin Modelinin Canlı Sisteme Yüklenmesi	112

TABLOLAR LİSTESİ

Tablo 1.1 Rezervasyon İptal Çalışmalarının Manuel Analizleri	19
Tablo 3.1 Örnekler Arası Uzaklık Tanımlama Yaklaşımları	50
Tablo 3.2 Sınıflandırma Algoritmalarının Karşılaştırılması.....	69
Tablo 3.3 Güçlendirme Algoritmaları Arasındaki Farklar	70
Tablo 3.4 Karışıklık Matrisi ve Değerlendirme Metrikleri	70
Tablo 3.5 Sınıflandırma Algoritmalarının Karşılaştırılma Metrikleri	71
Tablo 4.1 Değişken Adı ve Açıklamaları	79
Tablo 4.2 Sayısal Değişkenler için İstatistik Özeti.....	80
Tablo 4.3 Kategorik Değişkenler için İstatistik Özeti	80
Tablo 4.4 Azınlık Sınıfın Yüzdesine Göre Dengesiz Veri Seti Tespiti.....	85
Tablo 4.5 Naive Bayes için Karışıklık Matrisi	90
Tablo 4.6 KNN için Karışıklık Matrisi.....	92
Tablo 4.7 DVM için Karışıklık Matrisi	93
Tablo 4.8 LR için Karışıklık Matrisi	94
Tablo 4.9 Karar Ağaçları için Karışıklık Matrisi	96
Tablo 4.10 YSA için Karışıklık Matrisi	97
Tablo 4.11 RO için Karışıklık Matrisi.....	99
Tablo 4.12 AdaBoost için Karışıklık Matrisi	100
Tablo 4.13 XGBoost için Karışıklık Matrisi	101
Tablo 4.14 CatBoost için Karışıklık Matrisi	103
Tablo 4.15 LightGBM için Karışıklık Matrisi.....	104
Tablo 4.16 Temel Model Karşılaştırma Tablosu.....	106
Tablo 4.17 Parametreleri Optimize Edilmiş Sınıflandırıcıların Karşılaştırılması	106
Tablo 4.18 ADASYN ile Sınıflandırıcıların Karşılaştırılması	107
Tablo 4.19 ROSE ile Sınıflandırıcıların Karşılaştırılması.....	108
Tablo 4.20 SMOTE ile Sınıflandırıcıların Karşılaştırılması	108
Tablo 4.21 SMOTE-ENN ile Sınıflandırıcıların Karşılaştırılması.....	109
Tablo 4.22 En İyi Performansa Sahip Sınıflandırıcılara ait Özet Karşılaştırma	110

KISALTMALAR LİSTESİ

AdaBoost	Uyarlamalı Güçlendirme (Adaptive Boosting)
ADASYN	Uyarlanabilir Sentetik (Adaptive Synthetic)
AUC	Eğri Altında Kalan Alan (Area Under the Curve)
DT	Karar Ağaçları (Decision Trees)
CatBoost	Categorical Boosting
CRISP-DM	Cross-Industry Standard Process for Data Mining
DVM	Destek Vektör Makinaları
FN	Yanlış Negatif (False Negative)
FNR	Yanlış Negatif Oranı (False Negative Rate)
FP	Yanlış Pozitif (False Pozitif)
FPR	Yanlış Pozitif Oranı (False Positive Rate)
KDD	Veri Tabanlarında Bilgi Keşfi (Knowledge Discovery in Databases)
KNN	K En Yakın Komşu (K Nearest Neighbors)
LDA	Gizli Dirichlet Ayrımı (Latent Dirichlet Allocation)
LightGBM	Light Gradient Boosting Machine
ML	Makine Öğrenmesi (Machine Learning)
NB	Naive Bayes
NPV	Negatif Kestirim Değeri (Negative Predictive Value)
RO	Rastgele Orman
ROC	Alıcı İşletim Karakteristiği (Receiver Operating Characteristic)
ROSE	Rastgele Aşırı Örnekleme (Random Over Sampling Examples)
SEMMA	Sample, Explore, Modify, Model, Assess
SMOTE	Synthetic Minority Over-sampling Technique
TN	Doğru Negatif (True Negative)
TNR	Doğru Negatif Oranı (True Negative Rate)
TP	Doğru Pozitif
TPR	Doğru Pozitif Oranı (True Positive Rate)
VM	Veri Madenciliği
XGBoost	Aşırı Gradyan Artırma (EXtreme Gradient Boosting)
YSA	Yapay Sinir Ağları

ÖZET

Sağlık kuruluşları ve seyahat acentalarının sağlık turizmi süreçlerini etkin bir şekilde yönetmeleri, işletmenin itibarı ve en fazla kârı elde etmesi açısından çok kritik bir öneme sahiptir. Sağlık turizmi paydaşları, sosyal medya platformları aracılığı ile yaptıkları etkileyici tanıtımlar sayesinde sağlık turizmine ilgiyi arttırmış ve güçlü bir rekabet ortamı oluşturmuştur. Sağlık turistleri, rakip sağlık turizmi firmalarının sundukları cazip tedavi paketlerinin etkisi ile, yapmış oldukları rezervasyonları iptal edebilmektedir. İptal edilen rezervasyonları ön göremeyen ve gerekli önlemleri alamayan sağlık turizmi paydaşları, kaynaklarının doğru bir şekilde yönetilmesinde sorun yaşamakta ve bu durum gelir kaybı ile sonuçlanmaktadır. Doktora araştırması kapsamında, makine öğrenmesi teknikleri kullanılarak ve rezervasyon esnasında sağlık turistinin verilerini göz önünde bulundurarak rezervasyon iptal tahmini yapan “Sağlık Turizmi Yönetim Sistemi (STYS)” geliştirilmiştir. Kişisel veri içermeyen, anonim hale getirilmiş 5.060 adet sağlık turisti rezervasyon verisi, on iki farklı makine öğrenmesi yöntemi (ZeroR, Naive Bayes, K-En Yakın Komşu, Destek Vektör Makinaları, Lojistik Regresyon, Karar Ağaçları, Yapay Sinir Ağları, Rastgele Orman, XGBoost, CatBoost, AdaBoost ve LightGBM) ve on farklı model değerlendirme metriği (doğruluk, kesinlik, hassasiyet, F1-skor, özgüllük, FPR, FNR, NPV, ROC ve hız) kullanılarak sağlık turizmi rezervasyon iptal tahmin modelleri geliştirilmiştir. Veri sınıfı dengesizliklerinin önüne geçebilmek, tahmin sonuçlarını iyileştirebilmek ve daha tutarlı tahmin sonuçları elde edebilmek için ADASYN, ROSE, SMOTE ve SMOTE-ENN teknikleri ile veri seti dengeli hale getirilmiş ve sonuçlar karşılaştırılmıştır. Geliştirilen tahmin modelleri, sağlık turistlerinin rezervasyon iptallerinin başarılı bir şekilde tahmin edilebileceğini ve iptale neden olan bağımsız değişkenlerin tespit edilebileceğini göstermiştir. Tahmin modellerinin rezervasyon sistemleri içerisinde gerçek zamanlı olarak çalışabileceğini göstermek amacıyla bir prototip geliştirilmiştir. Sağlık turizmi paydaşları geliştirilen prototip sayesinde, rezervasyonlarını iptal edecek sağlık turistlerini tespit edebilmekte ve rezervasyon iptalinin azalmasını sağlayan gerekli aksiyonları alabilmektedir.

Anahtar Kelimeler: Sağlık Turizmi, Veri Madenciliği, Makine Öğrenimi, Rezervasyon İptali, Yapay Zekâ

ABSTRACT

**PREDICTION OF RESERVATION CANCELLATIONS IN HEALTH TOURISM
WITH DATA MINING TECHNIQUES**

Effectively managing health tourism processes by healthcare institutions and travel agencies is of critical importance for the reputation and maximum profitability of the business. Health tourism stakeholders have increased interest in health tourism and created a strong competitive environment through impressive promotions on social media platforms. Health tourists can cancel their reservations due to the attractive treatment packages offered by rival health tourism firms. Stakeholders in health tourism who fail to anticipate canceled reservations and take necessary precautions experience difficulties in managing their resources correctly, resulting in revenue loss. Within the scope of a doctoral research, a "Health Tourism Management System (HTMS)" has been developed using machine learning techniques and considering health tourist data during reservation to predict reservation cancellations. Anonymized data of 5,060 health tourist reservations, devoid of personal data, were utilized along with twelve different machine learning methods (ZeroR, Naive Bayes, K-Nearest Neighbor, Support Vector Machines, Logistic Regression, Decision Trees, Artificial Neural Networks, Random Forest, XGBoost, CatBoost, AdaBoost, and LightGBM) and ten different model evaluation metrics (accuracy, precision, recall, F1-score, specificity, FPR, FNR, NPV, ROC, and speed) to develop health tourism reservation cancellation prediction models. Techniques such as ADASYN, ROSE, SMOTE, and SMOTE-ENN were employed to balance the dataset, overcome data class imbalances, improve prediction results, and obtain more consistent prediction outcomes, and the results were compared. The developed prediction models have demonstrated that health tourist reservation cancellations can be successfully predicted, and independent variables causing cancellations can be identified. A prototype has been developed to demonstrate that prediction models can work in real-time within reservation systems. Health tourism stakeholders can identify health tourists who will cancel their reservations and take necessary actions to reduce reservation cancellations through the developed prototype.

Keywords: Health Tourism, Data Mining, Machine Learning, Reservation (Booking) Cancellation, Artificial Intelligence

TEŞEKKÜR

Akademik eğitim hayatının en önemli süreci olan doktora aşamasında desteğini her zaman yanımda hissettiğim, insanlığı ve akademisyenliği ile saygı duyduğum değerli hocam, tez danışmanım Prof. Dr. Emre İPEKÇİ ÇETİN'e teşekkürlerimi sunuyorum. Doktora eğitimim esnasında desteklerini hiçbir şekilde esirgemeyen ve öğrendiklerimi uygulamaya dönüştürmek için beni teşvik eden tüm Ekonometri Bölümü öğretim üyelerine teşekkür ederim. Tez süreci boyunca, tez izleme komitelerinde olumlu tavırları ve yapıcı eleştirileri ile tez çalışmamın ilerlemesine ve gelişim göstermesine katkı sağlayan değerli hocalarım Prof. Dr. Sezgin IRMAK ve Prof. Dr. Mehmet MERT'e teşekkürlerimi sunuyorum.

Başta Ekonometri Anabilim dalında doktora çalışmaları yapmak için yönlendiren ve cesaretlendiren Prof. Dr. Özcan ASILKAN'a, doktora yeterlilik ve tez süreci boyunca verdiği desteklerden dolayı Yusuf Ali DANIŞ'a, tez çalışmasında kullanılan "Sağlık Turizmi Yönetim Sistemi" yazılımı geliştirme sürecinde her türlü desteği sağlayan Talya Bilişim A.Ş. CEO'su Kemal ORAL'a, her daim desteklerinden dolayı öğrencilerim Ahmet Faruk HACIBALGİL, Samet Can SEZER ve Oğuzhan YALÇIN'a, akademisyenlik hayatına adım atmamda bana önder olan, desteğini her zaman yanımda hissettiğim ve beni her zaman cesaretlendiren andam Gökhan ÖZÇAYAN'a, badim Mehmet BOZ'a, Meltem Işık AŞKAR'a eşim Gülcan CANBAZOĞLU, annem Kadriye CANBAZOĞLU, abim Hakan CANBAZOĞLU ve ailesine, oğlum MEHMET YİĞİT CANBAZOĞLU, kızım ALYA CANBAZOĞLU ve tüm hayatım boyunca desteğini yanımda hissettiğim rahmetli babam Latif CANBAZOĞLU'na, şükranlarımı sunuyorum.

ÖNSÖZ

Sağlık turizmi alanında yapılan literatür taramaları ve piyasa araştırmalarının yanında sağlık turizmi yetki belgeli sağlık kuruluşları ve seyahat acentalarının sayısında da artış yaşanması, sağlık turizmine yönelim olduğunu göstermektedir. Sağlık turizmine duyulan ilginin son yıllarda artması, işletmelerin sosyal medya platformlarında etkin reklam ve tanıtım faaliyetleri ile sıkı bir rekabet ortamının oluşmasına yol açmıştır. Rakip firmaların cezbedici fiyat ve şartlarda tatil ve tedavi seçenekleri sunmaları, rezervasyonunu tamamlayan sağlık turistlerinin fikirlerinin değişmesine ve yapmış oldukları rezervasyonları iptal etmelerine yol açmaktadır. Rezervasyon iptallerinin artması, sağlık turizmi paydaşlarının gelir kaybına ve kaynaklarını etkin yönetmede zorluk yaşamalarına neden olmaktadır.

Doktora çalışması kapsamında, tüm hizmet sektörlerinde rezervasyon iptallerine yönelik yapılan akademik çalışmaları tespit edebilmek amacıyla otomatik literatür tarama yöntemi kullanılmıştır. Birinci bölüm içerisinde otomatik literatür tarama sürecinin tüm adım ve açıklamalarına detaylı bir şekilde yer verilmiştir. Rezervasyon iptal tahmini çalışmalarında CRISP-DM (Veri Madenciliği için Endüstriler Arası Standart Süreç) tercih edilmiş ve tüm süreç adım, görev ve çıktılarına ait teorik bilgiler ikinci bölümde detaylı bir şekilde ele alınmıştır. Veri madenciliği sınıflandırma problemlerinde kullanılan tahmin yöntemlerine ait teorik bilgiler üçüncü bölümde detaylı bir şekilde açıklanmıştır. Son bölümde, sağlık turizmi kapsamında rezervasyon iptallerinin tahmini konusunda CRISP-DM süreci tüm adımlarıyla işletilmiş ve tahmin modellerinin çıktıları listelenmiştir.

Sağlık turizmi rezervasyon verilerinin güçlüklerle temin edilebilmesi, elde edilen veri setinin dengeli bir veri seti olmaması, ve çok sayıda sınıflandırma tahmin modeli ve karşılaştırma metriğinin bulunması, bu çalışmanın önemli zorlukları arasında yer almıştır. Sağlık turizmi rezervasyon verilerinin temin edilebilmesi için “Sağlık Turizmi Yönetim Sistemi (STYS)” tasarlanmıştır. Dengesiz veri setine ADASYN, ROSE, SMOTE, SMOTE-ENN yöntemleri uygulanmış ve veri seti dengeli hale getirilmiştir.

Türkiye, modern sağlık kuruluşları, alanında uzman personelleri ve uygun fiyata sahip tedavi imkânları ile küresel çapta sağlık turizminde lider konuma gelebilecek yüksek potansiyele sahip ülkelerin başında gelmektedir. Ülkelerin kültürlerini, tarihi ve turistik mekanlarını tanıtmaya, yurt dışından döviz girişini sağlama ve ekonomiye can suyu sağlama konularında sağlık turizmi çok önemli bir sektördür. Rezervasyon iptal tahminlerini başarılı bir şekilde yapabilen sağlık turizmi paydaşları, karmaşık ve zorlu sağlık turizmi süreçlerini ve işletme kaynaklarını etkin bir şekilde yönetme imkanına sahip olmaktadır.

GİRİŞ

Hastalıkların tedavi edilmesi, bireylerin sađlıđının muhafaza edilmesi ve geliştirilmesi için bireylerin ikamet ettiđi yerden başka bir ülkeye en az bir gün süre kalmak için seyahat etmesi sađlık turizmi olarak tanımlanmaktadır (Kantar ve Işık, 2014: 15). Ülkelerin sađlık sistemlerinde ortaya çıkan uzun bekleme süresi, yüksek maliyet, yetersiz hizmet kalitesi, sigortaların tedavi masraflarını karşılamamaları ve bireylerin bilinçlenmesi sonucu sađlık turizmine yönelim artmakta ve sađlık turizmi hızla büyüyen bir sektör haline gelmektedir (Yılmaz ve Yılmaz, 2022: 65). Hizmet ihracatında her geçen gün ağırlığını hissettiren sađlık turizmi sayesinde sađlık kuruluşları ve seyahat acentaları önemli gelir elde etmektedir.

Gelir yönetimi, yöneylem araştırmasının en başarılı uygulama alanlarından biri olarak kabul edilmektedir (Talluri ve van Ryzin, 2005). Bir işletmenin müşterilerine, kapasite bilgisini doğru zaman ve fiyatta verilebilmesi için teknoloji ve stratejilerden faydalanması gelir yönetimi olarak tanımlanmaktadır (Kimes ve Wirtz, 2003). Gelir yönetimi, belirli bir kapasiteye, deđişken ve belirsiz bir talebe, hassas bir envantere, yüksek maliyetli sabit bir yapıya ve fiyatlara karşı farklı hassasiyetlere sahip müşterilere sahip endüstrilerde geleneksel olarak uygulanmaktadır (Kimes ve Wirtz, 2003). Turizm endüstrisinde gelir yönetimi, konaklanacak odayı, zamanında, uygun kanalla, doğru fiyata erişilebilir kılma olarak tanımlanmaktadır (Mehrotra ve Ruttley, 2006: 10). Hizmet sektöründe maksimum kâr elde edebilmek için sınırlı sayıdaki kaynakların rezervasyon sistemleri aracılığıyla müşterilere erişilebilir olması gerekmektedir. Rezervasyon, müşteri ile hizmet sađlayıcı arasında yapılan bir anlaşmadır. Bu anlaşma ile müşteri, belirlenen tarihler içerisinde işletmeden hizmet alacağını, aynı şekilde işletme belirlenen tarihlerde müşterinin talep ettiđi hizmeti sađlayacağını garanti etmektedir. İşletme, rezervasyon aracılığıyla işletme riskleri minimize etmekte, müşteri ise almak istediđi hizmeti sorun ile karşılaşmadan almaktadır.

Sađlık turizmi paydaşları olan sađlık kuruluşları ve seyahat acentaları, iptal edilen rezervasyonlar için uygun stratejileri uygulamadıđı takdirde gelir kaybına uğramaktadır. İşletmeler, rezervasyon iptallerinden kaynaklanan riski azaltmak için sıkı iptal politikaları geliştirmekte veya kapasitelerinin üzerinde rezervasyon alma (overbooking) işlemleri uygulamaktadır. Rezervasyon iptallerinin önüne geçebilmek için uygulanan katı iptal politikaları rezervasyon sayısında azalmalara yol açmaktadır (Smith vd., 2015). İşletmeler kâr maksimizasyonu yapabilmek için veri madenciliđi (VM) çalışmaları ve veri analizleri yaparak müşteri davranışlarını tahmin etmeye çalışmaktadır. Antonio vd. (2017b), veri bilimi kullanarak rezervasyon iptallerinin olumsuz etkilerinden kurtulmayı sađlayan rezervasyon

iptal tahmin modeli önermiştir. Bu tahmin modeli sayesinde iptal edilecek bir rezervasyonun olasılığı hesaplanarak, karar vericilere fayda sağlama durumu belirtilmiştir. Ayrıca, iptal edilme olasılığı yüksek olan rezervasyonlar tespit edilerek strateji oluşturma, eylem planı hazırlama ve aksiyon alma sağlanmaktadır (Antonio vd., 2017a). Rezervasyon iptal tahmini önemli bir çalışma konusu olmasına rağmen, yeterli sayıda çalışma yapılmamıştır.

Günümüzün rekabetçi, müşteri merkezli ve hizmet odaklı ortamında, işletmeleri besleyen en değerli ham madde, verilerden elde edilen bilgilerdir. Geçen her saniye, yapılan her türlü işlem neticesinde veri üretilmesine rağmen, veri bilimi kullanılmadığından dolayı verilerden faydalı bilgiler elde edilememektedir. Veri içinde gizli kalmış, kullanıcılar tarafından standart işlemlerle ortaya çıkarılamayan ve atıl durumda kalan yararlı bilgiler, veri madenciliği teknikleri kullanarak ortaya çıkarılmaktadır. Günümüzde tüm sektörlerde yoğun bir şekilde kullanım alanına sahip olan veri madenciliği, sağlık ve turizm sektörlerinde de sıklıkla kullanılmakta, strateji oluşturma ve karar alma süreçlerinde karar vericilere destek olmaktadır. Bu doktora araştırmasında, yurt dışında ikamet eden, sağlık turizmi amacıyla Türkiye'ye gelmeyi planlayan sağlık turistlerinin rezervasyon verileri üzerinde veri madenciliği çalışmaları yapılmış ve sağlık turizmi rezervasyon iptal tahminlerini başarı ile tahmin edebilen modeller geliştirilmiştir.

Birinci bölümde; doktora araştırma alanı olan sağlık turizm kavramına yönelik gerekli tanımlar ve açıklamalar yapıldıktan sonra sağlık turizminin önemi vurgulanmıştır. Rezervasyon iptallerine yönelik yapılmış olan önceki çalışmaları etkin bir şekilde inceleyebilmek için manuel, hibrit ve otomatik işlemleri içeren literatür tarama metodolojisi geliştirilmiştir. Literatür tarama metodolojisi kapsamında, Web of Sciences (WoS) ve Scopus veri tabanlarında yer alan çalışmaları tarama, veri seti birleştirme, aynı çalışmaların kaldırılması, anlamsız kelimelerin temizlenmesi, meta veri analizleri, sonuç daraltma işlemleri, alan dışı çalışmaların çıkartılması, doküman kümeleme, konu modelleme, sonuçların yorumlanması ve tam metin okuma sonucunda elde edilen çıktılar verilmiştir.

İkinci bölümde; veri madenciliği kavramının ve araştırmanın daha iyi anlaşılabilmesi için veri, enformasyon, bilgi, veritabanı, karar destek sistemi vb. terimlerin açıklaması yapılmıştır. Veri madenciliği tanımı, veri madenciliğinin kullanıldığı alanlar ve veri madenciliğinde kullanılan yazılımlar verildikten sonra CRISP-DM veri madenciliği sürecinin adımları olan işi anlama (problemin tanımlanması), veriyi anlama, verinin hazırlanması, modelleme, değerlendirme ve uygulama detaylı bir şekilde açıklanmıştır.

Üçüncü bölümde; veri madenciliğinde kullanılan sınıflandırma yöntemlerinden, ZeroR, Naive Bayes (NB), K-En Yakın Komşu (KNN), Lojistik Regresyon (LR), Destek

Vektör Makineleri (DVM), Karar Ağaçları (DT), Yapay Sinir Ağları (YSA), Rastgele Orman (RO), AdaBoost, XGBoost, LightGBM ve CatBoost yöntemlerine ait teorik bilgiler verilmiştir. Bu bölüm içerisinde sınıflandırma yöntemlerinin doğruluk, öğrenme hızı, sınıflandırma hızı, eksik değerlere direnç, ilgisiz değişkenlere direnç, fazla sayıda değişkenlere direnç, bağımlılığı yüksek değişkenlere direnç, kesikli ve sürekli değişkenlere yaklaşım, gürültüye direnç, aşırı öğrenme ile başa çıkma, artımlı öğrenme denemeleri, açıklama yeteneği ve model parametre işleme kriterlerine göre karşılaştırılmaları verilmiştir. Ayrıca, model başarımlarını karşılaştırmak amacıyla doğruluk (accuracy), kesinlik (precision), hassasiyet (sensitivity), F-ölçütü, özgüllük (specificity), AUC (Area Under Curve), FNR (False Negative Rate), FPR (False Positive Rate), NPV (Negative Predictive Value) ve hız parametrelerinin açıklamaları ve hesaplamaları üçüncü bölüm altında verilmiştir.

Dördüncü bölümde; doktora araştırma konusu olarak sağlık turizmi rezervasyon iptal tahmini probleminin seçilme amacı, araştırma soruları, sağlık turizminde rezervasyon iptalleri tahmini yapabilen sınıflandırma modellerinin geliştirilmesinin önemi, araştırmanın varsayım ve kısıtları, araştırma evren ve örnekleme, araştırmanın literatüre ve diğer alanlara katkıları açıklanmıştır. Üçüncü bölümde teorik bilgileri verilen CRISP-DM veri madenciliği sürecinin uygulama adımları, sağlık turizmi rezervasyon iptal tahmini araştırma konusu için baştan sona kadar tüm adımları ile işletilmiş ve sınıflandırıcı tahmin model sonuçları karşılaştırmalı olarak hazırlanmıştır.

Tartışma ve sonuç bölümünde, doktora araştırma soruları ve çıktıları, literatürde yer alan çalışmalarla tartışılmış ve alandaki araştırmacılara yönelik önerilere yer verilmiştir.

BİRİNCİ BÖLÜM

SAĞLIK TURİZMİ REZERVASYON İPTAL TAHMİNİ

Geçmişten günümüze sağlık kavramına yönelik çok farklı tanımlar yapıldığı ve değişik anlamlar yüklendiği görülmektedir. Bu farklılıkların temelinde bireylerin kendilerini sağlıklı görme kriterlerinin değişmesi ve tıbbi paradigmalarda meydana gelen değişimler etkili olmaktadır. Farklı anlamlara sahip sağlık kavramının geniş bir tanımı yapıldığında, objektif ve sübjektif boyutlarının olduğu görülmektedir (Kaplan, 2016: 12). Kişinin, kendi sağlığına dair fiziksel, zihinsel ve sosyal hissiyatı sübjektif boyutunu oluştururken, tıbbi teşhis ve testler aracılığıyla elde edilen sağlık durumu ise objektif boyutunu oluşturmaktadır.

Sağlık kavramı, Dünya Sağlık Örgütü (DSÖ) tarafından “sadece hastalığın veya rahatsızlığın bulunmamasının dışında, fiziksel, zihinsel, sosyal ve psikolojik olarak tam bir iyilik halde bulunma” şeklinde tanımlanmaktadır. DSÖ, sağlık sektörünün amacını, nüfusa en iyi hizmeti sunmak ve en yüksek sağlık düzeyini korumak olarak tanımlamaktadır. DSÖ tarafından yapılan tanımın dışında, sağlık kavramının tanımı ve içeriği ile ilgili olarak son yıllarda pek çok çalışma ortaya konulmuştur. Sağlık, bir organizmanın hücre bütünlüğünü koruma ve daha iyi organize olmuş bir düzende başarılı bir şekilde yaşama girişimi olarak tanımlanmaktadır (Ulutürk, 2015: 49).

İnsanlar belli bir süreliğine bile olsa, değişik tatlar, kültürler, insanlar ve ortamlar keşfetmeyi ve böylelikle bedenlen ve ruhen dinlenmeyi tercih etmektedir. (Öztürk ve Bayat, 2011: 136). Bu isteklerin sonucu olarak, hızlı bir şekilde gelişen ve büyük bir sektörün ortaya çıkmasına neden olan turistik hareketlilik faaliyetleri turizm olarak isimlendirilmektedir. Geçici olarak ikamet ettiği yerden, iş amacı olmadan seyahat amacıyla ayrılan, konaklayan ve harcama yapan kişiler turist olarak adlandırılmaktadır. Zaman, mekân ve bireylerin tercihlerine bağlı olarak sürekli şekillenen, dinamik bir yapıya sahip olan turist hareketliliği ile turizm faaliyetleri teşvik edilmekte ve hizmet çeşitlendirmesi ile turistik aktiviteler ve yatırımlar artmaktadır (Edwards vd., 2008: 1033).

1.1. Sağlık Turizmi

Sağlık turizmi, hastanın sağlığına kavuşması veya sağlıklı kişilerin yaşam kalitelerinin artırılması için tedavi amacıyla başka ülkelere seyahat etme eylemidir (Üzümcü ve Çimen, 2019). Sağlık turizmi, insanların ikamet yerlerinden uzakta, kötüye giden sağlıklarını iyileştirmek için bir günden az ve bir yıldan fazla olmamak koşuluyla, başka bir ülkeye seyahat etmeleridir (Akdeve vd., 2013). Sağlık Bakanlığı’nın tanımına göre sağlık turizmi;

sağlık ve turizm imkanlarından yararlanarak sağlık durumunu korumak ve iyileştirmek için başka bir ülkeye gitmek olarak tanımlanmıştır. Bu bağlamda seyahat eden kişiler, Sağlık Bakanlığı tarafından “Sağlık Turisti” olarak tanımlanmıştır. Sağlık turizmi; organ nakli, diş tedavileri, ameliyat ve operasyonlar, göz tedavileri, teşhis hizmetleri gibi medikal tedavileri; estetik operasyonlar, sağlıklı yaşam, wellness vb. medikal olmayan tedaviler için de ikamet edilen ülkeden farklı bir ülke veya bölgeye yapılan seyahatleri ifade etmektedir. Sağlık turistlerinin tatil imkanlarından yararlanma durumları olmasına rağmen, temel amaç sağlık hizmetlerinden yararlanmaktır (Bookman ve Bookman, 2007).

Son yıllarda dünya genelinde büyük bir gelişme gösteren, yeni turistik ürün ve hizmetlerden bir tanesi de sağlık turizmidir. Sağlık turizmi, farklı hastalıklara ya da sağlık boyutuna sahip kozmetik, infertilite, organ ve doku nakli, ortopedi, obezite, estetik cerrahi vb. çeşitli taleplere sahip kişilerin, fiziksel ve zihinsel tatminine yardımcı olmak amacıyla organize edilen turistik faaliyetler bütünüdür. Sağlık turizmi, bir hastanın sağlığını tedavi etme veya en azından sağlığının düzene girmesi maksadıyla, 24 saat ile 1 yıl süreler arasında olmak koşuluyla, çalışma amacı olmaksızın başka bir bölgeye seyahat etmek olarak da tanımlanmaktadır (Aslanova, 2013: 130). Doktorunun tavsiyesi veya hastanın kendi isteği ile tıbbi tedavi için geçici süreliğine başka bir ülkeye seyahat etmesi sağlık turizmi olarak tanımlanmaktadır (Öztürk ve Bayat, 2011: 137). Son yıllarda iletişim ve ulaşım imkanlarının artması, teknolojinin gelişmesi ve ucuzlaması, tıbbi tesislerin yaygınlaşması ve tıbbi hizmetlerin ticarileşmesi gibi faktörler neticesinde sağlık turizmi, turizm sektörünün önemli bileşeni haline gelmiştir. İlk zamanlarında sadece kaplıca seyahati veya kaplıca turizmi ile sınırlı kalan sağlık turizmi, günümüzde her çeşit tıbbi tedavileri içermektedir (Özer ve Songur, 2013: 71).

Sağlık turizmi kapsamında tedavi amacıyla gelen sağlık turistleri, gitmiş oldukları ülkelerde yer alan müzeleri, tarihi eserleri, doğal kaynakları veya bölgenin kültürel özelliklerini görme ve faaliyetlere katılma imkânı bulmaktadır. Yüksek bir potansiyele sahip sağlık turizminde, özgün çeşitlendirme politikalarının hayata geçirilmesiyle, turizm hareketliliğinden çok daha fazla paya sahip olması sağlanabilmektedir (Özkurt, 2007: 152). Sağlık turizmi yapılan bölgelerde, doğru sağlık turizmi politikalarının uygulanması, sağlık turistlerine uygun turizm etkinliklerinin sunulması, turizm hedeflerine daha hızlı ulaşılmasını ve bölgenin uzun vadede cazibe merkezi olmasını sağlamaktadır. Filipinler'deki Cebu Doktorlar Üniversite Hastanesinin, hastalarına bypass cerrahisi paketinin bir parçası olarak güzel bir tatil beldesinde dinlenme fırsatı sunması örnek olarak verilebilir. Wockhard Hospital USA, hastaları çekmek için Tac Mahal turları düzenlemekte, Ekvador ise hastalara uygun

fiyata diř tedavileri yanında ülkenin aktif volkan ve yağmur ormanlarını görme imkânı sunmaktadır. Bu nedenle, saėlık turizminde hizmet kalitesi, ürün çeřitliliėi ve uygun fiyat politikaları vurgulanmaktadır (Canapero ve Bertinato, 2024). Saėlık turizminde sunulan hizmet ve ürünleri çeřitlendirmek, sürekliliėi saėlayıcı bir rekabet unsuru olarak kabul edilmektedir.

Döviz kazandırıcı yapısından dolayı önemi sürekli artan saėlık turizmi, günümüzde ülkeler açısından çok farklı bir boyuta gelmiřtir. Önemli bir alternatif turizm seçeneėi olarak yaygınlařan saėlık turizminin, saėlık turistlerinin fiziksel ve psikolojik ihtiyaçlarının doğrudan veya dolaylı olarak karřılandığı en aktif etkinliklerden biri olduėu kabul edilmektedir. Saėlık turistlerinin, tedavi ve tıbbi operasyonlarını, modern yöntemlerle ve daha uygun fiyatlara yaptırmak için Türkiye’yi tercih ettikleri gözlemlenmektedir (Aktepe, 2021: 171). Türkiye’nin sahip olduėu doğal güzellikleri, misafirperverliėi, coėrafi konumu ve saėlık hizmetleri alanında sunduėu çeřitlilik nedeniyle, saėlık turizmi alanında önde gelen ülkelerden biri olma potansiyeline sahip olduėu görölmektedir.

Ülkelerin, saėlık turizminde, küresel ölçekte lider bir konuma gelebilmesi, kaliteli hizmet ve tedaviyi uygun fiyata sunmasına baėlıdır. Saėlık turizmi kapsamında sunulan saėlık hizmetlerinin geliştirilmesi amacıyla uluslararası kalite ve akreditasyon sertifikası (Joint Commission International, JCI) bulunmaktadır. 10.03.2024 tarihi itibari ile Birleřik Arap Emirlikleri (218), Suudi Arabistan (111), Brezilya (73), Tayland (64), Çin (41) olmak üzere, dünya genelinde 1018 JCI sertifikasyona sahip saėlık kuruluřu bulunmaktadır. Türkiye, 40 JCI sertifikasına sahip saėlık kuruluřu ile 6. sırada yer almaktadır (Joint Commission International, 2024). Saėlık turizminde, kaliteli ve uygun fiyata sunulan saėlık hizmetleri, hekim kadrosu, teknolojik ekipmanlar ve saėlık kuruluřlarının reklam ve tanıtımına yönelik faaliyetler çok önemlidir. Küresel çapta rekabet içinde olan saėlık kuruluřları, iřletme hedeflerine ulařmak için ürün ve hizmetlerini tanıtma faaliyetlerine ara vermeden devam etmektedir. Saėlık turizmi alanında hizmet veren kuruluřların, diėer hizmet alanlarında yer alan kuruluřlarla aynı tanıtım faaliyetlerine sahip olmaları gerekmektedir (řengöl ve Bulut, 2019: 46). Bu kapsamda Türkiye’de yer alan saėlık turizmi paydařlarının, yurt dıřına yönelik yabancı dil ile yapmıř oldukları reklam ve tanıtım faaliyetleri desteklenmektedir.

Teřvikler, devletin karřılıklı veya karřılıksız, doğrudan veya dolaylı olmak üzere iřletmelere saėladığı desteklerdir. Teřvik verilerek desteklenen faaliyetler, teřvik saėlanmayan diėer faaliyetlere göre daha hızlı gelişmektedir (Topal, 2006: 16). Devlet sübvansiyonları, ekonomide meydana gelen daralmaları ortadan kaldırmak ve ekonomiyi canlandırmak amacıyla řirketlere sunulan desteklerdir (Martin ve Scott, 2000: 438). Devletler,

küresel ölçekte rekabet avantajı oluşturabilmek amacıyla, sağlık turizmi hizmeti sunan paydaşlarını desteklemektedir. Singapur, Hindistan, Tayland, Malezya gibi birçok ülke, sağlık turizmi sektörünün gelişimini desteklemek için mali destek programları uygulamaktadır (Şahin, 2021: 985).

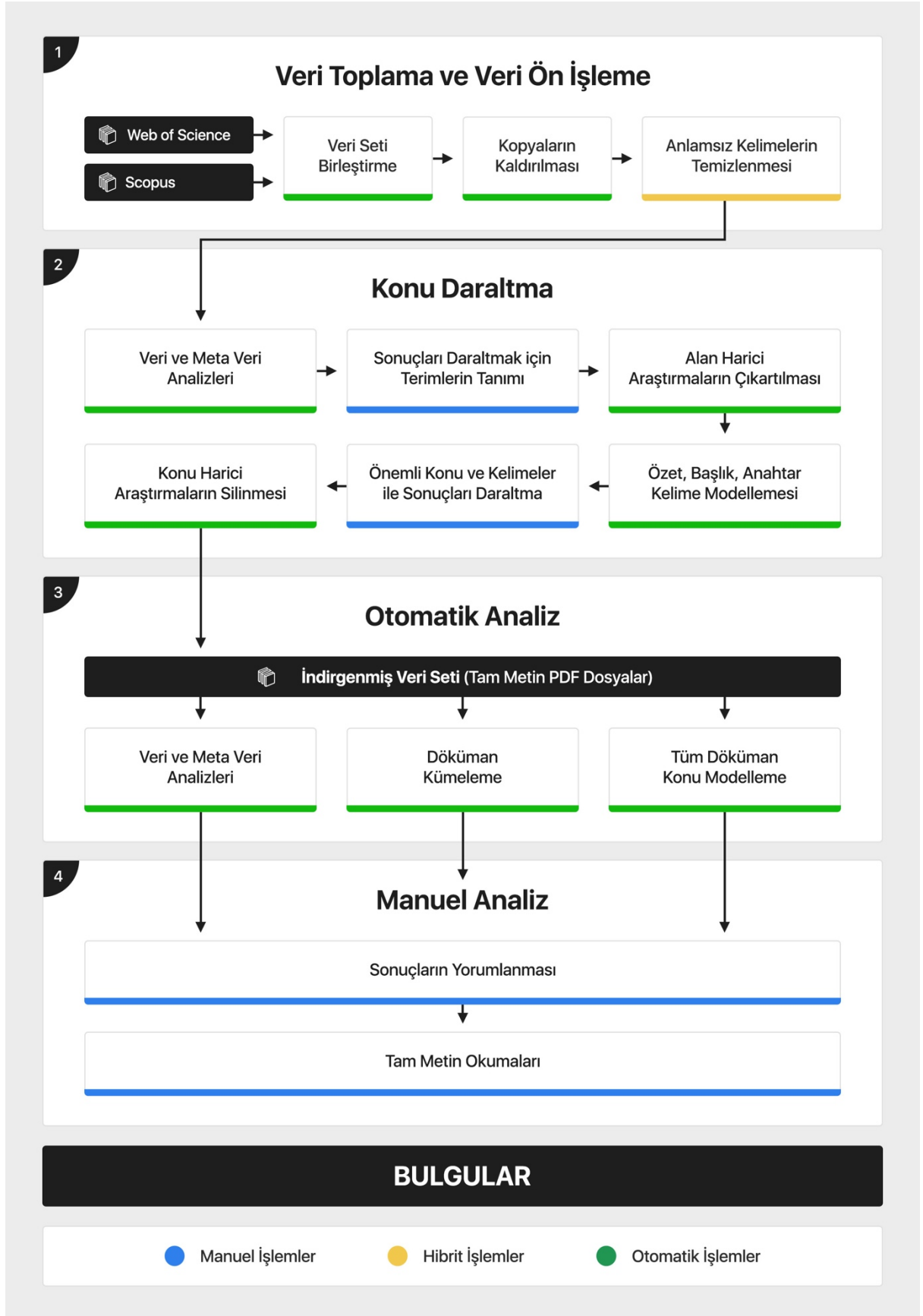
Türkiye, sağlık turizmi alanında uluslararası pazarda lider konuma gelmek, dış pazarlara açılarak döviz kazandırıcı faaliyetleri arttırmak ve sağlık turizminde dünya çapında bir marka olmak amacıyla, sağlık turizmi paydaşlarının sağlık turizmi faaliyetleri ile ilgili giderlerinin bir kısmını Ticaret Bakanlığı tarafından sübvans edilmesi uygulamasına gitmiştir (Tengilimoğlu, 2021: 9).

Sağlık Hizmetleri Genel Müdürlüğü'ne bağlı Sağlık Turizmi Daire Başkanlığı web sitesinde “Yetkili Sağlık Tesisleri ve Aracı Kuruluşlar” listelenmektedir. 17 Şubat 2021 tarihinde ilan edilen, “Uluslararası sağlık turizmi yetki belgesi alan sağlık tesisleri ve aracı kuruluşlar listesi” içerisinde 1097 sağlık kuruluşu ve 170 seyahat acentası yer almaktaydı. 26 Şubat 2024 tarihinde ilan edilen güncel listeye göre 1059 yetkilendirilmiş seyahat acentası ve 648 hastane, 179 tıp merkezi, 2096 muayenehane ve 1124 diğer sağlık tesisi olmak üzere toplamda 4047 yetkilendirilmiş sağlık kuruluşu yer almaktadır. Sağlık turizmi yetki belgeli paydaş sayısındaki artış, sağlık turizmine ilgiyi gösteren en önemli göstergeler arasındadır.

Sağlık turistlerinin sağlık turizmine yönelik ilgilerinin en üst seviyede olması ve devlet teşvikleri sayesinde işletmelerin sağlık turizmine yönelmesi, küresel anlamda rekabetin artmasına yol açmaktadır. Sağlık turistlerinin tedavi olabilecekleri alternatif ülke ve sağlık kuruluşu sayısındaki artış, karar alma ve rezervasyon sürecini zorlaştırmaktadır. Sağlık turistleri, rezervasyonlarını tamamladıktan sonra bile daha iyi fiyat ve kalite için sağlık kuruluşu arayışına devam etmekte ve rezervasyonlarını iptal edebilmektedir. Sağlık turistlerinin rezervasyon iptallerinin başarılı bir şekilde tahmin edilmesi ve gerekli aksiyonların alınması, sağlık turizmi paydaşları açısından kritik öneme sahiptir.

1.2. Sistematik Literatür Taraması

“Veri Madenciliği Teknikleri ile Sağlık Turizminde Rezervasyon İptallerinin Tahmini” doktora araştırması, sağlık turizmi ve veri madenciliği olmak üzere iki alanı kapsamaktadır. Doktora araştırmasının Ekonometri ana bilim dalında yapılmasından dolayı, literatür tarama çalışmaları, rezervasyon iptallerini tahmin etme üzerine yapılmıştır. Doktora araştırması kapsamında yapılan, “Veri Toplama ve Veri Ön İşleme”, “Konu Daraltma”, “Otomatik Analiz” ve “Manuel Analiz (elde edilen çalışmaların tam metin okumaları) adımlarından oluşan sistematik literatür tarama metodolojisi Şekil 1.1’de verilmiştir.



Şekil 1.1 Literatür Tarama Metodolojisi

Rezervasyon iptal tahmini yapmak için iki farklı problem türü bulunmaktadır. Eğer çalışmanın amacı; (1) yalnızca bir rezervasyonun iptal edilme oranını tahmin etmek ise, o zaman problem bir tahmin problemi olarak kabul edilmekte, (2) bir rezervasyonun iptal edilme durumunun “Rezervasyon İptal” veya “Rezervasyon İptal Değil” şeklinde belirlenmesi ise, o zaman bir sınıflandırma problemi olarak değerlendirilmektedir. Bu tez çalışmasında ele alınan sağlık turizmi rezervasyon iptal tahmini, bir sınıflandırma problemi olarak ele alınmaktadır.

Literatür tarama metodolojisinde, tamamen kullanıcının müdahale etmesi gereken manuel işlemler, otomatik işlemler ve hibrit işlemler olmak üzere üç tür işlem bulunmaktadır. Yapılan işlemin hangi tür işlem olduğu Şekil 1.1’in en altında yer alan renkler ile belirtilmiştir. Literatür tarama işlemleri, veri analizini kolaylaştırmak için tasarlanmış, çok sayıda pakete sahip, güçlü bir istatistiksel araç olan R programlama dili kullanılarak gerçekleştirilmiştir. Literatür taramasında kullanılan yol haritası, veri ve meta veri analizi, tam metnin veri madenciliği paketleri ile ele alınması ve görselleştirilmesi, en sık kullanılan unigramlar ve bigramlar içeren kelime bulutları ve konu modelleme çalışmaları, literatür taramalarında yapılan çalışmalar, António (2019) doktora araştırmasından örnek alınarak hazırlanmıştır.

Bir konuda araştırma yapmanın gerekliliği, ilgili literatür taramalarının detaylı analizi ile tespit edilmektedir. Akademik çalışma sayısının her geçen gün artması, literatür taramasını zorlaştırmakta ve bu durum bazı araştırmacıların bilişim teknolojilerinden faydalanarak literatür tarama işlemlerini otomatikleştirmesine yol açmaktadır. Otomatik literatür taramaları, daha az enerji ve zaman harcayarak, daha hızlı ve etkin literatür taramaları yapmaya imkân sağlamaktadır.

Otomatik literatür tarama ile araştırmacılar, arama stratejisi tasarlama, araştırma alanı ile ilgili arama yapma, farklı veri kaynaklarında yer alan kopya çalışmaları ayıklama, araştırma özetlerini tarama, tam metin makaleler üzerinde modelleme yapma, veri dönüştürme, veri çıkarma ve sentezleme imkanlarına sahip olmaktadır.

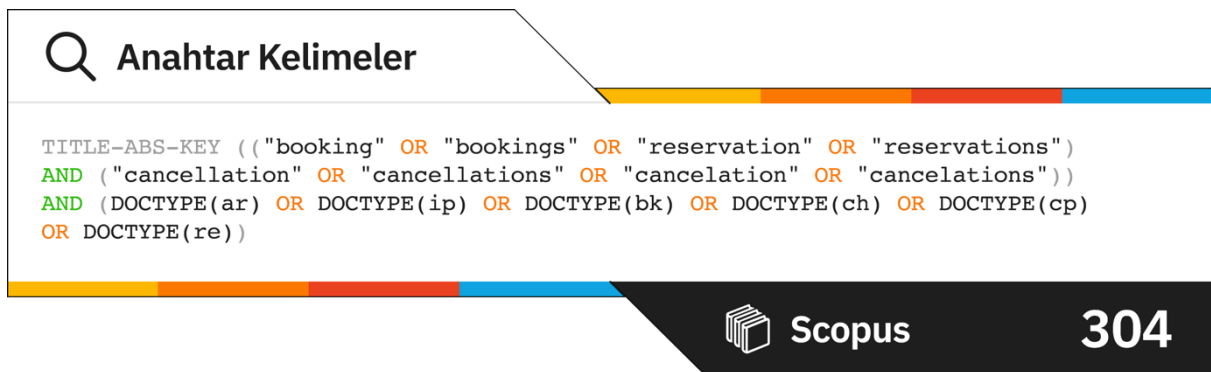
Literatür taramalarını, gelişmiş metin madenciliği yöntemleri ve bilişim teknolojileri kullanarak otomatikleştirmek veya en azından yarı otomatik bir hale getirmek, araştırmacılara büyük kolaylıklar sunmaktadır. Sistemik literatür taraması; veri erişimini kolaylaştırma, meta verilere dayalı olarak dergi, yazar, kuruluş bazlı kategorize etme, yeni kavram veya ilişkileri ortaya çıkarma, araştırma ana konu ve alt konularını belirleme, konular ve üst veriler arasındaki ilişkileri belirleme konularında katkı sağlamaktadır.

1.2.1. Veri Toplama ve Veri Ön İşleme

Doktora araştırması literatür taraması kapsamında, Scopus ve Web of Science (WoS) olmak üzere iyi bilinen iki veri tabanına yönelik kaynak taramaları yapılmıştır. Seçilen iki veri tabanı, turizm ve seyahat endüstrileri ile ilgili araştırmaların çoğunu içermektedir. Etkin bir otomatik literatür taraması yapabilmek için veri toplama ve ön işleme aşamalarının uygulanması gerekmektedir.

Veri tabanlarından araştırma elde etme işleminde, arama ifadesini doğru yapılandırabilmek için gerekli anahtar kelime seçimleri yapılmalıdır. Makale arama motorları arasındaki farklılıklardan kaynaklanabilecek sorunlar göz önünde bulundurularak, her iki veri tabanında basit bir sorgu yürütülmüş ve aramayı araştırma hedefine daraltmak için sonuçlar filtrelenmiştir. Arama dizilerinde, belgelerin başlığında veya anahtar kelimelerinde “reservation” ve “cancellation” kelimelerinin ve varyasyonlarının aynı anda kullanıldığı araştırmalar taranmıştır. Seçilen anahtar kelimeler Scopus veri tabanının özet bölümünde aranabilmesine rağmen, bu işlevselliğe sahip olmadığı için WoS veri tabanının özet kısmında aranmamıştır.

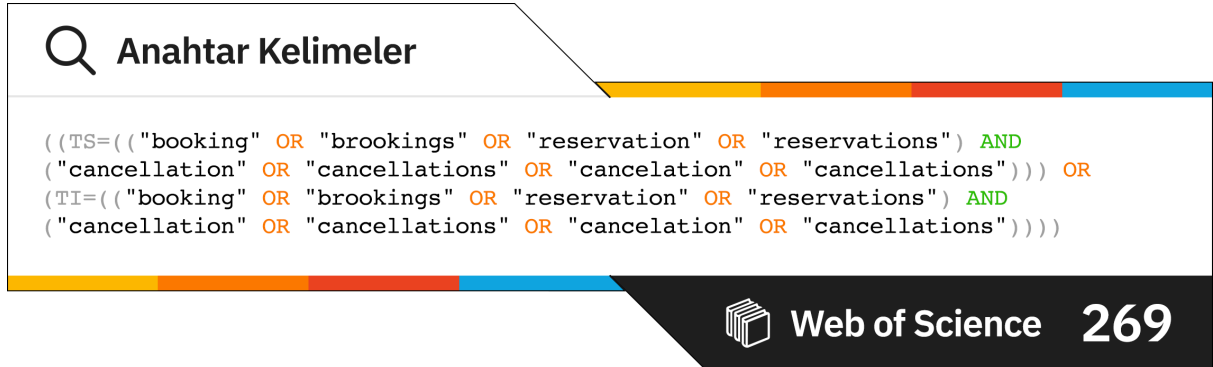
Veri madenciliği otomatik literatür taramasını etkin yapabilmek amacıyla arama, *Ekim 2022* tarihine kadar *İngilizce* hazırlanmış çalışmalarla sınırlandırılmıştır. Scopus ve WoS veri tabanlarının kendilerine özgü belge sınıflandırma kategorileri olduğundan, her bir veri tabanında seçilen belge türleri farklı olmaktadır. Scopus için belge türleri: makale, basında çıkan makale, kitap, kitap bölümü, konferans bildirisi ve incelemedir. Scopus veritabanında arama yapmak için kullanılan arama terimi Şekil 1.2’de gösterilmiştir.



Şekil 1.2 Scopus Arama Dizisi ve Çalışma Sayısı

Scopus veri tabanı arama sonucu, Scopus dışı aktarma işlevi (export) kullanılarak virgülle ayrılmış değerlerin saklandığı CSV formatındaki bir dosyaya aktarılmıştır. Scopus arama sonucunda 211 makale, 83 konferans bildirisi, 9 kitap bölümü ve 1 adet inceleme araştırması olmak üzere toplam 304 araştırma bulunmuştur.

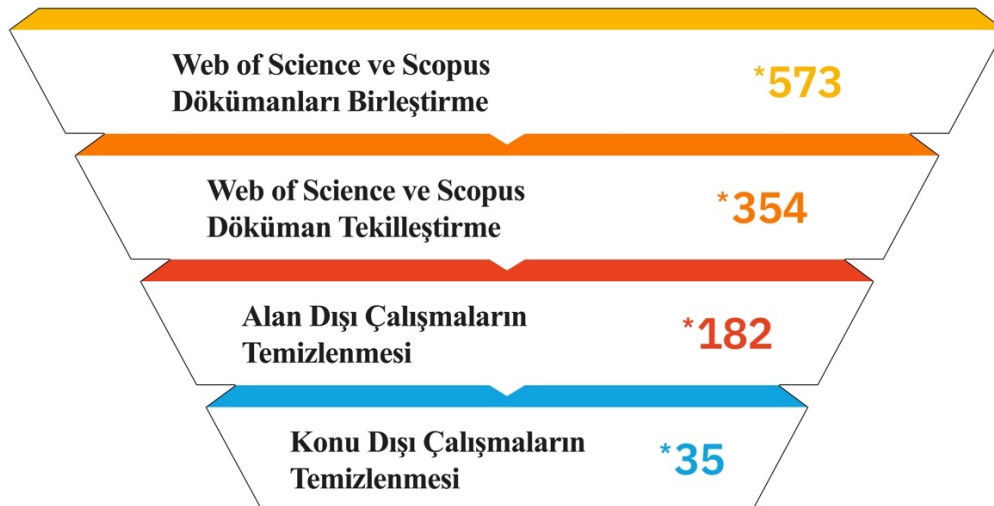
WoS için belge türleri; makale, kitap, kitap bölümü, bildiri ve inceleme türündedir. Web of Science veri tabanında arama yapmak için kullanılan arama terimi Şekil 1.3'te gösterilmiştir.



Şekil 1.3 Web of Science Arama Dizisi ve Çalışma Sayısı

WoS arama sonucu, sekmeyle ayrılmış değerlerden oluşan TSV formatındaki bir dosyaya aktarılmıştır. Arama sonucunda 187 makale, 64 konferans bildirisi, 9 erken erişim, 2 kitap bölümü ve 1 adet inceleme araştırması türüne ait toplam 269 araştırma bulunmuştur. Arama sonucunun geçerliliğini değerlendirmek ve arama terimlerinin dahil edilip edilmediğini belirlemek için rastgele seçilen sayıda araştırma kontrol edilmiştir.

Veri Seti Birleştirme: Farklı veri tabanı kaynaklarından gelen birden çok araştırmanın birleştirilmesi, belge yapıları, veri biçimleri ve veri kalitesindeki farklılıklar nedeniyle her zaman zorluk oluşturmaktadır. Bu sorunlara rağmen Scopus ve WoS veri tabanındaki benzer alanların birleştirilmesi ile tek ve benzersiz bir veri seti oluşturulmuştur. Şekil 1.4'te yer alan Veri Seti İndirgeme Adımları şeklinin en üst katmanında Scopus ve WoS veri tabanlarından elde edilen iki veri setinin birleştirilmesi gösterilmektedir. Scopus veri tabanından elde edilen 304 araştırma ve WoS veri tabanından elde edilen 269 araştırmanın birleştirilmesi ile toplam 573 araştırma elde edilmiştir.



Şekil 1.4 Veri Seti İndirgeme Adımları

Kopyaların Temizlenmesi: Scopus ve WoS veri tabanlarının ikisinde yer alan tek ortak alan tanımlayıcısı DOI (Digital Object Identifier) numarası olmasına rağmen, tüm araştırmaların veri tabanında kayıtlı bir DOI numarası bulunmamaktadır. Araştırmaların başlıkları ön işlemeye tabi tutulmuştur. Ön işleme, analizlerin iyileştirilmesi, hesaplama performansının artırılması için kelimelerin normalleştirilmesi, tüm metinlerin daha küçük ve belirli özelliklere dönüştürülmesi, noktalama işaretlerinin kaldırılması, sayıların kaldırılması, veri seti üzerinde tüm metnin küçük harfe dönüştürülmesi ve kelimelerin tek tip bir formata dönüştürülmesi işlemlerini içeren bir süreçtir. Metin ön işlemleri R programlama dili ve paketleri kullanılarak gerçekleştirilmiştir. Ön işlem uygulanmış araştırmaların DOI numaralarının karşılaştırılmasından sonra 219 mükerrer araştırma kaldırılmış ve Şekil 2.4'ün “Doküman Tekilleştirme” katmanında gösterildiği gibi toplam araştırma sayısı 354 tekil araştırmaya indirgenmiştir. Mükerrer araştırmaların kaldırılması ile elde edilen 354 araştırma; 261 (%74) makale, 83 (%23) konferans bildirisi, 9 (%3) kitap bölümü ve 1 adet incelemeden oluşmaktadır.

Anlamsız Kelimelerin Temizlenmesi: Anlamsız kelimelerin temizlenmesi işlemi manuel ve otomatik işlemlerin birlikte yapılması gereken hibrit bir adımdır. Mükerrer kayıtların ortadan kaldırılması ve tekil çalışmalara indirgenmiş veri seti üzerinde kullanıcı tarafından belirlenen “the”, “a” vb. anlam ifade etmeyen harf ve kelimelerin temizlenmesi işlemleri bu aşamada gerçekleştirilmiştir.

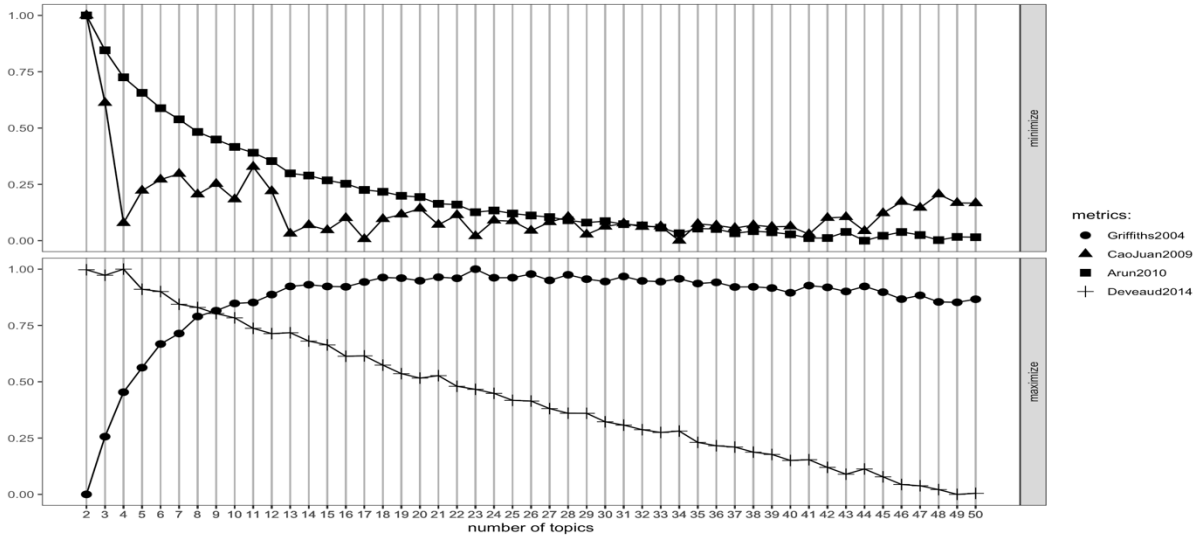
1.2.2. Konu Daraltma

Literatür taramasında incelenen rezervasyon iptal tahmini konusunda yapılan önceki araştırmaların tamamı, doktora araştırmasının amacı ile dolaylı olarak ilgili olmasına rağmen, sağlık turizmi alanı ile benzerlik göstermeyen araştırma sayısı daha fazladır. Doktora araştırma alanı ile ilgili olmayan yöneylem araştırması, konaklama yönetimi, tıp ve sağlık, ulaşım ve lojistik yönetimine ait tüm dokümanların veri setinden çıkartılmasının, literatür tarama verimliliğini arttıracak öngörülmüş ve sağlık turizmi alanı ile ilgili araştırmaların taranmasında kullanılacak terimleri belirlemek için metin madenciliği çalışması yapılmıştır.

Veri ve Meta Veri Analizleri: Veri setinde yer alan 354 araştırmanın başlık, anahtar kelime ve özet bilgilerinde; sayı ve terimlerin kaldırılması ve kök ayıklanması sonucunda veri setinin ön işlemesi tamamlanmış sürümü oluşturulmuştur. Metin madenciliğinde “n” kelime sayısını göstermek üzere, terimler “n-gram” olarak da adlandırılmaktadır. Tek kelimelik bir terim “unigram”, iki kelimelik terimler “bigram” olarak isimlendirilmektedir (Welbers vd.,

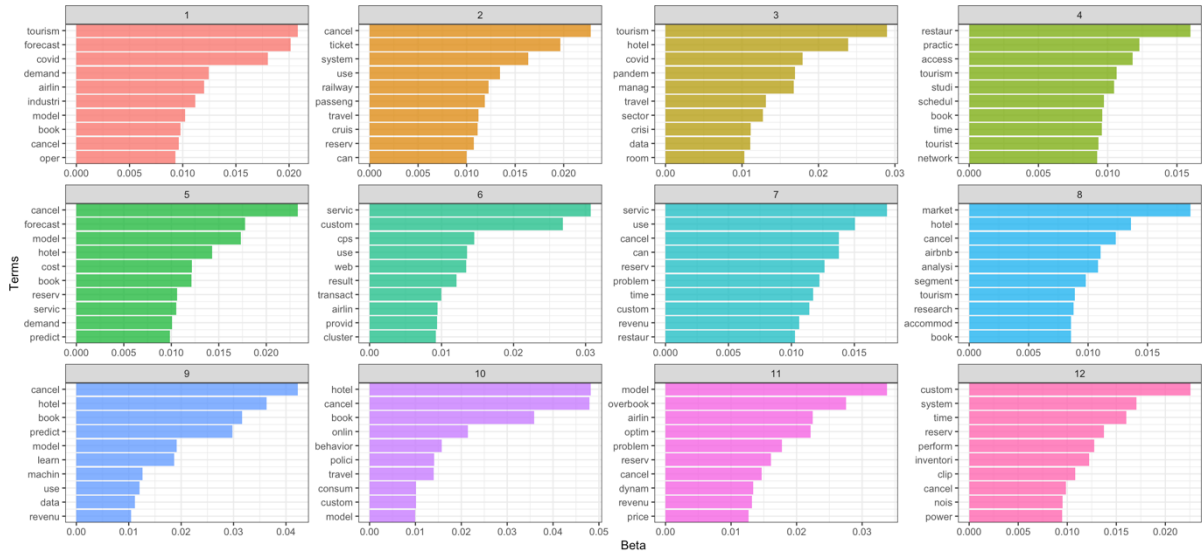
koleksiyonları verisinde yer alan konuları otomatik olarak keşfetmek için kullanılan istatistiksel bir yöntemdir (Blei, 2012: 77). Konu modelleme yöntemi (Latent Dirichlet Allocation, LDA), 2003 yılında geliştirilmiştir (Blei vd., 2003). Gizli Dirichlet Tahsisi veya Gizli Dirichlet Ayrımı olarak da bilinen LDA, her belgenin olasılıksal anlam dizinini ve benzer konuların kombinasyonunu ortaya koymak maksadıyla ortaya atılan olasılık tabanlı yaklaşımdır (Blei ve Lafferty, 2009). LDA, metin belgesinin birden fazla konuyu içerdiğini varsayarak ve belirli bir konu veya belgeye uzaklığını hesaplayarak sınıflandırmayı amaçlayan bir istatistik modelidir. LDA kullandığı yaklaşım farklılığıyla en basit, hızlı boyut küçültme ve sınıflandırma modelidir (Lane vd., 2019).

Terim sıklığı analizi ile alan harici araştırmaların çıkartılmasından sonra geriye kalan 182 araştırmanın hangi konuları kapsadığını anlamak için, R programlama dili “topicmodels” paketi kullanılarak LDA yöntemi uygulanmıştır. LDA yönteminin kullanımı aşamasında en önemli parametre, modellenen konu sayısıdır. Modellenen ideal konu sayısını tespit etme aşamasında “ldatuning” R paketi kullanımı büyük kolaylık sağlamaktadır (Nikita ve Chaney, 2020). Ldatuning paketi, modellenen konu sayısına karar vermeye yardımcı olmak amacıyla dört farklı seçenek (Griffiths2004, CaoJuan2009, Arun2010, Deveaud2014) sunmaktadır. Şekil 1.6’daki grafiğe göre, modelleme konu sayısı 12 olarak belirlenmiştir.



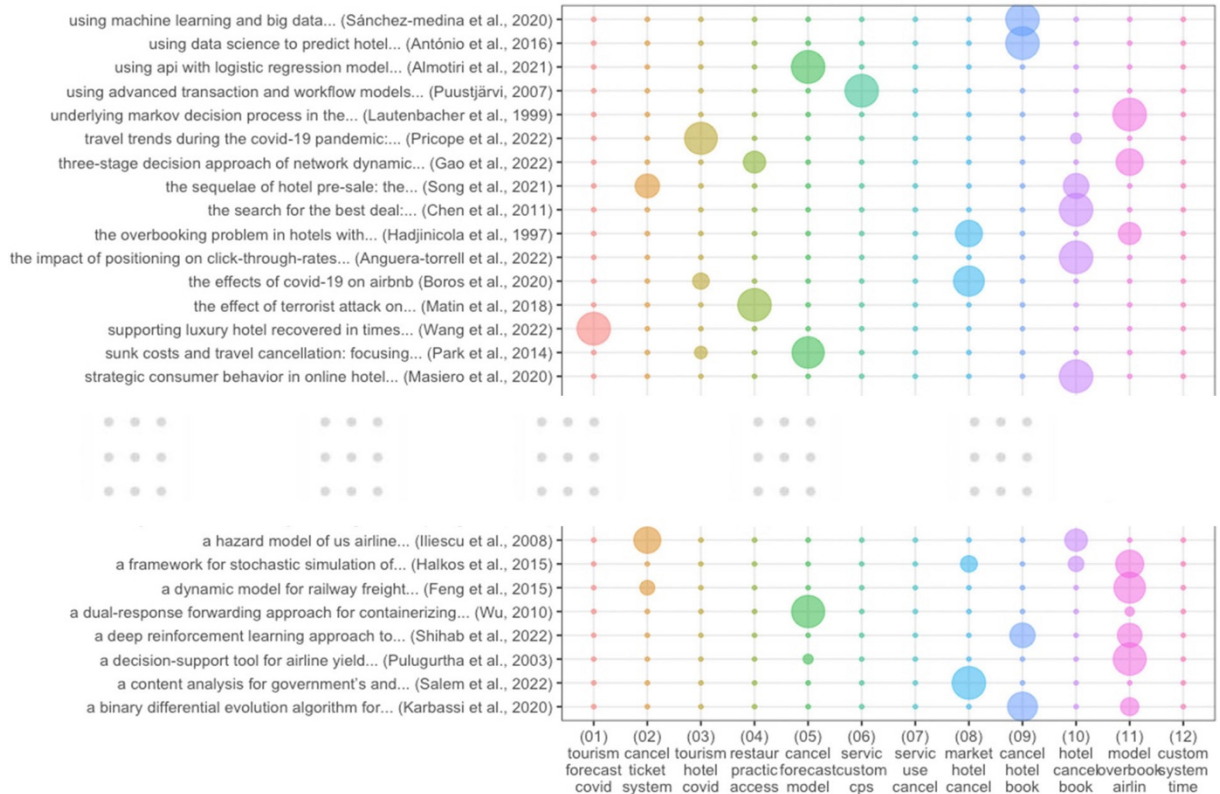
Şekil 1.6 Ldatuning Paketi ile LDA Konu Sayısına Karar Verme

Alan ile ilgili 182 araştırmanın başlık, anahtar sözcükleri, dizin anahtar sözcükleri ve özetinin ön işleme tabi tutulmuş versiyonlarına, modelleme konu sayısı 12 olacak şekilde LDA yöntemi uygulanmıştır. Her bir konunun ilk 10 terimine LDA analizi yapılmış olup, konuların anahtar kelime başına dağılımları Şekil 1.7’de gösterilmiştir.



Şekil 1.7 Tüm Dokümanlara ait Konu-Kelime Dağılımı

Konu 1'in "turizm, tahmin ve covid" ile; konu 2'nin "iptal, bilet, demiryolu ve sistem" ile; konu 3'ün "turizm, otel ve pandemi" ile; konu 4'ün "restoran ve ulaşılabilirlik" ile; konu 6 ve 7'nin "servis, işlem ve gelir" ile; konu 8'in "market, otel ve airbnb" ile; ve konu 12'nin ise "sistem, zaman ve envanter" ile ilgili olduğu ve bu konuların doktora araştırmasıyla ilgili olmadığı görülmektedir. Buna karşılık, konu 5, 9, 10 ve 11'in "iptal, tahmin ve model" konularını daha yoğun içerdiği ve doktora çalışma alanı ile benzer konulara sahip olduğu görülmektedir.



Şekil 1.8 Tüm Araştırmaların Konu Başlıklarına Göre Dağılımları

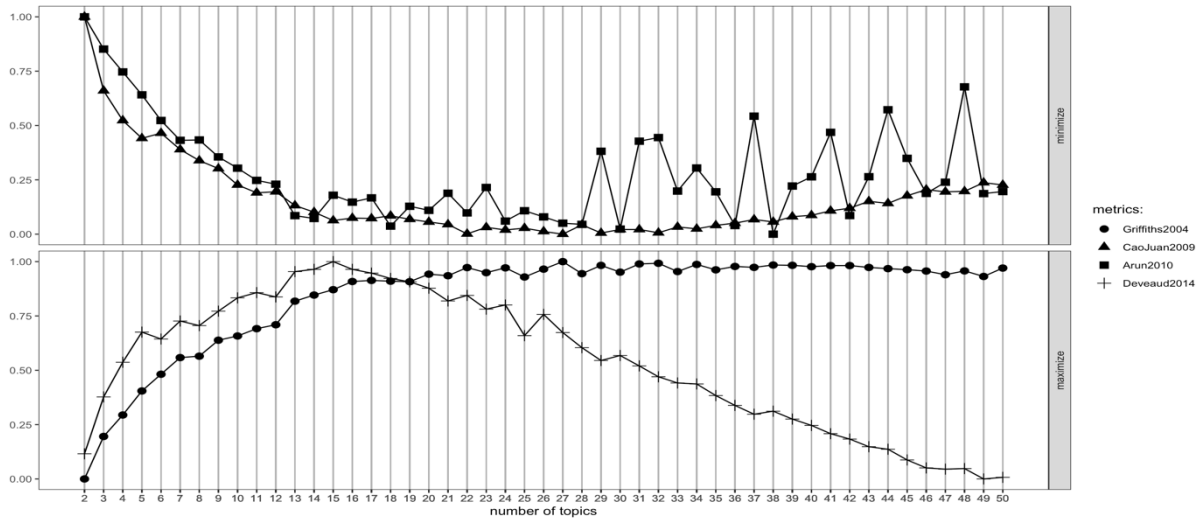
Önemli Konu ve Kelimeler ile Sonuçları Daraltma: Araştırmanın bir konuyu kapsama olasılığını tespit etmek için LDA yönteminin gama dağılımı kullanılmaktadır. LDA yöntemi ile araştırmaların hangi konuları kapsadığını belirlemek, Şekil 1.8’deki gibi mümkündür. Başlıkların ve içeriklerin analizi, bu tez araştırmasının rezervasyon iptal tahmini ve modellemesi konusunda olup olmadığını ortaya çıkarmaktadır. Toplam 182 adet araştırmanın, tam metin okuma öncesinde daha fazla daraltılması gerekmektedir.

Konu Harici Araştırmaların Silinmesi: Araştırmaların başlık, anahtar sözcük veya özetlerinde “prediction” veya “forecast” terimlerinin, kökten türetilmiş bir versiyonuna sahip olmayan araştırmalar üzerinde otomatik terim araması yapılmış ve araştırma sayısı Şekil 1.4’te belirtildiği gibi 35 araştırmaya indirgenmiştir.

1.2.3. Otomatik Analizler

Araştırma konusu ile ilgili 35 araştırmanın tam metin PDF dosyaları, yayıncı web sitelerinden veya bilimsel havuzlardan temin edilmiştir. Her PDF dosyasının tam metninde; büyük/küçük harflerin küçültülmesi, sayıların, noktalama işaretlerinin ve önceden tanımlanmış bilgilendirici olmayan terimlerin kaldırılması gibi ön işleme operasyonları uygulanmıştır.

Veri ve Meta Veri Analizleri: İndirgenmiş Veri Seti üzerinde otomatik analiz öncesinde LDA analizi için modelleme konu sayısı Şekil 1.9’da belirtilen grafik kullanılarak 15 olarak seçilmiştir.



Şekil 1.9 İndirgenmiş Son Veri Seti için LDA Analizi

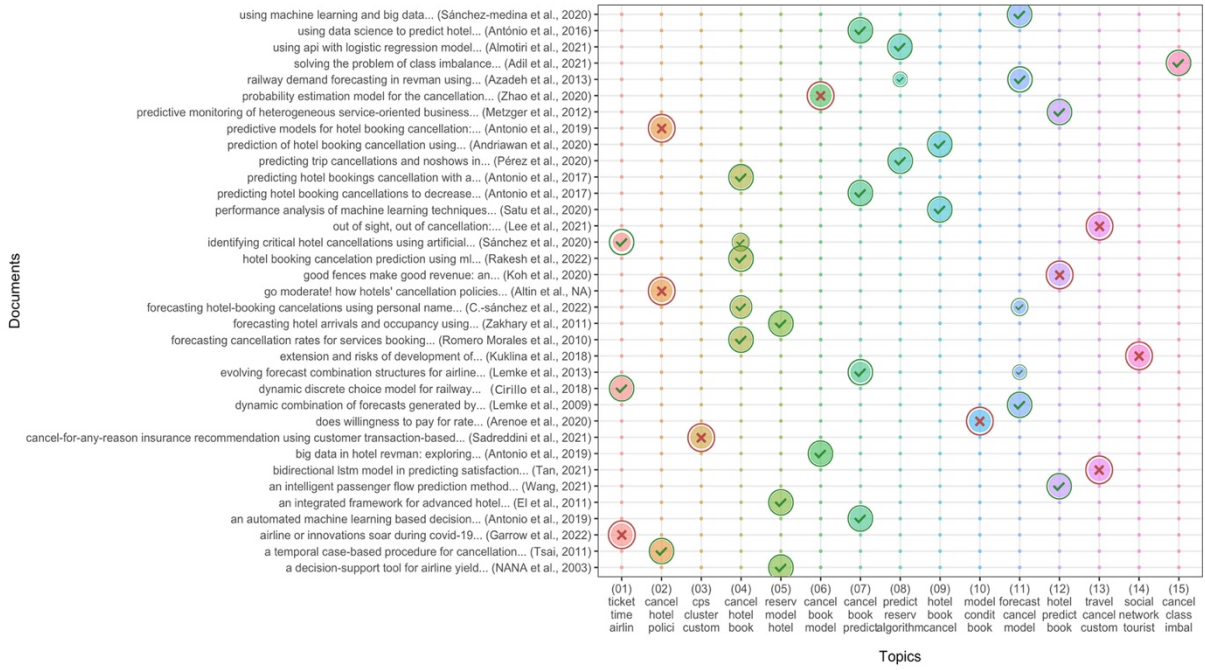
15 konudan her birinin ilk on teriminin Dirichlet analizi yapılmış olup, konuların anahtar kelime başına dağılımları Şekil 1.10’da gösterilmiştir.



Şekil 1.10 İndirgenmiş Son Dokümanlara ait Konu- ve Kelime Dağılımları

Doküman Kümeleme: Konu 1’in “zaman, bilet ve havalimanı”; konu 2’nin “iptal, otel ve politika”; konu 3’ün “kümeleme, algoritma, ücret, gelir ve cps”; konu 4’ün “iptal, otel, rezervasyon, model ve tahmin”; konu 5’in “rezervasyon, model, otel, tahmin ve talep”; konu 6’nın “gemi, slot, doğruluk ve tahmin”; konu 7’nin “iptal, tahmin ve model”; konu 8’in “tahmin, rezervasyon, algoritma ve gezi”; konu 9’un “otel, rezervasyon, iptaller, metot ve özellikler”; konu 10’un “model, rezervasyon ve iptaller”; konu 11’in “tahmin, iptal, model ve yaklaşım”; konu 12’nin “yolcu, endüstri, ücret ve gelir yönetimi”; konu 13’ün “servis, uzay, havacılık ve yolcu”; konu 14’ün “sosyal, ağ, turist ve risk”; ve konu 15’in ise “iptal, sınıflandırma, tahmin ve makine öğrenmesi” ile ilgili olduğu görülmektedir.

Tüm Doküman Modelleme: 35 araştırmanın tam metinlerinin okunması sonucunda bazı çalışmaların doktora araştırma konusu olan rezervasyon iptal tahmini veya tahmin modellerinin geliştirilmesi ile ilgili olmadığı görülmüştür. Doktora araştırma konusu ile ilgili 25 araştırma üzerinde veri ve meta veri analizi, belge kümelemesi, tam metin konu modellemesi ve tam metin okunmasından elde edilen çıkarımlar analiz edilmiş ve yorumlanmıştır.



Şekil 1.11 İndirgenmiş Dokümanların Konu Başlıklarına Göre Dağılımları

Doktora Araştırması ile Doğrudan İlişkisi Olmayan Konular: Konu 3: “kümeleme, algoritma, ücret, gelir ve cps”; konu 10: “model, rezervasyon ve iptaller”; konu 13: “servis, uzay, havacılık ve yolcu”; konu 14: “sosyal, ağ, turist ve risk”.

Doktora Araştırmasını Kısmi İçeren Konular: Konu 1: “zaman, bilet ve havalimanı”; konu 2: “iptal, otel ve politika”; konu 6: “gemi, slot, doğruluk ve tahmin”; konu 12: “yolcu, endüstri, ücret ve gelir yönetimi”.

Doktora Araştırması ile İlgili Konular: Konu 4: “iptal, otel, rezervasyon, model ve tahmin”; konu 5: “rezervasyon, model, otel, tahmin ve talep”; konu 7: “iptal, tahmin ve model”; konu 8: “tahmin, rezervasyon, algoritma ve gezi”; konu 9: “otel, rezervasyon, iptaller, metot ve özellikler”; konu 11: “tahmin, iptal, model ve yaklaşım”; ve konu 15: “iptal, sınıflandırma, tahmin ve makine öğrenmesi” ile ilgili olduğu görülmektedir.

1.2.4. Manuel Analiz

Araştırma konusu ile ilgili ilk belge 2003 yılına ait olup doktora araştırma konusunun popülerliği zaman ilerledikçe sürekli artmaktadır. Rezervasyon iptal tahmini modellemesine özgü ilk araştırma 2009 yılında yayınlanmıştır. 2009 yılından günümüze kadar 2014, 2015 ve 2016 yılları hariç araştırma sayısı istikrarlı bir şekilde artmıştır. Genel talep tahmini modellemesi ve belirli bir rezervasyon iptal tahmini modellemesi konularında 5 havayolu verisi, 4 demiryolu verisi, 2 restoran verisi ve 2 otel verisi olmak üzere toplam 35 araştırma bulunmasına rağmen, literatür taramalarında sağlık turizmi ile ilgili bir araştırmaya rastlanmamıştır. Manuel analiz sonucunda elde edilen bulgular Tablo 1.1’de verilmiştir.

Tablo 1.1 Rezervasyon İptal Çalışmalarının Manuel Analizleri

Yazar ve Yıl	Çalışmanın Konusu
Pulugurtha ve Nambisan, 2003	Gelir yönetimi probleminde her ücret sınıfına tahsis edilen koltuk sayısını tahmin etmek için karar destek aracı geliştirmeyi hedeflemiştir. Araştırma problemi, doğrusal bir tam sayılı programlama modeli olarak formüle edilmiştir. Çözüm tekniği olarak genetik algoritmalar (GA'lar) kullanmış ve geliştirdikleri karar destek aracı, zamana bağlı talebin, bilet iptallerinin ve çifte rezervasyon politikasının etkisini dikkate almaktadır.
Lemke vd., 2009	Rezervasyon iptal tahmini sonuçlarını iyileştirmek amacıyla havayolu endüstrisinden çok boyutlu verilere bakmaktadır. Bireysel tahminlerin üretilmesi için tek üstel düzleştirme, Brown üstel düzleştirme ve bir regresyon yaklaşımının kombinasyonu olan üç farklı yöntem kullanılmıştır.
Morales ve Wang, 2010	Yolcu Adı Kaydı (PNR) tabanlı veriler üzerinde, veri madenciliği yöntemleri ile iptal oranı tahmin modelleri çalışılmıştır. Gerçek dünya veri kümesinin kullanıldığı çalışmada, rezervasyonun farklı aşamalarında iptal davranışını tanımlayan değişkenler kümesinin nasıl çok farklı olduğu gösterilmiştir. Çalışmada, problemin dinamik yönü doğrulanmakla kalmamış, aynı zamanda yöneticilerin iptallere neden olan sebepleri daha iyi anlamasına yardımcı olunmuştur.
Gayar vd., 2011	Mevcut gelir yönetim sistemlerindeki eksiklikleri adresleyen bir gelir yönetim modeli sunulmaktadır. Önerilen tam sayılı programlama modeline dayalı optimizasyon modülü ile otel rezervasyon sürecinde, grup rezervasyon, konaklama süresi, mevsimsellik, eğilim vb. stokastik parametrelerin birbirleri ile olan ilişkisi, gerçek otel verileri üzerinde zamanda ileriye doğru modelleme ve simülasyon ile test edilmiş ve başarılı sonuçlar elde edilmiştir.
Zakhary vd., 2011	Otele geliş ve doluluk tahmini problemi için otel rezervasyon sürecini tarihsel veriler kullanarak zaman içinde ileriye doğru simüle eden yeni bir Monte Carlo simülasyon yaklaşımı önerilmektedir. Bu yaklaşım ile tarihsel veriler kullanılarak rezervasyon süreci, rezervasyon varışları, iptaller, konaklama süresi, grup rezervasyonları, mevsimsellik, trend vb. parametrelerin tahmini için bir yaklaşım önerilmiştir.
Metzger vd., 2012	Nakliye ve lojistik endüstrisinde tahmine dayalı izlemenin potansiyeli gösterilmektedir. Yaygın olarak kullanılan standartlara ve gerçek operasyonel verilere dayanarak, ulaşım rezervasyonlarının geç iptalleri ve gecikmiş teslimatlar gibi, o endüstri sektöründe karşılaşılan temel sorunlara çözüm olarak, bulut ve hizmet tabanlı bir iş birliği ve entegrasyon platformu tasarlanmıştır. Bu platforma dayanarak, nakliye ve lojistik endüstrisinde belirlenen sorunları proaktif olarak yönetmeye ve azaltmaya olanak tanıyan kısa vadeli tahmin yetenekleri geliştirilerek, iş verimliliğinin ve sürdürülebilirliğinin artırılması planlanmaktadır.

Tablo 1.1 (Devam)

Yazar ve Yıl	Çalışmanın Konusu
Azadeh vd., 2013	Büyük bir Avrupa demiryolu operatörünün toplam rezervasyonlarına ve kalkıştan önceki iptallerine ilişkin doğru tahminler üretmek için sinir ağlarının kullanımı analiz edilmektedir. Araştırmada kullanılan tahmin modeli, yolcu sayısı ile bu sayıyı etkileyen faktörler arasındaki ilişkiyi tarihsel verilere dayalı olarak tanımlayan geliştirilmiş çok katmanlı bir ağıdır. Öğrenmeyi daha verimli hale getirmek için ilgili ön işleme yaklaşımları kullanılmıştır. Ağın genelleştirilmesi, gerçek demiryolu verileri kullanılarak gelecekteki rezervasyon ve iptal eğilimleri için regresyon modelinin doğruluk tahminini değerlendirmek için test edilmiştir. Sonuçlar, düşük tahmin hatası ile demiryolu talep tahmininde umut verici bir yaklaşım olduğunu göstermektedir.
Lemke vd., 2013	Bir havayolu şirketi için gelir yönetimi yazılımı projesinin bulgularını sunmakta olup, rezervasyon tahmini doğruluğunu artırmak için bileşenlerinden birini değiştirme üzerine yapılan bir çalışmadır. Mevcut üç geleneksel bireysel algoritma setini tamamlayan ek bir yöntem sunulmuş ve yöntem havuzuna eklenmiştir. Bireysel tahminler için model parametrelerinin çeşitlendirilmesi ve öğrenme düzeyi tartışılmış, tahmin kombinasyon yapılarının evrimi araştırılmış ve bir havayolu veri setinde faydalı olduğu gösterilmiştir.
(António vd., 2016)	Dört resort otelden alınan veri setlerini kullanarak ve bir sınıflandırma problemi olarak ele alarak, rezervasyon iptallerini %90'ın üzerinde doğrulukla tahmin etmeye yönelik modeller oluşturma mümkün olduğunu göstermektedir. Sonuçlar, otel yöneticilerinin iptal olasılığı yüksek rezervasyonlar üzerinde harekete geçmesine ve ilgili gelir kayıplarını kontrol altına almasına, daha iyi net talep tahminleri üretmesine, fazla rezervasyon/iptal politikalarını iyileştirmesine ve daha iddialı fiyatlandırma ve envanter tahsis stratejilerine sahip olmasına olanak tanıdığını göstermiştir.
Antonio vd., 2017a	Otel rezervasyon iptallerinin etkisini azaltmak için otel yönetim sistemi verileri ve makine öğrenimi tabanlı bir sistem prototipi kullanılarak, hangi rezervasyonların "iptal edilmesi muhtemel" olduğunun tahmin edilmesine ve bununla birlikte net talebin hesaplanması için her gün bir sınıflandırma modeli eğitilmiştir. İki otelde devreye alınan ve "iptal edilmesi muhtemel" olarak tahmin edilen rezervasyonlar üzerine alınan aksiyonların etkisinin ölçülebilmesi için A/B testleri kullanılmıştır. Çalışma sonuçları, otellerin iletişime geçtiği rezervasyonların, iletişime geçilmeyen rezervasyonlardan daha az iptal edildiğini göstermiştir.
Antonio vd., 2017b	Dört resort otelden alınan veri kümesi ve veri bilimi kullanılarak bir rezervasyon iptal tahmini sınıflandırma problemi ele alınmış ve rezervasyon iptallerinin %90'ın üzerinde doğrulukla tahmin edilebilmesi için modeller oluşturma mümkün olduğu gösterilmiştir. Bu çalışma, rezervasyonun iptal edilme olasılığının yüksek doğrulukla tahmin edilmesinin mümkün olduğunu göstermekte, dahası, otel yöneticilerinin net talebi doğru bir şekilde tahmin etmesine, daha iyi tahminler oluşturmaya, iptal politikalarını iyileştirmesine, daha iyi çifte rezervasyon taktikleri tanımlamasına ve böylece daha iddialı fiyatlandırma ve envanter tahsis stratejileri kullanmasına imkân sunmaktadır.

Tablo 1.1 (Devam)

Yazar ve Yıl	Çalışmanın Konusu
Cirillo vd., 2018	İnternetin büyük bir bilet dağıtım kanalı olarak artan kullanımı, yolcuların ücret politikasında daha stratejik hale geldiği, yolcuların daha düşük ücretli bilet almak için bileti önceden ayırtmaya ve daha sonra seyahat programından emin olduklarında biletlerini almaya teşvik ettiği belirtilmiştir. Bu çalışmada, demiryolu yolcuları için bir bilet iptali ve değişimi seçim modeli önermekte olup, ücret ve yolculuk programı belirsizliğine yanıt olarak bilet değişimi veya iptalinin meydana geldiği zamanlamayı tahmin etmek için dinamik bir ayrık seçim modeli (DDCM) uygulanmıştır. Problem optimal durma problemi olarak formüle edilmiş ve dinamik programlama problemine yaklaşmak için iki aşamalı, ileriye bakma politikası benimsenmiştir. Yaklaşım, şehirlerarası demiryolu yolculukları için gerçek bilet rezervasyon verilerine uygulanmıştır. Tahmin sonuçları, DDCM'nin çok terimli logit (MNL) modellerine kıyasla daha sezgisel sonuçlar sağladığını göstermektedir. Ayrıca doğrulama sonuçları, DDCM'nin MNL'den daha iyi tahmin yeteneğine sahip olduğunu göstermektedir. Demiryolu gelir yönetimi için değişim ve geri ödeme politikaları bağlamında burada geliştirilen yaklaşım, esnek geri ödeme politikaları altında faaliyet gösteren diğer endüstrilere genişletilebilir ve uygulanabilir olduğunu göstermiştir.
Antonio, 2019	Otel endüstrisinde rezervasyon iptallerinden kaynaklanan belirsizliğin üstesinden gelinmesi ve bir rezervasyon iptalinin tahmin edilebilmesi için modeller geliştirilmiştir. Otellerin rezervasyon sistemlerinden gelen veriler, diğer kaynaklardan gelen verilerle birleştirilerek; veri sınıfı dengesizliğine ve veri setindeki sorunlara rağmen, rezervasyon iptallerinin doğru tahmin edilmesinin mümkün olduğu gösterilmiştir. Ayrıca, iptal faktörlerinin neler olabileceğinin daha iyi anlaşılmasına yardımcı olmuştur. Modellerin gerçek koşullar altında değerlendirilmesi için, rezervasyon iptallerini tahmin eden otomatik makine öğrenme sisteminin, otel sistemlerine nasıl entegre edilebileceğini gösteren bir prototip geliştirilmiş ve saha testleri için kullanılmıştır. Modelin gerçek ortamdaki performansı, iş üzerindeki etkisi de dahil olmak üzere değerlendirilmiştir. Prototip uygulaması, gerçek bir ortamda etkili bir şekilde çalışabilmeleri için modellerde yapılacak ayarlamaların anlaşılmasını sağlamış ve aynı zamanda yeni bir performans değerlendirme ölçüsünün oluşturulmasına katkı sağlamıştır. Prototip, otelcilerin belirlenen rezervasyonlara göre hareket etmelerini ve iptalleri etkin bir şekilde azaltmalarını sağlamıştır.
Antonio vd., 2019a	Otel rezervasyon iptallerinden kaynaklanan belirsizliğin üstesinden gelmek için, sekiz otelin yönetim sisteminden gelen verilerle diğer kaynaklardan gelen veriler birleştirilmiştir. Rezervasyon iptal tahmininin yapılabilmesi için makine öğrenimi ve algoritmalar kullanılarak oteller için tahmin modelleri hazırlanmıştır. Gerçek bir üretim ortamında, bu modellerin kullanımı ile tahmin doğruluğunun iyileşeceği, otelcilerin iptal sayısını azaltacağı ve dolayısıyla talep yönetimi kararlarına olan güvenin artmasını sağlayacağı gösterilmiştir. Bu çalışma, talep tahmininin iyileştirilmesinin otelcilerin net taleplerini ve tahmini iptalleri daha iyi anlamalarını sağlamaktadır. Ayrıca, iptal etkenlerini ortaya çıkararak, otelcilerin rezervasyon iptal modellerini daha iyi anlamalarına ve bir otelin iptal politikalarının ve fazla rezervasyon taktiklerinin rezervasyonlarının özelliklerine göre ayarlanmasına yardımcı olduğunu göstermektedir.

Tablo 1.1 (Devam)

Yazar ve Yıl	Çalışmanın Konusu
Antonio vd., 2019b	Günümüzün hesaplama gücü ve gelişmiş makine öğrenimi algoritmaları ile rezervasyonların iptal olasılığını tahmin etmek için modeller oluşturmanın mümkün olduğu, ancak bu modellerin etkinliğinin hiçbir zaman gerçek bir ortamda değerlendirilmediği belirtilmiştir. Bu boşluğu doldurmak için bu modellerin bir karar destek sisteminde nasıl uygulanabileceğini ve talep yönetimi kararları üzerindeki etkisini araştırmak için bir prototip oluşturulmuş ve iki otelde devreye alınmıştır. Sürekli öğrenmek için tasarlanmış otomatik bir makine öğrenimi sistemine dayanan prototip, zaman içinde iptal kalıplarındaki değişiklikleri yakalamak ve önceki günlerin tahmin isabetlerinden ve hatalarından ders çıkarmak için tasarlanmış bir eğitim yöntemi ve ağırlıklandırma mekanizmasının geliştirilmesine katkı sağlamıştır. Ayrıca, tahminlerin zaman içindeki kesinliğini ölçmek için minimum frekans gibi yeni bir ölçü konusunda katkı sağlamıştır. Prototip, doğrulukta %84, hassasiyette %82 ve Eğri Altındaki Alanda (AUC) %88 sonuçlarla etkinliğini göstermiştir. Sistem, otellerin net taleplerini tahmin etmelerine ve böylece hangi rezervasyonların kabul edilip reddedileceğine, fiyatlamasının doğru yapılmasına ve kaç odanın satılacağına ilişkin daha iyi kararlar vermelerine olanak sağlamıştır. İptal olasılığı yüksek olan rezervasyonların sistematik olarak tahmin edilmesi, otellerin iptallerini önlemek için harekete geçerek iptallerin yüzde 37 oranında düşmesini sağlamıştır.
Andriawan vd., 2020	CRISP-DM ve veri bilimi yardımıyla bir otel rezervasyonunun iptal edilip edilemeyeceğini tahmin etmek ve otelin kayıp kârlarını en aza indirebilmek için ağaç tabanlı algoritma kullanılmıştır. Rastgele Orman modelinin %87 doğruluk değeri ile en iyi değere sahip olduğu ve rezervasyonun yapıldığı zaman ile varış zamanı arasındaki farkın iptal düzeyini tahmin etmede en etkili özellik olduğu gösterilmiştir.
Pérez vd., 2020	Paratransit işlemlerini ilgilendiren iki tahmin problemi için makine öğrenmesi (ML) algoritmalarının uygulaması incelenmiştir. İlk problemde operatör sadece hangi rezervasyonların seyahatle sonuçlanıp hangilerinin olmayacağını tahmin etmekle ilgilenirken, ikinci tahmin probleminde operatör ikiden fazla rezervasyon sonucuyla ilgilenmektedir. Lojistik regresyon, rastgele orman, gradyan artırma ve aşırı gradyan artırma makine öğrenme algoritmalarının dışında, yolculuk rezervasyonlarına sonuç olasılıklarını atamak için kümeleme tabanlı bir yaklaşım geliştirilmiştir. Porto Riko Büyükşehir Otobüs Kurumu tarafından sağlanan seyahat rezervasyon verileri kullanılarak, seçilen algoritmaların tahmin doğruluğunu incelemek için testler yapılmış ve gradyan artırma ve aşırı gradyan artırma algoritmalarının, sınıflandırma testlerinde en iyi performans gösteren yöntemler olduğu gösterilmiştir. Algoritmaların bir uygulamasını göstermek için oluşturulan talep tahmin modellerinin, paratransit sistemlerinde günlük gezileri tahmin etmek için umut verici bir yaklaşım olduğu gösterilmiştir. Makine öğrenimi algoritmaları tarafından oluşturulan talep tahminlerini en uygun şekilde birleştiren regresyon modeli, çalışmada en iyi performans gösteren yöntem olmuştur.

Tablo 1.1 (Devam)

Yazar ve Yıl	Çalışmanın Konusu
Sánchez vd., 2020	Hizmet saatine yakın yapılan iptallerin, otellere en çok zarar verenler iptaller olduğu belirtilmiştir, PNR kullanımının bu alanda yeni yaklaşımlara yol açtığı fakat bu yeni araştırma alanına rağmen hizmet zamanına yakın yapılan bireysel otel iptallerinin tahminine odaklanan araştırmaların bulunmadığı üzerinde durulmuştur. Bu boşluğu doldurmak amacıyla, Yapay Zekâ teknikleri kullanılarak kısa bir süre içerisinde iptal edilecek olan rezervasyonlar PNR verileri üzerinden belirlenmeye çalışılmış ve 7 gün öncesinden yapılan iptaller %80 doğrulukla tahmin edilmiştir.
Sánchez-Medina ve C-Sánchez, 2020	13 değişken kullanarak otellere yapılan rezervasyon iptal tahminini yapan bir araç önerilmiştir. Makine öğrenimi tekniklerinin (RO, DVM, C5.0) yanı sıra genetik algoritmalarla optimize edilmiş yapay sinir ağları uygulanarak yaklaşık %98 tahmin başarıları yakalanmıştır. Önerilen metodoloji, yalnızca iptal oranlarının tahminine değil, aynı zamanda hangi müşterinin iptal etme olasılığının olduğunu belirlemeye de olanak tanımaktadır.
Satu vd., 2020	Otel rezervasyon iptali sürecinde farklı makine öğrenmesi yöntemlerinin etkileri araştırılmıştır. Kaggle veri havuzundan otel rezervasyonlarına ait elde edilen veri seti üzerinde öznitelik dönüştürme teknikleri uygulanmış ve dönüştürülmüş veri setleri oluşturulmuştur. Ayrıca, öznitelik seçim yöntemleri kullanılarak önemsiz değişkenler azaltılmış ve birincil ve oluşturulan alt kümelerde çeşitli sınıflandırıcılar kullanılmıştır. Makine öğrenimi yöntemlerinin etkileri bu adımda değerlendirilip en iyi yaklaşımlar araştırılmıştır. Tüm bu yöntemler arasında, XGBoost'un bu veri kümelerini analiz etmek için en sık kullanılan yöntem olduğu ve bilgi kazanımı öznitelik seçim yöntemi için bireysel sınıflandırıcıların en yüksek sonucu ürettiği görülmüştür.
Adil vd., 2021	Otel iptallerini tahmin etmek için önceki çalışmaların DVM, YSA ve karar ağacı modelleri gibi çeşitli makine öğrenimi yaklaşımlarını kullandığına, ancak otel iptallerinin tahmininde var olan sınıf dengesizlik sorununu henüz çözemediğine değinilmiştir. Bu çalışmada, otel rezervasyon iptallerini daha iyi tahmin etmek için makine öğrenimi algoritmaları ile sınıf dengesizliği sorunlarını ele almak için bir aşırı örnekleme tekniği sunulurken bu boşluk ele alınmıştır. Sınıf dengesizliği sorununu çözmek için sentetik azınlık aşırı örnekleme tekniği ve düzenlenmiş en yakın komşular (SMOTE-ENN) algoritmasının bir kombinasyonu önerilmiştir. Araştırmada, sınıf dengesizliği sorunu ele alındıktan sonra, bir makine öğrenimi sınıflandırıcısının performansının önemli ölçüde arttığı gösterilmiştir.
Almotiri vd., 2021	Bir rezervasyonun iptal edilip edilmeyeceğini tahmin etmek için işlev sağlayabilecek bir API oluşturulması amaçlanmaktadır. Bu API, her bir rezervasyon sürecini aynı parametrelerle değerlendirmek için otel yönetim sistemleriyle entegre olabilecektir. Bunun için Chi testi kullanılarak faktörlerin tanımlanması ve etkili değişkenlerin bulunması için korelasyon ve lineer regresyon kullanılmıştır. API için bir puanlama işlevi oluşturmak için lojistik regresyon modelinden kesme ve katsayılar kullanılmıştır.

Tablo 1.1 (Devam)

Yazar ve Yıl	Çalışmanın Konusu
Wang, 2021	Bir otelin gelecekteki talebini tahmin etmenin önemine değinilmiş ve otel yolcu akışının; hava durumu, çevre, mevsim, tatiller, ekonomi ve acil durumlar gibi birçok faktörden etkilendiği ve karmaşık doğrusal olmayan dalgalanma özelliklerine sahip olduğu vurgulanmıştır. Mevcut talep tahmin yöntemlerinden doğrusal tahmin yöntemlerinin, çevrenin ve zaman serilerinin kararlılığına dayandığı, bu nedenle otel yolcu akışının karmaşık doğrusal olmayan dalgalanma özelliklerini tam olarak simülasyon edemediği belirtilmiştir. Geleneksel doğrusal olmayan tahmin yöntemlerinin tahmin doğruluğunu iyileştirmesi gerektirdiğine ve artan otel yolcu akışı verileriyle başa çıkmanın zorluğuna değinilmiştir. Bu makalede, gerçek aylık varış rezervasyonlarının sayısını tahmin amacıyla Uzun Kısa Süreli Bellek yöntemi ile derin öğrenme uygulanmıştır. Gerçek aylık varış rezervasyonlarının sayısı, bir otelin gerçek aylık yolcu akışını yansıtabilir. Tahmin modeli, çeşitli nedenlerle rezervasyonların iptal edilmesi veya gelmemesinden kaynaklanan kayıpları etkin bir şekilde azaltabildiğini, LSTM modelinin tahmin yeteneğini etkin bir şekilde geliştirebildiğini ve otelin gelecekteki fiyatlandırma kararı ve işletme modu düzenlemesi için gerekli referansı sağlayabildiğini göstermektedir.
Rakesh vd., 2022	Tatil oteli ve şehir oteli olmak üzere iki otel ele alınmış ve otel yönetimi tarafından gerçekleştirilen belirli eylemlerin otel gelirleri ve iptaller üzerindeki etkisini tahmin etmek için ML modelleri kullanılmıştır. Bu modellerin, yönetimin politika ve kararları hakkında yeniden düşünmesini sağlayacağı ve makine öğrenmesi modellerinin ve meydana gelebilecek iptallerin sayısının tahmin edilmesine yardımcı olacağı belirtilmiştir.
C.-Sánchez vd., 2022	Bir otel ortağı tarafından toplanan PNR verileri kullanılarak otel rezervasyon iptallerini tahmin etmek için bir model tasarlanmış ve %74 doğrulukla tahmin gerçekleştirilmiştir. Bu çalışma, önceki çalışmalara kıyasla daha az sayıda ve yalnızca 13 farklı bağımsız değişkeni dikkate almaktadır. Ayrıca dahil edilen değişkenler, otel rezervasyon sürecinde müşterilerden talep edilenlerle örtüşmektedir. Bu değişkenlerin genellikle mevcut olan tek değişken olduğu göz önüne alındığında, bu durum konaklama işletmeleri için bir avantaj sağlamaktadır. Geliştirilen yöntem, iptal oranını iyi bir doğrulukla tahmin etmekte ve aynı zamanda rezervasyonlarını iptal etme olasılığı olan müşterileri de belirlemektedir. Bu yaklaşımın, gelen turistlerle ilgili onların eylem protokollerini geliştirmelerine yardımcı olacağı için kuruluşlara bir değer katacağı belirtilmiştir.

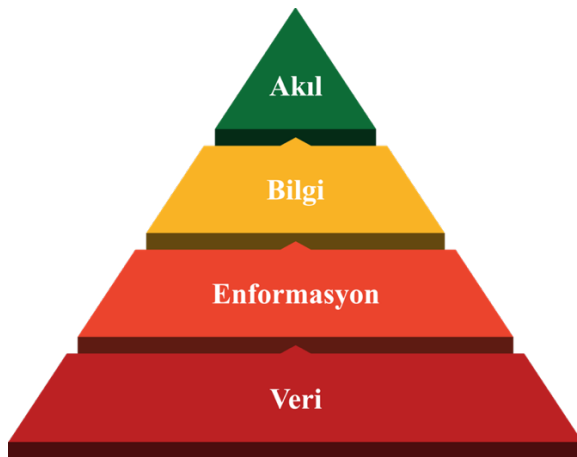
İKİNCİ BÖLÜM

VERİ MADENCİLİĞİ

İkinci bölümde veri madenciliği çalışmalarında sık kullanılan veri, enformasyon, bilgi, bilgelik, veri tabanı, karar destek sistemi, makine öğrenimi terimlerinin açıklamaları, veri madenciliği tanımı, veri madenciliğinin amacı, veri madenciliğinin gelişimi, veri madenciliğinin kullanıldığı alanlar, veri madenciliği işlemlerinde kullanılan yazılımlar ve veri madenciliği süreçlerinden CRISP-DM süreci tüm adımları ile detaylı bir şekilde ele alınmıştır.

2.1. Veri Madenciliği Terimleri

Veri madenciliğinde sık kullanılan terimler literatürde farklı şekillerde ifade edilmektedir. Bu terimler arasında veri (data), enformasyon (information), bilgi (knowledge), veritabanı (database), karar destek sistemi (decision support system), bilgi madenciliği (knowledge mining), veri tabanlarında bilgi keşfi (knowledge discovery in databases), gelişmiş veri analizi (advanced data analysis), iş analitikleri (business analytics), makine öğrenmesi (machine learning), örüntü analizi (pattern analysis) ve iş zekâsı (business intelligence) yer almaktadır (Altunkaynak, 2019: 13; Mitchell, 1999: 1). Veri madenciliği ile ilgili literatür taramalarında, bazı araştırmacılar veriden enformasyona, enformasyondan da bilgiye ulaşabileceğini ifade ederek bilgi hiyerarşisini üç katman olarak tanımlamaktadır (Chadha, 2023). VEBB piramidi olarak bilinen, *Veri - Enformasyon - Bilgi - Bilgelik (Akıl)* terimleri arasında Şekil 2.1’de gösterildiği gibi hiyerarşik bir ilişki yer almaktadır (Demir ve Dinçer, 2020: 5). Veri madenciliğinin temelini oluşturan veri, enformasyon, bilgi, veritabanı, karar destek sistemi ve makine öğrenmesi kavramları bu bölümde açıklanmıştır (Kelleher ve Tierney, 2022: 48).



Şekil 2.1 VEBB Piramidi

Kaynak: Demir ve Dinçer, 2020: 5

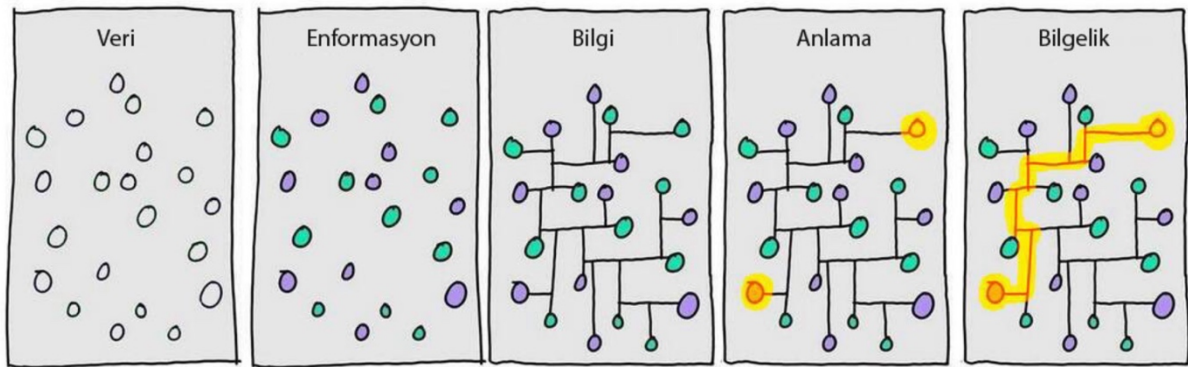
Veri: Latince “vermek” anlamına gelen “dare” eyleminin geçmiş zaman kipinde kullanımı olan “datum” isminin çoğul halini ifade etmektedir (Akpınar, 2017: 15). Veri madenciliğinin temelini oluşturan veri, yapılan gözlemler ve deneyler sonucu elde edilen, tek başına anlamsız hammadDEDİR (Gökçen, 2007: 4). Bilgi hiyerarşisine göre ilk olarak elde olan veri, bir sistemde tutulan sayısal veya alfa nümerik değerlerdir. Veri, işlenmemiş bilgilere verilen bir isimdir. Veriye örnek olarak; hastanın yaşı, rezervasyon tarihi, siteyi ziyaret sayısı, tedavi kategorisi, tedavi bedeli, hastanın uyruğu, konaklayacağı gece sayısı, medeni durumu, çocuk sayısı, rezervasyon ile otele giriş arasında geçen gün sayısı verilebilir. Veri tek başına anlamı olmayan fakat nesnenin özellik, olay ve ilişkilerini tanımlayan sembollerdir. Müşterinin rezervasyonunu neden iptal ettiğİ veya müşterinin genelde hangi aylarda tedavi olduđu, verilere bakılarak anlaşılamaz. Karar vericiler, işlenmemiş ham verilere bakarak, operasyonel ve yönetsel karar alamazlar. Verinin enformasyona dönüşebilmesi için bir araya getirilmesi, düzenlenmesi ve anlam ifade etmesi gerekmektedir. Bilgisayarların veri saklama hız ve kapasitelerinin gelişmesiyle birlikte veritabanında saklanan veri yığını ve veri alan sayısı da artmaktadır. Veritabanında saklanan veriler tek başlarına anlam ifade etmemesine rağmen, belirli bir amaç doğrultusunda işlenen, doğru yer ve zamanda kullanılan veriler, karar vericiler için gittikçe önemli hale gelmekte ve bir değer kazanmaktadır. (Savaş vd., 2012: 1).

Enformasyon: Bir amaç doğrultusunda önceden düzenlenen ve biçimlendirilen, kullanıcıların kendi bakış açılarıyla doğruluğuna karar vermesi ve anlam yüklemesi gereken veriler “enformasyon” olarak tanımlanmaktadır (Davenport ve Prusak, 2000: 25; Laudon ve Laudon, 2014: 46). Verilere kıyasla belirgin bir çerçeveye sahip olan çeşitli çıkarım ve işlemleri içermektedir. Enformasyon, diğEr bir ifadeyle malumat, belirli bir amaç doğrultusunda işlenen ve filtrelenen verilerdir. Örneğın, bir sağık kuruluşundaki günlük operasyon sayısı, hasta ile toplam iletişim kurma sayısı, bir tedaviye ait gelen ortalama başvuru sayısı ve uzmanlık bazlı en fazla rezervasyon sayısı, sağık kuruluşu için enformasyondur.

Bilgi: Enformasyon ve bilgi kavramları, sınırlarının net olmaması ve bu sözcüklerin İngilizce karşılığı olan “knowledge” ve “information” kelimelerinin Türkçe karşılığının “bilgi” kelimesi olması sebebiyle sıklıkla karıştırılmaktadır (Yılmaz, 2009). Günümüzde kullanılan “bilgi çağında yaşıyoruz” ifadesinin hatalı bir kullanım olduđu ve “veri çağında yaşıyoruz” şeklinde değışmesi gerektiğİ de kavramların karıştığının bir göstergesidir. Albert Einstein’ın “Information is not knowledge.” sözü ile, enformasyonunun bilgi olmadığı,

bilginin enformasyondan çok daha değerli olduğu, bilginin analiz edilmiş, yorumlanmış, zenginleştirilmiş, özümsemiş olduğu ve karar vericilerin işletme faaliyetleri için bilgiye ihtiyaç duyduğunu ifade etmektedir. Enformasyonun anlam kazanmış hali bilgi olarak isimlendirilmektedir (Bhatt, 2001: 69). Bilgi; enformasyonun incelenmesi, sentezlenmesi, değerlendirilmesi ve karar verilmesi sürecinin sonucunda elde edilmektedir. Bilgi bir konuyla ilgili deneyim ve gözlemler sonucu fikir sahibi olma, öğrenme ve özümsemedir. Bilginin nasıl tanımlandığı, bilginin nasıl yönetileceğini belirlemektedir (Allee, 1997: 2). Bilgi, karar sürecine katkı sağlayan yüksek değere sahip enformasyondur (Mårtensson, 2000: 208). İşletmeler, sahip oldukları bilgiyi kullanamadıklarında hiçbir fayda sağlayamazken, bilgiyi kullanmaları ölçüsünde pazarda rekabet avantajı sağlayıp avantajlı konuma ulaşabilmektedir (Kefe, 2013: 20). “Obezite en fazla rezervasyon iptaline sahip tedavidir.” ve “diş tedavisi için en çok başvuruyu İngiliz hastalar yapmaktadır” ifadeleri, bilgiye örnek olarak verilebilir.

Bilgelik: Verinin sentez ve analiz edilerek birleştirilmesi, bütünleştirilmesi, sindirilmesi ve anlamlı hale getirilmesin bilgelik (wisdom) denilmektedir. İlk aşamada ham olarak kaşımıza çıkan veri, son aşamada bilgelik kavramına dönüşmekte ve karar destek faaliyetlerini Şekil 2.2’deki gibi desteklemektedir (Intezari vd., 2017).



Şekil 2.2 Veriden Bilgelik Yolculuğu

Kaynak: Kaya, 2018

Veri tabanı: Verilerin bilgisayar ortamında belirli ilişkiler temelinde sistematik bir yapıya getirilmiş biçimine veri tabanı denilmektedir (Burma, 2009: 12). Veri tabanı, veri koleksiyonu olarak da tanımlanmıştır (Vidhya, 2016: 1). Veri tabanı, verileri gereksiz tekrarlardan arındırılmış ve düzenli bir yapıda saklamaktadır. Veri tabanları, veriye ulaşmak isteyen kullanıcılar arasında aracı görevi gören araçlar olarak kabul edilmektedir. Veri tabanları depolanmış veriyi düzenleyen ve ona erişimi sağlayan sistemlerdir (Abiteboul vd., 1995: 3). Günlük yaşamda yoğun bir şekilde ortaya çıkan veri, fiziksel olarak veri tabanında saklanmaktadır. Veriler, sayısal ortamlarda yığınlar şeklinde kayıt altına alınmakta ve veri

tabanlarında bilgiye dönüştürülmeyi beklemektedir (Silahtaroglu, 2013: 11). Veri tabanı yazılımları karar destek sistemlerinin önemli bir unsuru olarak değerlendirilmektedir.

Karar Destek Sistemi (KDS): Peter Keen ve Charles Stabell, 1950'lerin sonlarında yaptığı örgütsel karar vermenin teorik çalışmalarında ve 1960'larda gerçekleştirilen etkileşimli bilgisayar sistemleri üzerindeki teknik çalışmalarda karar destek kavramını tanımlamıştır (Keen ve Morton, 1978). 1961'de Michael S. Scott Morton, "Yönetim Karar Sistemleri: Karar Verme için Bilgisayar Tabanlı Destek" çalışmasını yayınlamıştır. 1968-1969 yıllarında kritik karar vermede bilgisayarların ve analitik modellerin etkisini incelemiştir. Morton'un bu araştırması KDS hareketinin başlatılmasında kilit rol oynamıştır (Rashidi vd., 2017). KDS, probleme çözüm üretme, sıralı ve bağımlı kararlara destek verme, farklı stratejiler ve koşul analizleri ortaya koyma, yanlış karar maliyetini azaltma, zaman tasarrufu, objektif ve tutarlı karar verme gibi birçok fayda sağlamaktadır (Kapanoğlu, 2016).

Makine Öğrenmesi: Veri madenciliği, yapay zekâ ve istatistiksel yöntemlerin bir arada kullanıldığı ileri düzey analiz yöntemidir (Ersöz, 2019: 14). Makine öğreniminin çalışma prensibinde, yazılımların insanın öğrenme biçimini taklit ederek veriden öğrenimi söz konusudur. Veriler analiz edildikten sonra ortaya çıkan değerli bilgiler, insanların bilinmeyenleri keşfetmelerini sağlamaktadır. Veri madenciliği yöntemleri, literatürde makine öğrenmesi teknikleri olarak geçmektedir. Makine öğrenmesinde; KNN, NB, DVM, YSA, DT yöntemleri sıklıkla tercih edilmektedir (Ikonomakis, vd., 2005: 6).

2.2. Veri Madenciliği Tanımı

Veri madenciliği, literatürde farklı şekillerde tanımlanmaktadır. Veri madenciliğine ait tanım ve açıklamalar aşağıdaki gibi kronolojik sırada verilmiştir.

Veri madenciliği, genellikle veri tabanlarında bilgi keşfi bağlamında yer bulmaktadır ve yapay zekâ araştırma alanından çıkmıştır (Hand vd., 2001: 3). Sigortacılık, bankacılık, sağlık ve perakende sektörlerinde işletme giderlerini azaltmak ve satışları arttırmak amacıyla veri madenciliğinden faydalanılmaktadır (Seifert, 2004: 463). Veri madenciliği, büyük veri setlerindeki modelleri ve ilişkileri keşfetme süreci olarak tanımlanmaktadır (Hardin ve Chhieng, 2007: 44). Veri tabanlarından bilgi keşfi ile veri madenciliği karıştırılan iki kavram olarak karşımıza çıkmaktadır. Veri tabanlarında bilgi keşfi, veriden faydalı bilginin keşfedilmesi sürecini ifade ederken, veri madenciliği ise veri içerisinde var olan örüntülerin çeşitli algoritmalar vasıtasıyla ortaya çıkarılması olarak tanımlanmaktadır (Aydın, 2007).

Veri madenciliği, özellikle bağımlılık yapıları olan çok boyutlu ve çok değişkenli veri setlerinin araştırılması anlamına gelmektedir. Veri madenciliğinde kilit nokta, verilerin, çok

büyük veri setleriyle yapılan analizler sonucunda belirsizliğinin giderilerek bilgiye dönüştürülebilmesidir (Clarke, 2009: 1-2). Veri madenciliği alanında gerekli donanım ve beceriye sahip örgütlerin, çağımızın rekabetçi ortamında bir avantaj sağlayacağı iddia edilmektedir (Weiss ve Davison, 2010: 1).

Veri madenciliği, tanımlı bir veri koleksiyonundan özet, değer ve model keşfetme sürecidir (Kantardzic, 2011: 6). Veri madenciliği terimi, kürekle değerli madenlerin çıkarılabildiği madencilğe benzerliğinden dolayı veri madenciliği ismini almıştır. Bilgi madenciliği, bilgi analizi, bilgi çıkarımı, bilgi arkeolojisi ve bilgi taranması, veri madenciliğine benzeyen terimlere örnek olarak gösterilebilir. Veri madenciliğinde tanımlama yapma ve tahmin etme olmak üzere iki amaç bulunmaktadır. Tanımlama, insanlar tarafından yorumlanabilecek verileri tanımlayan kalıplar bulmaya odaklanmaktadır. Tahmin ise, ilgili değişkenlerin bilinmeyen değerlerini tahmin etmek için veri kümesindeki bazı değişkenlerin kullanılmasını içermektedir (Kantardzic, 2011: 2). Veriler işlenmediği zaman anlamı olmayan veri yığınları olarak kalmaktadır. Bu yüzden literatür taramada veri madenciliği; veri seçimi, veri temizleme, veri dönüşümü, veri entegrasyonu, eğilim ve kalıp değerlendirme ve bilgi sunumu süreçlerini içeren veriden bilgi keşfi olarak tanımlanmaktadır (Han vd., 2012: 6-7).

Veri madenciliği 1960'lı yıllarda, analizler öncesi aykırı verilerin tespit edilmesi ve ayıklaması için tercih edilen “Veri Avlama” ve “Veri Tarama” terimlerinden oluşmaktadır (Pektaş, 2013: 99). Veri madenciliğinin amacı; günümüzün rekabetçi koşullarında rakiplere karşı fark oluşturmak amacıyla veri içinde geleneksel yöntemlerle bulunamayan, fark yaratacak bilgileri bulmaktır. Veri madenciliği işlemi, gerçek bir madencinin, değerli bir madeni bulma ve işleme faaliyetleri ile benzerlik göstermektedir (Özkan ve Boran, 2014: 60). Veri madenciliği; sigortacılık, borsa, bankacılık, pazarlama, elektronik ticaret, satış, sağlık, genetik bilimi, biyoloji ve tıp, telekomünikasyon, coğrafi bilgi sistemleri, görüntü tanıma, sinyal işleme, sanayi, robot görüş sistemleri, inşaat, eğitim, zekâ, mühendislik, meteoroloji, sosyal bilimler, sosyal medya ve web verileri analizi alanlarında uygulanmaktadır (Yalçın, 2016: 37). Veri madenciliği; veri tabanı teknolojisi, istatistik, görselleştirme, yapay zekâ ve makine öğreniminden yararlandığı için disiplinler arası bir özellik göstermektedir (Cemaloğlu ve Duykuluoğlu, 2020: 6; Ersöz, 2019: 14).

2.3. Veri Madenciliği Yazılımları

Veri madenciliği işlemlerinde kullanılabilecek çeşitli programlama dilleri ve araçları bulunmaktadır. Bu bölüm altında veri madenciliği süreçlerinde sıklıkla tercih edilen yazılım ve programlama dillerinden bahsedilecektir.

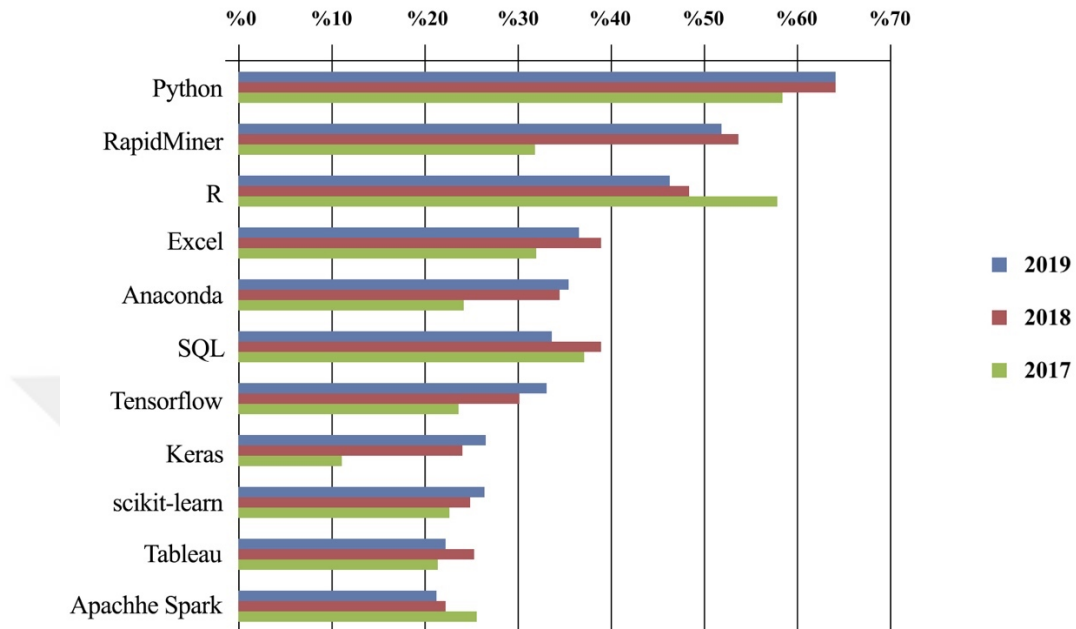
R: Veri madenciliği araçları içerisinde en kapsamlı istatistiksel analiz paketi olan *R*, veri manipülasyonu, programlama ve görselleştirmeleri bir arada sunan entegre bir yazılım ve programlama dilidir. En büyük avantajı ücretsiz olmasıdır. Ayrıca, başvurulacak kaynakların çeşitliliği, sık sık güncellenen yapısı ve veri madenciliği için sunduğu zenginlikler sebebiyle büyük veri analitiğinde birçok disiplin tarafından yaygın bir şekilde kullanılmaktadır. Açık kaynak kodlu olması da *R* dilinin hızlı gelişimine katkı sağlamaktadır. *R* kullanımı dünyada her geçen gün yaygınlaşmasına rağmen, ülkemizde akademik çalışmalar dahil olmak üzere kısıtlı seviyelerdedir. 2015 yılında yapılan araştırmada *R* en popüler kullanılan programlama dilleri arasında altıncı sıradadır (Pektaş, 2013: 1-5).

Knime: 2004 yılında Konstanz Üniversitesi tarafında geliştirilmeye başlanan, 2006 yılında ilaç endüstrisinde kullanılmak üzere ilk versiyonu piyasaya sürülen *Knime* yazılımı, Java programlama dilinde geliştirilmiş açık bir platformdur. *R* Project, Weka, LIBSVM, ImageJ gibi açık projelerle entegre yapıya sahip olması ve ücretsiz erişim imkânı sunulması, *Knime* kullanımını yaygınlaştırmaktadır. Sadece belirli hizmetlerden ücret talep edilmektedir (Köse, 2018: 59). *Knime*, kullanıcılarına geliştirme kiti sunarak kendi modüllerini oluşturma imkânı sağlayan tek veri madenciliği yazılımıdır. Ayrıca diğer veri madenciliği yazılımlarına göre kurulum şartı olmadan çalışabilen, en zengin görselleştirmelere sahip yazılım olarak ön plana çıkmaktadır (Kaya ve Özel, 2014: 49).

RapidMiner: Dortmund Üniversitesi yapay zekâ kürsüsünde “Yale” adı altında 2001 yılında geliştirilmiştir. 2010 yılından itibaren RapidAnalytics olarak piyasaya sürülen veri madenciliği yazılımıdır. *RapidMiner*, analizleri işlem görünümünde yapılandığı bir arayüze sahip olup, ön işleme ve öğrenme adımlarına sahiptir. Her adımın çıktıları operatör tarafından kullanıcılara gösterilmektedir. Operatörler giriş verilerini alma, işleme ve çıktıları sunma işlemleri için giriş ve çıkış portlarına sahiptir (Land ve Fischer, 2012: 1-2). Diğer veri madenciliği yazılımlarına göre oldukça kapsamlıdır. 22 adet farklı dosya formatı üzerinde işlem yapabilme imkânı sağlayan *RapidMiner*, çok sayıda veri madenciliği algoritmalarına sahip ücretli bir yazılımdır. Sahip olduğu algoritmalar ve desteklediği veri tabanlarının çeşitliliği, *RapidMiner* yazılımını veri madenciliği sürecinde ilk sıralara doğru taşımaktadır. Demo sürümlerinde kullanıcılara 10.000 satırlık bir işlem yapma imkânı vermektedir. Pahalı olması ve demo sürümünün kısıtlılığı sebebiyle büyük veri analizlerinde bireysel kullanımlar için uygun değildir.

Python: Büyük verilerden inceleme yaparak, metin örgüleri üzerinde makine öğrenimine dayalı sonuçlar çıkarılması ve elde edilen bilgilerin görselleştirilmesi çalışmanın ana gereksinimini oluşturmaktadır. Bu kapsamda, Google’ın yürütücülüğünde faaliyet

gösteren ve veri dünyasının büyük platformlarından biri olan Kaggle isimli sitede, makine öğrenimi temelinde kullanılan programlama dillerinin tercihiyle ilgili bir araştırma yapılmış, bunun sonucunda 36.824 oy (%89,6) ile Python dilinin birinci sırada ve 4.267 oy (%10) ile R programlama dilinin ikinci sırada tercih edildiği görülmüştür (Ateş, 2021).



Şekil 2.3 Veri Biliminde Tercih Edilen Teknolojiler

Kaynak: KDnuggets, 2023

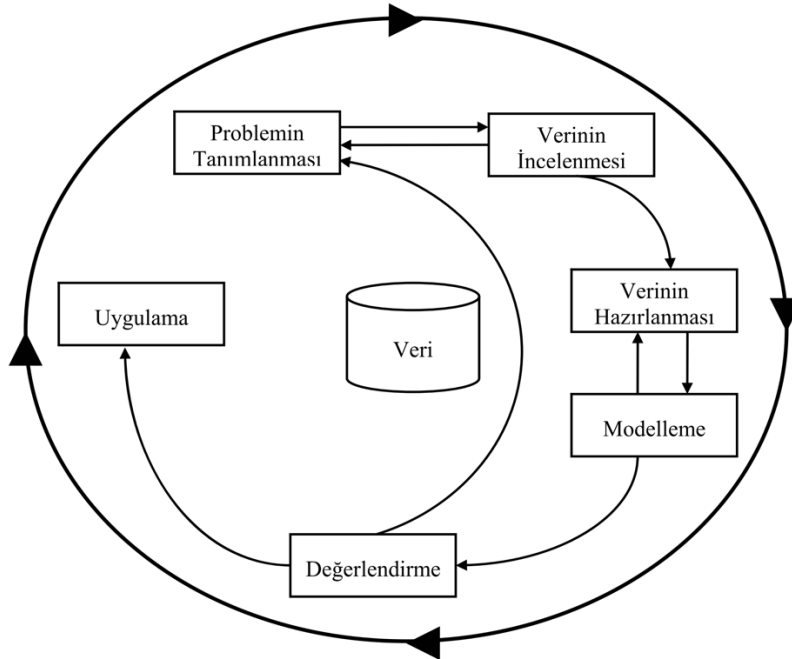
Veri madenciliği konusunda en büyük platformlardan olan KDnuggets'in 20. yıl anketi 1.800 katılımcıya uygulanmıştır. “Gerçek bir proje için son 12 ayda kullandığınız Analitik, Büyük Veri, Veri Bilimi ve Makine Öğrenimi yazılımı nedir?” sorusu yöneltilmiş olup, elde edilen cevaplarda 2017 yılında %59, 2018 yılında %65,6 ve 2019 yılında %65,8 oranları ile Python programlama dilinin en çok tercih edilen programlama dili olduğu Şekil 2.3'te gösterilmiştir.

Genel kullanım oranlarına bakıldığında sık kullanımı ile dikkat çeken Python, çoklu programlama paradigmalarını destekleyen genel amaçlı bir programlama dili olup, içinde sınıflar, nesneler, veriler ve yöntemler dahil olmak üzere birçok nesne yönelimli programlama kavramı da bulunmaktadır. Python, veri tabanı yönetimi, makine öğrenmesi, derin öğrenme, doğal dil işleme ve web geliştirme gibi amaçlarla aktif kullanılmaktadır (Sarkar, 2019). Bu tez çalışmasının veri madenciliği süreçlerinde de Python dili tercih edilmiştir.

2.4. CRISP-DM Veri Madenciliği Süreci

Veri madenciliği süreçleri için literatürde, CRISP-DM (Cross-Industry Standard Process for Data Mining), SEMMA (Sample, Explore, Modify, Model, Assess), KDD (Knowledge Discovery in Databases) vb. çeşitli yaklaşımlar bulunmaktadır.

Doktora araştırması kapsamında, CRISP-DM veri madenciliği süreç metodolojisi tercih edilmiştir. Şekil 2.4'te adımları gösterilen CRISP-DM sürecine ait görev ve çıktılar bu bölüm altında detaylı olarak açıklanmıştır.



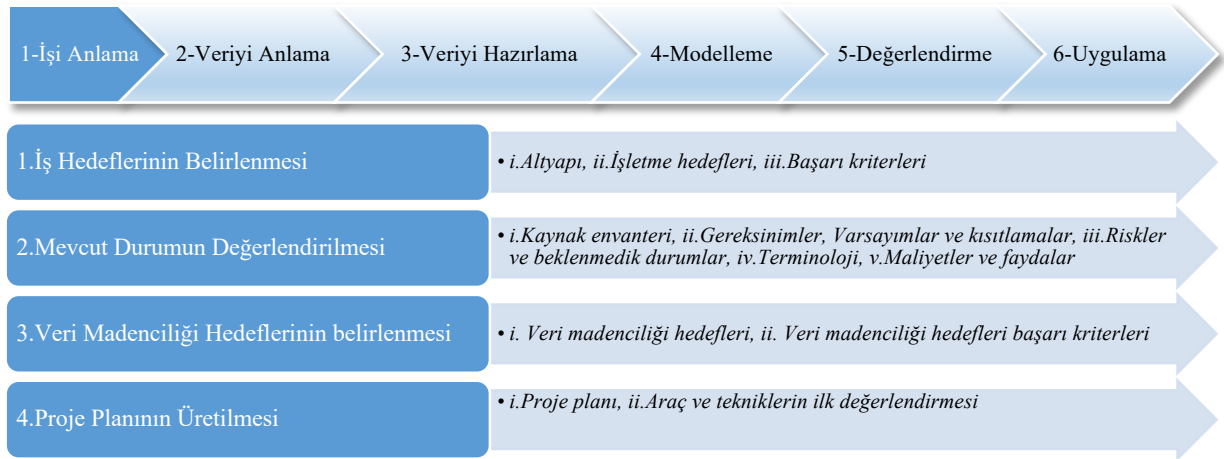
Şekil 2.4 CRISP-DM Veri Madenciliği Süreci

Kaynak: Irmak, 2009:14

2.4.1. İş Anlama

Veri madenciliğinin birinci ve en önemli ön şartı, iş problemini ve verileri iyi anlamaktır. Etkin ve başarılı bir veri madenciliği yapabilmek için hedeflerin net bir şekilde tanımlanması gerekmektedir (Two Crows Corporation, 1999: 22). İşin tanımlanması (iş anlama) aşaması, veri madenciliğinin hangi amaç ve ihtiyaçtan yapıldığının iyi anlaşılması, veri madenciliği problemine dönüştürülmesi ve hedeflere ulaşılabilmesi için tasarlanmış bir ön proje planına dönüştürülme işlemlerine odaklanmaktadır (Wirth ve Hipp, 1998: 33). Problemin çıkış sebebi, çözüm önerileri, problemle ilgili veri kaynakları ve verinin akışı gibi hususlar belirlenmelidir (Şeker, 2018: 12).

CRISP-DM rehberinin birinci adımı Şekil 2.5'te gösterildiği gibi “Problemin Tanımlanması”, diğer bir ifade ile “İş Anlama” adıımıdır. Proje amaç ve ihtiyaçlarının iş perspektifi açısından anlaşılması ve veri madenciliği problem tanımına dönüştürülmesi bu basamakta yapılmaktadır. İş hedeflerinin belirlenmesi, mevcut durumun değerlendirilmesi, veri madenciliği hedeflerinin belirlenmesi ve proje planının üretilmesi bu basamağın alt adımlarıdır (Cios vd., 2007: 13).



Şekil 2.5 CRISP-DM Birinci Basamak: İş Anlama

i. İş Hedeflerinin Belirlenmesi: İş perspektifi bakımından müşterinin başarmak istediği konu tam olarak anlaşılmalıdır. Çoğu zaman müşterinin uygun şekilde ele alınması gereken birçok hedefi ve kısıtlaması vardır. Projenin sonucunu etkileyebilecek önemli faktörler ortaya çıkarılmalıdır. Bu görevi atlamanın muhtemel sonucu, yanlış sorulara yanıtlar aramak için büyük çaba harcamak olmaktadır. Bu aşamada temel seviyede de olsa çalışmaya ait bir bütçe belirlenmelidir. İş hedeflerinin belirlenmesi görevine ait altyapının incelenmesi, işletme hedeflerinin tespiti ve başarı kriterlerinin belirlenmesi çıktıları gerçekleştirilmektedir (Olson ve Delen, 2008: 11).

Altyapı: Altyapı çıktısı altında projenin başlangıcında kuruluşun iş durumu hakkında bilinen bilgiler kaydedilmektedir (Chapman vd., 2000: 16). Proje parametreleri, kaynaklar ve sınırlılıklar en doğru biçimde tespit edilmelidir. Amaç ve gereksinimler doğru anlaşıldıktan sonra, veri madenciliği araçları ile en iyi çözümün nasıl sağlanacağına dair çalışmalar gerçekleştirilmektedir (Gökmen, 2014: 42).

İşletme hedefleri: İşletme hedefleri çıktısı altında, işletmenin birincil amacı iş perspektifi açısından tanımlanmaktadır. Temel iş hedefine ek olarak, işletmenin yanıtlamak isteyeceği diğer ilgili iş soruları da dahil edilmektedir. Örneğin, “Sağlık turistinin rezervasyon verileri incelenerek randevusunu iptal etmesi tahmin edilebilir mi?” sorusuna cevap aramak işletmenin birincil iş hedefidir. “Sağlık turistinin başvuru yaptığı tedavi türünün rezervasyon iptal tahminine anlamlı bir katkısı var mı?” sorusu da örnek olarak verilebilir.

Başarı kriterleri: Projenin başarılı veya yararlı olma kriterlerinin iş açısından açıklandığı çıktıdır. Spesifik bir çıktı olabileceği gibi “rezervasyon iptallerinin belirli bir düzeye indirilmesi” şeklinde objektif veya “rezervasyon iptal tahminleriyle ilgili faydalı bilgiler verilmesi” şeklinde subjektif değerler de olabilmektedir. İkinci durumda subjektif yargıyı kimin yaptığı belirtilmelidir (Chapman vd., 2000: 17).

ii. Mevcut Durumun Değerlendirilmesi: Proje planının belirlenmesi ve veri analizi hedefinde dikkate alınması gereken varsayımlar, kısıtlamalar, kaynaklar ve diğer faktörler konusunda ayrıntılı bilgi toplama faaliyetlerini kapsamaktadır. Önceki görevde durumun özüne hızlıca ulaşılmasına karşın, bu görevde ayrıntılar detaylandırılmaktadır. Mevcut durumun değerlendirilmesi görevine ait beş adet çıktı bulunmaktadır.

Kaynak envanteri: Projede yer alan personel (veri ve iş uzmanları, veri madenciliği ve teknik destek personeli), veriler (canlı, depolanmış veya operasyonel verilere erişim), bilgi işlem kaynakları (donanım ve yazılım) vb. mevcut kaynakların listelendiği çıktıdır.

Gereksinimler, varsayımlar ve kısıtlamalar: Tamamlanma takvimi, anlaşılabilirlik, sonuçların kalitesi, güvenlik ve yasal konular da dahil olmak üzere projenin tüm gerekliliklerinin listelendiği çıktıdır. Bu çıktının bir parçası olarak verilerin kullanım izinin verilmiş ve proje varsayımların listelenmiş olması gerekmektedir. Bu varsayımlar, veri madenciliği sırasında kontrol edilebilecek verilere ilişkin varsayımlar olabileceği gibi, projenin dayandığı işe ilişkin kontrol edilemeyen varsayımları da içerebilmektedir. Kaynakların kullanılabilirliğine ilişkin kısıtlamalar veya modelleme için kullanılması pratik olan veri boyutu gibi teknolojik kısıtlamaları da içeren proje kısıtlamaları listelenmektedir (Chapman vd., 2000: 17).

Riskler ve beklenmedik durumlar: Projeyi geciktirecek, başarısızlığa neden olabilecek riskler veya olaylar bu çıktıda listelenmektedir. Risklerin gerçekleşmesi durumunda ne gibi önlemlerin alınacağı ve acil durum planları da bu çıktıda belirtilmektedir.

Terminoloji: Projede, mevcut iş anlayışının bir parçasını oluşturan yararlı bir bilgi ve eğitim uygulaması elde etmek için, iş terminolojisi sözlüğü ile söz konusu iş sorunuyla ilgili projeyi geciktirecek veya başarısızlığa neden olabilecek riskleri örneklerle gösteren veri madenciliği terminolojisi sözlüğünü içeren bir terminoloji sözlüğü hazırlanmaktadır.

Maliyetler ve faydalar: Proje maliyetlerini, projenin başarılı olması durumunda işletmeye sağlayacağı potansiyel faydayla mümkün olduğunca spesifik karşılaştıran bir maliyet ve fayda analizi oluşturulmaktadır.

iii. Veri Madenciliği Hedeflerinin Tespiti: Veri madenciliği hedefi, proje hedeflerinin teknik terimler kullanılarak belirtilmesidir (Chapman vd., 2000: 18). Örneğin, iş hedefi "sağlık turisti rezervasyon iptallerinin azaltılması" iken, veri madenciliği hedefi "sağlık turistinin rezervasyon bilgilerini (ülke, cinsiyet, tedavi, tarih, otel ve oda tercihi vb.) göz önünde bulundurarak rezervasyon iptal etme durumunu tahmin etmek" olmaktadır. Veri madenciliği hedeflerinin tespiti görevinde, veri madenciliği başarı kriterleri ve veri madenciliği hedef çıktıları bulunmaktadır. Veri madenciliği hedeflerinin belirlenmesi aynı

zamanda örgütsel açıdan problemler ve çözüm önerileri üzerine de bir değerlendirme yapılmasını gerekli kılmaktadır (Laurinen, 2006: 20-21).

Veri madenciliği hedefleri: İş hedeflerine ulaşılmasını sağlayan projenin amaçlanan çıktıları tanımlanmaktadır.

Veri madenciliği başarı kriterleri: Projenin başarılı bir sonuca ulaşması için gerekli kriterler teknik terimlerle tanımlanmaktadır. Belirli bir düzeyde “rezervasyon iptal tahmin doğruluğu”, başarı kriterine örnek olarak verilebilir. İş başarısı kriterlerinde olduğu gibi, veri madenciliği başarı kriterlerinin de subjektif terimlerle tanımlanması gerekli olabilmektedir. Bu durumda subjektif değerlendirmeyi yapan kişi veya kişilerin kimliğinin belirlenmesi gerekmektedir.

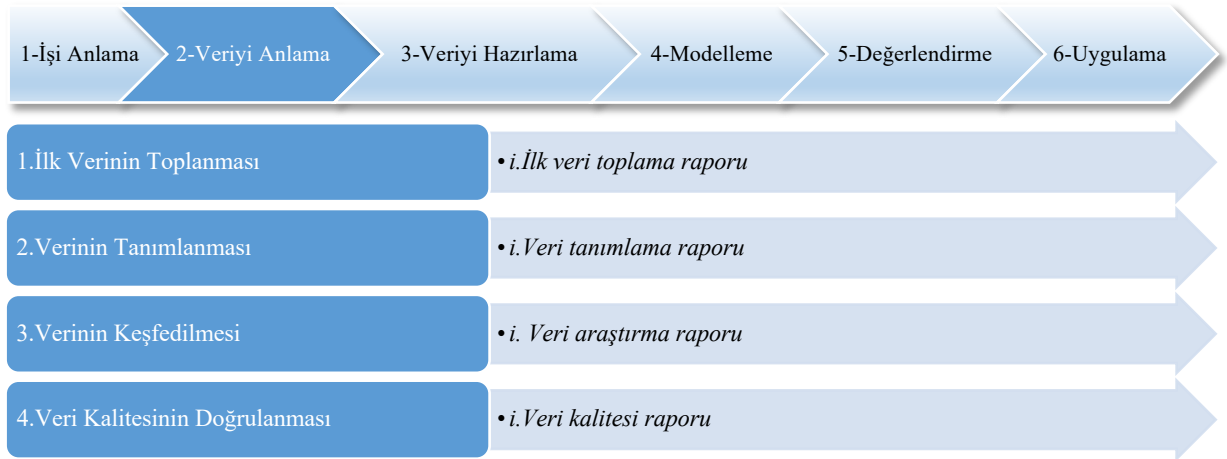
iv. Proje Planının Üretilmesi: Veri madenciliği ve iş hedeflerine ulaşmak için proje planının üretilmesi gerekmektedir. Proje planında, araçların ve tekniklerin seçimi de dahil olmak üzere, projenin geri kalanı boyunca gerçekleştirilmesi beklenen adımları belirtilmektedir. Proje planı, araç ve tekniklerin ilk değerlendirme çıktıları yer almaktadır.

Proje planı: Projede yürütülecek olan aşamaların süreleri, projede gerekli olan kaynaklar, girdiler, çıktılar ve bağımlılıklarla birlikte her aşama için ayrıntılı olarak listelenmektedir. Proje planında, modelleme ve değerlendirme aşamaları gibi büyük ölçekli yinelemeler mümkün olduğunca açıkça belirtilmektedir. Proje planında zaman çizelgesi ile riskler arasındaki bağımlılıkların analiz edilmesi de önemlidir. Bu analizlerin sonuçları, risklerin ortaya çıkması halinde eylem ve önerilerle birlikte proje planında açıkça işaretlenmektedir. Proje planı, her aşamanın sonunda ilerleme ve başarıların gözden geçirildiği ve ilerleme durumuna göre proje planının güncellendiği dinamik bir belgedir.

Araç ve tekniklerin ilk değerlendirmesi: Problemin tanımlanması (işin anlaşılması) aşamasının sonunda, projede yer alan araç ve tekniklerin ilk değerlendirmesi gerçekleştirilmektedir. Araç ve tekniklerin seçimi tüm proje gidişatını etkileyebileceğinden, bunların sürecin başında ön değerlendirilmesi önemlidir (Chapman vd., 2000: 19).

2.4.2. Veriyi Anlama

Veriyi anlama adımı veri gereksinimleri ve veriyi anlama göz önünde bulundurulmaktadır (Olson ve Delen, 2008: 9) CRISP-DM’in ikinci adımı olan “Veriyi Anlama” aşaması Şekil 2.6’da gösterildiği gibi ilk verinin toplanması, verinin tanımlanması, verinin keşfedilmesi ve veri kalitesinin doğrulanması görev ve çıktıları içerir (Kumdereli, 2012: 61).



Şekil 2.6 CRISP-DM İkinci Basamak: Veriyi Anlama

Veri anlama süreci, konu ile ilgili verinin tanınması ve toplanması ile başlamakta, veri kalitesi sorunlarının belirlenmesi ve varsa veri toplama sırasında çıkan farklı tespitlerin de değerlendirilerek, hipotezler oluşturulmasını konu almaktadır (Deshpande vd., 2010). Veri anlama aşaması, verilere aşina olmak, veri kalitesi sorunlarını belirlemek, verilere ilişkin öngörülerde bulunmak veya ortaya çıkmamış bilgiler için hipotezler kurmak üzere farklı alt kümeleri tespit etmek için yapılan faaliyetlerle devam etmektedir (Chapman vd., 2000: 14). Veriyi anlama aşamasının başarısı, birinci adımın başarısına bağlıdır. Aksi takdirde elde edilen veriler gereksiz yere işlenmiş olacaktır. Verideki gürültü ve eksik veri gibi sorunların tespiti gerçekleştirildikten sonra veri tanımlanmaktadır (Şeker, 2018: 12). Diğer aşamalara geçmeden önce verideki problemlerin tespit edilmesi, sorunların çözümlenmesi ve veri kalitesinin ölçülmesi gerekmektedir. İş anlama ve veriyi anlama birbirini içine geçmiş alt süreçlerdir. İş anladıkça farklı verilere ihtiyaç duyulabileceği gibi, verilere baktıkça işe farklı bakış açıları kazandırılmaktadır. Bu sürece devam edildikçe çalışmada kullanılacak veriler de netlik kazanmaktadır (Yurdakul, 2015: 24). Probleme uygun veri setindeki değişkenler, bu değişkenlerin açıklamaları ve veri tipleri açıkça ifade edilerek verinin anlaşılması sağlanmaktadır (Balaban ve Kartal, 2015: 39).

i. İlk Verinin Toplanması: Proje kaynağında listelenen verilere erişerek veriyi elde etme görevidir. İlk veri toplama, veriyi anlamak için özel araçlara veriyi yükleme işlemidir. İlk verinin toplanması görevinde ilk veri toplama raporu çıktısı bulunmaktadır (Chapman vd., 2000: 20).

İlk veri toplama raporu: Elde edilen veri seti veya setlerinin proje içindeki konumlarını elde etmek için kullanılan yöntemlerin ve karşılaşılan sorunların listelendiği çıktıdır. Bu proje ve benzer projelerin gelecekte karşılaşılan benzer sorunlara yardımcı olabilmesi için sorunlar ve elde edilen çözümler kaydedilmektedir.

ii. Verinin Tanımlanması: Elde edilen veri, yüzeysel veya tüm özellikleri ile incelenmekte ve sonuçlar raporlanmaktadır. Verinin tanımlanması görevinde, verinin tanımlama raporu çıktısı bulunmaktadır.

Veri tanımlanma raporu: Verinin formatı, tablo sayısı, tablodaki alan ve kayıt sayısı ve bu alanların özellikleri, keşfedilen verilerin diğer özelliklerinin tanımlandığı veri toplama raporu içerisinde yer almaktadır.

iii. Verinin Keşfedilmesi: Sorgulama, görselleştirme ve raporlama kullanarak önemli özniteliklerin dağılımı ve aralarındaki ilişkiler, önemli alt veri setlerinin özellikleri ve istatistiksel incelemeleri, bir tahmin görevinin hedef niteliği vb. çözülebilecek veri madenciliği soruları ele alınmaktadır. Bu analizler doğrudan veri madenciliği hedeflerine hitap edebileceği gibi veri tanımına, kalite raporlarına, daha ileri analiz için gerekli dönüşüm ve diğer veri hazırlıklarına katkıda bulunabilmektedir (Chapman vd., 2000: 21). Verinin keşfedilmesi görevinde veri araştırma raporu çıktısı bulunmaktadır.

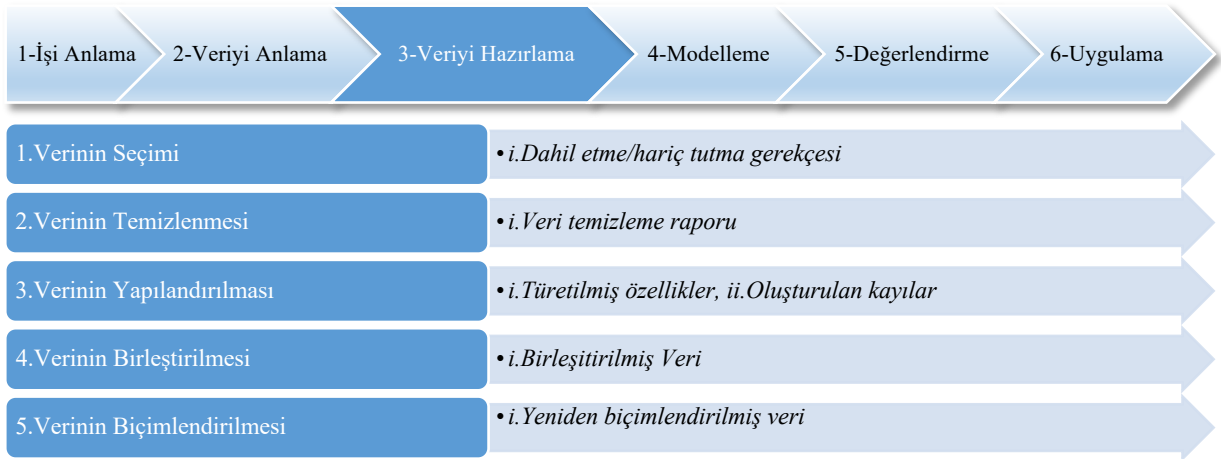
Veri araştırma raporu: Başlangıç hipotezi, ilk bulgular, veri inceleme sonuçları ve bunların projenin geri kalanı üzerindeki etkilerinin açıklandığı, veri özelliklerini gösteren ve daha ileri inceleme için ilginç veri alt kümelerini içeren, uygun grafikleri ve veri karakteristiklerini yansıtacak çizimleri içeren “Veri Araştırma Raporu” hazırlanmaktadır.

iv. Veri Kalitesini Doğrulama: “Veriler gerekli tüm durumları kapsıyor mu? Veriler doğru mu? Veride hatalar var mı? Hatalar ne kadar yaygın? Verilerde eksik değerler var mı? Eksik veriler nasıl temsil ediliyorlar? Eksik veriler nerede ortaya çıkıyor? Eksik veriler ne kadar yaygın?” gibi soruları yanıtlayarak verilerin kalitesi incelenmekte ve doğrulanmaktadır. Veri kalitesi doğrulama çalışmalarında veri kalitesi raporu çıktısı bulunmaktadır.

Veri kalitesi raporu: Veri kalitesi problemleri hem veri, hem de iş bilgisinden kaynaklanmaktadır. Veri Kalitesi Raporu, veri kalitesinin doğrulanma çalışmaları ve sonuçlarının listelendiği, kalite sorunları ve muhtemel çözümlerin belirtildiği rapordur (Chapman vd., 2000: 22).

2.4.3. Veriyi Hazırlama

Veri seti üzerinde gerçekleştirilecek işlemlere ve kullanılacak yöntemlere karar verilen aşamadır (Şeker, 2018: 42). Veriyi hazırlama aşamasında Şekil 2.7’de gösterildiği gibi veri seçimi, veri temizleme, veri yapılandırma, veri birleştirme, veri biçimlendirme görev ve çalışmaları gerçekleştirilmektedir.



Şekil 2.7 CRISP-DM Üçüncü Basamak: Veriyi Hazırlama

Bu adımda başlangıç verileri, çalışmalara temel oluşturacak final verilere dönüştürülmektedir. Veriyi dönüştürme adımları sabit ve belirli bir sıra veya tekrar sayısı takip etmeden birçok kez tekrarlanan işlemleri ifade etmektedir (Jackson, 2002: 281). Tahmin modelinin tasarlanması aşamasında tespit edilen sorunlardan dolayı, bu aşamaya sık sık geri dönülmekte ve yeniden düzenlenmektedir (Yurdakul, 2015: 26). Veri madenciliği uygulamalarında, kaynakların %80 civarında verilerin temizlenmesi ve ön işlenmesi esnasında harcandığı yaygın olarak kabul edilmektedir (Piramuthu, 2004: 483). Veri bilimcilerle 2016'da yapılan bir ankete göre, ana görevler için harcanan zaman olarak veri setlerini toplamak %19, veriyi temizleme ve düzenleme %60, eğitimde kullanılacak veri kümeleri inşa etmek %3, veriden örüntü aramak %9, algoritmaları iyileştirmek %4, diğer görevleri yürütmek %5 olarak belirtilmiştir (Kelleher ve Tierney, 2018: 65).

i. Verinin Seçimi: Analiz için kullanılacak verilere karar verilmektedir. Veri madenciliği hedeflerine uygunluğu, veri kalitesi, veri hacmi veya veri türleri gibi teknik kısıtlamaları içermektedir. Veri seçimi, bir veri tabanı tablolarında yer alan değişkenleri ve kayıtları kapsamaktadır. Verinin seçimi görevinde dahil etme ve hariç tutma gerekçesi çıktısı bulunmaktadır.

Dahil etme ve hariç tutma gerekçesi: Dahil edilecek ve hariç tutulacak verilerin, nedenleri ile listelendiği çıktıdır (Chapman vd., 2000: 24).

ii. Verinin Temizlenmesi: Veri kalitesinin, planlanan analiz tekniklerinin gerektirdiği veri kalitesi düzeyine yükseltilmesi, verinin temizlenmesi görevinde yapılmaktadır. Verinin temizlenmesi görevinde, veriye ait alt veri setlerinin seçilmesi, eksik verilerin doldurulması, uygun varsayılan değerlerin atanması gibi çalışmalar yer almaktadır. Veri temizleme, verilerdeki gürültü ve tutarsızlıkların giderilmesi için uygulanmaktadır. Verinin temizlenmesi görevinde veri temizleme rapor çıktısı bulunmaktadır.

Veri temizleme raporu: Veri kalitesi sorunlarını çözmek için hangi kararların alındığı ve hangi eylemlerin gerçekleştiğini açıklayan rapordur. Verilerin temizlik amacıyla dönüştürülmesi ve analiz sonuçlarına olası etkisi dikkate alınmaktadır.

iii. Verinin Yapılandırılması: Yeni değişkenlerin üretilmesi, tamamen yeni kayıtlar elde edilmesi veya mevcut değişkenlerin dönüştürülmesi gibi veri yapılandırıcı faaliyetleri içermektedir. Verinin yapılandırılması görevinde türetilmiş özellikler ve oluşturulan kayıtların çıktıları bulunmaktadır.

Türetilmiş özellikler: Aynı kayıta yer alan bir veya daha fazla mevcut özellikten hesaplanarak oluşturulan yeni özellikleri içermektedir. Türetilmiş özelliğe örnek olarak “toplam_ucret = gunluk_ucret * gece_sayısı” verilebilir.

Oluşturulan kayıtlar: Tamamen yeni kayıtların oluşturulmasını ifade etmektedir (Chapman vd., 2000: 24). Örnek olarak yıl içinde tedavi olmamış bir hastanın ham veride kaydı bulunmaz. Analiz yapabilmek adına veri tabanındaki bu hasta kaydı için hiç tedavi olmadığını ifade eden bir kayıt eklenmesi, oluşturulan kayıtlara örnek olarak verilebilir.

iv. Verinin Birleştirilmesi: Farklı veri kaynaklarından toplanan verilerde uyumsuzluklar sıklıkla görülmektedir. Uyumsuzlukların ortadan kaldırılması, veri birleştirilmesi ve verinin temizlenmesi işlemleri gerekmektedir (Gürsoy, 2009: 33). Verinin birleştirilmesinde, birden fazla veri kaynağından gelen veriyi kullanarak, yeni kayıt oluşturulmasını sağlayan yöntemler kullanılmaktadır. Verinin birleştirilmesi görevinde ise birleştirilmiş veri çıktısı bulunmaktadır.

Birleştirilmiş veri: Aynı nesneye ait farklı tablolarda bulunan verilerin bir araya toplanmasıyla birleştirilmiş veri oluşturulmaktadır. Bir rezervasyona ait kullanıcı bilgileriyle (cinsiyet, yaş, adı, soyadı vb.) tedavi bilgileri (uzmanlık, tedavi süresi, tedavi ücreti vb.) ayrı tablolarda yer almaktadır. Bu verilerin birleştirilmesi örnek olarak verilebilir. Birleştirilmiş veriler birden fazla kayıt ve tablodan alınan bilgilerin bir araya toplanmasıyla yeni değerlerin hesaplandığı işlemleri de kapsamaktadır. Örneğin, tüm tedaviler için bir kaydın bulunduğu tedaviler tablosunda, tedavi sayısı, ortalama tedavi tutarı gibi her hasta için bir kayıt bulunan yeni bir tabloya dönüştürülmesi vb. toplama faaliyetlerini de içermektedir.

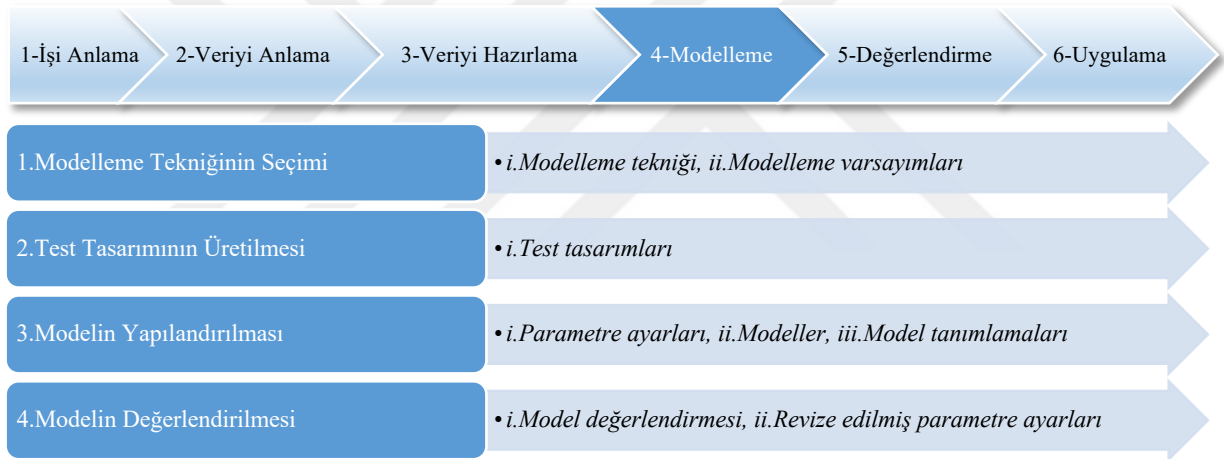
v. Verinin Biçimlendirilmesi: Verinin biçimlendirilmesi, verilerde yapılan ve verinin anlamını değiştirmeyen ancak modelleme aracının gerektirdiği söz dizimsel değişiklikleri ifade etmektedir. Verinin biçimlendirilmesi görevinde yeniden biçimlendirilmiş veri çıktısı bulunmaktadır.

Yeniden biçimlendirilmiş veri: Veri setindeki kayıtların sırasının değiştirilmesi bazı araçlar veya modeller için önemli olabilmektedir. İlk alanın her kayıt için benzersiz bir

tanımlayıcı olması veya son alanın modelin tahmin edeceği sonuç alanı olması gibi değişiklikler yeniden biçimlendirilmiş veriye örnek olarak verilmektedir. YSA gibi bazı modelleme araçlarında veri kümesindeki kayıtların sırasının değiştirilmesi, kayıtların sonuç niteliğinin değerine göre sıralanması veya kayıtların rastgele bir sırada sunulması önemli olabilmektedir. Ek olarak, belirli modelleme aracının gereksinimlerini karşılamak için yapılan tamamen söz dizimsel değişiklikler (virgülle ayrılmış veri dosyalarındaki metin alanlarından virgüllerin kaldırılması, tüm değerlerin maksimum belirli bir karaktere kısaltılması vb.) bulunmaktadır (Chapman vd., 2000: 25).

2.4.4. Modelleme

CRISP-DM'in dördüncü adımı olan "Modelleme" aşamasında Şekil 2.8'de gösterildiği gibi modelleme tekniğinin seçilmesi, test tasarımının üretilmesi, modelin yapılandırılması ve modelin değerlendirilmesi görevleri yer almaktadır (Chapman vd., 2000: 27).



Şekil 2.8 CRISP-DM Dördüncü Basamak: Modelleme Görev ve Çıktıları

i. Modelleme Tekniğinin Seçimi: Sağlık turizmi alanında yapılan rezervasyonların iptal edilme olasılığının hesaplanması ve sınıflandırılması tahminleri için çeşitli yöntemler bulunmaktadır. Modelleme aşamasının birinci adımı, kullanılacak modelleme tekniğinin seçilmesi görevidir. Bu görev sonunda iki adet çıktı elde edilmektedir.

Modelleme tekniği: Kullanılacak nihai modelleme tekniklerinin belirlendiği çıktıdır. İşi anlama aşamasının "Proje Planının Üretilmesi" görevinde bir araç belirlenmiş olmasına rağmen bu aşamada modelleme işlemlerinde kullanılacak belirli modelleme tekniklerine karar verilmesi gerekmektedir. Sınıflandırma tahmin modellerine ait bilgi üçüncü bölümün altında kapsamlı bir şekilde verilmiştir.

Modelleme varsayımları: Birçok modelleme tekniği, veriler üzerinde belirli varsayımlarda bulunmaktadır. Örneğin, tüm değişkenlerin tekdüze dağılımlara sahip olduğu,

eksik değerlere izin verilmediği, sınıf niteliğinin sembolik olması gerektiği vb. yapılan bu tür varsayımların kaydedildiği çıktıdır.

ii. Test Tasarımının Üretilmesi: Bir modelin kullanımından önce, model kalitesini ve geçerliliğini değerlendirmek için bir prosedür oluşturulmaktadır. Sınıflandırma modelleri için kalite ölçütü olarak hata oranlarının kullanılması yaygın olup genellikle veri, test ve eğitim seti olarak ikiye ayrılmaktadır. Modeli eğitim veri seti üzerinde oluşturduktan sonra, model kalitesi test seti üzerinde tahmin edilmektedir. Test tasarımının üretilmesi görevinde test tasarımları çıktısı bulunmaktadır.

Test tasarımları: Modellerin eğitimi, test edilmesi ve değerlendirilmesi için amaçlanan planı açıklayan çıktılarıdır. Planın birincil bileşeni; mevcut veri kümesinin, eğitim verileri, test verileri ve doğrulama veri kümelerine nasıl bölüneceğine karar vermektir.

iii. Modelin Yapılandırılması: Veri seti üzerinde modeller oluşturma ve çalıştırma görevidir. Modelin yapılandırılması görevinde parametre ayarları ve model tanım çıktıları yer almaktadır.

Parametre ayarları: Bütün modelleme araçlarında ayarlanabilecek çok sayıda parametre bulunmaktadır. Parametre ve seçilen değerlerin gerekçeleriyle birlikte listelendiği çıktıdır.

Modeller: Bu modeller bir rapor olmayıp, modelleme aracı tarafından üretilen gerçek modelleri ifade etmektedir.

Model tanımlamaları: Oluşturulan modellerle ve bunlarla ilgili karşılaşılan zorlukların birlikte açıklandığı ve yorumlandığı çıktıdır (Chapman vd., 2000: 28).

iv. Modelin Değerlendirilmesi: Veri madenciliği mühendisinin alan bilgisine göre oluşturulan modellerin, başarı kriterlerine bakımından değerlendirildiği görevdir. Bu görev yalnızca model değerlendirmesini amaçlamaktadır. Sonraki görev olan “Değerlendirme” aşamasında, oluşturulan modeller değerlendirilmeyip, proje süresince üretilen tüm diğer sonuçlar dikkate alınmaktadır. Veri madenciliği mühendisi, modelleme ve keşif tekniklerinin uygulanmasının başarısını daha teknik olarak değerlendirirken, veri madenciliği sonuçlarını iş bağlamında tartışmak için daha sonra iş analistleri ve alan uzmanlarıyla iletişime geçmektedir.

Bir teknik birden fazla uygulamakta veya teknikler birleştirilerek alternatif veri madenciliği sonuçları üretilmektedir. Modeller uygulamaya koyulmadan önce elde edilen tüm model sonuçları belirlenen kriterlere göre karşılaştırılmakta ve arzu edilen seviyeye ulaşma durumu değerlendirilmektedir (Larose, 2005: 7). Model değerlendirme kriterleri üçüncü bölümün altında detaylı bir şekilde ele alınmıştır. Modelin değerlendirilmesi görevi içerisinde, model değerlendirmesi ve revize edilmiş parametre ayarları çıktıları yer almaktadır.

Model değerlendirme: Modeli değerlendirme görev ve sonuç özetlerinin yer aldığı çıktıdır. Oluşturulan modellerin doğruluk, kesinlik, hatırlama, özgüllük, FPR, TPR, NPV, F1-değeri, hız vb. değerlendirme kriterleri açısından sıralandığı listeleri içermektedir.

Revize edilmiş parametre ayarları: Model değerlendirme sonuçları dikkate alınarak parametre ayarları düzenlenmektedir. Modelin değerlendirilmesi ve yapılandırılması görevleri, en iyi sonucun elde edildiğine emin olunana kadar sürdürülmektedir. Yapılan tüm revizyon ve çalışmalar bu aşamada raporlanmaktadır (Chapman vd., 2000: 29).

2.4.5. Değerlendirme

Modelin iş hedeflerini karşılama durumu değerlendirilmekte ve modelin kusurlu olmasına neden olan iş nedenleri belirlenmeye çalışılmaktadır. Zaman ve bütçe olanaklarının imkân verdiği ölçüde model gerçek uygulama üzerinde test edilmektedir. CRISP-DM'in beşinci adımı olan "Değerlendirme" aşamasında, Şekil 2.9'da gösterildiği gibi sonuçların değerlendirilmesi, gözden geçirme süreci ve gelecek aşamalarının belirlenmesi görevleri yer almaktadır (Chapman vd., 2000: 30).



Şekil 2.9 CRISP-DM Beşinci Basamak: Değerlendirme Görev ve Çıktıları

i. Sonuçların Değerlendirilmesi: Veri madenciliği çıktılarının yorumlanması ve özümsemesi, sonuçların değerlendirilmesi aşamasında gerçekleştiği için bu aşama kritik öneme sahiptir. Elde edilen bilgilerin iş değeri olarak nasıl algılanacağına ve sonuçların görselleştirilmesi için kullanılacak araçlara karar verilmesi gerekmektedir. Histogramlar, dağılım, pasta ve kutu grafikleri gibi çok sayıda görselleştirme aracı bulunmaktadır (Olson ve Delen, 2008: 18).

Veri madenciliği sonuçlarının iş başarısı kriterlerine göre değerlendirilmesi: Değerlendirme sonuçları, projenin başlangıçtaki iş hedeflerini karşılama durumuna ait nihai bir açıklama da dahil olmak üzere iş başarısı kriterleri açısından özetlenmektedir.

Onaylanan modeller: İş başarısı kriterlerine göre model değerlendirmesi yapıldıktan sonra, seçilen kriterleri karşılayan modellerin onaylı modeller haline geldiği çıktıdır.

ii. Gözden Geçirme Süreci: Gözden geçirme süreci, kalite güvence konuları açısından önemli faktör ve görevlerin gözden geçirildiği bir süreçtir. Elde edilen modellerin tatmin edici olup olmadığı ve iş ihtiyaçların karşılanıp karşılanmadığı ele alınmaktadır. Ayrıca, modelin doğruluğu ve kullanma izni verilen değişkenlerin kontrolü sağlanmaktadır.

Sürecin gözden geçirilmesi: Sürecin gözden geçirilmesi çıktısında, tekrarlanması gereken veya ihmal edilen görevler raporlanmaktadır.

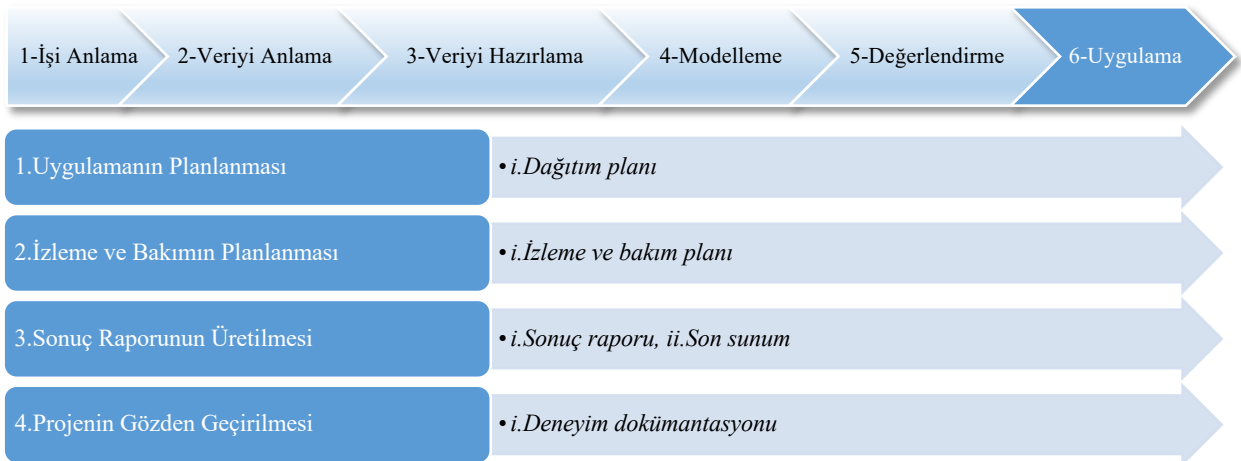
iii. Gelecek Aşamaların Belirlenmesi: Değerlendirme sonuçlarına ve süreç incelemesine göre projenin tamamlanma durumu, dağıtımına geçme veya daha fazla yineleme kararları ve yeni veri madenciliği projeleri hazırlama kararları alınmaktadır. Kararları etkileyen kalan kaynakları ve bütçe analizlerini içermektedir.

Olası eylemlerin listesi: Olumlu ve olumsuz tüm nedenleriyle beraber gerçekleşmesi olası faaliyetlerin, tüm seçenekler için listelendiği çıktıdır.

Karar: Kararlar ve nasıl gerçekleştirileceklerinin dayanakları ile birlikte tanımlandığı çıktıdır (Chapman vd., 2000: 31).

2.4.6. Uygulama

CRISP-DM'in altıncı adımı olan "Uygulama" aşamasında Şekil 2.10'da gösterildiği uygulamanın planlanması, izleme ve bakımın planlanması, sonuç raporunun üretilmesi ve projenin gözden geçirilmesi görevleri yer almaktadır.



Şekil 2.10 CRISP-DM Altıncı Basamak: Uygulama Görev ve Çıktıları

i. Uygulamanın Planlanması Uygulamanın planlanması görevinde değerlendirme sonuçları alma ve bunları işletmeye dağıtmak için bir dağıtım stratejisi oluşturulmaktadır. İlgili modelleri oluşturmak için genel bir prosedür tanımlanmışsa, bu prosedür daha sonra devreye alınmak üzere burada belgelenmektedir.

Dağıtım planı: Gerekli adımlar ve bunların nasıl gerçekleştirileceği de dahil olmak üzere dağıtım stratejisi bu çıktıda özetlenmektedir (Chapman vd., 2000: 32).

ii. İzleme ve Bakımın Planlanması: Veri madenciliği çıktılarının, işletmenin günlük faaliyetlerinde kullanılmaya başlanması ile izleme ve bakımın planlanması önemli bir görev haline gelmektedir. Bakım stratejisinin kapsamlı bir şekilde hazırlanması, veri madenciliği çıktılarının uzun süreler kullanılabilmesini sağlamakta ve hatalı kullanımları önlemeye yardımcı olmaktadır. Veri madenciliği sonuçlarının dağıtımını izlemek için projenin izleme sürecine ilişkin ayrıntılı bir plana ihtiyaç duyulmaktadır.

İzleme ve bakım planı: İzleme ve bakım stratejisindeki gerekli olan adımları ve bu adımların nasıl gerçekleştirileceğini özetleyen çıktıdır.

iii. Sonuç Raporunun Üretilmesi: Proje lideri ve ekibi, proje sonunda dağıtım planına bağlı olarak projenin ve deneyimlerinin özeti olan veya veri madenciliği sonuçlarının kapsamlı bir sunumu olan nihai bir sonuç raporu hazırlamaktadır.

Sonuç raporu: Veri madenciliği sürecindeki tüm aşamalarda yer alan çıktıların tümünün içerildiği ve sonuçlarının özetlendiği, kısaca, veri madenciliği sözleşmesinin nihai raporunun yer aldığı çıktıdır.

Son sunum: Projenin sonunda veri madenciliği projesinin sahibi olan müşteriye proje sonuçlarının sözel olarak sunulduğu bir toplantı bu çıktıda yapılmaktadır (Chapman vd., 2000: 33).

iv. Projenin Gözden Geçirilmesi

Projenin gözden geçirilmesi görevinde, projede iyi ve kötü giden noktaların, iyi yapılan ve iyileştirmesi gereken noktaların değerlendirilmesi yapılmaktadır.

Deneyim dokümantasyonu: Benzer durumlarda kullanılmak üzere, en uygun veri madenciliği tekniklerini seçmeye yönelik ipuçları, tuzaklar ve yanıltıcı yaklaşımlar bu çıktıda yer almaktadır (Chapman vd., 2000: 34).

ÜÇÜNCÜ BÖLÜM

VERİ MADENCİLİĞİ SINIFLANDIRMA YÖNTEMLERİ

Sınıflandırma, veri madenciliği çalışmalarında sıklıkla karşılaşılan uygulamalar arasında yer almaktadır. Sınıflandırma yöntemlerinde, veri tabanında kayıt altında tutulan veriler önceden belirlenmiş bir sınıf ya da kategori altında toplanmaktadır (Refaat, 2007: 7-8). Sınıflandırma, nesnelerin karşılıklı olarak birbirlerini dışta tutan sınıflara ayrılmasını, nesnelerin sadece bir sınıf altında yer alabileceğini öngören bir analiz biçimidir. Sınıflandırma temel olarak etiketlenmemiş bir veri örneğine ait sınıfın tahmin edilmesine yönelik analizi ifade etmektedir (Zaki, 2014: 467). Bir ürünü alan ya da almayan müşteriler, öğrencilerin başarılı ya da başarısız olarak sınıflandırmaları, bir bölgedeki yağış beklentisinin düşük, orta ve yüksek biçimde sınıflandırılması, belirli bir hastalığa yakalanma riskleri düşük, orta ve yüksek düzeyde olan bireyler sınıflandırmaya örnek olarak verilmektedir (Bramer, 2020: 21). Sınıflandırma analizi, makine öğrenme kuramları arasında denetimli öğrenme (supervised learning) yöntemleri arasında yer almaktadır. Veri madenciliği sürecinde sınıflandırma büyük miktarda veriden yararlı verilerin ortaya çıkarılması anlamında önemli işlevler görmektedir.

3.1. Veri Madenciliği Sınıflandırma Yöntemleri

Araştırmanın bu bölümünde veri madenciliği sınıflandırma işlemlerinde kullanılan popüler yöntemler ve sınıflandırma performanslarını karşılaştırmada kullanılan değerlendirme metrikleri açıklanmıştır.

3.1.1. ZeroR

Veri madenciliği sınıflandırma algoritmaları arasında en basit ve temel algoritmadır. ZeroR (Zero Rule), bir sınıflandırma probleminde veri setinde çoğunluğa sahip sınıfı göz önünde bulundurarak sınıflandırma yapmaktadır (Arvanitakis vd., 2019: 43). ZeroR çalışma mantığını örnek ile ifade etmek gerekirse; rezervasyon başvurusunda bulunmuş 100 kişiden, 70 kişinin “rezervasyonunu iptal etmeyen”, medeni halinin “evli” ve tedavi masraflarının “200.000TL ve üzeri” olduğu, geri kalan 30 kişinin ise medeni durumunun “bekar” ve tedavi masraflarının “200.000 TL’den az” olduğu kabul edilsin. Birinci grupta yer alan evli ve tedavi tutarı yüksek olanların, yapmış oldukları rezervasyonları iptal etmedikleri, diğer grubun ise rezervasyonları iptal ettikleri varsayılın. Yeni rezervasyon yapacak bir kişinin medeni durumu bekâr, tedavi tutarının ise 100.000 TL olduğu örnekte, bu rezervasyonun ikinci gruba, yani “rezervasyon iptal” şeklinde sınıflandırılmasını belirtmesine rağmen, ZeroR

sınıflandırması farklı olmaktadır. Toplam 100 kişilik veri setinde çoğunluğa sahip sınıfın, 65 kişi ile rezervasyonu iptal etmeyecek sınıf olduğu göz önünde bulundurulduğunda, ZeroR algoritması yeni gelen rezervasyonu, medeni hali ve tedavi tutarını dikkate alınmaksızın “rezervasyonunu iptal etmeyecek” kişi olarak değerlendirecektir. Bu durum ZeroR’ın sınıflandırma yaparken verilerin sahip olduğu özelliklere ve karmaşık çıkarımlara bakmadan baskın sınıfı dikkate alarak tahmin gücü bulunmadan sınıflandırma yaptığını göstermektedir (Lavesson vd., 2011: 296).

ZeroR algoritması çalışma prensibi göz önünde bulundurulduğu zaman ilk aşamada anlamsız gelmesine rağmen, ZeroR sınıflandırma başarımı, diğer algoritma performanslarını değerlendirmede referans olarak kabul edilmektedir (Cotha ve Sokolova, 2015: 42). Diğer sınıflandırma algoritmalarının, tahmin gücü bulunmayan ve basit bir mantık ile çalışan ZeroR algoritmasından daha iyi bir sınıflandırma başarısı elde etmeleri gerekmektedir (Hussain vd., 2013: 263). ZeroR, bir algoritmanın ne derece başarılı veya sıradan bir sınıflandırma başarısı ortaya koyabildiğini net bir şekilde göstermektedir (Murgia vd., 2018: 545).

3.1.2. Naive Bayes

Bayes Kuramı, İngiliz matematikçi ve rahip Thomas Bayes’in (1702-1761) felsefi öğelerin de yer aldığı “Şans Doktrindeki Bir Sorunu Çözmeye Yönelik Bir Deneme” isimli eserinde açıklanmış, Fransız matematikçi Pierre-Simon Laplace (1749-1827) tarafından çağdaş şekline kavuşturulmuştur. 1930’lı yıllarda öznel ve bilim dışı olarak damgalanan kuramın, II. Dünya Savaşı sırasında Alan Turing ve Jack Good’un Almanların meşhur şifre makinesi Enigma’yı çözmekte kullanması tekrar popüler hale gelmesini sağlamıştır (Akpınar, 2017: 237).

Naive Bayes (NB) sınıflandırıcısı, veri örneklerinin etiketlerine göre gruplandırılması gereken sınıfı belirlerken özellikler arasında bir bağımlılık olmadığı varsayımına dayandığı için "naive" (saf) olarak adlandırılır. Veri örneklerinin etiketlerine göre gruplandırılması gerektiği sınıflandırma problemlerinde olasılık hesaplaması yaptığı için Bayes teoremine dayanmaktadır. Bayes teoremi veri setinin belirli bir sınıfa ait olduğu varsayımına dayanmakta olup hipotezin doğru olma olasılığı hesaplanmaktadır (Bhatia, 2019: 110; Bramer, 2020: 22).

Sınıflandırma işlemini yapabilen Bayes teoremi, olasılık hesaplamada çok önemli bir konuma sahiptir (Özkan, 2016: 156). Matematiksel olarak koşullu olasılıklara dayanmaktadır. Belirli özniteliklere sahip yeni bir örnek verildiğinde, NB algoritması, verilen örneğin

özniteliklerini temel alarak her bir sınıfın ortaya çıkma olasılığını tahmin etmektedir (Zhou, 2020: 109).

NB sınıflandırma aşamasında sınıf değerlerine karar verirken, eldeki veri kümesindeki nitelik vektörünün her bir niteliğini, sınıf niteliğinde hangi değer olacağına bir kanıt olarak kabul etmekte ve sınıf değerinin hangi olasılıkla hangi değer olacağını, diğer bir ifadeyle koşullu olasılıkları hesaplamaktadır. Bu nedenle Bayes sınıflandırıcılarda sınıflara atanmış olasılık değerleri vardır (Köse, 2018: 248). Bayesci yöntemler tahmin için dikkate alınması gereken düşük etkiye sahip özellikleri de kullanmaktadır. Göreceli olarak küçük katkıları olan çok sayıda özelliğin bir arada kullanılmasının güçlü sınıflandırıcılar oluşturacağı varsayımına dayanmaktadır (Albon, 2018: 279).

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \quad (1)$$

Denklem (1)'de verilen Bayes teoremi için;

- $P(A|B)$: B olayının gerçekleşmesi durumunda A olayının gerçekleşme olasılığını,
- $P(B|A)$: B olayının gerçekleşmesi durumunda A olayının gerçekleşme olasılığını,
- $P(A)$: A olayına ilişkin ön olasılık değerini ve
- $P(B)$: B olayına ilişkin ön olasılık değerini

ifade etmektedir (Oğuzlar ve Kızılkaya, 2019: 45).

Bayes sınıflandırıcılar teşhis sorunlarında etkili sonuçlar üretebilmektedir. Sınıflandırma problemleri içerisinde gösterilen belirsizlik ve gürültü durumlarında oldukça etkili bir yapıya sahiptir. Basitlik ifadesine karşın çalışmalar karmaşık tümevarım algoritması ile rekabetçi olduğunu göstermektedir (John ve Langley, 1995: 338).

NB algoritmasının kullanımında dikkat edilmesi gereken durumlar bulunmaktadır. Birincisi, Naive Bayes sınıflandırıcı sağlıklı sonuçlar elde edebilmek için çok sayıda kayda ihtiyaç duymaktadır. İkincisi, eğitim verisi içinde, özellikle önem arz eden açıklayıcı sınıfa ait kayıt olmadığı durumda, bu sınıfa ait önceki verilerden yola çıkarak yeni bir kayıt karşısında hedef değişken üzerinde sıfır olasılığı olduğu varsayılacaktır (Shmueli vd., 2016: 182). Bayes sınıflandırıcısının olumlu ve olumsuz yanları bulunmaktadır (Köse, 2018: 246-250).

Olumlu özellikler: Naive Bayes sınıflandırıcısının avantajı; sadeliği, hesaplama etkinliği, sınıflandırmadaki iyi performansı ve kategorik değişkenleri doğrudan işleme yeteneğidir. Aslında, bağımsız değişkenlerin altında yatan varsayım doğruluktan uzak olsa bile, çoğunlukla daha karmaşık sınıflandırıcılardan daha iyi performans göstermektedir. Bu üstünlük, özellikle açıklayıcı sayısı çok fazla olduğunda belirgindir. Model oluşturulduktan sonra çalışma hızının yüksek olması gerçek zaman serilerinde kullanımı kolaylaştırmaktadır.

Veri seti çeşitliliğinin yüksek olduğu yani çoklu sınıflandırma olan veriler üzerinde etkilidir. Veri setindeki değişkenlerin birbiriyle olan bağımlılık düzeyi düşükse diğer yöntemlere göre daha başarılı sonuçlar üretmektedir. Sayısal değişkenlere göre kategorik olanlar daha başarılıdır. Basit bir tasarımı olmasına rağmen gelişmiş performans yapısı ile çok yaygın bir kullanıma sahiptir. Sınıflandırma algoritmalarını kullanan araştırmalarda Bayes sınıflandırıcılar performans bakımından YSA ve karar ağacı algoritması ile kıyaslanmaktadır. Bayes sınıflandırıcılar büyük veri ile çalışırken doğruluk ve yüksek hız göstermektedir (Han vd., 2012: 350). Naive Bayes sınıflandırıcısı, pratik uygulamalarda şaşırtıcı bir şekilde iyi performans sergilemektedir (Dangeti, 2017: 203).

Olumsuz Özellikleri: Naive Bayes algoritmasının bazı dezavantajları da bulunmaktadır. Veri örneklerinde bulunan özellikler arasındaki ilişkileri önemsemeyen yapısı nedeniyle algoritmanın performansı diğer algoritmalara göre düşük olabilmektedir. Veri örneklerinin etiketlerine göre gruplandırılması gereken sınıfın çok fazla çeşitlilik göstermesi durumunda algoritmanın doğruluk oranı düşebilmektedir. Veri setlerinde yer alan değişkenlerin birbiriyle olan bağımlılık düzeyleri gerçek hayatta çok yüksek olmasına rağmen Naive Bayes, değişkenlerin birbirinden bağımsız olduğu varsaymaktadır (Köse, 2018: 246-250). Olasılıklar göz önünde bulundurularak model oluşturulduğu için eğitim veri setinde yer almayan bir kategorik değer sıfır olasılıkla tanımlanmaktadır. Bayes sınıflandırıcılarda sıfır değer sorunu olarak anılan, olasılıklardan birinin sıfır olması durumunda, çarpımda bulunan diğer tüm olasılıkların etkilerini yok ettiği ve sonucun anlamını yitirdiği bir sorun söz konusudur. “Laplacian düzeltme” ya da “Laplace tahmini” olarak adlandırılan teknikten yararlanarak bu sorunun çözümüne gidilmektedir (Özkan, 2016: 163-164). Naive Bayes algoritması sürekli veriler ile çalışması için sürekli veriler kategorik veriye dönüştürülmelidir.

3.1.3. K-En Yakın Komşu

En yakın komşu algoritması (KNN – k Nearest Neighbour) 1951 yılında Hodges ve Fix tarafından parametrik olmayan yöntem olarak tanıtılmıştır (Silverman ve Jones, 1989: 233). KNN, makine öğrenmesi algoritmaları içerisinde en basit sınıflandırma yöntemlerinden birisidir (Akpınar, 2017: 274). Cover ve Hart (1967) tarafından geliştirilen KNN, büyük eğitim veri setlerinde yoğun kaynak gerektirmesi nedeniyle bilgisayarların hesaplama gücünde gelişme kaydedilene kadar yaygın bir kullanıma sahip olamamıştır. Günümüze kadar örüntü tanıma alanında yaygın olarak kullanılmaktadır (Han vd., 2012: 423). KNN, sınıflandırma yapılmamış örneği, kendisine en yakın karakteristiğe sahip sınıfa atamaktadır (Gaganis vd., 2007: 30). Yeni veri setlerini daha önce depolanmış eğitim veri setleri ile

kıyaslayarak en benzer kayıtlara göre sınıflandırma yapan örnek tabanlı bir öğrenme algoritmasıdır (Larose, 2005: 96). Algoritma etkinliği açısından sınıflandırmada aktif rol alan k değeri, komşuların sayısını ifade etmektedir (Akpınar, 2017: 274). KNN, test verilerinin hangi sınıfta yer alacağına karar verirken, önceden sınıflandırılmış verilerden k tanesinin yakınlığına bakarak karar vermektedir. K değeri, sınıflandırma modelinin test veri seti üzerinde ne kadar iyi genelleme yapabileceğinin en önemli göstergelerinden bir tanesidir (Filiz ve Öz, 2017: 122). Algoritmanın başarılı bir şekilde çalışması, k parametresine uygun bir değerin belirlenmesine bağlıdır (Balaban ve Kartal, 2015: 61). K değeri pozitif bir tamsayı olup, genellikle küçük ve tek sayı değerleri almaktadır. Algoritma çalışırken k değeri ilk başta seçilmektedir. Yüksek seçilen k değeri benzerlik bulunmayan verilerin bir arada toplanmasına, düşük seçilen k değeri ise aynı sınıfta yer alması gereken verilerin farklı sınıflara atanmasına yol açmaktadır (Khan vd., 2001: 3). İdeal k değeri iki uç nokta arasında bulunmaktadır (Bishop, 2006: 125). K değerinin tek sayı olarak atanmasında büyük fayda bulunmaktadır (Ledolter, 2013: 117). Örneğin, k değerinin çift değer olarak belirlendiği durumda ve en yakın komşularından yarısının A sınıfına, diğer yarısının ise B sınıfına dahil olması durumunda, KNN sınıflandırma yapamamaktadır (Lantz, 2013: 72).

Analiz aşamasında eğitim veri setinin modele önceden sunulmasından dolayı KNN yöntemi bellek temelli sınıflandırma (Memory-based Classification) yöntemi olarak da isimlendirilmektedir. KNN tekniğinde tüm örnekler test ve eğitim veri seti olmak üzere iki gruba ayrılmaktadır. Eğitim veri setindeki bir örneğin, test veri setindeki bir örneğe uzaklığı hesaplanmakta ve en kısa uzaklık, en yakın komşu olarak sınıflandırılmaktadır (Cunningham ve Delany, 2020: 1; Jabbar vd., 2013: 86; Khamis, vd., 2014: 122). NB yöntemine göre daha basit bir sınıflandırıcı olan KNN, sınıflandırmayı gerçekleştirme sürecinde ihtiyaç duyduğu tek unsur, gözlem çiftleri arasındaki mesafedir. En yakın mesafede olan veri setleri aynı sınıflarda yer almaktadır (Oğuzlar ve Kızılkaya, 2019: 47). Bu sebeple en basit makine öğrenme algoritması olarak kabul edilmektedir (Cemaloğlu ve Duykuluoğlu, 2020: 483-484).

KNN, eğitim verilerini öğrenme yerine ezberlemekte ve bundan dolayı tembel öğrenme algoritması olarak değerlendirilmektedir. Tembel öğrenme algoritmalarında eğitim aşaması daha hızlı tamamlanmasına rağmen test verilerinin sınıflandırıldığı tahmin aşaması daha uzun sürmektedir (Boschetti ve Massaron, 2016: 209-210). Tembel öğrenme algoritması diğer öğrenme algoritmaları gibi bir model oluşturmamaktadır. Eğitim verileri üzerinden bir öğrenme gerçekleşmediği ve bir parametre öğrenimi olmadığı için sınıflandırma algoritmasının veriyi nasıl sınıflandırdığının anlaşılması zorlaşmaktadır (Lantz, 2013: 75).

KNN, iki örnek arasında yer alan uzaklığı hesaplamak için bir uzaklık fonksiyonu kullanmaktadır (Murphy, 2012: 17). İki nesne arasındaki uzaklık değeri ne kadar küçük ise, iki nesne arasındaki benzerlik o ölçüde büyük olmaktadır (Akpınar, 2017: 177). Sık kullanılan uzaklık ölçme yaklaşımları Tablo 3.1’de verilmiştir.

Tablo 3.1 Örnekler Arası Uzaklık Tanımlama Yaklaşımları

Uzaklık Yöntemi	Formül
Minkowsky	$D(x, y) = \left(\sum_{i=1}^m x_i - y_i ^r \right)^{1/r}$
Manhattan	$D(x, y) = \sum_{i=1}^m x_i - y_i $
Chebychev	$D(x, y) = \max_{i=1}^m x_i - y_i $
Euclidean	$D(x, y) = \left(\sum_{i=1}^m x_i - y_i ^2 \right)^{1/2}$
Camberra	$D(x, y) = \sum_{i=1}^m \frac{ x_i - y_i }{ x_i + y_i }$
Kendall’s Rank Correlation	$D(x, y) = 1 - \frac{2}{m(m-1)} \sum_{i=j}^m \sum_{j=1}^{i-1} \text{sign}(x_i - x_j) \text{sign}(y_i - y_j)$

Kaynak: Kotsiantis, 2007: 259

KNN uzaklığı hesaplamak için, genel olarak öklid mesafesi kullanılmaktadır (Murphy, 2012: 17-18). Öklid mesafesi kuş bakışı olacak şekilde en kısa rota olarak da tanımlanmaktadır. Yüksek değerlere sahip özellikler, diğer özellik değerlerine göre uzaklık hesaplamasında ağır basmaktadır. Özelliklerin yeniden ölçeklendirilerek uzaklık hesaplamalarının daha dengeli ve etkin yapılması sağlanmaktadır. KNN öncesi, özellikler genellikle standart bir değer aralığına dönüştürülmektedir (Ledolter, 2013: 116). Yeniden ölçeklendirme işleminde kullanılan ve özelliklere ait tüm değerlerin 0-1 aralığına indirgendiği min-maks normalizasyon hesaplaması Denklem (2)’de verilmiştir (Berry ve Linoff, 2004: 275).

$$X_{yeni} = \frac{X - \min(X)}{\max(X) - \min(X)} \quad (2)$$

Yaygın olarak kullanılan diğer bir dönüştürme işlemi ise Denklem (3)’ de verilen z-skor standardizasyonudur (Lantz, 2013: 73).

Algoritma Adımları: KNN algoritması adımları, gözlem değerleri içeren bir küme için aşağıda verilmiştir (Özkan, 2016: 142):

- i. K parametresi tespiti.
- ii. Sınıflandırma yapılmak istenilen noktanın, en yakın komşuları arasında bulunan mesafenin ayrı ayrı hesaplanması.
- iii. Hesaplanmış mesafelere göre satırların sıralanması ve en küçük değere sahip olan k seçimi.
- iv. Seçili satırların sınıflarının tespiti ve en çok tekrarlanan kategorinin tespiti.
- v. Seçilen kategorinin, tahmin edilmek istenen değer sınıfı olarak kabul edilmesi.

Olumlu Yönleri: KNN algoritmasında hesaplamalar basit ve eğitim hızlıdır. Eğitim verisinin gürültülü olduğu durumlarda bile algoritma verimli ve dirençlidir (Balaban ve Kartal, 2015: 60). Veri setinde yer alan özelliklerin karmaşık ilişki içinde olduğu ve hedef sınıf sayısının fazla olduğu uygulamalarda KNN kullanılabilir (Lantz, 2013: 66). Sınıflandırma problemlerinde eğrinin fonksiyonunun belirlenmediği ve lineer ayrılmadığı sınıflarda başarılı sonuçlar elde etmektedir (Zacccone, 2016: 81).

Olumsuz Yönleri: Veri kümesinin büyük olduğu durumlarda algoritmanın performansı düşmektedir. Sınıflandırılmak istenilen grupların net bir ayrım göstermediği durumlarda KNN algoritması uygun olmamaktadır (Lantz, 2013: 66). KNN algoritmasının kullanılacağı durumlar belirli bir veri kümesine özel olarak değerlendirilmelidir. KNN algoritmasının, k komşu sayısı parametresine bağlı olması, büyük miktarda bellek gereksinimi ve ilgisiz nitelikler ile hataya kolay düşebilmesi, olumsuz yönleri olarak sayılmaktadır (Balaban ve Kartal, 2015: 60).

3.1.4. Destek Vektör Makineleri

Destek Vektör Makineleri (DVM) algoritmasının temeli Vapnik-Chervonenkis teorisine dayanmakta olup, Boser, Guyon ve Vapnik tarafından 1992 yılında geliştirilmiştir (Akpınar, 2017: 291). DVM, istatistiksel öğrenme teorisi bağlamında geliştirilen bir yöntemdir (Vercellis, 2008: 262). Bu sebeple kontrollü sınıflandırma algoritması olarak da tanımlanmaktadır. Başlangıçta doğrusal verilerin iki sınıflı gruplandırılması için kullanılırken zaman içerisinde çok sınıflı ve doğrusal olmayan verilerin sınıflandırılmasında kullanılmaya başlanmıştır (Cemaloğlu ve Duykuluoğlu, 2020: 475).

DVM, etikete sahip eğitim verisinden girdi ve çıktı haritalama çıktıları oluşturan denetimli öğrenme yöntemidir. DVM, genelleştirilmiş doğrusal modeller ailesine aittir. Tıbbi tanı koyma, biyoinformatik, yüz tanıma, görüntü işleme vb. çok sayıda çalışma alanında

tatmin edici performans sergilemektedir. DVM, karmaşık sistemlerin ve süreçlerin modellenmesinde iyi sonuçlar üretmektedir (Dangeti, 2017: 220; Huang vd., 2006: 209). DVM ile yapılan çeşitli araştırmalar, yöntemin diğer algoritmalar gibi performans sergilediğini ve çok güçlü olduğunu göstermiştir (Cristianini ve Taylor, 2000: 7).

DVM, sınıflandırma yapılmak istenilen verileri kategorilere ayırmak için çok boyutlu alanda aşırı düzlemler oluşturabilmek için haritalama yapmaktadır (Vajirachorn ve Chongwatpol, 2017: 81). Basit ve son derece etkili bir sınıflandırma yöntemi olan DVM ile doğrusal bir alanda bulunan iki grubu birbirinden ayırmak oldukça kolaydır. Harita üzerinde yer alan iki bölgenin en uygun şekilde ayrılabilmesi için sınır çizgisinin en doğru yerden çizilmesi gerekmektedir. İki grup arasında çizilen sınır, iki gruba da en uzak noktada yer almakta olup DVM söz konusu sınırın çizilmesinde kullanılmaktadır (Oğuzlar ve Kızılkaya, 2019: 49-50).

Destek vektör makineleri terminolojisinde iki bölge arasındaki boşluklar “sınıflar arasındaki marj” olarak isimlendirilmektedir. DVM sınıflandırma yaklaşımı olarak istatistiksel dağılımların tahmin edilmesi yerine maksimum marjı aramaktadır (Melgani ve Bruzzone, 2004). Veriler doğrusal bir şekilde ayrılabilir şekilde dağılmadıklarından onları düz bir çizgiyle ayırmak çok zordur. Bu kapsamda kullanılacak verileri sınıflandırmak için farklı bir yöntem kullanılarak, yeni değişkenler oluşturulmaktadır. Böylelikle sınıflar hiper bir düzlemde doğrusal olarak ayrılabilir (Williams, 2011: 293).

Girdi verilerinin doğrusal olarak sınıflandırmaya mümkün olmadığı durumlarda doğrusal, radyal tabanlı fonksiyon, sigmoid ve polinom çekirdek işlevleri kullanılarak dönüşümler gerçekleştirilmekte ve girdi verileri daha ayrılabilir duruma gelmektedir (Siering vd., 2018: 57; Vajirachorn ve Chongwatpol, 2017: 81).

Algoritma Adımları: Destek vektör makinalarına ait çalışma adımları aşağıda sıralanmıştır (Hazim, 2018: 43).

- i. Veri setinin yüklenmesi.
- ii. Ardışık işlem farklılıkları, işlem sınıf ve zamanına göre verilerin yeniden sıralanması.
- iii. Her işlemin iki alanın vektörü olduğu veri formuna dönüştürülmesi.
- iv. Negatif ve pozitif işlem grupları olmak üzere iki ayrı veri kümesi oluşturulması.
- v. Çekirdek fonksiyon seçimi.
- vi. Destek vektör makinasının eğitilmesi
- vii. Uygulama sonrası sınıflandırma performansının kaydedilmesi.
- viii. Mevcut işlem değerlendirmesi.

- ix. Uygulamanın geçerli işlem verileri için birinci adımdan üçüncü adıma kadar yeniden başlatılması. Kaydedilmiş sınıflandırıcı ve üretilen vektörün yer değiştirilmesi.
- x. Elde edilen kararın kabul edilmesi.

DVM olumlu ve olumsuz yönleri aşağıdaki gibidir.

Olumlu Yönleri: DVM kullanılarak karmaşık veri setlerinde bulunan ilişkiler ortaya çıkarılmakta ve çeşitliliği yüksek veri tipleri sınıflandırılabilir (Cemaloğlu ve Duykuluoğlu, 2020: 476-477). DVM'nin en önemli katkısı, sınıflandırma problemini kareli optimizasyona sahip probleme dönüştürmesidir. Böylelikle problem çözme aşamasında hesaplama karmaşıklığı ortadan kalkmakta ve daha hızlı sonuç elde edilmektedir (Osowski vd., 2004: 149). DVM, doğrusal olmayan sınıflandırmada yüksek başarıya ve güvenilirliğe, sahip bir yöntemdir (Akpınar, 2017: 291).

Olumsuz Yönleri: DVM'nin iki boyuta sahip sınıflandırma problemleri için tasarlanmış olması ve daha fazla sınıfa sahip üyeliklerin olasılıklarını modelleyememesi olumsuz yönüdür (Canty, 2014: 228). Karmaşık ve zaman gerektiren hesaplamalar içeren yapısından dolayı eğitim ve test sürecinde hız ve boyut sınırlaması bulunmaktadır. Veri kümesinin büyük ve veri örnekleri arasındaki benzerliklerin önemli ölçüde değişiklik göstermediği durumlarda DVM performansı düşmektedir. DVM, geniş ve kapsamlı görevlerde karmaşık bir algoritmik yaklaşıma ve geniş bellek gereksinimine ihtiyaç duymaktadır. DVM'nin kategorik veri işlenmesine yönelik sınırlamasının olması diğer olumsuz yönüdür. Tüm olumsuz yönlerine rağmen, kuvvetli teorik temele dayanması ve ürettiği çözümlerden dolayı veri madenciliğinde tercih edilen popüler yöntemler arasında yer almaktadır (Olson ve Delen, 2008: 122).

3.1.5. Lojistik Regresyon

Lojistik Regresyon (LR), bir hastalığın oluşma ihtimali karşısında uygulanacak tıbbi uygulamaların başlangıcı için 1944, 1953 ve 1955 yıllarında Berkson tarafından kullanılmıştır. Finney (1972), LR analizinin probit analizine alternatif olarak kullanılmasını önermiştir. Truett, Cornfield, Kannel (1967) ve Halperin, Blackwelder, Verter (1971) ise LR modelinin normallik varsayımının sağlanmadığı durumlar için diskriminant analizine alternatif olabileceğini öne sürmüştür. 1970'li yıllarda, lojistik regresyon kullanılarak üstel dağılım ailesinden poisson regresyonu ve genelleştirilmiş doğrusal modeller geliştirilmiştir (Nelder ve Wedderburn, 1972: 380).

Regresyon, bağımsız değişkenler arasında var olan fonksiyonel ilişkileri araştırmak için kavramsal açıdan basit bir analiz yöntemidir (Balaban ve Kartal, 2015: 78). Söz edilen

fonksiyonel ilişki, bağımlı değişken ile bir veya daha fazla belirleyici veya açıklayıcı değişkeni, bağlayıcı bir denklem veya bir model şeklinde ifade etmektir (Chatterjee ve Hadi, 2012: 1). Regresyon kendi içerisinde çoklu regresyon, doğrusal regresyon, probit, LR vb. çeşitlere sahiptir. Regresyon modeline, bağımlı değişkenin türüne göre karar verilmekte ve hatalı model seçimi çıktıları yanlış olmasına yol açmaktadır (Arı ve Önder, 2013: 168). Bağımlı ve bağımsız değişken arasındaki ilişkiyi inceleyen çoklu regresyon analizinde, değişkenlerin metrik özellikli olması gerekmektedir. Bağımlı değişkenin kategorik olması durumunda çoklu regresyon analizi uygun olmamaktadır (Altunışık vd., 2012: 249).

Kategorik değişkenler, bir regresyonda bağımsız değişkenler olarak kullanıldıklarında etkileri kukla değişkenler tarafından açıklanabilmesine rağmen bağımlı değişkenler olarak kullanıldıklarında özel bir değerlendirmeye ihtiyaç duyulmaktadır. Sürekli bağımlı değişkenler için tasarlanmış olan klasik regresyon modellerinde kategorik bağımlı değişkenler kullanıldıklarında, özellikle tahmin konusunda önemli sorunlar ortaya çıkmaktadır. Doğrusal bir regresyon modelini kategorik bağımlı değişkene uydurmaya çalışmak, tahminlerin bağımlı değişken için makul aralığın dışında kalmasına neden olmaktadır (Powers ve Xie, 1999: 37). LR analizi, bağımlı değişkenin iki kategorili veya ikiden fazla kategorili sıralanabilir veya sırasız olması durumlarında uygulanabilmektedir (Balaban ve Kartal, 2015: 79).

Finansal başarı veya başarısızlıkları değerlendirmede, gerçekleşme veya gerçekleşmeme durumlarında, evet veya hayır cevaplarını kapsayan bağımlı değişkenli anketler gibi pek çok alanda ikili LR sık kullanılan bir modeldir (Özdemir, 2004: 589; Budak ve Erpolat, 2012: 26). Günümüzde LR, bilimin tüm alanlarında yaygın olarak kullanılmaktadır (Önder ve Cebeci, 2002: 106).

Bağımsız ve bağımlı değişkenler arasındaki ilişkileri matematiksel olarak modelleyen denetimli öğrenme yöntemleri arasında yer almaktadır (Hosmer vd., 2013: 1). Logit modelinin katsayı tahmininde, tahmincisi tutarlı olan ve büyük örneklerde normal dağılım gösteren “en çok olabilirlik” yöntemi kullanılmaktadır (Stock ve Watson, 2007: 398). En çok olabilirlik yaklaşımında, katsayıların hesaplanması için iteratif işlemler ile katsayıların en olası değerleri aranmaktadır (Altunışık vd., 2012: 252). LR, odds oranının logaritması olan logit fonksiyonuna dayanmaktadır (Mert, 2016: 180).

LR denkleminin doğrusal olmamasından dolayı, bağımsız değişkenler ile bağımlı değişken arasındaki ilişkinin doğrudan yorumlaması zordur. Buna rağmen, lehine üstünlük (odds oranı) kullanılarak, dolaylı olarak yorumlanabilmektedir (Anderson vd., 2010: 688). Odds oranı, bir olayın meydana gelme olasılığının (P_i), meydana gelmeme olasılığına ($1 - P_i$) oranı ($P_i/(1 - P_i)$) olarak tanımlanmaktadır (Mertler ve Reinhart, 2017: 308). Odds

oranı, beta katsayılarının üstel değeri olup bağımsız değişkende meydana gelen bir birimlik değişimin bağımlı değişkeni nasıl etkilediğini ifade etmektedir. Odds oranı, logaritmik dönüşüm gerektirmeyen ve bağımsız değişkenin kategorik olması durumlarında dahi yorumlanması çok kolaydır (Field, 2018: 1489). Odds oranı, bağımsız değişkendeki bir birimlik artışın üstünlük üzerindeki etkisini ölçmektedir (Anderson vd., 2010: 688). Üstünlükte alt sınır olasılıkta olduğu gibi sıfır iken üst sınır söz konusu değildir. Üstünlüğün dört olması, bir olayın meydana gelme olasılığının, meydana gelmeme olasılığından dört kat daha fazla olması anlamına gelmektedir (Balaban ve Kartal, 2015: 80). LR modeli tek ve birden fazla bağımsız değişken içermesi durumunda sırasıyla Denklem (4) ve Denklem (5)'teki gibi ifade edilmektedir. LR modelindeki tahmin edilen olasılık değerleri ise Denklem (6)'daki gibi gösterilmektedir.

$$\text{Logit}(Y) = \ln \frac{P_i}{1 - P_i} = b_0 + b_i X_i + e_i \quad (3)$$

$$\text{Logit}(Y) = \ln \frac{P_i}{1 - P_i} = b_0 + b_1 X_1 + b_2 X_2 + \dots + b_p X_p + e_i \quad (4)$$

$$P = \frac{e^{b_0 + b_1 X_i}}{1 + e^{b_0 + b_1 X_i}} = \frac{1}{1 + e^{-(b_0 + b_1 X_i)}} \quad (5)$$

LR, nitel bağımlı değişkeni, bağımsız değişkenlere göre bir logit değişkenine dönüştürdükten sonra maksimum olabilirlik tahmini uygulamaktadır. Bu şekilde, LR belirli bir olayın gerçekleşme olasılığını tahmin etmektedir. Denklem (7), açıklayıcı değişkenlerin bir fonksiyonu olarak doğrusal olarak değişmektedir (Albon, 2018: 259).

$$\log(odds) = \log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n \quad (6)$$

β_0 : Sabit

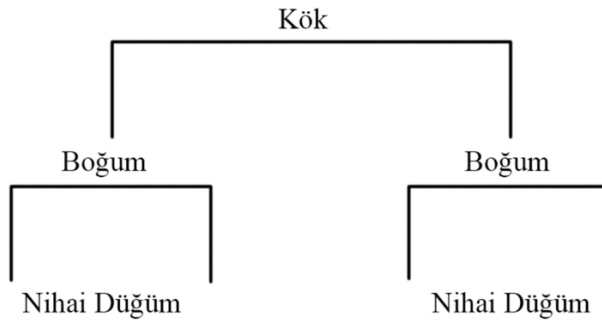
$\beta_{1,2,\dots,n}$: Değişkenlere ait regresyon katsayıları

3.1.6. Karar Ağaçları

Karar ağaçları (DT) kullanarak, mevcut veri kümesinden öğrenme yoluyla, karar verici model oluşturan algoritmalarla karar ağacı algoritmaları ismi verilmektedir (Köse, 2018: 207). DT “regresyon ağaçları” veya “sınıflandırma ağaçları” olarak da adlandırılmaktadır (Williams, 2011: 205). DT değişkenlerin seçim tekniği olarak da kabul edilmektedir (Ripley, 2002: 36). Makine öğrenmesi algoritmalarının ve veri madenciliğinin yapı taşları olan DT, 1980’lerde ilk geliştirildiklerinden günümüze kadar, makine öğrenmelerine dayalı veri madenciliği modellerinin en önemli unsurlarından biri durumuna gelmiştir. Yaygın

biçimde kullanılmalarının en temel nedeni olarak; karar ağaçlarında model kurulumunun kolay olması, model çıktılarının kolay yorumlanabilmesi, veri tabanı sistemi ile kolay entegrasyonu ve tatmin edici düzeyde güvenilirlik sağlamasıdır (Ersöz, 2019: 70). DT modelinin oluşturulmasında alan bilgisi gerekliliğinin olmaması, gizli verileri işleyebilme kabiliyeti ve adımların basit ve hızlı olması da tercih edilmesini olumlu etkileyen faktörlerdir (Patel vd., 2012: 8). DT tahmin ve sınıflandırma işlemleri için kullanılmaktadır (Refaat, 2007: 21). Birçok veri madenciliği yöntemi parametrik iken, karar ağaçları kural tabanlı bir yöntemdir (Zhou, 2020: 125). DT, çok sayıda vaka ve değişken içeren veri setlerine uygulanabilen, aykırı değerlere karşı son derece dayanıklı, yoğun hesaplama gerektiren ve parametrik olmayan etkili bir veri madenciliği tekniğidir (Steinberg ve Colla, 1995: 23-24).

DT oluşturulmadan önce verinin sınıflandırma için uygunluk analizi yapılmalıdır. DT modeli oluşturma sürecinde sınıflandırılacak düzeyde ayıklanmış ve iyi tanımlanmış veri olması, basit bir analiz ya da sorgulama yöntemi ile çözülemeyecek bir problemin olması gibi belirli varsayımların karşılanması gerekmektedir (Patel vd., 2012: 8).



Kaynak: Williams, 2011: 206

Şekil 3.1 Temel Karar Ağacı Yapısı

Temel bir karar ağacı örneği Şekil 3.1’de gösterilmiştir (Williams, 2011: 206). Bir karar ağacı kök veya kök boğumundaki bir merkez noktası ile başlamaktadır (Han vd., 2012: 330). Karar ağacının en son elemanı “yaprak”, en üst yapısı “kök” ve bu elemanlar arasında kalan yapılar ise “dal” olarak isimlendirilmektedir. Kök boğumundan başlayan dallarda karar boğumları yer alır. Karar boğumları kökten aşağıya doğru bir yapıda sonuç değişkenine doğru bir yol izlemektedir. Kök boğumlarındaki dallar sonuca ulaşmaya kadar budaklanmaktadır (Cemaloğlu ve Duykuluoğlu, 2020: 384). Karar ağaçlarında hedef değişkenine ilişkin nihai bir karara varılması öngörülmektedir. Bu bağlamda kökten dallarla ayrılan karar boğumları, nihai değişkene ulaşmaya kadar devam eder (Larose, 2005: 107). DT son aşamasında yer alan ve sınıf etiketlerini ortaya koyan boğumlara nihai boğum veya uç boğum adı verilmektedir (Han vd., 2012: 330). Karar ağacının son şekli Eğer/Sonra/Başka (If/Then/Else) kural yapısına çevrilmektedir (Bramer, 2020: 42; Refaat, 2007: 21).

DT oluřturma yntemlerinin hedefi homojenlik elde etmektir. Homojenlik saęlanana kadar veri seti yinelemeli olarak blnmektedir. DT modeli oluřturmak iin; inřa etme, budama ve iřleme adımları olmak zere  adım uygulanmaktadır. İnřa etme adımı, eęitim veri seti kullanılarak DT oluřturulmaktadır. Sonuca ulařıncaya kadar blme iřlemi devam etmektedir. Budama adımı, performansın dřmesine neden olan bazı dalların ve dęmlerin kaldırılmasını iermektedir. İřleme ise anlařılabilirlięi arttırmak iin budama iřleminden geen karar aęacının iřlenmesi ařamasıdır. DT algoritmalarının byk oęunluęu bu  adımı barındırmasına raęmen, bazı durumlarda ilk iki ařama tek bir srete ele alınmaktadır (Pujari, 2008: 160).

Tercih edilen algoritmaya gre DT řekil deęiřtirdięinden ve farklı sınıflandırma sonucu verdięinden dolayı, karar aęacı oluřturulurken kullanılacak algoritma tercihi nemlidir (Silahtaręlu, 2013: 12). DT oluřturmak iin CART, ID3, C4.5, SLIQ, SPRINT, CHAID ve QUEST gibi birok algoritma kullanılmaktadır (Bell ve Mgbemena, 2017: 202). Karar dęmlerinin etkinlięini lmek iin bir takım lme kriterleri kullanılmaktadır. DT oluřturulma sırasında ama, her bir dęmn bir sonraki blme noktasındaki (o dęmdeki znitelięin deęeri) nesnelerin optimal saflıęını (homojenlięini) elde etmektir (Gorunescu, 2011: 165). En ok kullanılan ayırma indeksleri; Gini indeksi, entropi ve χ^2 (ki-kare) olarak belirtilmektedir (Bramer, 2020).

Gini İndeksi: Gini indeksi, nesnelerin arasındaki eřitlik veya eřitsizlik seviyesini lmek iin kullanılır. Gini indeksi, 0-1 arasında deęer almaktadır. Gini indeksi bir dęmde “0” olduęunda tm verilerin tek bir sınıfta toplandıęını, yani tek bir sınıfın olduęu, “1” olduęunda ise bir deęiřkendeki tm deęerlerin aynı olduęunu ifade etmektedir. Bu nedenle bir znitelik Gini indeksi yksek olduęunda daha fazla bilgi tařımaktadır (Bhatia, 2019: 90-91; Bramer, 2020: 67).

Entropi (Entropy): Entropi, genel olarak, bir sistemdeki dzensizlik veya belirsizlik miktarının llmesidir. Sınıflandırmada, bir blm dřk entropiye sahipse daha az dzensizlik vardır. Bir blm daha fazla dzensizlięe sahipse daha yksek entropi sz konusudur (Zaki ve Meira, 2020: 479). Entropi, bir eęitim kmesinde birden fazla olası sınıflandırmanın olduęu belirsizlik lsdr (Bramer, 2020: 57).

Bilgi Kazanımı: Tahmin edilecek hedef deęiřken iin bir girdi nitelięinin “iyilięini” lmektedir. En yksek bilgi kazanımına sahip olan zellik bir sonraki blme iřlemi iin seilen deęiřken olmaktadır. Bilgi kazanımı en byk olan deęer en kk entropiye sahip olan deęerdir. Bir olayın sonucu kesinse, dięer bir ifade ile bir olayın olasılıęı 1 ise, bilgi

kazancı 0'dır (en iyi değerdir). Böylece bu değişken, karar ağacının bir sonraki bölünme noktası olmaktadır (Bhatia, 2019: 72).

Olumlu Yönleri: Karar ağaçları ile basit ve anlaşılabilir sınıflandırma kuralları oluşturulabilir. Hem nicel ve sürekli hem de kategorik değişkenler üzerinde analiz ve sınıflandırma yapabilmektedir. DT, hangi niteliğin sınıflandırma için en uygun bilgiler ortaya koyduğunu sunmaktadır (Pujari, 2008: 159). Diğer istatistiksel yöntemlerde olduğu gibi özellik puanlarının dağılımına, özelliklerin birbirinden bağımsız olmasına ilişkin varsayımlar DT'de yoktur. Farklı analiz yöntemleri arasında DT, istatistiksel yöntemlere göre daha güçlü bir analiz yöntemi olarak kabul edilmektedir. Bölünme değerlerine ilişkin hesaplamalar gerçekleştirildiğinde DT oluşturulması oldukça kolay ve hızlıdır (Kantardzic, 2011: 192).

Olumsuz Yönleri: Bazı DT'ler ikili değer alabilen hedef değişkenler ile analiz yapabilmektedir. Bazı DT ise, isteğe bağlı olarak belirlenen sınıflara kayıt ya da gözlemleri atayabilmektedir. Buna karşın her sınıfa ilişkin eğitim veri setinde yer alan kayıt ya da gözlemler azaldığından DT hataya açık hale gelmektedir. Bu durum çok fazla düzeyi ya da alt kategorisi bulunan ağaçlarda daha hızlı gerçekleşmektedir. DT oluşturulma süreci hesaplamalar açısından ele alındığında oldukça yorucu bir süreçtir. Her bir aday bölünme sürecindeki her bir çoğum için en iyi bölünme değeri ya da indeksi hesaplanarak bölünme işlemi gerçekleştirilmektedir (Pujari, 2008: 159). DT, sürekli bir iyileştirme sürecinin temel alınarak örnek verilerin yoğunlaştığı sınıflara atanarak bölünmenin gerçekleştiği sınıflandırma araçlarıdır. İyi bir sınıflandırıcı oluşturmak için gerekli eğitim veri setinin sayısı, belirlenen alt bölgelerin sayısı ile doğru orantılıdır. Karmaşık sınıflandırmalar daha fazla kural ile açıklanmakta ve daha karmaşık DT ile temsil edilmektedir. Bu durum daha iyi bir sınıflandırma elde etmek için gerekli eğitim veri setlerinin sayısını arttırmaktadır. Karar kuralları dikey bir düzlemde ifade edilebilen verileri içermektedir. Karar kurallarında sınıflandırma düzlemleri dikey değil de doğrusal ya da doğrusal olmayan bir nitelik taşırsa, kural temelli bir modelleme için sorun teşkil edebilmektedir. Çok fazla sınıflandırma kuralı mevcut olduğunda, sınıflandırma oldukça karmaşık bir yapıya bürünebilir ve bu durum hata oranını yükselten bir etmendir (Kantardzic, 2011:193).

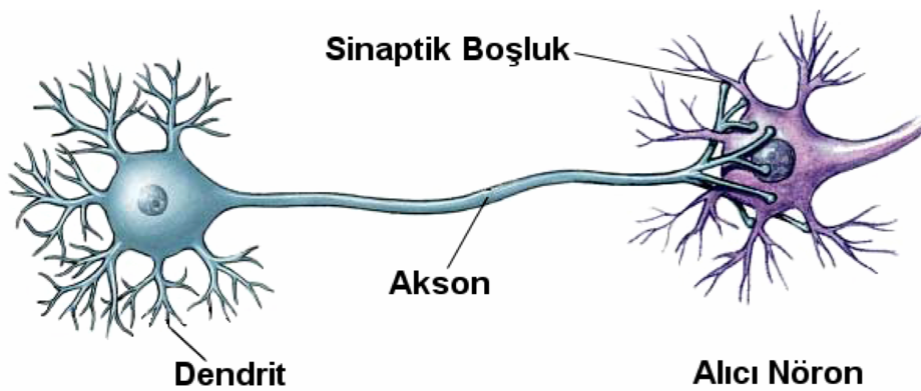
3.1.7. Yapay Sinir Ağları

Yapay Sinir Ağları (YSA), nöroloji uzmanları McCulloch ve Pitts (1943) tarafından insan beyninin hesaplama yeteneğinden yola çıkılarak geliştirilmiştir (Geron, 2019: 278). İnsan sinir sisteminden ilham alınarak tasarlanan YSA, birbirine bağlı yapay sinir hücrelerinden oluşmaktadır. Biyolojik sinir hücrelerine benzer bir yapıda çalışan yapay sinir

hücreleri, girdi şeklinde iletilen verileri değerlendirmekte ve bu girdileri çıktı nöronuna iletmektedir.

3.1.7.1. Biyolojik Sinir Hücresi

İnsan beyinde gerçekleşen; bilgileri hatırlama, düşünerek karar verme ve geçmiş deneyimleri yeni eylemlere uygulama yeteneklerini sağlayan yaklaşık 100 milyar sinir hücresi (nöron) bulunmaktadır. Her bir nöron diğer nöronlar ile 200 bin kadar farklı bağlantı kurabilmektedir. İnsan beyindeki öğrenme, nöronlar arasındaki bu bağlantılar ile mümkün olabilmektedir. Nöronlar, pek çok alt sistem ve kontrol mekanizması bulunan karmaşık bir yapıya sahiptir (Çavuşoğlu, 2021: 50).



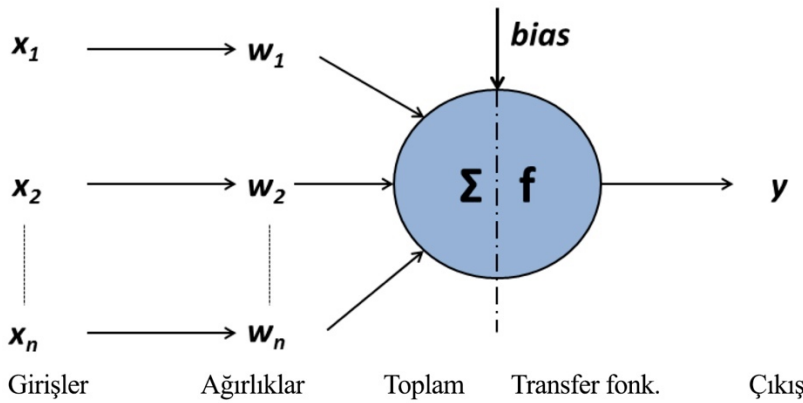
Şekil 3.2 Sinir Hücresi Yapısı

Kaynak: (Öğücü, 2006:22)

Sinir hücreleri, Şekil 3.2’de gösterildiği gibi sinaps, akson, soma ve dendritlerden meydana gelmektedir. Dışarıdan gelen sinyaller dendritlerden hücrenin aksonuna doğru karmaşık işlemler ve doğrusal olmayan bir biçimde iletilmektedir. Ulaşan sinyaller başka sinir hücrelerine sinaps ile gönderilmektedir. Sinir hücrelerinden gelen elektrokimyasal sinyali somaya ulaştıran parçalar dendrit olarak isimlendirilmektedir. Soma sinir hücresinin gövdesini ifade etmekte olup alınan bilgilerin birleştirildiği, anlam kazandığı, işlendiği ve elde edilen çıktıların aksona iletildiği mikro işlem parçalarıdır. Aksonda ise sinir hücrelerinin sağladığı sinyalleri çıktı elemanlarına ve diğer sinir hücrelerine gönderen elemandır. Aksonun, diğer sinir hücrelerinin dendritlerine bağlanmasını sağlayan birleşme noktaları sinapsler tarafından sağlanmaktadır (Öztemel, 2006: 47). Alıcı hücrenin gövdesi içindeki elektrik potansiyelini yükseltme ve azaltma ile etki sağlanmakta olup, potansiyel bir eşiğe ulaştığında aksona bir titreşim gönderilerek hücre ateşlenmektedir (Abraham, 2005: 901).

3.1.7.2. Yapay Sinir Hücresi Yapısı

Yapay sinir ağları gerçek sinir hücre yapısının basitleştirilmiş halini sunmaktadır. Gerçek nöronlarda olduğu gibi yapay sinir ağları modelinde girdi kısmı, hücre gövdesi toplama ve transfer fonksiyonlarını içeren hücre gövdesi ve diğer yapay nöronlara bağlanan çıkış dizini bulunmaktadır (Pujari, 2008: 203-204). YSA'nın en temel elamanı olan yapay sinir hücresinin yapısı Şekil 3.3'te yer almaktadır. Çevre ve diğer sinir ağlarından alınan girdiler ağırlıklarla çarpıldıktan sonra ilgili nöronun toplama fonksiyonu kullanılarak toplandıktan sonra transfer fonksiyonuna iletilmektedir. Ayrıca her bir sinir hücresine bias adı verilen eşik değeri eklenerek transfer fonksiyonuna gelen girdilerde kayıp olmaması sağlanmaktadır.



Şekil 3.3 Yapay Sinir Hücresi Yapısı

Kaynak: (Bilgiç vd., 2016:10)

$$y(k) = F\left(\sum_{i=0}^m w_i(k) * x_i(k) + b\right) \quad (7)$$

$x_i(k)$: k ayrık zamandaki girdi değeri

$w_i(k)$: ayrık zamanda ağırlık değeri

b : yanlılık

F : transfer fonksiyonu

$y(k)$: ayrık zamandaki çıkış değeri

Yapay sinir hücresi modeline ait Denklem (8)'de verilmiş olup, modeldeki en büyük bilinmeyen değişken transfer fonksiyonudur. Transfer fonksiyonu yapay sinir hücresinin özelliklerini tanımlayan bir matematiksel fonksiyondur. Bu fonksiyon yapay sinir ağını çözmesi gereken problemin doğrusal, doğrusal olmayan veya adım işlevine bağlı olmaktadır. Adım işlevinin yalnızca iki olası çıkış değeri bulunmaktadır. Burada giriş değeri belirli bir eşiği karşılırsa çıkış değeri tek bir değerle sonuçlanır ancak çıkış değeri belirli bir eşiği

karşılamazsa farklı bir çıkış değeriyle sonuçlanmaktadır. Bu durum Denklem (9) ile ifade edilmektedir (Suzuki, 2011: 6).

$$y = \begin{cases} 1, & w_i x_i \geq \text{eşikdeğer} \\ 0, & w_i x_i < \text{eşikdeğer} \end{cases} \quad (8)$$

Yapay nöronlarda kullanılan transfer fonksiyonu yapay sinir ağının son katmanında bulunmakta, yapay nöron algılayıcısı olarak isimlendirilmekte ve çok katmanlı algılayıcı sınıflandırma problemlerini çözmek için kullanılmaktadır. Fonksiyonun doğrusal olması durumunda yapay nöron, ağırlıklı girdi ve yanlılığın toplamı üzerinde basit doğrusal bir dönüşüm uygulanır. Doğrusal olmayan bir fonksiyon tercih edildiğinde ise en yaygın kullanılan sigmoid fonksiyonudur (Suzuki, 2011: 6).

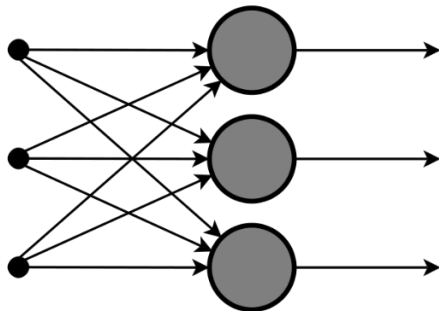
YSA'da öğrenme örnek verilerin ağa sunulması ile gerçekleştirilmektedir. En iyi YSA öğrenmesi için en doğru YSA modeli, ağın topolojisi, toplama fonksiyonu, aktivasyon fonksiyonu, öğrenme stratejisi ve öğrenme kuralı seçilmelidir. YSA katman sayılarına (tek katmanlı ve çok katmanlı), öğrenme stratejilerine (öğretmenli, öğretmensiz ve destekleyici) ve yapay sinir hücreleri arasındaki bağlantı yapılarına (ileri beslemeli ve geri beslemeli) göre sınıflandırılmaktadır (Hamzaçebi, 2011: 19).

3.1.7.3. Yapay Sinir Ağı Genel Yapısı

YSA katman sayısını, çözülmeye çalışılan problemin yapısı ve karmaşıklık düzeyi belirlemektedir. YSA tek katmanlı veya çok katmanlı yapıya sahip olmakla birlikte genellikle yapay sinir ağları ile çözülmek istenilen problemler çok katmanlı sinir ağlarının kullanılmasını gerektiren yapıdadır.

3.1.7.3.1. Tek Katmanlı Yapay Sinir Ağları

Tek katmanlı YSA girdi ve çıktı katmanı olmak üzere iki katmandan meydana gelmektedir. Tek katmanlı yapay sinir ağlarından ara veya gizli katman bulunmamaktadır. Şekil 3.4'te tek katmanlı YSA örnek verilmiştir.



Şekil 3.4 Tek Katmanlı YSA

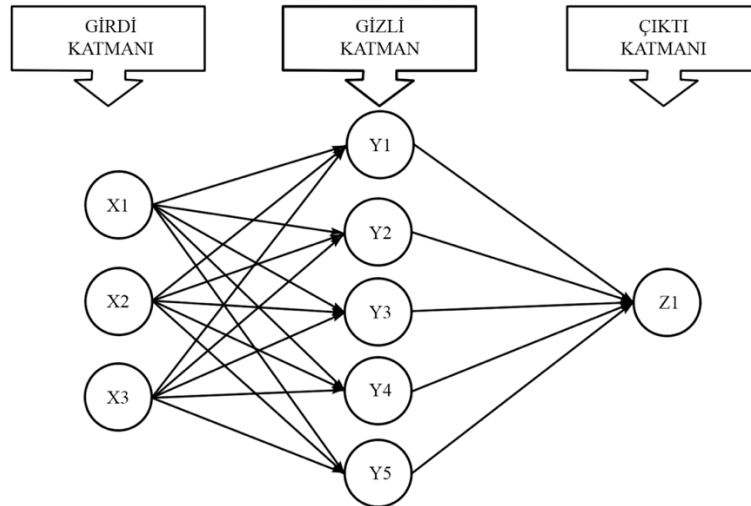
Kaynak: (WikiMedia Commons, 2023)

Tek katmanlı YSA doğrusal problem çözümlerinde başarı gösterirken, doğrusal olmayan problemlerin çözümlerinde ise yeterli bulunmamaktadır (Duman, 2016: 26). Basit sınıflandırma problemlerinin çözümünde tek katmanlı yapay sinir ağları ile etkili sonuçlar alınabilmektedir. Tek katmanlı yapay sinir ağları ile çözülemeyecek düzeyde karmaşık problemlerle karşılaşılması durumunda ise çok katmanlı yapay sinir ağları tercih edilmektedir.

3.1.7.3.2. Çok Katmanlı Yapay Sinir Ağı Genel Yapısı

YSA, paralel çalışan ve birbirlerine hiyerarşik bağlı yapay sinir hücrelerinden oluşmaktadır. Sinir ağının en temel görevi, girdi setine karşı çıktı seti belirlemektir. Bu görev için veri örnekleri kullanılarak eğitilmekte ve genelleme yeteneğine ulaştırılmaktadır. İki katmanlı yapay sinir ağları ile çözülemeyen problemlerde giriş, çıkış katmanlarına ek olarak ağın karmaşıklığını ve depolama kapasitesini arttıran katman eklenmektedir.

YSA; girdi, ara ve çıktı katmanlarının paralel olarak bir araya gelmesi ile oluşmuştur. Girdi katmanı, dışarıdan alınan verileri ara katmanlara transfer etmektedir. Ara katmanlar, girdi katmanından gelen verileri işleyerek çıktı katmanına göndermektedir. Çıktı katmanı, ara katmandan gelen verileri işleyerek girdi setine uygun çıktı üretmektedir. YSA, Şekil 3.5'te gösterildiği gibi yalnızca kendinden önceki katmandan giriş almaktadır.



Şekil 3.5 Çok Katmanlı YSA

Kaynak: (Akkaya vd., 2009: 193)

Çok katmanlı YSA; birleştirme, aktivasyon, hata fonksiyonları ve öğrenme algoritması gibi temel bileşenlerden oluşmaktadır. Aktivasyon fonksiyonu, ağa girdi olarak alınan, ağda sürekli olarak ileriye dönük bir yapıda ilerleyen ve çıktı katmanında ürüne dönüşen fonksiyonları ifade etmektedir. Bu fonksiyonun ana işlevi hücreye gelen girdiyi işlemek ve bu girdiye karşılık ortaya koyacağı çıktıya karar vermektir. Aktivasyon fonksiyonunu şekilde

adlandırılması, çıktı verebilmesi için geçtiği her sinir hücresi için sinyalin ölçülmesi ve her sinir hücresi için bir ağırlık değerinin atanmasıdır (Haykin, 2009: 125).

Yapay sinir ağlarına, öğrenmesi için sunulan örneklerle ağına ürettiği ürün arasındaki fark hata olarak adlandırılmaktadır. Bu durumda “hata fonksiyonu” çıktı katmanında ortaya çıkan ürün üzerinden değerlendirilmektedir. Ağda geriye doğru gidilene bu fark, ağa farklı örnekler sunulursa en aza indirgenmeye çalışılmaktadır. Ayrıca ağda ağırlıkların belirlenmesi hata fonksiyonunun en aza indirgenmesi ile sağlanmaktadır (Hamzaçebi ve Kutay, 2014: 229). Bu kapsamda yapay sinir ağında etkili sonuçların elde edilebilmesinin ağda ortaya çıkan hata fonksiyonunun en aza indirgenmesi ile mümkün olmaktadır.

3.1.7.4. Nöronlar Arası Bağlantı Yapılarına Göre YSA

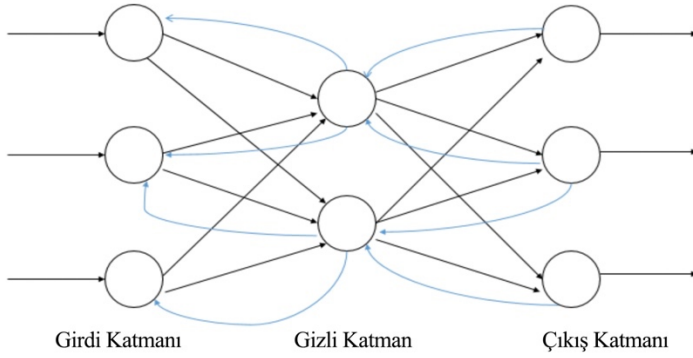
Nöronlar arası bağlantı yapıları ile yapay sinir ağlarından bilginin işleyişine yönelik önemli farklılıklar oluşmaktadır. Nöronlar arası bağlantı yapılarına göre “ileri beslemeli” ve “geri beslemeli” olmak üzere yapay sinir ağları iki ana grup altında incelenmektedir (Cemaloğlu ve Duykuluoğlu, 2020: 513).

3.1.7.4.1. İleri Beslemeli YSA

Şekil 3.4’te tek ara katmana sahip ileri beslemeli YSA’da gösterildiği gibi, girdiden çıktıya doğru tek yönlü bağlantı bulunmaktadır (MacKay, 2003: 505). Bu durum ileri beslemeli yapay sinir ağında girdi olarak alınan nesneler için geri beslemenin olmadığını anlamına gelmektedir. Her katmanda çözülmesi amaçlanan probleme bağlı olarak farklı sayıda sinir hücreleri yer almaktadır (Hamzaçebi ve Kutay, 2014: 229). İleri beslemeli ağlarda ara katmanların sayısı, problemin ya da ağa sunulan örneğin karmaşıklığına bağlı olarak değişmektedir. Çalışma problemi ne kadar karmaşık ise, yapay sinir ağında o kadar fazla gizli ya da ara katmanın yer alacağı ileri sürülmektedir (Cemaloğlu ve Duykuluoğlu, 2020: 513).

3.1.7.4.2. Geri Beslemeli YSA

Geri beslemeli YSA’da bilgi işleme sistemi, ileri beslemeli YSA’dan farklılık göstermektedir. Bu tür ağlarda sinir hücreleri kendilerinden önce gelen veya kendi katmanındaki herhangi bir sinir hücresine bağlanabilmekte ve bu sinir hücrelerine girdi sağlayabilmektedir (Zurada, 1992: 42-43). Şekil 3.6’da geri beslemeli yapay sinir ağına örnek verilmiştir.



Şekil 3.6 Geri Beslemeli YSA

Kaynak: (Can, 2022: 54)

Geri beslemeli yapay sinir ağlarında bilginin hücreler arası aktarımı çift yönlü olarak gerçekleşmektedir. Bu yapıda katmanlarda yer alan nöronlar sadece kendilerinden sonra gelen katmanlarda yer alan nöronlarla değil, kendi katmanlarında diğer nöronlara ya da kendinden önce gelen katmanda yer alan nöronlarla da bağlantılı olabilmektedir.

Olumlu özellikleri: Bilgi işleme süreçlerinin ağına tamamına yayılarak saklanması nedeniyle klasik programlamadan farklılık göstermektedir. Veriler ağına tamamına yayılmış halde olduklarından bazı nöronların işlevsiz kalması herhangi bir bilgi kaybı oluşturmamaktadır. Örnekler kullanılarak gerçekleştirilen öğrenme sayesinde ağ eğitilmekte ve hedeflenen çıktıya uygun biçimde eğitim sürecinin gerçekleşmesi sağlanmaktadır. Yapay sinir ağı eğitimi için gönderilen veri, ağı tamamen temsil ediyorsa çıktı başarısı artmaktadır. Yapay sinir ağları daha önce karşılaşılmayan veriler üzerinden de sonuç üretebildikleri için algılama süreçlerinde sıklıkla tercih edilmektedirler. Örüntüleri tanımlama ve gruplandırma işlevini gerçekleştirebilen yapay sinir ağları, eğitim verisi üzerinden örüntü tanımlayarak mevcut örüntülerle ilişkilendirme yapabilmektedir. Eksik verilerin tamamlanması konusunda oldukça önemli olan örüntü tanımlama fonksiyonu sayesinde eksik veri seti yapay sinir ağı tarafından tamamlanabilmektedir (Tolon ve Güneri Tosunoğlu, 2008: 252). Ancak eksik bilginin önemi yüksekse tahmin doğruluğunda düşüş gözlemlenmektedir. Sahip oldukları hata toleransı nedeniyle ağ içerisinde yer alan bazı hücre fonksiyonları kaybolduysa da yapay sinir ağı çalışmaya devam edebilmektedir. Dağınık bellek yapıları olan yapay sinir ağlarında, nöronlardaki bağlantı ve ağırlık dereceleri ağdaki bilgiyi göstermektedir. Sınırsız sayıdaki değişken ve parametre ile çalışma yeterliliği bulunan yapay sinir ağları karmaşık problem çözümlerinde oldukça büyük başarı göstermektedir (Şerbetli, 2008: 98).

Olumsuz özellikleri: Yapay sinir ağlarının etkin biçimde çalışabilmesi için paralel işlemcilere gereksinimi duymakta ve donanım yapısına bağımlılık göstermektedir. Eğitim ve öğrenme de kullanılacak veri sayısının belirlenmesi için nihai bir kriter bulunmamaktadır. Problemin çözümü için gerekli olan veri sayısı değişebilmektedir. Probleme uygun ağ yapısı

tespit ederken sabit bir kural bulunmamakta ve deneme yanılma ile ağ topolojisi belirlenmektedir. Yapay sinir ağlarında öğrenme katsayısı, nöron ve katman sayısı gibi ağ parametrelerinin belirlenmesi için standart bir kural bulunmamaktadır. Bu gibi değerler belirlenirken herhangi bir standart bulunmaması her problem için farklı bir yaklaşım kullanılmasına neden olmaktadır. Yapay sinir ağları sayısal verilerle çalışabildiklerinden dolayı, problem verileri ağa tanıtılmadan önce sayısal değerlere dönüştürülmelidir. Yapay sinir ağının eğitilmesi uzun zaman alabileceği için maliyeti de yüksek olmaktadır (Şerbetli, 2008: 34).

3.1.8. Topluluk Öğrenmesi Teknikleri

1980’li yıllarda altın çağını yaşayan karar ağaçları, algoritmaların olgunluğa ulaşmaları neticesinde, 1990’lı yılların sonlarında çalışmalar “topluluk öğrenmesi” (ensemble learning) gibi sonuçların etkinliğini artıracak alanlara kaymıştır (Akpınar, 2017: 242). Topluluk öğrenmesi, sınıflandırıcıların birleşiminden oluşan bileşik bir modeldir. Sınıflandırıcıların teker teker kendi kararlarını verdiği ve kararların toplamına bağlı olarak topluluk tarafından bir sınıf tahmini geri döndürülmektedir. Topluluk öğrenmesi, bileşiminde yer alan sınıflandırıcılardan daha hatasız olması beklenmektedir. Torbalama (Bagging: Bootstrap Aggregating), Takviyeli Sınıflandırma veya hızlandırma (Boosting) ve Rastlantısal Ormanlar (Random Forests) topluluk öğrenmesine algoritmalarıdır (Han vd., 2012: 377-378). Topluluk öğrenmesi, birkaç modeli daha iyi sonuçlar elde etmek için birleştiren tekniklerdir. Topluluk öğrenmesi teknikleri genellikle tek bir modelle kıyaslandığında daha iyi tahmin performansı ürettiği için Kaggle vb. makine öğrenimi yarışmalarında derece almaktadır. Topluluk öğrenmesi yöntemleri, karmaşık algoritmalar tasarlayıp yüksek doğrulukta sonuçlar üreterek makine öğrenmesi yarışmalarını kazanmada, sağlık hizmetleri gibi en küçük bir başarımlar artışının bile çok kıymetli olduğu çalışma alanlarında tercih edilmektedir. Diğer yandan yorumlanabilirliğin daha önemli olduğu çalışmalarda tercih edilmemektedir.

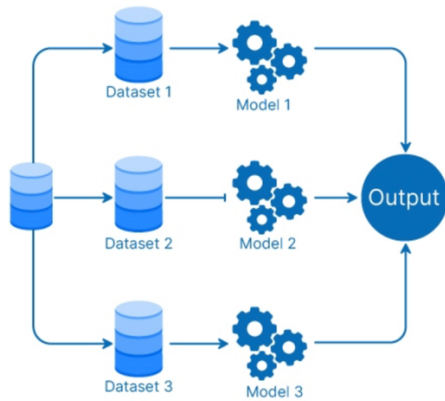
Makine öğreniminde topluluk yöntemlerini uygulamanın çeşitli yöntemleri bulunmaktadır. Oylama ve ortalama alma esaslı topluluk yöntemleri, torbalama ve güçlendirme en popüler topluluk yöntemleridir. Breiman (1996) tarafından geliştirilen torbalama yönteminde, birbirini izleyen ağaçların önceki ağaçlara bağlı değildir. Her bir ağaç veri kümesinde yeniden örnekleme örneği kullanılarak bağımsız olarak oluşturulmakta ve işlem sonunda tahmin için basit çoğunluk oyu alınmaktadır (Liaw ve Wiener, 2001: 18).

Yeniden örnekleme yöntemi ile veri kümesinden eğitim için çekilen veri gruplarını değiştirerek yenileme yoluyla homojen biçimde örnekleme yapmaktadır. Bu sayede, bir veri

kümesinden çok sayıda farklı eğitim verisi elde edilmektedir. Süreç içinde kullanılan veri grubunun veri setine dahil edilmesi ve yeni bir eğitim grubunun çekilmesinden dolayı eğitim veri kümesinde tekrarlı kayıtlar yer alabilmektedir. Bir veri grubunun tekrar seçilme ve eğitim kümesine yeniden eklenme olasılığı eşit olmaktadır. Eğitim verisinde yer almayan gözlemler test verisini oluşturmaktadır. Yeniden örnekleme yöntemleri içinden “0,632 yeniden önyükleme” yaygın olarak tercih edilmektedir. Veri setinin %63,2’si yeniden örnekleme için kullanılırken, %36,8’i test verisi için kullanılmaktadır. Elde edilen kümelerin her biri için sınıflandırma modeli doğruluk değerleri elde etmekte bu değerlerin ortalaması alınarak modelin doğruluğu hesaplanmaktadır (Özkan, 2016: 90).

Oylama ve ortalama alma: Makine öğreniminde topluluk öğreniminin en kolay uygulanan ve anlaşılan örnekleri oylama ve ortalama alma olup, oylama sınıflandırma uygulamaları için, ortalama ise regresyon için kullanılmaktadır. Sınıflandırma işlemleri için ağırlıklı ve çoğunluklu oylama kullanılmaktadır. Ağırlıklı sınıflandırmada, her sınıflandırıcının sonucu bir ağırlık değerine sahiptir. Ağırlık hesabına göre sonuç elde edilmektedir. Çoğunluklu oylamada ise en çok oya sahip sonuç çıktıyı oluşturmaktadır.

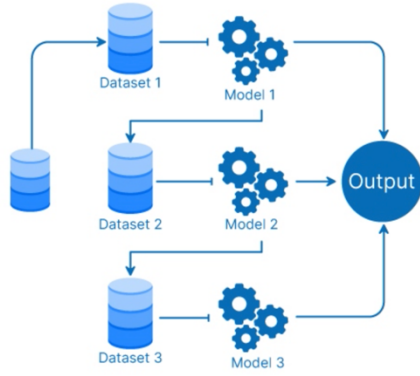
Torbalama: Rastgele seçmiş olduğu veri setleri ile model eğitme işlemlerini Şekil 3.7’de gösterildiği gibi paralel bir şekilde yapan ve oylama ile birleştiren yöntemlerdir. Rastgele orman (RO) torbalama yöntemini kullanmaktadır.



Şekil 3.7 Torbalama

Kaynak: Shiksha Online, 2023

Güçlendirme: Eğitim setlerinin rastgele seçildiği, her model eğitme işlemi sonunda eğitim veri setinin iade edildiği ve modellerin Şekil 3.8’de gösterildiği gibi sıralı olarak eğitildiği yöntemdir. Her modelin bir önceki modelden ders aldığı, yanlış ve doğru sınıflanan örneklerin ağırlığının değiştirildiği, yanlış sınıflanan örneğin hata değerinin artırıldığı, doğru sınıflanan örneklerin ağırlıklarının düşürüldüğü ve hatalı sınıflanan verilerin eğitilmesine ağırlık verildiği sınıflandırma yöntemidir. AdaBoost, CatBoost, XGBoost güçlendirme yöntemini kullanmaktadır.



Şekil 3.8 Güçlendirme

Kaynak: Shiksha Online, 2023

3.1.8.1. Rastgele Orman

Breiman (2001), torbalamaya ek olarak rastgelelik katmanı ekleyerek Rastgele Ormanı (RO) önermiştir. Verilerin farklı bir önyükleme örneğini kullanan RO, sınıflandırma veya regresyon ağaçlarının oluşturulma şeklini de değiştirmektedir. Standart ağaçlarda her düğüm, tüm değişkenler arasındaki en iyi bölünme kullanılarak bölünmektedir. Rastgele bir ormanda, her düğüm, o düğümde rastgele seçilen bir tahmin alt kümesi arasından en iyisi kullanarak bölünmektedir. RO, sahip olduğu sıra dışı sayesinde, diskriminant analizi, DVM ve YSA dahil diğer birçok sınıflandırıcıyla karşılaştırıldığında çok iyi performans gösterdiği ve aşırı uyum sağlamaya karşı dayanıklı olduğu ortaya çıkmıştır. Yalnızca iki parametreye sahip RO, kullanıcı dostu ve parametrelerine karşı çok duyarlı değildir (Liaw ve Wiener, 2001: 18).

3.1.8.2. AdaBoost

Schapire ve Freund'un geliştirdiği AdaBoost algoritması, birden fazla hipotez oluşturmak için eğitim örneklerini yönetmektedir (Freund ve Schapire, 1997). Her modelin bir önceki modelin hatasını düzelttiği ağırlıklandırma prensibine göre çalışmaktadır (Chengsheng vd., 2017). AdaBoost yöntemde eğitim verileri üzerinde bir olasılık dağılım değeri tutulmakta ve bu değere göre varyasyonlu örnekleme yaparak tekrarlı olarak m boyutlu bir eğitim seti üretilmektedir. Sınıflandırıcının hata oranı eğitim verileri kullanılarak hesaplanmaktadır. Hata değeri kullanılarak yanlış sınıflandırılan veri noktalarına daha fazla ağırlık verilmekte modelde hata düzeltilmektedir (Qi vd., 2016).

3.1.8.3. XGBoost

XGBoost, Tianqi Chen ve Carlos Guestrin tarafından geliştirilen bir artırma algoritmasıdır (Chen ve Guestrin, 2016). Tüm veri türlerini hızlı ve güvenilir bir şekilde işleyebilmek için Sebebe Dayalı Karar Ağacı ve Gradyan Arttırma Makinesini yöntemlerini

etkili tek bir yöntemde birleştirmektedir. Çok sayıda özellik ve büyük veri kümelerini analiz etmek ve etkili tahmin modelleri oluşturmak ve için kullanılmaktadır (Shehadeh vd., 2021). XGBoost'un yetenekleri aşağıdaki şekilde sıralanmaktadır (Ma vd., 2019):

- i. XGBoost, Lasso (L1) ve Ridge (L2) düzenleme yöntemlerine sahiptir.
- ii. En yüksek derinliğe göre bölmeler oluşturan ve kazanç sağlamayan bölümleri budayan derinlik öncelikli bir yaklaşıma sahiptir.
- iii. Önbelleği maksimum düzeyde kullanabildiği için hızlı çalışmaktadır.
- iv. XGBoost, en doğru bölünme noktalarını bulabilmek için ağırlıklı çeyrek çizim algoritmasını kullanmaktadır.

3.1.8.4. CatBoost

Kategorik artırma, temelleri gradyan artırma algoritmasına dayanan, kategorik verileri daha kolay ele alabilmek için Yandex'in ortaya çıkardığı bir algoritmadır. CatBoost algoritması kategorik değişkenleri kodlama yeteneğinden dolayı, veri ön işleme aşamasında çok avantajlı olmaktadır. Tahmin kaymaları problemini çözmek için sıralı hedef istatistikleri tekniği ile sıralı yükseltmeyi kullanmaktadır (Prokhorenkova vd., 2018).

3.1.8.5. LightGBM

Microsoft ve Pekin Üniversitesi'nden araştırmacılar, gradyan artış ve XGBoost algoritmalarında büyük boyuta sahip veri kullanıldığında karşılaşılan verimlilik ve ölçeklenebilirlik sorunlarını aşmak için LightGBM'yi oluşturmuştur. LightGBM, Gradyan Tabanlı Tek Yönlü Örnekleme ve Özel Özellik Paketleme gibi yeni tekniklerle veri miktarı ve değişken sayısı ile ilgilenmektedir. Bu iki yöntem ile yükseltme algoritması daha verimli ve ölçeklenebilir hale gelmektedir (Ke vd., 2017). Bu entegre özellikler, veri tarama, örnekleme, kümeleme ve sınıflandırma işlemlerinin kısa sürede, sorunsuz ve doğru bir şekilde gerçekleştirilmesine olanak tanımaktadır. Bellek tüketimi, işlem süresi ve aritmetik hızın tümü dikkate alındığında LightGBM, daha hızlı eğitim, verimlilik, optimum bellek kullanımı, kabul edilebilir doğruluk, paralellik ve büyük ölçekli veri işleme için iyi bir çözüm haline gelmektedir (Shehadeh vd., 2021).

3.2. Veri Madenciliği Yöntemlerinin Karşılaştırılması

Veri madenciliği alanında kullanılan denetimli makine öğrenmesi sınıflandırma tekniklerine yönelik literatürde yapılan karşılaştırmalar Tablo 3.2'de verilmiştir. En iyi performans (****) ile, en kötü performans (*) ile ifade edilmiştir (Kotsiantis, 2007: 262-263).

Tablo 3.2 Sınıflandırma Algoritmalarının Karşılaştırılması

İşlem Adı	Karar Ağaçları	YSA	NB	KNN	DVM
Genel olarak doğruluk	**	***	*	**	****
Öğrenme hızı	***	*	****	****	*
Sınıflandırma hızı	****	****	****	*	****
Eksik değerlere direnç	***	*	****	*	**
İlgisiz değişkenlere direnç	***	*	**	**	****
Fazla sayıda değişkene direnç	**	**	*	**	***
Bağımlılığı yüksek değişkene direnç	**	***	*	*	***
Kesikli, sürekli değişkenlere yaklaşım	****	***	***	***	**
Gürültüye direnç	**	**	***	*	**
Aşırı öğrenme ile başa çıkma	**	*	***	***	**
Artımlı öğrenme denemeleri	**	***	****	****	**
Açıklanma yeteneği	****	*	****	**	*
Model parametre işleme	***	*	****	***	*

Genel olarak, DVM çok boyutlu ve sürekli değişken özniteliklerle çok daha iyi performans gösterme eğilimindedir. DVM, maksimum tahmin doğruluğu elde etmek için büyük örneklem boyutuna ihtiyaç duyarken, NB nispeten küçük veri setine ihtiyaç duymaktadır. KNN'nin, ilgisiz özelliklere çok duyarlı olduğu konusunda genel bir fikir birliği bulunmaktadır. Yanlılık, farklı veriler üzerinde eğitildiğinde, sınıflandırıcının merkezi eğiliminin hataya katkısını ölçmektedir. Varyans, merkezi eğilimden sapma hatalarına olan katkının bir ölçüsüdür. NB gibi, yüksek yanlılık profiline sahip öğrenme algoritmaları, genellikle veri dalgalanmalarına karşı oldukça duyarsız olan basit, oldukça kısıtlı modeller üretmektedir. DVM gibi yüksek varyans profiline sahip algoritmalar, veri değişimlerine daha kolay uyan sağlayabilen, rastgele karmaşık modeller üretebilmekle birlikte, bu model sınıflarının bariz tuzağı aşırı öğrenmedir. DVM, çoklu bağlantı mevcut olduğunda, giriş ve çıkış öznitelikleri arasında doğrusal olmayan bir ilişki olduğunda iyi performans göstermektedir. NB, çok hızlı bir şekilde eğitilmekte ve eksik değerlere karşı dayanıklıdır. KNN, çalışabilmek için eksiksiz kayıtlara ihtiyaç duymaktadır. KNN genel olarak gürültüye tolerans göstermemektedir. Kullanıcı tarafından ayarlanacak model veya çalışma zamanı parametrelerinin sayısı, bir algoritmanın kullanım kolaylığının bir göstergesidir. DVM algoritmaları diğer tekniklere göre daha fazla parametreye sahiptir. KNN, genellikle ayarlanması nispeten kolay olan sadece tek bir k parametresine sahiptir. Açıklanma yeteneği, bir öğrenme algoritmasının çıktıların ne derece kolay yorumlanabildiğinin bir göstergesi olup, YSA açıklanma yeteneği en zor algoritmadır. Şeffaflık ise, yöntemin ilkesinin kolayca anlaşılıp anlaşılmadığını ifade etmektedir. DVM ve KNN algoritmalarının yorumlama yeteneği zayıfken, aynı zamanda KNN ve NB şeffaf yöntemlerdir.

Topluluk öğrenmesi algoritmaları altında yer alan; Rastgele Orman, AdaBoost, XGBoost, CatBoost ve LightGBM sınıflandırma algoritmalarının üretim yılı, kategorik değişkenler ile çalışma durumları, hız, ölçeklenebilirlik durumları, hafıza kullanımları, düzenlileştirme, paralel işleme, CPU destekleri ve özellik önem kıyaslamaları Tablo 3.3'te verilmiştir (GeekforGeeks.org, 2023).

Tablo 3.3 Güçlendirme Algoritmaları Arasındaki Farklar

Algoritma	Rastgele Orman	AdaBoost	XGBoost	CatBoost	LightGBM
Tasarım Yılı	2001	1995	2014	2017	2017
Kategorik Değişken Ele Alma	Hayır	Hayır	Hayır	Otomatik	Hayır
Hız/Ölçeklenebilirlik	İlımlı	Hızlı	Hızlı	İlımlı	Hızlı
Hafıza Kullanımı	Yüksek	Düşük	İlımlı	Yüksek	Düşük
Düzenlileştirme	Evet	Hayır	Evet	Evet	Evet
Paralel İşleme	Evet	Hayır	Evet	Evet	Evet
GPU Desteği	Evet	Hayır	Evet	Evet	Evet
Özellik Önemi	Mevcut	Mevcut	Mevcut	Mevcut	Mevcut

3.3. Model Değerlendirme Kriterleri

Makine öğrenimi alanında ikili sınıflandırıcılar “Evet”/“hayır”, “spam”/“spam değil”, “pozitif”/“negatif”, “Rezervasyon iptal”/“Rezervasyon iptal değil” iki olası sonuca göre karar vermemize yardımcı olan güçlü araçlardır. İkili sınıflandırıcılar, e-posta spam filtreleri, tıbbi teşhis sistemleri, sahtekarlık tespit algoritmaları ve görüntü tanıma teknolojisi gibi birçok gerçek dünya uygulamasında kullanılmaktadır. İkili sınıflandırıcıları kullanırken, güvenilir tahminler yapmalarını sağlamak için performanslarını doğru bir şekilde değerlendirmek önemlidir. İkili sınıflandırma problemlerinde tahmin modellerin değerlendirilmesi için Tablo 3.4'te örneği gösterilen Karışıklık Matrisinden (Confusion Matrix) ve doğrulama metriklerinden faydalanılmaktadır. Karışıklık matrisi ve doğrulama metrikleri, sınıflandırıcının etkinliğini ölçmek için açık ve sezgisel bir yol sağlamaktadır.

Tablo 3.4 Karışıklık Matrisi ve Değerlendirme Metrikleri

		Tahmin Değeri			
		Pozitif (PP) 1	Negatif (PN) 0		
Gerçek Değer	Pozitif (P) 1=İptal	TP	FN (Tip 2 Hata)	$TPR = \frac{TP}{P}$ Sensitivity, Recall	$FNR = \frac{FN}{P}$ False Negative Rate
	Negatif (N) 0=İptal değil	FP (Tip 1 Hata)	TN	$TNR = \frac{TN}{N}$ Specificity	$FPR = \frac{FP}{N}$ False Positive Rate
		$PPV = \frac{TP}{PP}$ Precision	$NPV = \frac{TN}{PN}$	$Accuracy = \frac{TP+TN}{P+N}$	$F1 = \frac{2*TPR*PPV}{TPR+PPV}$

Doğru Pozitif (TP: True Positive) değeri, gerçek sınıf pozitif olduğunda doğru tahminlerin sayısını yani gerçekte “Rezervasyon iptal” verilerinin kaç tanesinin “Rezervasyon iptal” şeklinde tahmin edildiğini,

Doğru Negatif (TN: True Negative) değeri, gerçek sınıf negatif olduğunda doğru tahminlerin sayısını yani gerçekte “Rezervasyon iptal değil” verilerinin kaç tanesinin “Rezervasyon iptal değil” şeklinde tahmin edildiğini,

Yanlış Pozitif (FP: False Positive) değeri Tip 1 Hata olarak geçen ve gerçek sınıf negatif olduğunda tahminin pozitif olarak yapılmasını, yani gerçekte “Rezervasyon iptal değil” verilerinin kaç tanesinin “Rezervasyon iptal” şeklinde tahmin edildiğini,

Yanlış Negatif (FN: False Negative) değeri Tip 2 Hata olarak geçen ve gerçek sınıf pozitif olduğunda tahminin negatif yapılmasını yani gerçekte “Rezervasyon iptal” verilerinin kaç tanesinin “Rezervasyon iptal değil” şeklinde tahmin edildiğini göstermektedir.

Sınıflandırıcının doğruluğu ve olumlu/olumsuz örnekleri tanımlama becerisi hakkında değerli bilgiler veren doğruluk, kesinlik, hassasiyet, F1 Puanı, özgüllük, FPR, FNR, NPV, ROC ve hız gibi değerlendirme ölçütleri Tablo 3.5’te verilmiştir.

Tablo 3.5 Sınıflandırma Algoritmalarının Karşılaştırılma Metrikleri

Ölçütler	Formülü	Değerlendirme Odağı
Doğruluk (Accuracy)	$\frac{TP + TN}{TP + FP + FN + TN}$	Doğru tahmin sayısının toplam tahmin sayısına oranı.
Kesinlik (Precision)	$\frac{TP}{TP + FP}$	Doğru tahminlerin tüm tahminler arasındaki oranı.
Hassasiyet, Duyarlılık (Recall, Sensitivity)	$\frac{TP}{TP + FN}$	Sınıflandırıcının pozitif etiketleri belirleme etkinliği
F ₁ -Skor (F1-Score)	$\frac{2 * Hassasiyet * Kesinlik}{Hassasiyet + Kesinlik}$	Kesinlik ve hassasiyet ölçütlerinin harmonik ortalaması
Özgüllük, Seçicilik (Specificity)	$\frac{TN}{TN + FP}$	Sınıflandırıcının negatif etiketleri ne kadar etkili tanımlaması
Yanlış Pozitif Oranı (FPR)	$\frac{FP}{TN + FP}$	Hatalı şekilde pozitif sınıflandırılan negatif örneklerin oranı
Yanlış Negatif Oranı (FNR)	$\frac{FN}{FN + TP}$	Hatalı şekilde negatif sınıflandırılan pozitif örneklerin oranı
Negatif Tahmin Değeri (NPV)	$\frac{TN}{TN + FN}$	Negatif tahminlerin doğruluğu
ROC - AUC	$\frac{1}{2} \left(\frac{TP}{TP + FN} + \frac{TN}{TN + FP} \right)$	Sınıflandırıcının yanlış sınıflandırma işleminden kaçınma yeteneği
Hız	-	Algoritma çalışma süresi (sn)

Doğruluk (Accuracy): Doğruluk bize sınıflandırıcının ne sıklıkta doğru tahminler yaptığına dair genel bir fikir vermektedir. Doğru tahmin sayısının toplam tahmin sayısına oranıdır (Siering vd., 2018: 58) Sınıflandırıcının %90 doğruluğu, tüm rezervasyonların %90'ının “Rezervasyon iptal” veya “Rezervasyon iptal değil” olarak doğru şekilde sınıflandırdığı anlamına gelmektedir.

Kesinlik (Precision): Modelin pozitif olarak tahmin ettiği örneklerin ne kadarının gerçekten pozitif olduğunu ölçmektedir. Doğru pozitiflerin sayısının (TP), pozitif sınıfa atanan tüm örneklerin sayısına (TP+FP) bölünerek hesaplanmaktadır. “Rezervasyon iptal” gerçek durumunu “Rezervasyon iptal” tahmininde ne kadar doğru olduğunu göstermektedir. Kesinlik değeri %80 olduğu durumda, sınıflandırıcının bir rezervasyonu, “Rezervasyon iptal” olarak tahmin etmesi durumunda bunun %80 oranında doğru olduğu anlamına gelmektedir. Kesinlik, tek başına bir performans metriği olarak yeterli olmayıp, hassasiyet gibi diğer metriklerle birlikte değerlendirilmektedir.

Hassasiyet (Sensitivity=Recall): Hassasiyet değeri, duyarlılık olarak da adlandırılmaktadır. Bir sınıflandırma modelinin doğru pozitiflerin ne kadarını doğru bir şekilde tahmin ettiğini ölçmektedir. Hassasiyet, sınıflandırıcının tüm olumlu örnekleri ne kadar iyi bulabildiğini anlamamıza yardımcı olmaktadır. Yüksek bir hassasiyet değeri, doğru pozitiflerin çoğunu yakaladığı anlamına gelmekte ve modelin o sınıfı kaçırma olasılığının düşük olduğunu göstermektedir (Puri vd., 2017: 241). Hassasiyet, $TP/(TP+FN)$ formülü ile hesaplanmakta ve “Sınıflandırıcının veri kümesindeki tüm olumlu örneklerin, kaç tanesini doğru bir şekilde tanımladı?” sorusuna yanıt aramaktadır. Hassasiyet ile gerçek değeri “Rezervasyon iptal” olan verilerden ne kadarının “Rezervasyon iptal” olarak doğru şekilde tahmin edildiğini göstermektedir. Örneğin, hassasiyetin %85 olması, sınıflandırıcının veri kümesindeki tüm gerçek “Rezervasyon iptal” değerlerinden %85’ini doğru bir şekilde “Rezervasyon iptal” olarak sınıflandırdığı anlamına gelmektedir.

Hassasiyet değerinin yüksek olması, modelin aynı zamanda yanlış pozitifler üretebileceği anlamına gelmektedir. Bu nedenle, birbirini tamamlayan iki metrik olan hassasiyet ve kesinlik birlikte değerlendirilmelidir. Kesinlik, modelin pozitif olarak tahmin ettiği örneklerin ne kadarının gerçekten pozitif olduğunu ölçerken, hassasiyet doğru pozitiflerin ne kadarının tespit edildiğini ölçmektedir. Belirli bir sınıf için farklı hassasiyet ve kesinlik kombinasyonları aşağıdaki anlamlara gelmektedir.

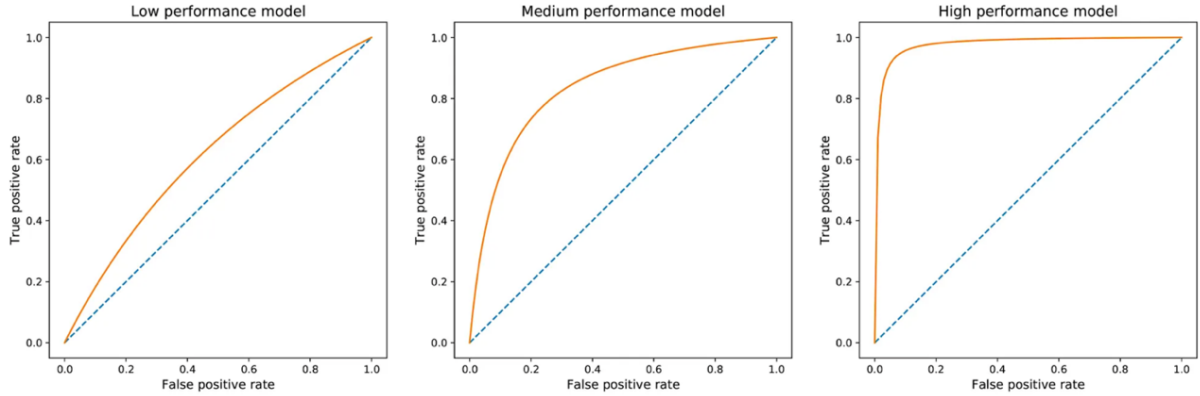
- Yüksek hassasiyet ve yüksek kesinlik olduğu durumlarda sınıfın, model tarafından mükemmel şekilde işlendiği,
- Düşük hassasiyet ve yüksek kesinlik olduğu durumlarda modelin sınıfı iyi bir şekilde tespit edemediği ancak tespit ettiğinde son derece güvenilir olduğu;
- Yüksek hassasiyet ve düşük kesinlik olduğu durumlarda sınıfın iyi tespit edildiği ancak model aynı zamanda diğer sınıfların noktalarını da içermekte olduğu,
- Düşük hassasiyet ve düşük kesinlik olduğu durumlarda ise sınıfın model tarafından yeterince ele alınamadığı sonucuna ulaşılmaktadır.

Modelin uygunluğunu iki perspektiften göstermesinden dolayı hassasiyet ve kesinlik genellikle ikili olarak raporlanmaktadır. Hassasiyet ile ölçülen “tip I hata” ve kesinlik ile ölçülen “tip II hata” olarak bilinmektedir. Hassasiyet ve kesinlik genellikle aralarında ters bir ilişki bulunmakta olup bir değeri azaltma pahasına diğer değeri arttırmanın mümkün olduğu durumlar bulunmaktadır. Örneğin, daha sıkı bir iptal politikası uygulanması durumunda, rezervasyonları iptal eden sayısında azalma olmasına rağmen, yapılacak olan toplam rezervasyon sayısında azalmaya yol açmaktadır. Tam tersi olan gevşek bir iptal politikası ile yapılacak toplam rezervasyon sayısı arttırmasına rağmen, rezervasyonlarını iptal etmeyenlerin azınlıkta olduğu ve çoğu rezervasyonun iptal edildiği bir durum oluşturmaktadır. Hassasiyet ve kesinlik metrikleri, modelin ne kadar iyi performansa sahip olduğunu gösteren tek bir metrik olan F1 ölçütü ile ifade edilebilmektedir.

F1 ölçütü: Kesinlik ve hassasiyet ölçümleri arasında bir denge sağlayan, bu iki metriğin harmonik ortalamasıdır. $2 * (\text{Kesinlik} * \text{Hassasiyet}) / (\text{Kesinlik} + \text{Hassasiyet})$ formülü ile hesaplanmaktadır. F1 Puanı, pozitif ve negatif sınıflar arasında dengesizlik olduğunda yararlı bir ölçümdür. F1 puanı hem kesinliği hem de hassasiyeti dikkate alan dengeli bir ölçümdür. F1 puanının yüksek olması, “Rezervasyon iptal” olan rezervasyonların doğru şekilde tanımlanması ile hatalı pozitif sonuçların önlenmesi arasında iyi bir denge olduğunu göstermektedir.

Özgüllük (Specificity): Sınıflandırıcının tüm olumsuz örnekleri tanımlama yeteneğini diğer bir ifadeyle yanlış negatifleri ne kadar doğru bir şekilde tahmin ettiğini ölçmektedir. Yanlış negatiflerin sayısı (TN), veri kümesindeki tüm negatif olayların sayısına (TN + FP) bölünerek hesaplanmaktadır. Yüksek bir özgüllük değeri, modelin yanlış negatifleri kaçırma olasılığının düşük olduğunu göstermektedir. Hassasiyet ve özgüllük birbirini tamamlayan iki metriktir. Hassasiyet, doğru pozitifleri ne kadar iyi tespit ettiğini ölçerken; özgüllük, yanlış negatifleri ne kadar iyi tespit ettiğini ölçmektedir. Bir modelin özgüllük değeri yüksekse, modelin genellikle yanlış negatifleri doğru bir şekilde sınıflandırdığı anlamına gelmekle birlikte, modelin doğru pozitifleri kaçırma olasılığını arttırabilmektedir. Özgüllük değerinin %95 olması, veri kümesindeki tüm gerçek “Rezervasyon iptal değil” rezervasyonların %95’inin “Rezervasyon iptal değil” olarak doğru bir şekilde tanımladığı anlamına gelmektedir. Özgüllük bize sınıflandırıcının tüm olumsuz örnekleri ne kadar iyi tespit edebildiğini göstermektedir.

Alıcı çalışma karakteristik eğrisi veya ROC Eğrisi: Pozitif vakaların doğru şekilde sınıflandırıldığı, Doğru Pozitif Oran (TPR) ile pozitif vakaların yanlış sınıflandırıldığı Yanlış Pozitif Oran (FPR) arasındaki denge Şekil 3.9’da gösterilmektedir.



Şekil 3.9 Alıcı Çalışma Karakteristik (ROC) Örnek Eğrileri

Kaynak: (Chhetry, 2023)

ROC eğrisini çizerken doğru pozitif oran (TPR), geri Çağırma veya hassasiyet değeridir. Y ekseninde Doğru pozitif oranı ve X ekseninde Yanlış pozitif oranına karşılık gelen tüm değerleri içeren grafik ROC eğrisini oluşturmaktadır. ROC eğrisinin altındaki alan AUC (Area Under Curve), Doğru ve Yanlış pozitif oranları arasındaki dengenin ölçüsüdür. AUC değerinin 1 olması, Doğru ve Yanlış pozitif oranları arasında bir denge olmadığını, en güvenilir eğriyi ve mükemmel bir sınıflandırıcı olduğu göstermektedir. AUC aşağıdaki gibi yorumlanmaktadır (Grzybowski ve Younger, 1997: 821).

- 0,90 ve 1,00 arası değerler için: Mükemmel
- 0,80 ve 0,90 arası değerler için: İyi
- 0,70 ve 0,80 arası değerler için: Orta
- 0,60 ve 0,70 arası değerler için: Zayıf
- 0,50 ve 0,60 arası değerleri için: Başarısız

Yanlış Pozitif Oranı (FPR-False Positive Rate): Bu ölçüm $FP/(FP+TN)$ formülü ile hesaplanmakta ve bir sınıflandırma modelinin yanlış negatifleri yanlış pozitif olarak tahmin etme oranını ölçmektedir. FPR, aynı zamanda "false alarm rate" veya "Type I error rate" olarak da adlandırılmaktadır. FPR'nin düşük olması, modelin yanlış negatifleri yanlış pozitiflere karşı daha iyi bir şekilde sınıflandırdığı, modelin daha az "yanlış alarm" ürettiği ve yanlış negatifleri daha iyi ayırt ettiği anlamına gelmektedir. FPR ve özgüllük (specificity) birbirine ters orantılı olup FPR'in azalması, özgüllüğün artması ve yanlış negatiflerin doğru bir şekilde sınıflandırılmasını göstermektedir.

Yanlış Negatif Oranı (FNR) veya Kaçırma Oranı: Bu ölçüm $FN/(FN+TP)$ formülü ile hesaplanmakta ve bir sınıflandırma modelinin doğru pozitifleri yanlış negatif olarak tahmin etme oranını ölçmektedir. Düşük bir FNR, modelin doğru pozitifleri kaçırma olasılığının

düşük olduğunu ve bu nedenle modelin doğru pozitifleri iyi bir şekilde tespit ettiğini göstermektedir.

Negatif Tahmin Değeri (NPV): $TN / (TN + FN)$ formülü ile hesaplanmakta ve bir sınıflandırma modelinin yanlış negatifleri ne kadar doğru bir şekilde tahmin ettiğini ölçmektedir. Yüksek bir NPV değeri, modelin yanlış negatifleri iyi bir şekilde ayırt ettiğini ve negatif sınıfı doğru bir şekilde tahmin ettiğini göstermektedir. NPV ve özgüllük (specificity) metriklerinin ikisi de yanlış negatifleri değerlendirmektedir. Ancak, NPV özellikle negatif sınıfın sınıflandırılmasına odaklandığı için, negatif sınıfın doğru tahmin edilme başarısını vurgulamaktadır.



DÖRDÜNCÜ BÖLÜM

VERİ MADENCİLİĞİ TEKNİKLERİ İLE ÇEVİRİM İÇİ KARAR DESTEK SİSTEMİ TASARIMI: SAĞLIK TURİZMİ UYGULAMASI

Dördüncü bölüm altında araştırma problemi, araştırmanın amacı, araştırmanın önemi, araştırmanın kısıtları, araştırmanın katkısı ve araştırma süreci yer almaktadır.

4.1. Araştırma Problemi

Türkiye’de sağlık turizmi alanında faaliyet gösteren, yetki belgesine sahip sağlık kuruluşları ve aracı kuruluşlar, yurt dışında bulunan sağlık turistlerine ulaşabilmek için sosyal medya ve arama motorları üzerinden yüksek bütçeler ile reklam vermektedir. Dijital mecralarda tedavi reklamlarını gören sağlık turistleri, sağlık kuruluşları ve seyahat acentalarının web sayfaları üzerinden çeşitli kanallar aracılığıyla iletişime geçmektedir. Sağlık turizmi paydaşları, çağrı merkezleri aracılığı ile tedavi, transfer, konaklama konularında bilgilendirme yaptıktan sonra, hastanın tercihi doğrultusundan rezervasyon işlemini tamamlamaktadır. Rezervasyon işlemi tamamlanan sağlık turisti, daha uygun bir hizmet bulabilmek amacıyla alternatif arayışlara devam edebilmektedir. Yeni bir hizmet arayışında olmayan sağlık turistleri bile, sosyal medyada karşısına çıkan rakip sağlık kuruluşlarının tedavi reklamlarından etkilenebilmekte ve yapmış oldukları rezervasyonları iptal edebilmektedir. Rezervasyon iptallerini öngöremeyen ve gerekli önlemleri alamayan sağlık kuruluşları ve seyahat acentaları gelir kaybına uğramakta ve kaynaklarını etkin bir şekilde yönetememektedir.

4.2. Araştırma Amacı

Veri madenciliği teknikleri ve makine öğrenimi algoritmaları ile sağlık turistlerinin rezervasyon verilerini kullanarak, rezervasyonlarını iptal edecek sağlık turistlerini tespit edebilen tahmin modellerinin geliştirilmesi, başarı kriterlerine göre model performanslarının karşılaştırılması, rezervasyon iptallerinde en çok etkiye sahip bağımsız değişkenlerin tespit edilmesi, doktora araştırmasının amacını oluşturmaktadır. Sağlık turizmi rezervasyon iptallerini başarılı bir şekilde tahmin eden modellerinin çevrimiçi karar destek sistemi içerisine entegre edilmesi, doktora araştırmasının diğer bir amacını oluşturmaktadır.

4.3. Araştırmanın Önemi

Sanayi ve Teknoloji Bakanlığı'nın sağlık turizmi alanına yönelik yeni teşvikler sunması ile Türkiye'de sağlık turizmine olan ilgi artmıştır. Sağlık Hizmetleri Genel Müdürlüğü'ne bağlı Sağlık Turizmi Daire Başkanlığının resmî web sitesinde 17.02.2021 tarihinde ilan edilen, “Uluslararası sağlık turizmi yetki belgesi alan sağlık tesisleri ve aracı kuruluşlar listesi” içerisinde 1097 sağlık kuruluşu ve 170 seyahat acentası yer almaktaydı. 26.02.2024 tarihinde ilan edilen güncel listede, 1059 yetkilendirilmiş seyahat acentası ve 3956 yetkilendirilmiş sağlık kuruluşu (648 hastane, 179 tıp merkezi, 2096 muayenehane ve 1124 diğer sağlık tesisi olmak üzere toplamda 4047 yetkilendirilmiş sağlık kuruluşu yer almaktadır. Son üç sene içerisinde yetki belgeli sağlık kuruluşu sayısının yaklaşık 3.7 katına, aracı kuruluş sayısı 6,2 katına çıkması sağlık turizmi alanına yönelik artan ilgiyi göstermektedir.

Sağlık Turizmi Yetki Belgesine sahip sağlık kuruluşu ve aracı kuruluşlar, sağlık turistleri tarafından yapılan rezervasyon iptallerini öngöremediği durumlarda, kaynaklarını etkin bir şekilde kullanamamakta ve gelir kaybına uğramaktadır. Sağlık turizmi kapsamında Türkiye'ye gelmeyi planlayan sağlık turistlerinin rezervasyonlarını iptal etme olasılıklarının tahmin edilmesi, sağlık paydaşlarının alacağı aksiyonlara karar verme noktasında fayda sağlamaktadır. Rezervasyon iptalleri konusunda yapılan literatür taramasında; otel, uçak, restoran, tur, ulaşım vb. alanlarda çalışmalar olmasına rağmen, sağlık turizmi rezervasyon iptallerini tahmin etmeye yönelik bir çalışma bulunmamaktadır. Doktora araştırmasının, sağlık turizmi rezervasyon iptallerinin tespiti ve elde edilen faydalı bilgiler açısından alanındaki ilk çalışma olması, tezin özgün değerini oluşturmaktadır.

4.4. Araştırma Kısıtları

Araştırma kapsamında Antalya ve İstanbul ilinde faaliyet gösteren sağlık turizmi yetki belgeli sağlık kuruluşları ve seyahat acentaları ile iş birliği yapılmıştır. Sağlık turizmi paydaşlarının 2022 yılına ait rezervasyon verileri anonimleştirilerek kullanılmıştır. Araştırma verileri diş, saç ekimi vb. estetik tedavilerine ait veriler ile sınırlı tutulmuştur.

Doktora araştırmaları kapsamında, yurt dışından tedavi amacı ile rezervasyon yaptıran sağlık turistlerinin rezervasyon iptallerinin tahmini üzerinde çalışmalar yapılacaktır. Türkiye'de tatilleri esnasında hastalanan misafirlerin tedavi ihtiyaçlarını ifade eden “Turist Sağlığı” araştırma kapsamında yer almamaktadır.

4.5. Araştırmanın Katkısı

Araştırmanın birinci bölümünde yapılan detaylı literatür taramasına göre, sağlık turizmi kapsamında rezervasyon iptallerini tahmin etmeye yönelik bir çalışmaya rastlanmamıştır. Doktora tezinin en önemli katkılarından biri, sağlık alanına özgü tedavi, uzmanlık, tedavi fiyatı verileri ve turizmi alanına özgü konaklama, kanal, tur verileri üzerine, makine öğrenimi algoritmaları ve veri madenciliği teknikleri kullanarak, rezervasyon tahmin modelleri geliştirmektir.

Doktora araştırması kapsamında, sağlık turizmi yetki belgeli sağlık kuruluşları ve seyahat acentalarının sağlık turizmi süreçlerini etkin bir şekilde yönetmelerini sağlamak amacıyla “Sağlık Turizmi Yönetim Sistemi (STYS)” yazılımı geliştirilmiştir. En iyi başarıma sahip sınıflandırma algoritmaları STYS yazılımına entegre edilerek, sağlık turistinin rezervasyon aşamasında girmiş olduğu verileri dikkate alarak, gerçek zamanlı olarak rezervasyon iptal tahmini yapılmaktadır. STYS kullanıcısı, rezervasyonunu iptal etme ihtimali yüksek olan sağlık turistine yönelik indirim tanımlama, promosyon tanımlama, ek hizmet hediye etme vb. stratejileri uygulama imkanına sahip olmaktadır.

4.6. Araştırma Süreci

Doktora araştırmasının tüm süreçlerinde, CRISP-DM metodolojisine ait işlem basamakları tüm adımları ile eksiksiz bir şekilde takip edilmiştir.

4.6.1. İşi Anlama

İşi anlama aşamasında önemli olan, problemin detayları ile tanımlanmasıdır. STYS projesinin ilk aşamasında, sağlık turizmi paydaşları ile yapılan kullanıcı görüşmeleri, fokus grup, kart sıralama vb. araştırma teknikleri sonucunda, sağlık turizmi süreçlerini takip etme ve yönetmede yaşanan zorluklar, sağlık turistlerinin rezervasyon iptallerini öngörececek bir sistemlerinin olmaması, tespit edilen genel problemler olarak sıralanabilir.

4.6.2. Veriyi Anlama

Veriyi anlama iş sürecinde; ilk verinin toplanması, verinin tanımlanması, verinin keşfedilmesi ve veri kalitesinin doğrulanması görevleri tamamlanmıştır.

i. İlk Verinin Toplanması: Doktora araştırma kapsamında, sağlık turizmi yetki belgesine sahip sağlık kuruluşları ve aracı kuruluşlar için STYS yazılımı geliştirilmiştir. STYS üzerinden sağlık turistlerine ait rezervasyon verileri, *Kişisel Verileri Koruma Kanunu* gözetilerek ve anonimleştirme işlemi yapılarak toplanmıştır.

ii. Verinin Tanımlanması: STYS yazılımında, sağlık turistlerine ait konaklama, tedavi, tur ve rezervasyon verileri farklı tablolardan alınmış ve tek bir Excel dosyasından birleştirilmiştir. Veri setine ait 39 değişken ve açıklamaları Tablo 4.1’de verilmiştir.

Tablo 4.1 Değişken Adı ve Açıklamaları

Değişken Adı	Gözlem	Veri	Değişken Açıklaması
otel_turu	5.060	object	Otel Tipi (Tatil Oteli, Şehir Oteli)
oda_sayisi	5.060	int64	Rezerve Edilmiş Oda Sayısı
yetiskin	5.060	int64	Konaklayacak Yetişkin Misafir Sayısı
cocuk	5.060	int64	2-12 Yaş Arası Konaklayacak Misafir Sayısı
bebek	5.060	int64	0-2 Yaş Arası Konaklayacak Misafir Sayısı
ozel_istek	5.056	float64	Misafirlerin Özel Talep Sayısı
toplam_fiyat	5.034	float64	Misafirlerin Toplam Konaklama Ücreti
uyruk	5.060	object	Misafirlerin Uyuşuğu
rezervasyon_durumu	5.060	object	Rezervasyonun Son Durumu (Çıkış, İptal, Gelmedi vb.)
para_birimi	5.060	int64	Ödemenin Yapıldığı Para Birimi
misafir_folyo	5.060	float64	Misafirlerin Otelde Yapmış Olduğu Harcama Tutarı
ucretli_kisi_sayisi	5.060	int64	Ücretli Kişi Sayısı
gece_sayisi	5.060	int64	Misafirlerin Konaklama Süresi
otel_onodeme	5.060	object	Misafirlerin varış öncesinde yaptığı ödeme miktarı
pan_tipi	5.004	object	Ücrete Dahil Yemek Tipi (Kahvaltı, Tam Pansiyon vb.)
oda_tipi	5.014	object	Misafir için rezerve edilmiş oda tipi
acenta	5.060	object	Rezervasyonu Yapan Seyahat Acentası
toplam_rez_say	5.060	int64	Misafirin Toplam Rezervasyon Sayısı
toplam iptal_say	5.060	int64	Misafirin Toplam Rezervasyon İptal Sayısı
iptal_durumu	5.060	int64	Rezervasyon İptal Durumu (İptal Değil = 0, İptal = 1)
gecelik_oda_fiyati	5.060	float64	Rezervasyon yapılmış otelin gecelik oda ücreti
uzmanlik_alani	5.060	object	İstenilen Tedavinin Uzmanlık Alanı (Saç, Diş vb.)
tedavi_tipi	5.060	object	Uzmanlığa Bağlı Tedavi Türü (Dental İmplant vb.)
tedavi_fiyati	5.060	int64	Tedavi İçin Ödenecek Toplam Tutar
uygulanan_indirim	5.060	float64	Tedavi Fiyatına Uygulanan İndirim
satis_fiyati	5.060	float64	Misafirin Tedavi İçin Ödediği Tutar.
tekrarlanan_misafir	5.060	int64	Misafirin Geçmiş Ziyaret Kaydı (Var = 1, Yok = 0)
gecmis_rez iptal edilmemis	5.060	int64	Geçmiş Rezervasyon İptalleri (Yok=1 , Var =0)
hasta_dogum_tarihi	5.060	object	Misafirin Doğum Tarihi
hasta_cinsiyet	5.060	object	Misafirin Cinsiyeti
hasta_yas	5.060	int64	Misafirin Yaşı
kanal	5.060	object	Misafirin Rezervasyonu Yaptığı Kanal
hasta_tercih_edilen_iletisim	5.060	object	Misafirin Tercih Ettiği İletişim Tipi
depozito_durumu	5.060	object	Depozito Verme Durumu (Vermedi=0, Verdi=1)
tur_durumu	5.060	object	Misafirin Tur Talebi (Tur Dahil=1, Tur Dahil Değil=0)
varis_tarihi	5.060	object	Misafirin Otele Varış Tarihi
kayit_tarihi	5.060	object	Rezervasyonun Oluşturulma Tarihi
varis_suresi	5.060	int64	Rezervasyon ve Varış Tarihi Arasındaki Gün Sayısı
cikis_tarihi	5.060	object	Konaklama Süresinin Bitiş Tarihi

Değişkenlerden 18 tanesi string ve kategorik verileri içeren “object” veri tipindedir. Tam sayı değerleri temsil eden “int64” veri türüne sahip 15 değişken bulunmaktadır. Ondalık sayıları temsil eden “float64” tipine sahip 6 adet değişken bulunmaktadır. Sayısal değişkenlere ait istatistik özeti Tablo 4.2’de verilmiştir.

Tablo 4.2 Sayısal Değişkenler için İstatistik Özeti

Değişken Adı	Ortalama	Std,	Min	25%	50%	75%	Maks
oda_sayisi	1,0339	0,5620	1,0000	1,0000	1,0000	1,000000	15,0000
yetiskin	1,9968	0,6904	1,0000	2,0000	2,0000	2,000000	6,0000
cocuk	0,0928	0,3386	0,0000	0,0000	0,0000	0,000000	3,0000
bebek	0,1247	0,4714	0,0000	0,0000	0,0000	0,000000	10,0000
toplam_fiyat	1534,55	1163,18	335,94	791,68	1196,52	1887,98	21675,42
misafir_folyo	141,2954	399,0462	0,0000	0,0000	0,0000	45,6600	4827,9000
ucretli_kisi_sayisi	2,0038	0,6949	1,0000	2,0000	2,0000	2,0000	6,0000
gece_sayisi	13,9579	7,3287	8,0000	9,0000	11,0000	17,0000	157,0000
toplam_rez_say	1,4294	1,8079	1,0000	1,0000	1,0000	1,0000	20,0000
toplam iptal_say	0,3428	1,3546	0,0000	0,0000	0,0000	0,0000	15,0000
gecelik_oda_fiyati	107,4055	46,3812	40,3825	71,9275	97,7925	133,4580	349,2216
tedavi_fiyati	2378,3241	1352,4508	50,0000	1500,0000	2100,000	3000,0000	6000,0000
satis_fiyati	2199,9884	1252,692470	45,00000	1350,0000	1900,000	2850,0000	5700,0000
hasta_yas	45,2701	7,3717	23,000	41,0000	46,0000	51,0000	58,0000
varis_suresi	82,0184	51,726781	30,0000	38,0000	65,0000	114,0000	236,0000

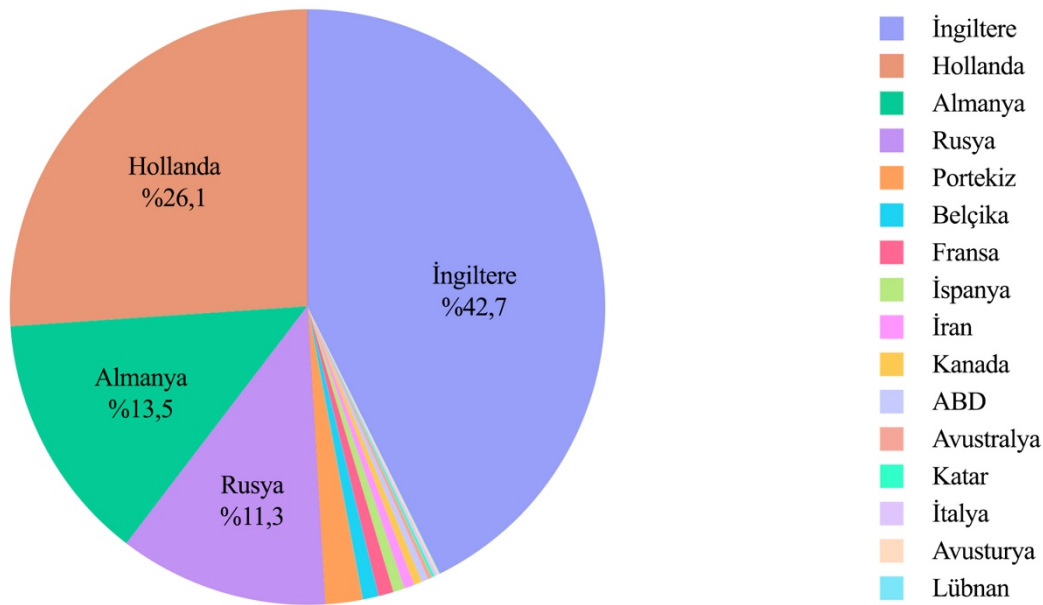
Rezervasyon yapan kişiler genellikle iki yetişkinden oluşmaktadır. Rezervasyon yapanların büyük çoğunluğunu bebek ve çocukları bulunmamaktadır. Tüm rezervasyon yapan kişilerden %16'sı rezervasyon sürecinde özel bir istekte bulunmuştur. Ortalama toplam fiyat 1.534 Euro'dur. Türkiye'ye sağlık turizmi kapsamında gelen sağlık turistleri ortalama iki hafta, en az 8 gece konaklamaktadır. Rezervasyon yapanlar genellikle bir tane rezervasyon yapmıştır. Ortalama tedavi fiyatı 2.378 Euro'dur. Sağlık turistlerinin ortalama hasta yaşı 45, en geç hasta 23 en yaşlı hasta ise 58 yaşındadır. Sağlık turistleri tedavi almak istediği sağlık kuruluşuna karar verdikten sonra ortalama 82 gün sonra Türkiye'ye gelecek şekilde plan yapmaktadır. Veri setinde yer alan kategorik değişkenlere ait istatistik özeti Tablo 4.3'te verilmiştir.

Tablo 4.3 Kategorik Değişkenler için İstatistik Özeti

Değişken Adı	Eşsiz	Değer	Frekans
otel_turu	2	Şehir Oteli	3584
uyruk	16	İngiltere	2177
rezervasyon_durumu	5	Çıkış	3755
otel_onodeme	1084	Yok	3630
pan_tipi	8	Kahvaltı	2266
oda_tipi	8	Standart	3054
acenta	16	Online	2723
uzmanlik_alani	7	Diş Tedavileri	3610
tedavi_tipi	35	Implant	1641
tekrarlanan_misafir	2	Hayır	4815
hasta_cinsiyet	2	Erkek	2597
kanal	5	Facebook	1428
hasta_tercih_edilen_iletisim	6	Whatsapp	2222
depozito_durumu	2	Ön ödemesiz	3631
tur_durumu	2	Dahil değil	4045

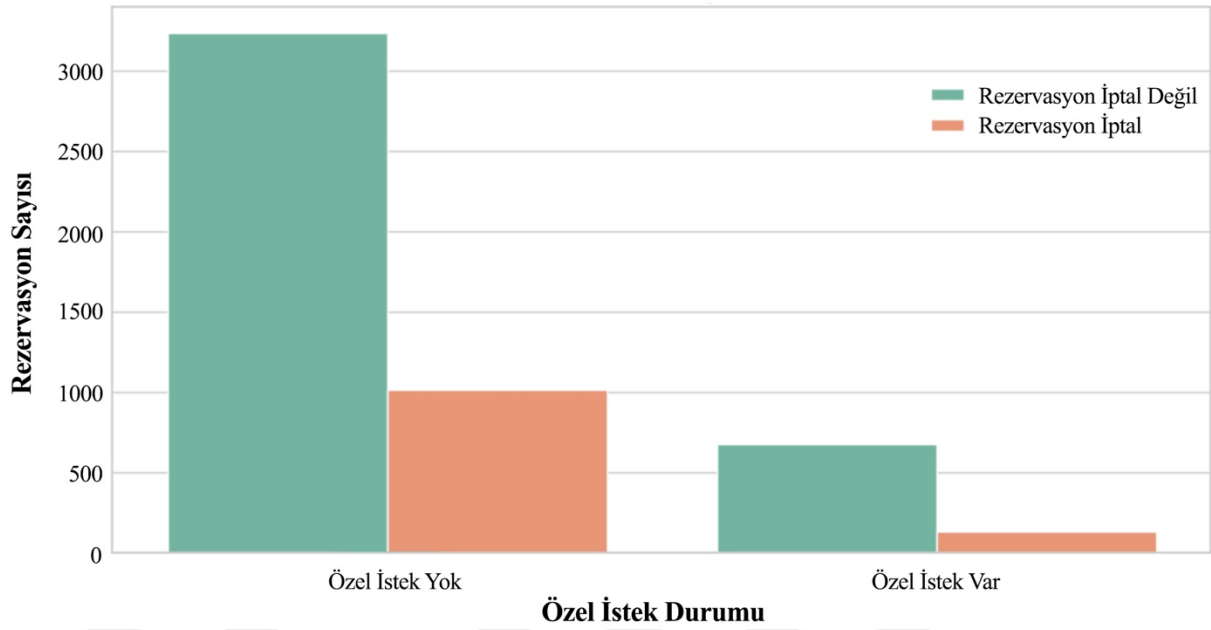
Sağlık turistleri en çok “Şehir” otellerini tercih etmektedir. 16 Farklı ülkeden sağlık turisti gelmiş olup en çok sağlık turistinin geldiği ülke “İngiltere”dir. Otel ödemeleri çoğunlukla ön ödemesiz olarak yapılmaktadır. 8 Farklı pansiyon tipi içerisinde ve çok “Kahvaltı” olarak tercih edilmektedir. 8 farklı oda tipi bulunmaktadır ve “Deluxe Room” en sık rastlanan oda tipidir. 16 Farklı acenta üzerinden rezervasyonlar yapılmış ve "Online" en çok tekrar eden değer olmuştur. “Dış Tedavi Girişimleri” yedi uzmanlık içerisinde en sık rastlanan uzmanlık alanıdır. “Dental İmplant” otuzbeş farklı tedavi tipi içerisinde en sık rastlanan tedavi tipidir. Tekrarlı olmayan sağlık turistlerine ait rezervasyonlar çoğunluktadır. Sağlık turistlerinin 2.597 tanesi erkek hastadır. "Facebook", iletişim kanallarından en sık kullanılan kanalıdır. Hastanın tercih edebileceği altı farklı iletişim tercihi bulunmaktadır. "WhatsApp" en sık tercih edilen hasta iletişim kanalıdır. Genel olarak sağlık turistlerinden ön ödeme almadan rezervasyonlar alınmaktadır. Sağlık turistleri tarafından tur hizmeti rezervasyon aşamasında genellikle tercih edilmemektedir.

iii. Verinin Keşfedilmesi: Verinin keşfedilmesi görevinde özniteliklerin dağılımı, görselleştirmesi ve raporlanması yapılmıştır. Verinin keşfedilmesi için yapılan çalışmalar, veri tanımına, kalite raporlarına ve analizler için gerekli dönüşümlere katkı sağlamaktadır.



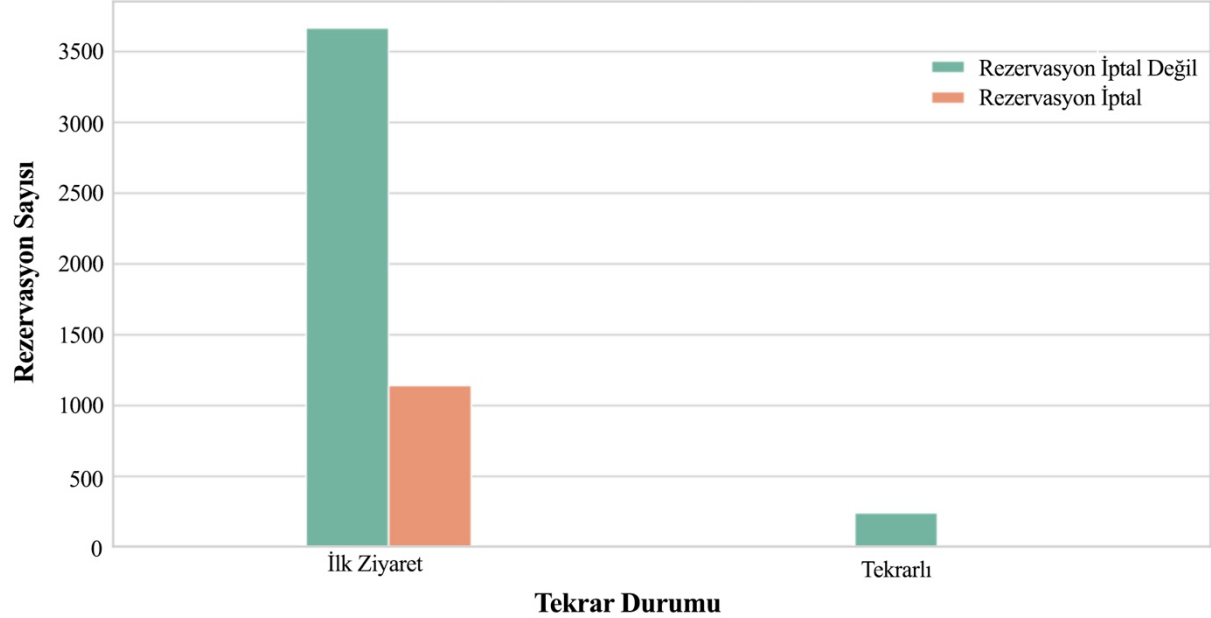
Şekil 4.1 Sağlık Turistlerinin Geldikleri Ülkelere Göre Dağılımları

Sağlık turistlerinin geldikleri ülkelere göre dağılımları Şekil 4.1’de verilmiştir. Sağlık turistleri çoğunlukla İngiltere (%42,7), Hollanda (%26,1), Almanya (%13,5), Rusya (%11,3) kaynaklı olmasına rağmen, Portekiz, Belçika, Fransa, İspanya, İran, Kanada, Amerika Birleşik Devletleri, Avustralya, Katar, İtalya, Avusturya ve Lübnan olmak üzere dünyanın dört bir yanından sağlık turistleri Türkiye’ye tedavi amacıyla gelmektedir.



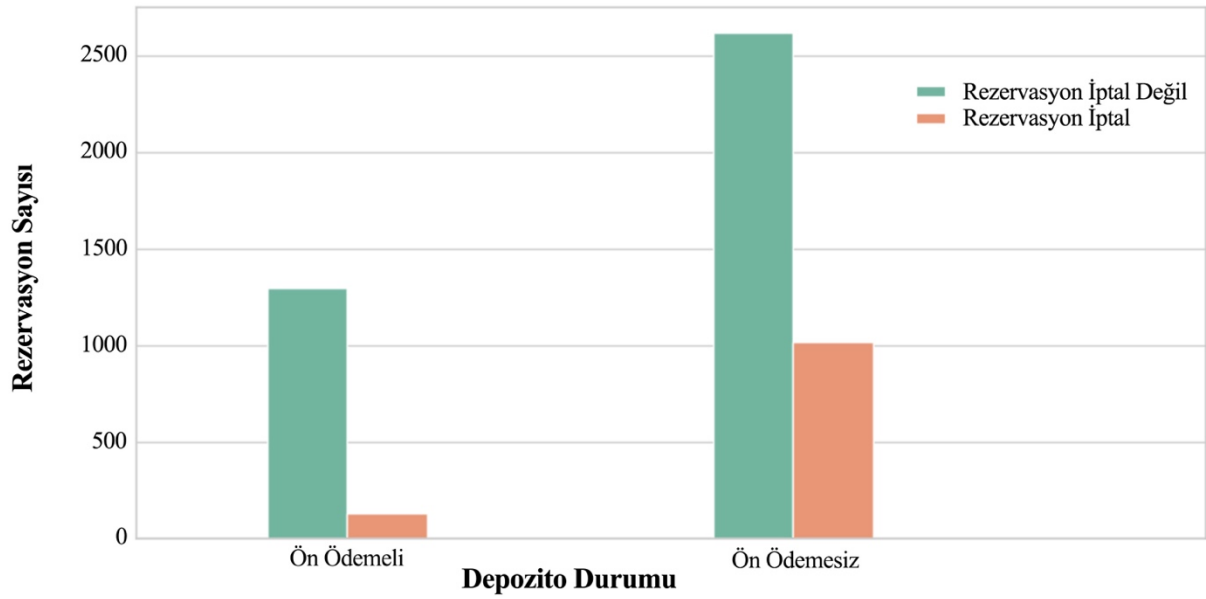
Şekil 4.2 Özel İsteklere Göre İptal Durumu

Sağlık turistlerinin yapmış olduğu 5.060 toplam rezervasyondan, 811 rezervasyonda sağlık turisti özel bir istekte bulunmuştur. Rezervasyon esnasında özel istekte bulunan 678 sağlık turisti rezervasyonunu iptal etmemiştir. Özel istekte bulunmayan 4.249 sağlık turisti içerisinde 1.015 sağlık turisti rezervasyonunu iptal ettiği Şekil 4.2’de gösterilmektedir.



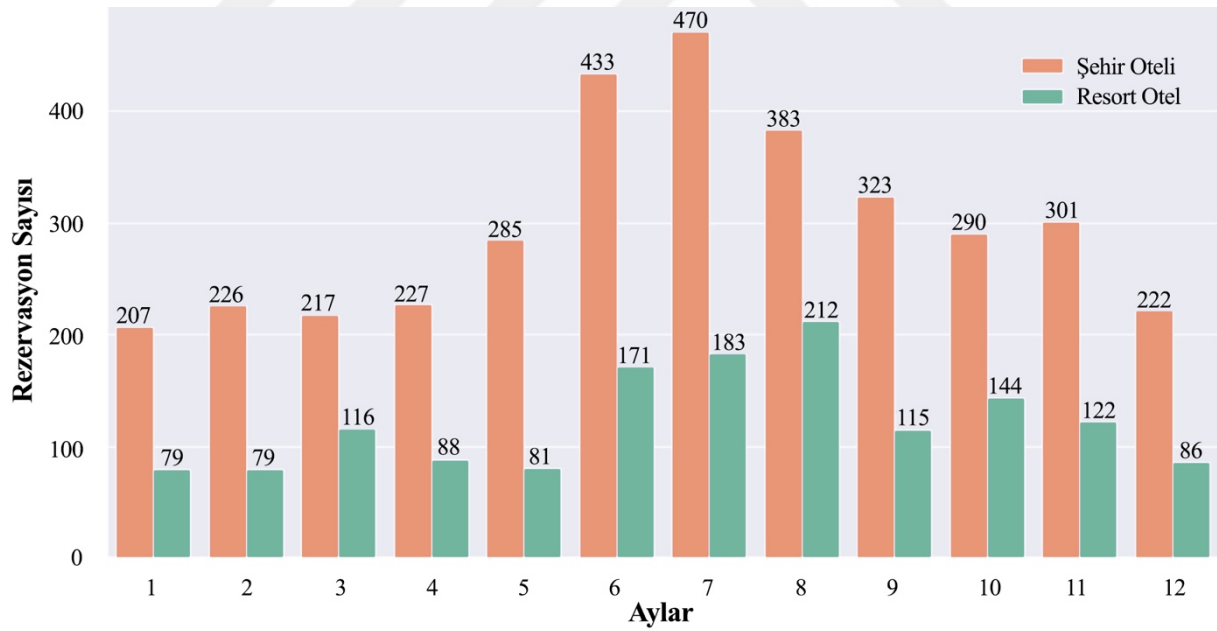
Şekil 4.3 Tekrarlı Sağlık Turistin Rezervasyon İptal Etme Durumu

Tedavi amacıyla rezervasyon yaptıran sağlık turistlerinden %95’i Türkiye’de daha önceden tedavi olmamış ve ilk kez rezervasyon yaptırmıştır. Türkiye’de daha önceden tedavi olmuş sağlık turistlerinin rezervasyonlarını iptal etmedikleri, Türkiye’de ilk defa tedavi olacak sağlık turistlerinin %24 oranında rezervasyonlarını iptal ettikleri Şekil 4.3’te gösterilmektedir.



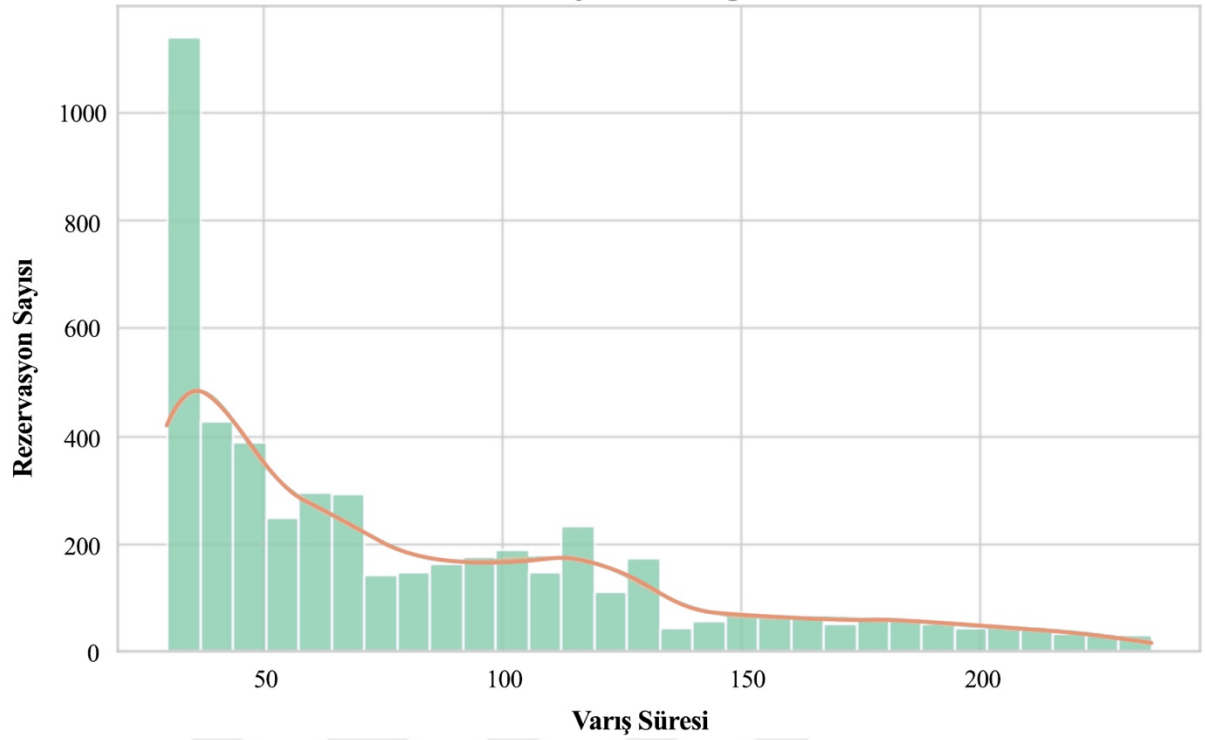
Şekil 4.4 Depozito Durumuna Göre İptal Durumu

Sağlık turizmi paydaşları, sağlık turistlerine güven verebilmek için rezervasyon sırasında genellikle ön ödeme almadan rezervasyon yapmaktadır. Sağlık turistlerinden ön ödeme alınmadan yapılan rezervasyonlar %28 oran ile iptal ettiği, ön ödeme alınan rezervasyonların %9 oranında iptal edildiği Şekil 4.4'te gösterilmektedir.



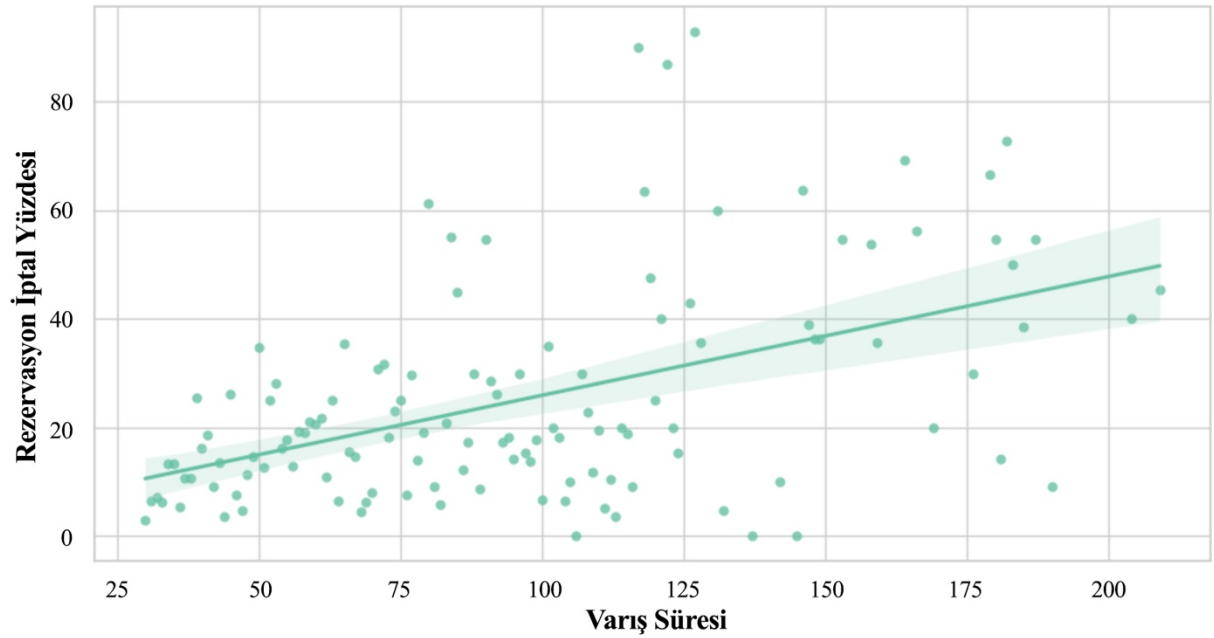
Şekil 4.5 Otel Türlerine Göre Aylık Sağlık Turisti Sayıları

Sağlık turistleri konaklama tercihini şehir otellerinden yana kullanmaktadır. Şehir otellerinin en çok tercih edildiği aylar haziran, temmuz ve ağustos ayları olmaktadır. Sağlık turizmi rezervasyonlarının aylara göre dağılımı incelendiğinde, rezervasyonların tüm aylara dağıldığı Şekil 4.5'te gösterilmektedir.



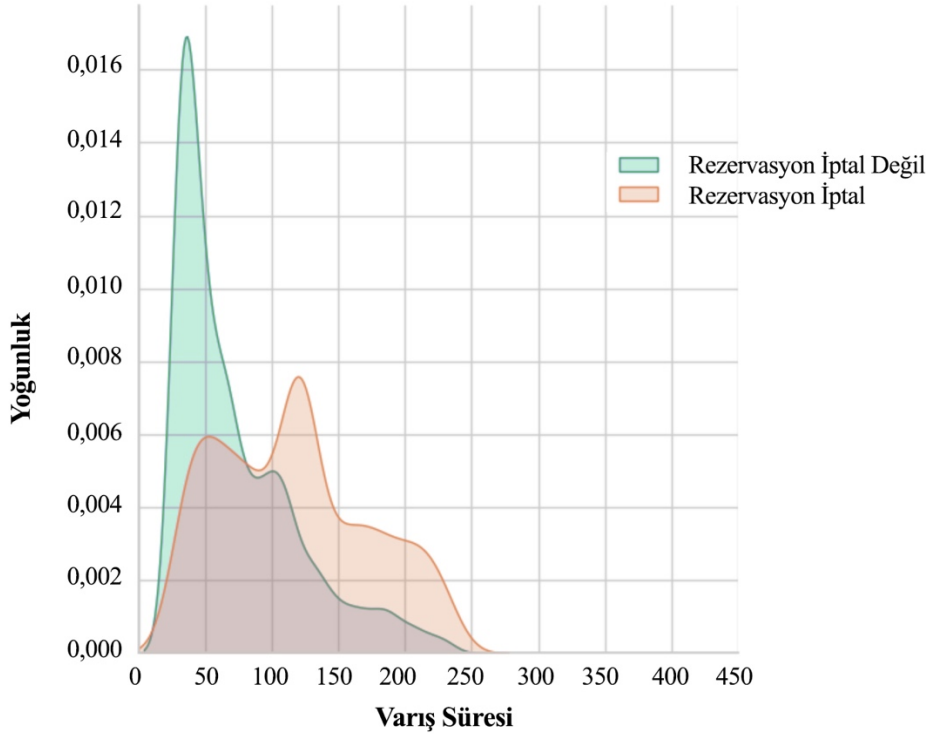
Şekil 4.6 Tedavi Tarihi ve Rezervasyon Tarihi Arasında Geçen Süre

Rezervasyonların tedaviden ne kadar süre önce yapıldığı Şekil 4.6’da gösterilmektedir. Sağlık turistleri rezervasyonlarını genellikle tedaviden üç ay önce yaptırımlarına rağmen bazı rezervasyonlar 200 gün önce bile yapılabilmektedir.



Şekil 4.7 Tedavi ve Varış Süresinin İptallere Etkisi

Tedavi olunması planlanan tarih ile rezervasyon tarihi arasındaki süre ne kadar uzun olursa, rezervasyonun iptal edilme olasılığı da Şekil 4.7’de gösterildiği gibi o kadar yüksek olmaktadır.



Şekil 4.8 Varış Süresinin Yoğunluk Dağılımı

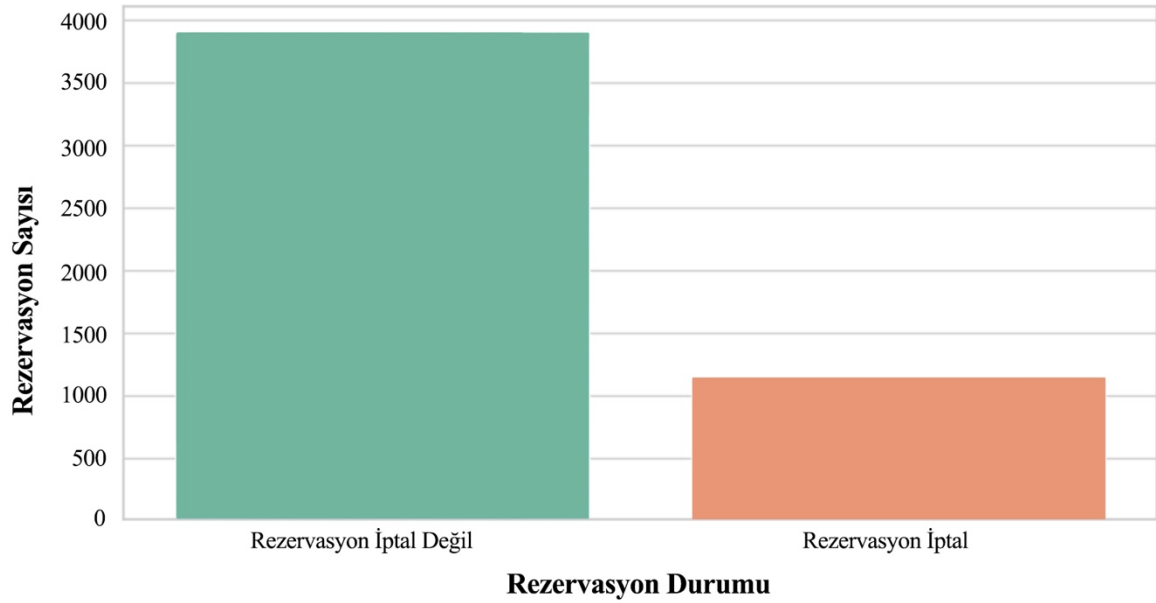
Tedavi olunması planlanan tarih ile rezervasyon tarihi arasındaki süre 60 günün fazla olduğu durumlarda, sağlık turistlerinin rezervasyonlarını iptal etme olasılıkları Şekil 4.8’de gösterildiği gibi artmaktadır.

iv. Veri kalitesinin doğrulanması: Sınıflandırma işlemleri yapan makine öğrenimi algoritmaları, hedef sınıfta yer alan değerlerin birbirine yakın olduğunu varsaymaktadır. Bazı niteliklere ait sınıflar çok daha fazla sayıda olduğu durumlarda, dağılım homojen olmamaktadır. Model, çoğunluğa sahip sınıf ile eğitildiği durumda, makine öğrenimi algoritmaları yanlış tahminler yapabilmektedir. Ayrıca, azınlık sınıf nadirliğine rağmen önemli bilgiler de içerebilmektedir. Bu nedenle, veri dengesizliklerini göz önünde bulunduran ve gerçek dünya problemlerinin çözümüne uygun tahminler yapabilen sistemler tasarlamak gerekmektedir.

Dengesiz veri setlerinde bazı nitelikler daha fazla veri içerirken, diğer nitelikler daha az veri içermektedir. Az veri içeren nitelikler, model tarafından yeterince öğrenilmemektedir. Tablo 4.4’te veri setinin sahip olduğu azınlık sınıfın yüzdesine göre veri setinin dengesizlik durumuna karar verilmektedir (Google Machine Learning, 2023).

Tablo 4.4 Azınlık Sınıfın Yüzdesine Göre Dengesiz Veri Seti Tespiti

Dengesiz Veri Seti	Azınlık Sınıfın Yüzdesi
İlımlı	Azınlık Sınıfın Toplam Veri Setinin %20 - %40 oluşturması
Orta	Azınlık Sınıfın Toplam Veri Setinin %1 - %20 oluşturması
Aşırı	Azınlık Sınıfın Toplam Veri Setinin < %1 oluşturması



Şekil 4.9 İptal Durumunun Veri Setindeki Dağılımı

Araştırmada kullanılan veri setinde iptal edilen ve iptal edilmeyen rezervasyonların dağılımlar Şekil 4.9’da gösterilmiştir. Sağlık turistlerinin yapmış olduğu 5.060 toplam rezervasyonun, %22,69 oran ile 1.148 rezervasyon iptal edilmiş, %77,31 oran ile 3.912 rezervasyon iptal edilmemiştir. Tablo 4.4’e göre, araştırma veri seti ılımlı bir dengesiz veri seti olarak, azınlık sınıfın yeterince iyi temsil edildiği değerlendirilmiştir. Veri setinde sınıfları dengelemek için aşırı örnekleme, alt örnekleme teknikleri uygulama zorunluluğu olmamasına rağmen ROSE, ADASYN, SMOTE ve SMOTE-ENN teknikleri ile veri seti homojen bir yapıya getirilmiş ve sonuçlar değerlendirme bölümünde verilmiştir.

4.6.3. Verinin Hazırlanması

Verinin hazırlanması iş sürecinde; verinin seçimi, verinin temizlenmesi, verinin yapılandırılması, verinin birleştirilmesi ve veri biçimlendirilmesi görevleri tamamlanmıştır.

i. Verinin Seçimi: Doğrudan ilişkili değişkenlerin kaldırılması ve ilgisiz değişkenlerin çıkartılması işlemleri ile veri seçimi görevi tamamlanmıştır.

Doğrudan İlişkili Değişkenlerin Kaldırılması: “rezervasyon_durumu” değişkeni bir rezervasyonun mevcut durumunu gösteren “İptal Edildi”, “Çıkış Yapıldı” ve “Gelmedi” verilerini saklayan, kategorik bir değişkendir. “rezervasyon_durumu” değişkeni, “iptal_durumu” hedef değişkeni ile doğrudan ilişkili olduğundan, bu bağımsız değişkeni tahmin modeline dahil etmek veri sızıntısına (data leakage) sebep olmaktadır. “varis_tarihi” değişkeni, misafirin otele giriş yaptığı tarihi ve “cikis_tarihi” değişkeni, misafirin otelden çıkış yaptığı tarihi saklamakta olup, rezervasyonun iptal edilmediğini dolaylı olarak göstermektedir. Tahminlerin tutarlı olması ve etkin bir model oluşturabilmek için

“tedavi_fiyati” değişkeni, “satis_fiyati” değişkeninin indirim uygulanmadan önceki değerini ifade etmektedir. Bu nedenle “tedavi_fiyati” değişkeni modele dahil edilmemiştir. “toplam_fiyat” değişkeni, “gecelik_oda_fiyati” değişkeni arasında yüksek korelasyon bulunduğu için “toplam_fiyat” değişkeni modele dahil edilmemiştir.

ii. Verinin temizlenmesi: Verinin temizlenmesi iş sürecinde eksik verilerin tespiti ve aykırı verilerin temizlenmesi görevleri tamamlanmıştır.

Eksik Verilerin Tespiti: Veri setinde yer alan “toplam_fiyat” ve “ozel_istek” bağımsız değişkenlerinde eksik veriler bulunmaktadır. “ozel_istek” değişkeninde 4.245 adet özel istek bulunmadığını ifade eden 0 değeri, 811 adet özel istek sayısını ifade eden 1 değeri bulunmaktadır. Tüm rezervasyon verileri içerisinde %0,0791 yüzde ile 4 adet “ozel_istek” verisi eksiktir. “ozel_istek” değişkenin aldığı değerlerin modu 0 olduğu için, “ozel_istek” değişkeni 0 değeri ile doldurulmuştur. “toplam_fiyat” değişkeninde %0,5138 oranında eksik veri yer almaktadır. “toplam_fiyat” değişkeninde eksik olan veriler, “gecelik_oda_fiyati” ve “gece_sayisi” değişkenlerinin çarpımı ile doldurulmuştur.

Aykırı Verilerin Temizlenmesi: “bebek_sayisi” değişkeninde 16, ve “oda_sayisi” niteliğinde 5 adet aykırı değer tespit edilmiş ve aykırı değerler veri setinden temizlenmiştir.

iii. Verinin Yapılandırılması: Yeni değişkenlerin üretilmesi, tamamen yeni kayıtlar elde edilmesi veya mevcut değişkenlerin dönüştürülmesi gibi veri yapılandırıcı faaliyetlerini bu aşamada gerçekleştirilmiştir. “varis_tarihi” değişkeni kullanılarak; “varis_gun”, “varis_ay”, “varis_yil”, “varis_tarihi_ayin_gunu”, “varis_gunu_hafta_numarasi”, “varis_gunu_numarasi” ve değişkenleri elde edilmiştir. “olusturulma_tarihi” değişkeni kullanılarak, “olusturulma_gun”, “olusturulma_ay”, “olusturulma_yil” değişkenleri elde edilmiştir. “cikis_tarihi” değişkeninden “cikis_gun”, “cikis_ay”, “cikis_yil” değişkenleri türetilmiştir. “varis_tarihi” değişkeni ile “olusturulma_tarihi” değişkeni arasında geçen sürenin gün olarak hesaplanarak değerlerin saklanması için “varis_suresi” değişkeni, “haftaici_gunler”, “haftasonu_gunler” değişkenleri oluşturulmuştur.

iv. Verilerin Birleştirilmesi: Farklı veri kaynaklardan toplanan verilerin birleştirilmesi aşamasında uyumsuzlukların ortadan kaldırılması ve verilerin birleştirilmesi gerekmektedir. Aynı rezervasyona ait birden fazla tabloda yer alan verilerin bir araya getirilmesi ile birleştirilmiş veri oluşturulmuştur. Bir rezervasyona ait cinsiyet, doğum tarihi, geldiği ülke vb. kullanıcı verilerinin yer aldığı kullanıcılar tablosu ile uzmanlık, tedavi süresi, tedavi ücreti vb. tedavi verilerinin yer aldığı tedaviler tablosundan gelen veriler birleştirilmiştir.

v. Verilerin Biçimlendirilmesi: Verilerde yapılan ve verinin anlamını değiştirmeyen ancak modelleme aracının gerektirdiği değişiklikler yapılmıştır. yer almaktadır. Yapay sınır

ağları, kategorik veriler için çalışmadığından dolayı, “otel_turu”, “uyruk”, “acentas”, “uzmanlik_alani”, “tekrarlanan_misafir”, “kanal”, “tur_durumu” vb. kategorik değişkenler One-Hot Encoding kullanılarak numerik hâle dönüştürülmüştür.

4.6.4. Modelleme

Verinin modelleme için hazır hale gelmesinden sonra, modelleme tekniğinin seçimi, test tasarımının üretilmesi, modelin yapılandırılması ve modelin değerlendirilmesi işlemleri yapılmaktadır.

i. Modelleme Tekniğinin Seçimi: Modelleme aşamasında veri madenciliği yöntemlerinden ZeroR, Naive Bayes, KNN, Karar Ağaçları, Lojistik Regresyon, Destek Vektör Makineleri, Yapay Sinir Ağları, Rastgele Orman, XGBoost, AdaBoost, CatBoost ve LightGBM olmak üzere 12 modelleme tekniği kullanılmıştır.

ii. Test Tasarımının Üretilmesi: Bir model oluşturmadan önce modelin kalite ve geçerliliğini test etmek amacıyla bir mekanizma veya prosedür oluşturulması gerekmektedir. Veri setinde bulunan 5.060 adet rezervasyon verisi %70 eğitim verisi (3.542) ve %30 test verisi (1.518) olmak üzere iki bölüme ayrılmıştır. Tahmin modellerini eğitmek için 3.542 adet rezervasyon verisi kullanılmıştır. Tahmin modelinin eğitilmesi aşamasında hiç göremediği 1.518 adet test verisi, modelin sınıflandırma performansını tespit etmek için kullanılmıştır. Veri seti, eğitim veri kümesi (training dataset) ve test veri kümesi (test dataset) olarak ikiye bölünürken, veri setinde mevcut dengesizlik oranında eğitim ve test kümelerine eşit dağılması gerekmektedir. Eğitim ve test veri setlerinde, sınıf oranlarının benzerliği, model performansının istikrarlı olması bakımından önemlidir. Bu bağlamda veri seti eğitim ve test veri setlerine bölünmesi işleminde kullanılan `train_test_split()` fonksiyonunda yer alan “stratify” parametresi ile katmanlama (stratification) yapılmıştır.

iii. Modelin Yapılandırılması: Eğitim veri seti üzerinde seçilen modelleme teknikleri varsayılan parametreleri ile çalıştırılmıştır. Modellerin varsayılan parametreleri ile çalıştırılması, tahmin modelinin performansını tam olarak gösteremeyeceğinden, KFold Cross Validation 5 parametresi kullanılarak uygulanmış ve en iyi başarıma sahip seçenek üzerinde grid search uygulanarak optimum parametreler tespit edilmiştir. Bu parametreler ile en yüksek başarıma sahip modellerin “tune” edilmiş versiyonları elde edilmiştir.

iv. Modelin Değerlendirilmesi: Tahmin modelleri test veri seti kullanılarak, karşılaştırma metriklerine göre değerlendirilmiştir. Modellerin karışıklık matrisleri, modele en çok katkı sağlayan bağımsız değişkenleri, her bir model altında verilmiştir. Bölüm sonunda tüm modellere ait karşılaştırma tablosu verilmiştir.

4.6.4.1. ZeroR

Tahmin gücü bulunmayan ve baskın sınıfa üye sayısının, tüm veri seti sayısına oranı ile doğruluk değeri hesaplana ZeroR algoritması diğer yöntemlerin kıyaslanması için referans değer oluşturmaktadır. Sağlık turistlerinin yapmış olduğu toplam rezervasyon sayısı 5.060 olup, %22,69 oran ile 1.148 rezervasyon iptal edilmiş, %77,31 oran ile 3.912 rezervasyon iptal edilmemiştir. ZeroR algoritmasının doğruluk değeri %77,31'dir.

4.6.4.2. Naive Bayes

Naive Bayes modelinin test verilerine ait karışıklık matrisi Tablo 4.5'te verilmiştir. 1.518 adet rezervasyon test verisi üzerinde yapılan tahminlere göre 298 adet gerçek "Rezervasyon iptal" değeri, "Rezervasyon iptal" şeklinde başarılı bir şekilde sınıflandırılmıştır. 594 adet gerçek "Rezervasyon iptal değil" değeri, "Rezervasyon iptal değil" şeklinde başarılı bir şekilde sınıflandırılmıştır.

580 adet gerçek "Rezervasyon iptal değeri" değeri, "Rezervasyon iptal" şeklinde hatalı tahmin edilmiştir. 46 adet gerçek "Rezervasyon iptal" değeri, "Rezervasyon iptal değil" şeklinde hatalı tahmin edilmiştir.

Tablo 4.5 Naive Bayes için Karışıklık Matrisi

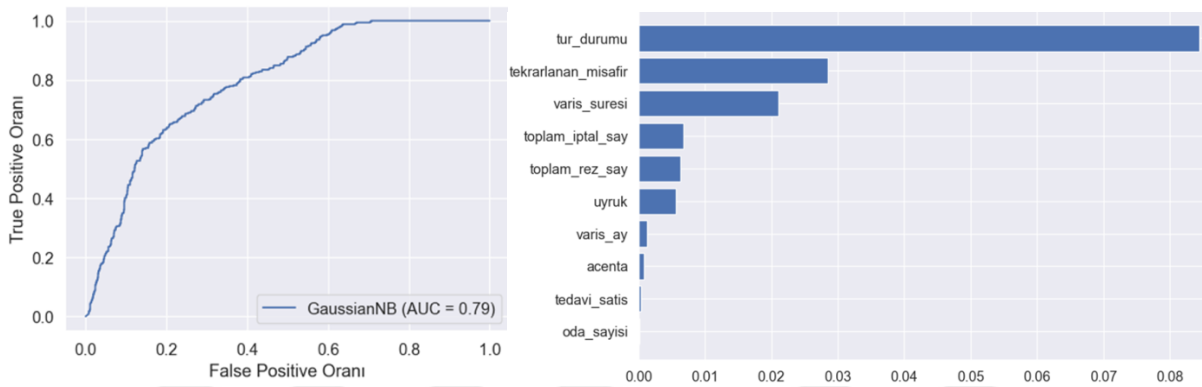
		Tahmin Değeri			
		İptal	İptal Değil		
Gerçek	İptal	TP = 298	FN = 46	TPR = 0,8663	FNR = 0,1337
	İptal Değil	FP = 580	TN = 594	TNR = 0,5060	FPR = 0,4940
		PPV = 0,3394	NPV = 0,9281	Doğruluk = 0,5876	F1 = 0,4877

Naive Bayes sınıflandırıcısının 0,5876 doğruluk değeri ile, tüm rezervasyonların %59 oran ile "Rezervasyon iptal" veya "Rezervasyon iptal değil" olarak doğru şekilde tahmin edilmektedir. Sınıflandırıcının 0,3394 kesinlik değeri, sınıflandırıcının bir rezervasyonu "Rezervasyon iptal" olarak tahmin etmesi durumunda %34 oran ile doğru olduğu anlamına gelmektedir. Başka bir ifadeyle, "Rezervasyon iptal" olarak sınıflandırılan rezervasyonlar %34 oran ile gerçekten "Rezervasyon iptal" değerine sahiptir.

NB sınıflandırıcısının 0,8663 hassasiyet değeri, veri kümesindeki tüm gerçek "Rezervasyon iptal" değerlerinin %87 oran ile doğru bir şekilde "Rezervasyon iptal" olarak tanımlandığını göstermektedir. NB sınıflandırıcısının 0,5060 özgüllük değeri, veri kümesinde yer alan tüm gerçek "Rezervasyon iptal değil" rezervasyonların %51 oran ile "Rezervasyon iptal değil" olarak doğru bir şekilde tanımlandığı anlamına gelmektedir.

Sınıflandırıcının F1-skorunun 0,4877 gibi bir düşük bir değer olması, “Rezervasyon iptal” olan rezervasyonların doğru şekilde tanımlanması ile hatalı pozitif sonuçların önlenmesi arasında ortalama bir denge olduğunu göstermektedir. Naive Bayes sınıflandırıcının FPR değerinin 0,494 olması, gerçek değeri “Rezervasyon iptal değil” olan rezervasyonların %50 oran ile hatalı bir şekilde “Rezervasyon iptal” olarak sınıflandırıldığı anlamına gelmektedir.

Sınıflandırıcının FNR değerinin 0,1337 olması, gerçek değeri “Rezervasyon iptal” olan rezervasyonların %1,33 oran ile hatalı bir şekilde “Rezervasyon iptal değil” olarak sınıflandırıldığı anlamına gelmektedir. Sınıflandırıcının NPV değerinin 0,9281 olması, “Rezervasyon iptal değil” tahmin edilen rezervasyonların %93 oranı ile gerçekten “Rezervasyon iptal değil” olduğunu göstermektedir.



Şekil 4.11 Naive Bayes ROC ve En Çok Katkı Sağlayan Değişkenler

Eğri altında kalan alanın (AUC) değeri Şekil 4.11’de gösterildiği gibi 0,79 olup, Naive Bayes sınıflandırıcısının doğru, güvenilir sayılabilecek bir test sonucunu ve test grubu için ortalama bir ayrımı ifade etmektedir. Naive Bayes tahmin modelinde en çok katkırı sağlayan üç özellik “tur_durumu”, “tekrarlanan_misafir” ve “varis_suresi” bağımsız değişkenleridir.

4.6.4.3. K-En Yakın Komşu

KNN sınıflandırıcısının test verilerine ait karışıklık matrisi Tablo 4.6’da verilmiştir. 1.518 adet rezervasyon test verisi üzerinde yapılan tahminlere göre 152 adet gerçek “Rezervasyon iptal” değeri, “Rezervasyon iptal” şeklinde başarılı bir şekilde sınıflandırılmıştır. 1.077 adet gerçek “Rezervasyon iptal değil” değeri, “Rezervasyon iptal değil” şeklinde başarılı bir şekilde sınıflandırılmıştır.

97 adet gerçek “Rezervasyon iptal değeri” değeri, “Rezervasyon iptal” şeklinde hatalı tahmin edilmiştir. 192 adet gerçek “Rezervasyon iptal” değeri, “Rezervasyon iptal değil” şeklinde hatalı tahmin edilmiştir.

Tablo 4.6 KNN için Karışıklık Matrisi

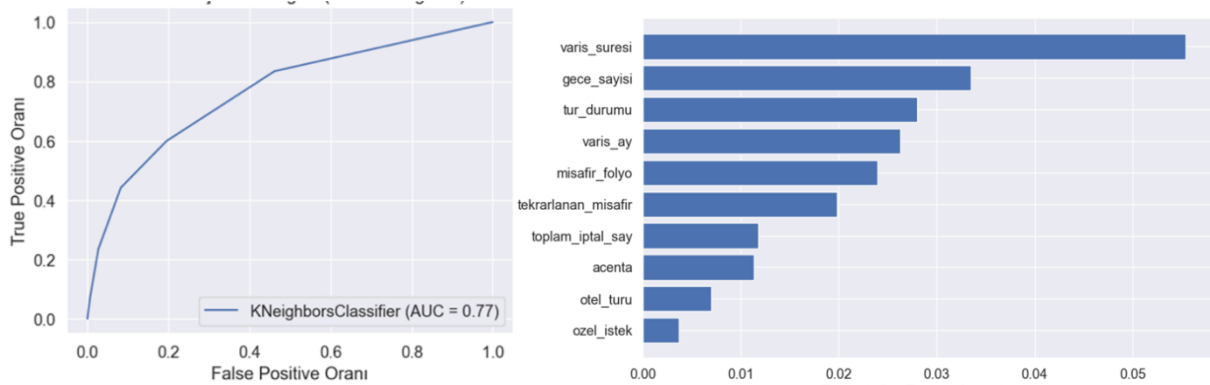
		Tahmin Değeri			
		İptal	İptal Değil		
Gerçek	İptal	TP = 152	FN = 192	TPR = 0,4419	FNR = 0,5581
	İptal Değil	FP = 97	TN = 1.077	TNR = 0,9174	FPR = 0,0826
		PPV = 0,6104	NPV = 0,8487	Doğruluk = 0,8096	F1 = 0,5126

KNN sınıflandırıcısının 0,8096 doğruluk değeri ile, tüm rezervasyonların %81 oran ile “Rezervasyon iptal” veya “Rezervasyon iptal değil” olarak doğru şekilde tahmin edilmektedir. KNN sınıflandırıcısının 0,6104 kesinlik değeri, sınıflandırıcının bir rezervasyonu “Rezervasyon iptal” olarak tahmin etmesi durumunda %61 oran ile doğru olduğu anlamına gelmektedir.

Sınıflandırıcının 0,4419 hassasiyet değeri, veri kümesindeki tüm gerçek “Rezervasyon iptal” değerlerinin %44 oran ile doğru bir şekilde “Rezervasyon iptal” olarak tanımladığını göstermektedir. KNN sınıflandırıcının özgüllük değerinin 0,9174 olması, veri kümesindeki tüm gerçek “Rezervasyon iptal değil” rezervasyonların %92 oran ile “Rezervasyon iptal değil” olarak doğru bir şekilde tanımlandığı anlamına gelmektedir.

Sınıflandırıcının F1 puanının 0,5126 gibi bir ortalama bir değer olması, “Rezervasyon iptal” olan rezervasyonların doğru şekilde tanımlanması ile hatalı pozitif sonuçların önlenmesi arasında ortalama bir denge olduğunu göstermektedir. KNN sınıflandırıcının FPR değerinin 0,0826 olması, gerçek değeri “Rezervasyon iptal değil” olan rezervasyonların %8 oran ile hatalı bir şekilde “Rezervasyon iptal” olarak sınıflandırıldığı anlamına gelmektedir.

Sınıflandırıcının FNR değerinin 0,5581 olması, gerçek değeri “Rezervasyon iptal” olan rezervasyonların %56 oran ile hatalı bir şekilde “Rezervasyon iptal değil” olarak sınıflandırıldığı anlamına gelmektedir. Sınıflandırıcının NPV değerinin 0,8487 olması, “Rezervasyon iptal değil” tahmin edilen rezervasyonların %85 oran ile gerçekten “Rezervasyon iptal değil” olduğunu göstermektedir.



Şekil 4.12 KNN için ROC ve En Çok Katkı Sağlayan Değişkenler

Eğri altında kalan alanın (AUC) değeri Şekil 4.12’de gösterildiği gibi 0,77 olup, KNN sınıflandırıcısının doğru, güvenilir sayılabilecek bir test sonucunu ve test grubu için ortalama bir ayrımı ifade etmektedir. KNN tahmin modelinde en çok katkıyı sağlayan özellikler “varis_suresi”, “gece_sayisi” ve “tur_durumu” bağımsız değişkenleridir.

4.6.4.4. Destek Vektör Makineleri

DVM sınıflandırıcısının test verilerine ait karışıklık matrisi Tablo 4.7’de verilmiştir. 1.518 adet rezervasyon test verisi üzerinde yapılan tahminlere göre 103 adet gerçek “Rezervasyon iptal” değeri, “Rezervasyon iptal” şeklinde başarılı bir şekilde sınıflandırılmıştır. 1.147 adet gerçek “Rezervasyon iptal değil” değeri, “Rezervasyon iptal değil” şeklinde başarılı bir şekilde sınıflandırılmıştır.

27 adet gerçek “Rezervasyon iptal değeri” değeri, “Rezervasyon iptal” şeklinde hatalı tahmin edilmiştir. 241 adet gerçek “Rezervasyon iptal” değeri, “Rezervasyon iptal değil” şeklinde hatalı tahmin edilmiştir.

Tablo 4.7 DVM için Karışıklık Matrisi

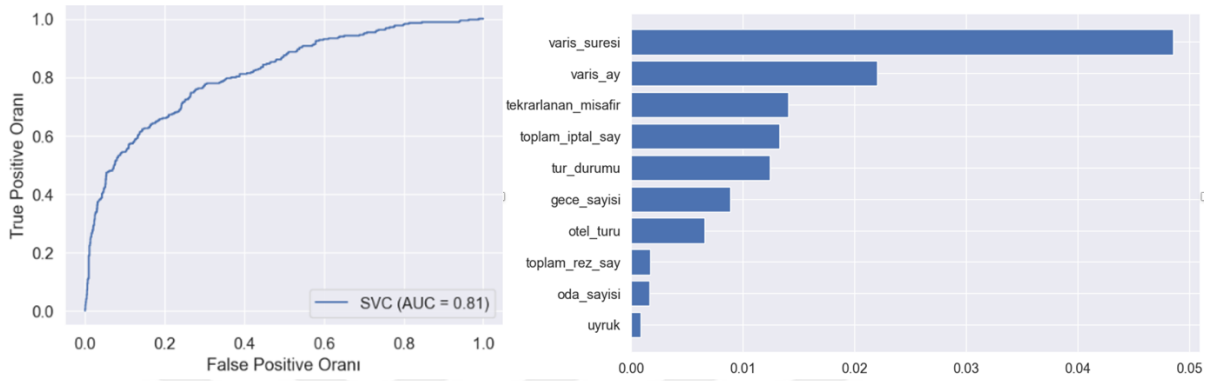
		Tahmin Değeri			
		İptal	İptal Değil		
Gerçek	İptal	TP = 103	FN = 241	TPR = 0,2994	FNR = 0,7006
	İptal Değil	FP = 27	TN = 1.147	TNR = 0,9770	FPR = 0,0230
		PPV = 0,7923	NPV = 0,8264	Doğruluk = 0,8235	F1 = 0,4346

DVM sınıflandırıcısının 0,8235 doğruluk değeri ile, tüm rezervasyonlar %82 oranında “Rezervasyon iptal” veya “Rezervasyon iptal değil” olarak doğru şekilde tahmin edilmektedir. Sınıflandırıcının 0,7923 kesinlik değeri, sınıflandırıcının bir rezervasyonu “Rezervasyon iptal” olarak tahmin etmesi durumunda %79 oran ile doğru olduğu anlamına gelmektedir. Başka bir ifade ile, “Rezervasyon iptal” olarak sınıflandırılan rezervasyonlar %79 oran ile gerçekten “Rezervasyon iptal” değerine sahiptir.

DVM sınıflandırıcısının 0,2994 hassasiyet değeri, veri kümesindeki tüm gerçek “Rezervasyon iptal” değerlerinden %30 oran ile doğru bir şekilde “Rezervasyon iptal” olarak tanımladığını göstermektedir. Sınıflandırıcının özgüllük değerinin 0,9770 olması, veri kümesindeki tüm gerçek “Rezervasyon iptal değil” rezervasyonların %98 oran ile “Rezervasyon iptal değil” olarak doğru bir şekilde tanımlandığı anlamına gelmektedir.

Sınıflandırıcının F1 puanının 0,4346 gibi bir ortalama bir değer olması, “Rezervasyon iptal” olan rezervasyonların doğru şekilde tanımlanması ile hatalı pozitif sonuçların önlenmesi arasında ortalama bir denge olduğunu göstermektedir. DVM sınıflandırıcının FPR değerinin 0,0230 olması, gerçek değeri “Rezervasyon iptal değil” olan rezervasyonların %2,3 oran ile

hatalı bir şekilde “Rezervasyon iptal” olarak sınıflandırıldığı anlamına gelmektedir. Sınıflandırıcının FNR değerinin 0,7006 olması, gerçek değeri “Rezervasyon iptal” olan rezervasyonların %70 oran ile hatalı bir şekilde “Rezervasyon iptal değil” olarak sınıflandırıldığı anlamına gelmektedir. Sınıflandırıcının NPV değerinin 0,8264 olması, “Rezervasyon iptal değil” tahmin edilen rezervasyonların %83 oran ile gerçekten “Rezervasyon iptal değil” olduğunu göstermektedir.



Şekil 4.13 DVM için ROC ve En Çok Katkı Sağlayan Değişkenler

Eğri altında kalan alanın (AUC) değeri Şekil 4.13’te gösterildiği gibi 0,81 olup, DVM sınıflandırıcısının doğru, güvenilir sayılabilecek bir test sonucunu ve test grubu için iyi bir ayrımı ifade etmektedir. DVM tahmin modelinde en çok katkısı sağlayan özellikler “varis_suresi”, “varis_ay” ve “tekrarlanan_misafir” bağımsız değişkenleridir.

4.6.4.5. Lojistik Regresyon

LR sınıflandırıcının test verilerine ait karışıklık matrisi Tablo 4.8’de verilmiştir. 1.518 adet rezervasyon test verisi üzerinde yapılan tahminlere göre 55 adet gerçek “Rezervasyon iptal” değeri, “Rezervasyon iptal” şeklinde başarılı bir şekilde sınıflandırılmıştır. 1.129 adet gerçek “Rezervasyon iptal değil” değeri, “Rezervasyon iptal değil” şeklinde başarılı bir şekilde sınıflandırılmıştır.

45 adet gerçek “Rezervasyon iptal değeri” değeri, “Rezervasyon iptal” şeklinde hatalı tahmin edilmiştir. 289 adet gerçek “Rezervasyon iptal” değeri, “Rezervasyon iptal değil” şeklinde hatalı tahmin edilmiştir.

Tablo 4.8 LR için Karışıklık Matrisi

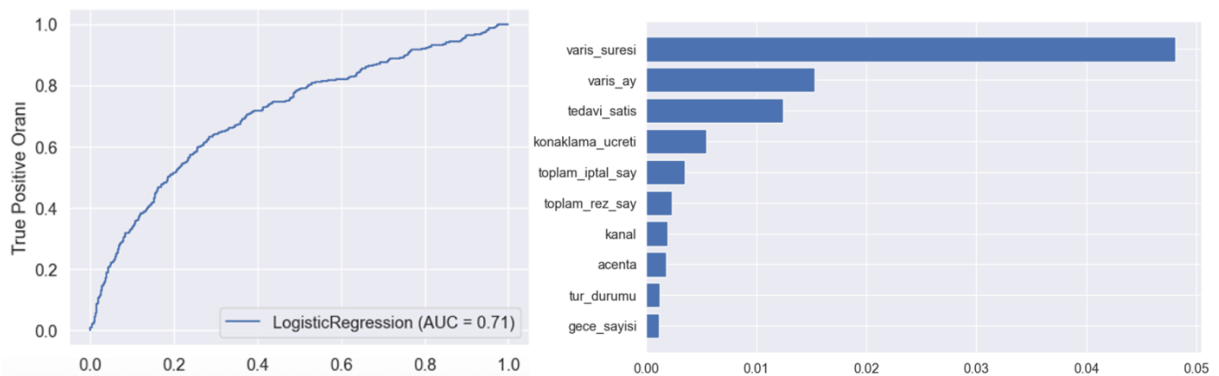
		Tahmin Değeri			
		İptal	İptal Değil		
Gerçek	İptal	TP = 55	FN = 289	TPR = 0,1599	FNR = 0,8401
	İptal Değil	FP = 45	TN = 1.129	TNR = 0,9617	FPR = 0,0383
		PPV = 0,5500	NPV = 0,7962	Doğruluk = 0,7800	F1 = 0,2477

LR sınıflandırıcısının 0,78 doğruluk değeri ile, tüm rezervasyonların %78 oran ile “Rezervasyon iptal” veya “Rezervasyon iptal değil” olarak doğru şekilde tahmin edilmektedir. Sınıflandırıcının 0,55 kesinlik değeri, sınıflandırıcının bir rezervasyonu “Rezervasyon iptal” olarak tahmin etmesi durumundan bunun %55 oran ile doğru olduğu anlamına gelmektedir. Başka bir ifade ile, “Rezervasyon iptal” olarak sınıflandırılan rezervasyonların %55 oran ile gerçekten “Rezervasyon iptal” değerine sahiptir.

Sınıflandırıcının 0,1599 hassasiyet değeri, veri kümesindeki tüm gerçek “Rezervasyon iptal” değerlerinden %16 oran ile doğru bir şekilde “Rezervasyon iptal” olarak tanımladığını göstermektedir. Sınıflandırıcının özgüllük değerinin 0,9617 olması, veri kümesindeki tüm gerçek “Rezervasyon iptal değil” rezervasyonların %96 oran ile “Rezervasyon iptal değil” olarak doğru bir şekilde tanımlandığı anlamına gelmektedir.

Sınıflandırıcının F1 puanının 0,2477 gibi bir düşük bir değer olması, “Rezervasyon iptal” olan rezervasyonların doğru şekilde tanımlanması ile hatalı pozitif sonuçların önlenmesi arasında kötü bir denge olduğunu göstermektedir. LR sınıflandırıcının FPR değerinin 0,0383 olması, gerçek değeri “Rezervasyon iptal değil” olan rezervasyonların %3,8 oran ile hatalı bir şekilde “Rezervasyon iptal” olarak sınıflandırıldığı anlamına gelmektedir.

Sınıflandırıcının FNR değerinin 0,8401 olması, gerçek değeri “Rezervasyon iptal” olan rezervasyonların %84 oran ile hatalı bir şekilde “Rezervasyon iptal değil” olarak sınıflandırıldığı anlamına gelmektedir. Sınıflandırıcının NPV değerinin 0,7962 olması, “Rezervasyon iptal değil” tahmin edilen rezervasyonların %80 oran ile gerçekten “Rezervasyon iptal değil” olduğunu göstermektedir.



Şekil 4.14 LR için ROC ve En Çok Katkı Sağlayan Değişkenler

Eğri altında kalan alanın (AUC) değeri Şekil 4.14’te gösterildiği gibi 0,71 olup, LR sınıflandırıcısının doğru, güvenilir sayılabilecek bir test sonucunu ve test grubu için kötü bir ayrımı ifade etmektedir. LR tahmin modelinde en çok katkıyı sağlayan özellikler “varis_suresi”, “varis_ay”, “tedavi satis” ve “konaklama_ucreti” bağımsız değişkenleridir.

4.6.4.6. Karar Ağaçları

Karar ağaçları sınıflandırıcısının test verilerine ait karışıklık matrisi Tablo 4.9’da verilmiştir. 1.518 adet rezervasyon test verisi üzerinde yapılan tahminlere göre 184 adet gerçek “Rezervasyon iptal” değeri, “Rezervasyon iptal” şeklinde başarılı bir şekilde sınıflandırılmıştır. 1.065 adet gerçek “Rezervasyon iptal değil” değeri, “Rezervasyon iptal değil” şeklinde başarılı bir şekilde sınıflandırılmıştır.

109 adet gerçek “Rezervasyon iptal değeri” değeri, “Rezervasyon iptal” şeklinde hatalı tahmin edilmiştir. 160 adet gerçek “Rezervasyon iptal” değeri, “Rezervasyon iptal değil” şeklinde hatalı tahmin edilmiştir.

Tablo 4.9 Karar Ağaçları için Karışıklık Matrisi

		Tahmin Değeri			
		İptal	İptal Değil		
Gerçek	İptal	TP = 184	FN = 160	TPR = 0,5349	FNR = 0,4651
	İptal Değil	FP = 109	TN = 1.065	TNR = 0,9072	FPR = 0,0928
		PPV = 0,6280	NPV = 0,8694	Doğruluk = 0,8228	F1 = 0,5777

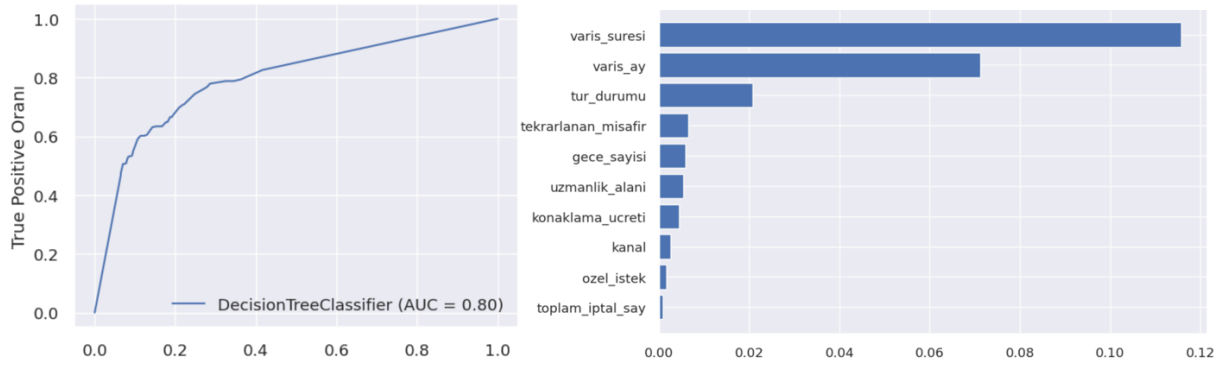
Karar ağaçları sınıflandırıcısının 0,8228 doğruluk değeri ile, tüm rezervasyonların %82 oran ile “Rezervasyon iptal” veya “Rezervasyon iptal değil” olarak doğru şekilde tahmin edilmektedir. Sınıflandırıcısının 0,6280 kesinlik değeri, sınıflandırıcısının bir rezervasyonu “Rezervasyon iptal” olarak tahmin etmesi durumundan %63 oran ile doğru olduğu anlamına gelmektedir. Başka bir ifade ile, “Rezervasyon iptal” olarak sınıflandırılan rezervasyonların %63 oran ile gerçekten “Rezervasyon iptal” değerine sahiptir.

Sınıflandırıcısının 0,5349 hassasiyet değeri, veri kümesindeki tüm gerçek “Rezervasyon iptal” değerlerinden %54 oran ile doğru bir şekilde “Rezervasyon iptal” olarak tanımladığını göstermektedir. Sınıflandırıcısının özgüllük değerinin 0,9072 olması, veri kümesindeki tüm gerçek “Rezervasyon iptal değil” rezervasyonların %91 oran ile “Rezervasyon iptal değil” olarak doğru bir şekilde tanımlandığı anlamına gelmektedir.

Sınıflandırıcısının F1 puanının 0,5777 gibi bir ortalama bir değer olması, “Rezervasyon iptal” olan rezervasyonların vasat şekilde tanımlanması ile hatalı pozitif sonuçların önlenmesi arasında ortalama bir denge olduğunu göstermektedir. Karar ağaçları sınıflandırıcısının FPR değerinin 0,0928 olması, gerçek değeri “Rezervasyon iptal değil” olan rezervasyonların %9 oran ile hatalı bir şekilde “Rezervasyon iptal” olarak sınıflandırıldığı anlamına gelmektedir.

Sınıflandırıcısının FNR değerinin 0,4651 olması, gerçek değeri “Rezervasyon iptal” olan rezervasyonların %47 oran ile hatalı bir şekilde “Rezervasyon iptal değil” olarak sınıflandırıldığı anlamına gelmektedir. Sınıflandırıcısının NPV değerinin 0,8694 olması,

“Rezervasyon iptal değil” tahmin edilen rezervasyonların %87 oran ile gerçekten “Rezervasyon iptal değil” olduğunu göstermektedir.



Şekil 4.15 Karar Ağaçları için ROC ve En Çok Katkı Sağlayan Değişkenler

Eğri altında kalan alanın (AUC) değeri Şekil 4.15’te gösterildiği gibi 0,80 olup, karar ağaçları sınıflandırıcısının doğru, güvenilir sayılabilecek bir test sonucunu ve test grubu için ortalama bir ayrımı ifade etmektedir. Karar ağaçları tahmin modelinde en çok katkıyı sağlayan özellikler “varis_suresi”, “varis_ay” ve “tur_durumu” bağımsız değişkenleridir.

4.6.4.7. Yapay Sinir Ağları

YSA sınıflandırıcısının test verilerine ait karışıklık matrisi Tablo 4.10’da verilmiştir. 1.518 adet rezervasyon test verisi üzerinde yapılan tahminlere göre 163 adet gerçek “Rezervasyon iptal” değeri, “Rezervasyon iptal” şeklinde başarılı bir şekilde sınıflandırılmıştır. 1.113 adet gerçek “Rezervasyon iptal değil” değeri, “Rezervasyon iptal değil” şeklinde başarılı bir şekilde sınıflandırılmıştır.

61 adet gerçek “Rezervasyon iptal değeri” değeri, “Rezervasyon iptal” şeklinde hatalı tahmin edilmiştir. 181 adet gerçek “Rezervasyon iptal” değeri, “Rezervasyon iptal değil” şeklinde hatalı tahmin edilmiştir.

Tablo 4.10 YSA için Karışıklık Matrisi

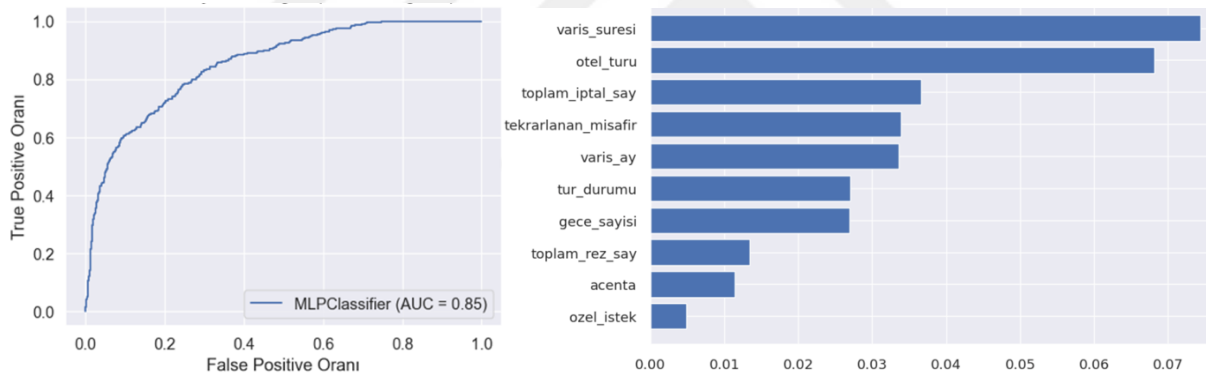
		Tahmin Değeri			
		İptal	İptal Değil		
Gerçek	İptal	TP = 163	FN = 181	TPR = 0,4738	FNR = 0,5262
	İptal Değil	FP = 61	TN = 1.113	TNR = 0,9480	FPR = 0,0520
		PPV = 0,7277	NPV = 0,8601	Doğruluk = 0,8406	F1 = 0,5739

YSA sınıflandırıcısının 0,8406 doğruluk değeri ile, tüm rezervasyonların %84 oran ile “Rezervasyon iptal” veya “Rezervasyon iptal değil” olarak doğru şekilde tahmin edilmektedir. Sınıflandırıcısının 0,7277 kesinlik değeri, sınıflandırıcısının bir rezervasyonu “Rezervasyon iptal” olarak tahmin etmesi durumunda %73 oran ile doğru olduğu anlamına gelmektedir.

Sınıflandırıcının 0,4738 hassasiyet değeri, veri kümesindeki tüm gerçek “Rezervasyon iptal” değerlerinden %47 oran ile doğru bir şekilde “Rezervasyon iptal” olarak tanımladığını göstermektedir. Sınıflandırıcının özgüllük değerinin 0,9480 olması, veri kümesindeki tüm gerçek “Rezervasyon iptal değil” rezervasyonların %95 oran ile “Rezervasyon iptal değil” olarak doğru bir şekilde tanımlandığı anlamına gelmektedir.

Sınıflandırıcının F1 puanının 0,5739 gibi bir ortalama bir değer olması, “Rezervasyon iptal” olan rezervasyonların doğru şekilde tanımlanması ile hatalı pozitif sonuçların önlenmesi arasında ortalama bir denge olduğunu göstermektedir. YSA sınıflandırıcının FPR değerinin 0,0520 olması, gerçek değeri “Rezervasyon iptal değil” olan rezervasyonların %5 oran ile hatalı bir şekilde “Rezervasyon iptal” olarak sınıflandırıldığı anlamına gelmektedir.

Sınıflandırıcının FNR değerinin 0,5262 olması, gerçek değeri “Rezervasyon iptal” olan rezervasyonların %53 oran ile hatalı bir şekilde “Rezervasyon iptal değil” olarak sınıflandırıldığı anlamına gelmektedir. Sınıflandırıcının NPV değerinin 0,8601 olması, “Rezervasyon iptal değil” tahmin edilen rezervasyonların %86 oran ile gerçekten “Rezervasyon iptal değil” olduğunu göstermektedir.



Şekil 4.16 YSA için ROC ve En Çok Katkı Sağlayan Değişkenler

Eğri altında kalan alanın (AUC) değeri Şekil 4.16’da gösterildiği gibi 0,85 olup, YSA sınıflandırıcısının doğru, güvenilir sayılabilecek bir test sonucunu ve test grubu için iyi bir ayrımı ifade etmektedir. YSA tahmin modelinde en çok katkıyı sağlayan özellikler “otel_tur”, “varis_suresi” ve “toplam iptal_say” bağımsız değişkenleridir.

4.6.4.8. Rastgele Orman

RO sınıflandırıcının test verilerine ait karışıklık matrisi Tablo 4.11’de verilmiştir. 1.518 adet rezervasyon test verisi üzerinde yapılan tahminlere göre 180 adet gerçek “Rezervasyon iptal” değeri, “Rezervasyon iptal” şeklinde başarılı bir şekilde sınıflandırılmıştır. 1.135 adet gerçek “Rezervasyon iptal değil” değeri, “Rezervasyon iptal değil” şeklinde başarılı bir şekilde sınıflandırılmıştır. 39 adet gerçek “Rezervasyon iptal

değeri” değeri, “Rezervasyon iptal” şeklinde hatalı tahmin edilmiştir. 164 adet gerçek “Rezervasyon iptal” değeri, “Rezervasyon iptal değil” şeklinde hatalı tahmin edilmiştir.

Tablo 4.11 RO için Karışıklık Matrisi

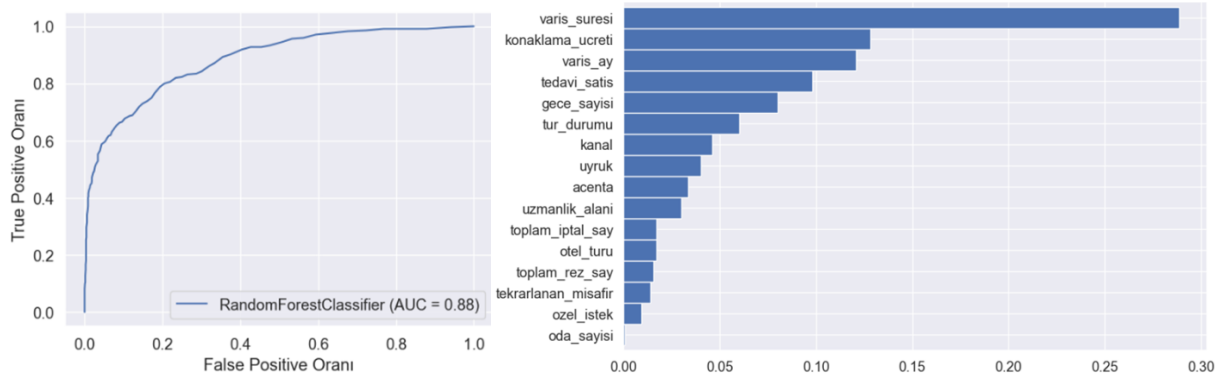
Gerçek		Tahmin Değeri			
		İptal	İptal Değil		
	İptal	TP = 180	FN = 164	TPR = 0,5233	FNR = 0,4767
	İptal Değil	FP = 39	TN = 1.135	TNR = 0,9668	FPR = 0,0332
		PPV = 0,8219	NPV = 0,8737	Doğruluk = 0,8663	F1 = 0,6394

RO sınıflandırıcısının 0,8663 doğruluk değeri ile, tüm rezervasyonların %87 oran ile “Rezervasyon iptal” veya “Rezervasyon iptal değil” olarak doğru şekilde tahmin edilmektedir. Sınıflandırıcının 0,8219 kesinlik değeri, sınıflandırıcının bir rezervasyonu “Rezervasyon iptal” olarak tahmin etmesi durumunda %82 oran ile doğru olduğu anlamına gelmektedir.

Sınıflandırıcının 0,5233 hassasiyet değeri, veri kümesindeki tüm gerçek “Rezervasyon iptal” değerlerinden %52 oran ile doğru bir şekilde “Rezervasyon iptal” olarak tanımladığını göstermektedir. Sınıflandırıcının özgüllük değerinin 0,9668 olması, veri kümesindeki tüm gerçek “Rezervasyon iptal değil” rezervasyonların %97 oran ile “Rezervasyon iptal değil” olarak doğru bir şekilde tanımlandığı anlamına gelmektedir.

Sınıflandırıcının F1 puanının 0,6394 gibi bir değer olması, “Rezervasyon iptal” olan rezervasyonların doğru şekilde tanımlanması ile hatalı pozitif sonuçların önlenmesi arasında iyi bir denge olduğunu göstermektedir. RO sınıflandırıcının FPR değerinin 0,0332 olması, gerçek değeri “Rezervasyon iptal değil” olan rezervasyonların %3 oran ile hatalı bir şekilde “Rezervasyon iptal” olarak sınıflandırıldığı anlamına gelmektedir.

Sınıflandırıcının FNR değerinin 0,4767 olması, gerçek değeri “Rezervasyon iptal” olan rezervasyonların %48 oran ile hatalı bir şekilde “Rezervasyon iptal değil” olarak sınıflandırıldığı anlamına gelmektedir. Sınıflandırıcının NPV değerinin 0,8737 olması, “Rezervasyon iptal değil” tahmin edilen rezervasyonların %87 oran ile gerçekten “Rezervasyon iptal değil” olduğunu göstermektedir.



Şekil 4.17 RO için ROC ve En Çok Katkı Sağlayan Değişkenler

Eğri altında kalan alanın (AUC) değeri Şekil 4.17’de gösterildiği gibi 0,88 olup, RO sınıflandırıcısının doğru, güvenilir sayılabilecek bir test sonucunu ve test grubu için iyi bir ayrımı ifade etmektedir. RO tahmin modelinde en çok katkıyı sağlayan özellikler “varis_suresi”, “toplam_fiyat” ve “varis_ay” bağımsız değişkenleridir.

4.6.4.9. AdaBoost

AdaBoost sınıflandırıcısının test verilerine ait karışıklık matrisi Tablo 4.12’de verilmiştir. 1.518 adet rezervasyon test verisi üzerinde yapılan tahminlere göre 164 adet gerçek “Rezervasyon iptal” değeri, “Rezervasyon iptal” şeklinde başarılı bir şekilde sınıflandırılmıştır. 1.095 adet gerçek “Rezervasyon iptal değil” değeri, “Rezervasyon iptal değil” şeklinde başarılı bir şekilde sınıflandırılmıştır.

79 adet gerçek “Rezervasyon iptal değeri” değeri, “Rezervasyon iptal” şeklinde hatalı tahmin edilmiştir. 180 adet gerçek “Rezervasyon iptal” değeri, “Rezervasyon iptal değil” şeklinde hatalı tahmin edilmiştir.

Tablo 4.12 AdaBoost için Karışıklık Matrisi

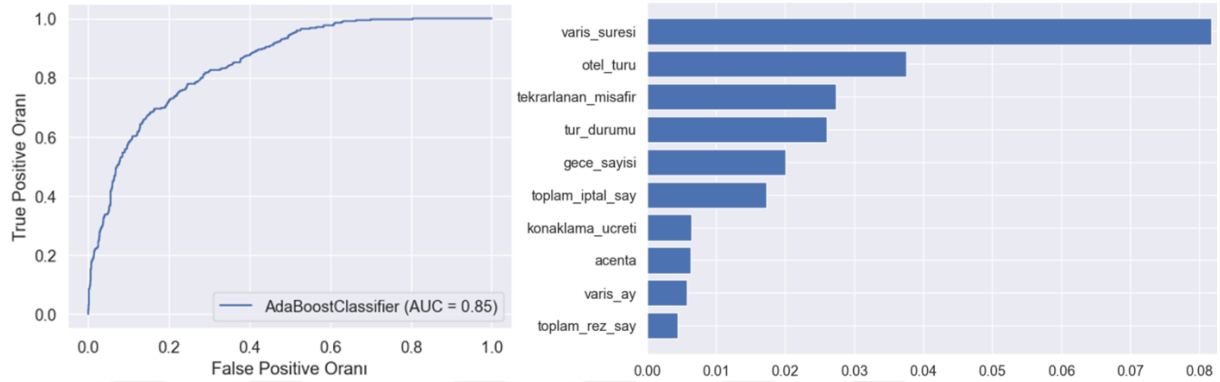
		Tahmin Değeri			
		İptal	İptal Değil		
Gerçek	İptal	TP = 164	FN = 180	TPR = 0,4767	FNR = 0,5233
	İptal Değil	FP = 79	TN = 1.095	TNR = 0,9327	FPR = 0,0673
		PPV = 0,6749	NPV = 0,8588	Doğruluk = 0,8294	F1 = 0,5588

AdaBoost sınıflandırıcısının 0,8294 doğruluk değeri ile, tüm rezervasyonların %83 oran ile “Rezervasyon iptal” veya “Rezervasyon iptal değil” olarak doğru şekilde tahmin edilmektedir. Sınıflandırıcısının 0,6749 kesinlik değeri, sınıflandırıcısının bir rezervasyonu “Rezervasyon iptal” olarak tahmin etmesi durumunda %68 oran ile doğru olduğu anlamına gelmektedir.

Sınıflandırıcısının 0,4767 hassasiyet değeri, veri kümesindeki tüm gerçek “Rezervasyon iptal” değerlerinden %48 oran ile doğru bir şekilde “Rezervasyon iptal” olarak tanımladığını göstermektedir. Sınıflandırıcısının özgüllük değerinin 0,9327 olması, veri kümesindeki tüm gerçek “Rezervasyon iptal değil” rezervasyonların %93 oran ile “Rezervasyon iptal değil” olarak doğru bir şekilde tanımlandığı anlamına gelmektedir.

Sınıflandırıcısının F1 puanının 0,5588 gibi bir ortalama bir değer olması, “Rezervasyon iptal” olan rezervasyonların doğru şekilde tanımlanması ile hatalı pozitif sonuçların önlenmesi arasında ortalama bir denge olduğunu göstermektedir. AdaBoost sınıflandırıcısının FPR değerinin 0,0673 olması, gerçek değeri “Rezervasyon iptal değil” olan rezervasyonların %7 oran ile hatalı bir şekilde “Rezervasyon iptal” olarak sınıflandırıldığı anlamına gelmektedir.

Sınıflandırıcının FNR değerinin 0,5233 olması, gerçek değeri “Rezervasyon iptal” olan rezervasyonların %52 oran ile hatalı bir şekilde “Rezervasyon iptal değil” olarak sınıflandırıldığı anlamına gelmektedir. Sınıflandırıcının NPV değerinin 0,8588 olması, “Rezervasyon iptal değil” tahmin edilen rezervasyonların %86 oran ile gerçekten “Rezervasyon iptal değil” olduğunu göstermektedir.



Şekil 4.18 AdaBoost için ROC ve En Çok Katkı Sağlayan Değişkenler

Eğri altında kalan alanın (AUC) değeri Şekil 4.18’de gösterildiği gibi 0,85 olup, AdaBoost sınıflandırıcısının doğru, güvenilir sayılabilecek bir test sonucunu ve test grubu için iyi bir ayrımı ifade etmektedir. AdaBoost tahmin modelinde en çok katkıyı sağlayan “varis_suresi”, “otel_turu” ve “tekrarlanan_misafir” bağımsız değişkenleridir.

4.6.4.10. XGBoost

XGBoost sınıflandırıcısının test verilerine ait karışıklık matrisi Tablo 4.13’te verilmiştir. 1.518 adet rezervasyon test verisi üzerinde yapılan tahminlere göre 202 adet gerçek “Rezervasyon iptal” değeri, “Rezervasyon iptal” şeklinde başarılı bir şekilde sınıflandırılmıştır. 1.101 adet gerçek “Rezervasyon iptal değil” değeri, “Rezervasyon iptal değil” şeklinde başarılı bir şekilde sınıflandırılmıştır. 73 adet gerçek “Rezervasyon iptal değeri” değeri, “Rezervasyon iptal” şeklinde hatalı tahmin edilmiştir. 142 adet gerçek “Rezervasyon iptal” değeri, “Rezervasyon iptal değil” şeklinde hatalı tahmin edilmiştir.

Tablo 4.13 XGBoost için Karışıklık Matrisi

		Tahmin Değeri			
		İptal	İptal Değil		
Gerçek	İptal	TP = 202	FN = 142	TPR = 0,5872	FNR = 0,4128
	İptal Değil	FP = 73	TN = 1.101	TNR = 0,9378	FPR = 0,0622
		PPV = 0,7345	NPV = 0,8858	Doğruluk = 0,8584	F1 = 0,6527

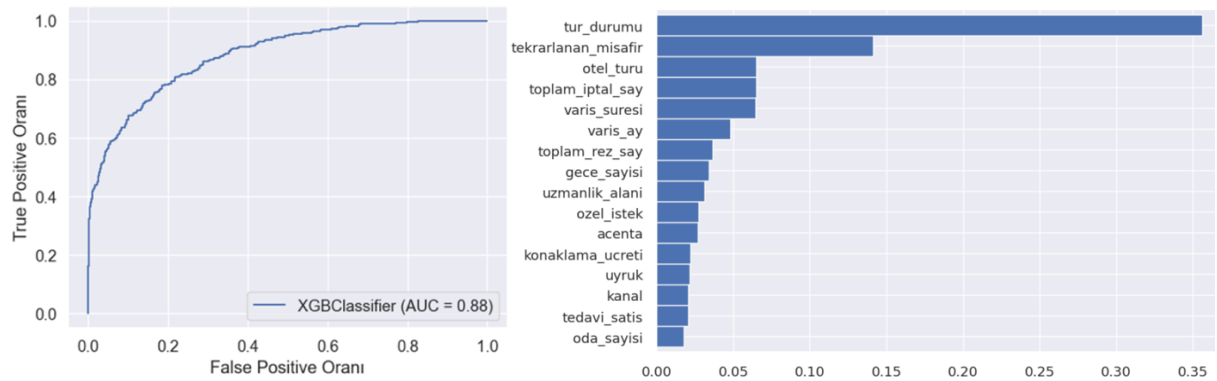
XGBoost sınıflandırıcısının 0,8584 doğruluk değeri ile, tüm rezervasyonların %86 ile “Rezervasyon iptal” veya “Rezervasyon iptal değil” olarak doğru şekilde tahmin edilmektedir.

Sınıflandırıcının 0,7345 kesinlik değeri, sınıflandırıcının bir rezervasyonu “Rezervasyon iptal” olarak tahmin etmesi durumunda %72 ile doğru olduğu anlamına gelmektedir.

Sınıflandırıcının 0,5872 hassasiyet değeri, veri kümesindeki tüm gerçek “Rezervasyon iptal” değerlerinden %59 oran ile doğru bir şekilde “Rezervasyon iptal” olarak tanımladığını göstermektedir. Sınıflandırıcının özgüllük değerinin 0,9378 olması, veri kümesindeki tüm gerçek “Rezervasyon iptal değil” rezervasyonların %94 oran ile “Rezervasyon iptal değil” olarak doğru bir şekilde tanımlandığı anlamına gelmektedir.

Sınıflandırıcının F1 puanının 0,6527 gibi bir ortalama üstü bir değer olması, “Rezervasyon iptal” olan rezervasyonların doğru şekilde tanımlanması ile hatalı pozitif sonuçların önlenmesi arasında iyi bir denge olduğunu göstermektedir. XGBoost sınıflandırıcının FPR değerinin 0,0622 olması, gerçek değeri “Rezervasyon iptal değil” olan rezervasyonların %6 oran ile hatalı bir şekilde “Rezervasyon iptal” olarak sınıflandırıldığı anlamına gelmektedir.

Sınıflandırıcının FNR değerinin 0,4128 olması, gerçek değeri “Rezervasyon iptal” olan rezervasyonların %41 oran ile hatalı bir şekilde “Rezervasyon iptal değil” olarak sınıflandırıldığı anlamına gelmektedir. Sınıflandırıcının NPV değerinin 0,8858 olması, “Rezervasyon iptal değil” tahmin edilen rezervasyonların %89 oran ile gerçekten “Rezervasyon iptal değil” olduğunu göstermektedir.



Şekil 4.19 XGBoost için ROC ve En Çok Katkı Sağlayan Değişkenler

Eğri altında kalan alanın (AUC) değeri Şekil 4.19’da gösterildiği gibi 0,88 olup, XGBoost sınıflandırıcısının doğru, güvenilir sayılabilecek bir test sonucunu ve test grubu için iyi bir ayrımı ifade etmektedir. XGBoost tahmin modelinde en çok katkıyı sağlayan özellikler “tur_durumu”, “tekrarlanan_misafir” ve “toplam iptal_say” bağımsız değişkenleridir.

4.6.4.11. CatBoost

CatBoost sınıflandırıcısının test verilerine ait karışıklık matrisi Tablo 4.14’te verilmiştir. 1.518 adet rezervasyon test verisi üzerinde yapılan tahminlere göre 196 adet gerçek “Rezervasyon iptal” değeri, “Rezervasyon iptal” şeklinde başarılı bir şekilde sınıflandırılmıştır. 1.129 adet gerçek “Rezervasyon iptal değil” değeri, “Rezervasyon iptal değil” şeklinde başarılı bir şekilde sınıflandırılmıştır.

45 adet gerçek “Rezervasyon iptal değeri” değeri, “Rezervasyon iptal” şeklinde hatalı tahmin edilmiştir. 148 adet gerçek “Rezervasyon iptal” değeri, “Rezervasyon iptal değil” şeklinde hatalı tahmin edilmiştir.

Tablo 4.14 CatBoost için Karışıklık Matrisi

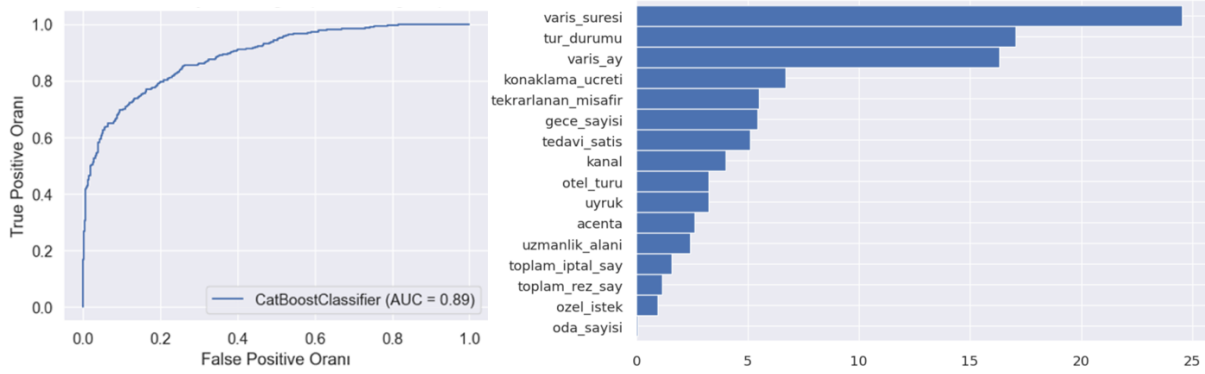
		Tahmin Değeri			
		İptal	İptal Değil		
Gerçek	İptal	TP = 196	FN = 148	TPR = 0,5698	FNR = 0,4302
	İptal Değil	FP = 45	TN = 1.129	TNR = 0,9617	FPR = 0,0383
		PPV = 0,8133	NPV = 0,8841	Doğruluk = 0,8729	F1 = 0,6701

CatBoost sınıflandırıcısının 0,8729 doğruluk değeri ile, tüm rezervasyonların %87 oran ile “Rezervasyon iptal” veya “Rezervasyon iptal değil” olarak doğru şekilde tahmin edilmektedir. Sınıflandırıcının 0,8133 kesinlik değeri, sınıflandırıcının bir rezervasyonu “Rezervasyon iptal” olarak tahmin etmesi durumunda %81 oran ile doğru olduğu anlamına gelmektedir.

Sınıflandırıcının 0,5698 hassasiyet değeri, veri kümesindeki tüm gerçek “Rezervasyon iptal” değerlerinden %57 oran ile doğru bir şekilde “Rezervasyon iptal” olarak tanımladığını göstermektedir. Sınıflandırıcının özgüllük değerinin 0,9617 olması, veri kümesindeki tüm gerçek “Rezervasyon iptal değil” rezervasyonların %96 oran ile “Rezervasyon iptal değil” olarak doğru bir şekilde tanımlandığı anlamına gelmektedir.

Sınıflandırıcının F1 puanının 0,6701 gibi bir yüksek bir değer olması, “Rezervasyon iptal” olan rezervasyonların doğru şekilde tanımlanması ile hatalı pozitif sonuçların önlenmesi arasında iyi bir denge olduğunu göstermektedir. CatBoost sınıflandırıcının FPR değerinin 0,0383 olması, gerçek değeri “Rezervasyon iptal değil” olan rezervasyonların %3 oran ile hatalı bir şekilde “Rezervasyon iptal” olarak sınıflandırıldığı anlamına gelmektedir.

Sınıflandırıcının FNR değerinin 0,4302 olması, gerçek değeri “Rezervasyon iptal” olan rezervasyonların %43 oran ile hatalı bir şekilde “Rezervasyon iptal değil” olarak sınıflandırıldığı anlamına gelmektedir. Sınıflandırıcının NPV değerinin 0,8841 olması, “Rezervasyon iptal değil” tahmin edilen rezervasyonların %88 oran ile gerçekten “Rezervasyon iptal değil” olduğunu göstermektedir.



Şekil 4.20 CatBoost için ROC ve En Çok Katkı Sağlayan Değişkenler

Eğri altında kalan alanın (AUC) değeri Şekil 4.20’de gösterildiği gibi 0,89 olup, CatBoost sınıflandırıcısının doğru, güvenilir sayılabilecek bir test sonucunu ve test grubu için iyi bir ayrımı ifade etmektedir. CatBoost tahmin modelinde en çok katkıyı sağlayan özellikler “varis_suresi”, “tur_durumu”, “varis_ay” ve “konaklama_ucreti” bağımsız değişkenleridir.

4.6.4.12. LightGBM

LightGBM sınıflandırıcısının test verilerine ait karışıklık matrisi Tablo 4.15’te verilmiştir. 1.518 adet rezervasyon test verisi üzerinde yapılan tahminlere göre 195 adet gerçek “Rezervasyon iptal” değeri, “Rezervasyon iptal” şeklinde başarılı bir şekilde sınıflandırılmıştır. 1.113 adet gerçek “Rezervasyon iptal değil” değeri, “Rezervasyon iptal değil” şeklinde başarılı bir şekilde sınıflandırılmıştır.

61 adet gerçek “Rezervasyon iptal değeri” değeri, “Rezervasyon iptal” şeklinde hatalı tahmin edilmiştir. 149 adet gerçek “Rezervasyon iptal” değeri, “Rezervasyon iptal değil” şeklinde hatalı tahmin edilmiştir.

Tablo 4.15 LightGBM için Karışıklık Matrisi

		Tahmin Değeri			
		İptal	İptal Değil		
Gerçek	İptal	TP = 195	FN = 149	TPR = 0,5669	FNR = 0,4331
	İptal Değil	FP = 61	TN = 1.113	TNR = 0,9480	FPR = 0,0520
		PPV = 0,7617	NPV = 0,8819	Doğruluk = 0,8617	F1 = 0,6500

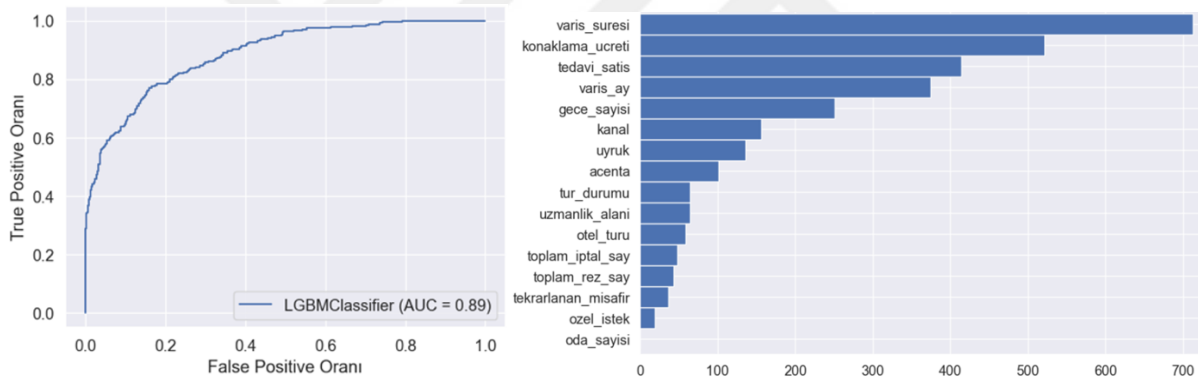
LightGBM sınıflandırıcısının 0,8617 doğruluk değeri ile, tüm rezervasyonların %86 oran ile “Rezervasyon iptal” veya “Rezervasyon iptal değil” olarak doğru şekilde tahmin edilmektedir. Sınıflandırıcısının 0,7617 kesinlik değeri, sınıflandırıcısının bir rezervasyonu “Rezervasyon iptal” olarak tahmin etmesi durumunda %76 oran ile doğru olduğu anlamına gelmektedir.

Sınıflandırıcısının 0,5669 hassasiyet değeri, veri kümesindeki tüm gerçek “Rezervasyon iptal” değerlerinden %57 oran ile doğru bir şekilde “Rezervasyon iptal” olarak tanımladığını

göstermektedir. Sınıflandırıcının özgüllük değerinin 0,9480 olması, veri kümesindeki tüm gerçek “Rezervasyon iptal değil” rezervasyonların %95 oran ile “Rezervasyon iptal değil” olarak doğru bir şekilde tanımlandığı anlamına gelmektedir.

Sınıflandırıcının F1 puanının 0,65 gibi bir ortalama üstü bir değer olması, “Rezervasyon iptal” olan rezervasyonların doğru şekilde tanımlanması ile hatalı pozitif sonuçların önlenmesi arasında iyi bir denge olduğunu göstermektedir. LightGBM sınıflandırıcının FPR değerinin 0,0520 olması, gerçek değeri “Rezervasyon iptal değil” olan rezervasyonların %5 oran ile hatalı bir şekilde “Rezervasyon iptal” olarak sınıflandırıldığı anlamına gelmektedir.

Sınıflandırıcının FNR değerinin 0,4331 olması, gerçek değeri “Rezervasyon iptal” olan rezervasyonların %43 oranında hatalı bir şekilde “Rezervasyon iptal değil” olarak sınıflandırıldığı anlamına gelmektedir. Sınıflandırıcının NPV değerinin 0,8819 olması, “Rezervasyon iptal değil” tahmin edilen rezervasyonların %88 oran ile gerçekten “Rezervasyon iptal değil” olduğunu göstermektedir.



Şekil 4.21 LightGBM için ROC ve En Çok Katkı Sağlayan Değişkenler

Eğri altında kalan alanın (AUC) değeri Şekil 4.21’de gösterildiği gibi 0,89 olup, LightGBM sınıflandırıcısının doğru, güvenilir sayılabilecek bir test sonucunu ve test grubu için iyi bir ayrımı ifade etmektedir. LightGBM tahmin modelinde en çok katkıyı sağlayan özellikler “varis_suresi”, “konaklama_ucreti” ve “tedavi_satis” bağımsız değişkenleridir.

4.6.5. Değerlendirme

Sağlık turizmi rezervasyon iptallerinin tahmini için, dengesiz veri seti üzerinde, parametre optimizasyonu yapılmamış CatBoost, Rastgele Orman, LightGBM, XGBoost, YSA, AdaBoost, DVM, Karar Ağaçları, K-En Yakın Komşu, Lojistik Regresyon, Naive Bayes ve ZeroR sınıflandırma algoritmalarına ait sonuçlar Tablo 4.16’da verilmiştir. Sınıflandırma modelleri doğruluk, F1-skor, FNR, FPR, NPV, kesinlik, ROC, hassasiyet, özgüllük ve süre olmak üzere 10 adet değerlendirme kriterine göre karşılaştırılmıştır.

Tablo 4.16 Temel Model Karşılaştırma Tablosu

Model	Doğruluk	F1	FNR	FPR	NPV	Kesinlik	ROC	Hatırlama	Özgüllük	Süre
CatBoost	0,873	0,670	0,430	0,038	0,884	0,813	0,888	0,570	0,962	6,834
RO	0,866	0,639	0,477	0,033	0,874	0,822	0,876	0,523	0,967	4,749
LightGBM	0,862	0,650	0,433	0,052	0,882	0,762	0,886	0,567	0,948	0,797
XGBoost	0,858	0,653	0,413	0,062	0,886	0,735	0,883	0,587	0,938	0,867
YSA	0,841	0,574	0,526	0,052	0,860	0,728	0,853	0,474	0,948	9,647
AdaBoost	0,829	0,559	0,523	0,067	0,859	0,675	0,850	0,477	0,933	1,474
DVM	0,824	0,435	0,701	0,023	0,826	0,792	0,813	0,299	0,977	30,918
DT	0,823	0,578	0,465	0,093	0,869	0,628	0,796	0,535	0,907	0,162
KNN	0,810	0,513	0,558	0,083	0,849	0,610	0,766	0,442	0,917	0,867
LR	0,780	0,248	0,840	0,038	0,796	0,550	0,709	0,160	0,962	1,173
ZeroR	0,773	-	-	-	-	-	-	-	-	-
NB	0,588	0,488	0,134	0,494	0,928	0,339	0,794	0,866	0,506	0,119

Dengesiz veri seti üzerinde, parametre optimizasyonu yapılmamış model karşılaştırmasına göre en etkin modeller CatBoost, Rastgele Orman, LightGBM, XGBoost ve Yapay Sinir Ağları algoritmalarıdır.

Model başarımlarını arttırabilmek ve en iyi sınıflandırma sonuçlarını elde edebilmek için parametre optimizasyon (grid search) işlemleri gerçekleştirilmiştir. Elde edilen en iyi parametreler ile sınıflandırma modelleri tekrar çalıştırılarak tahmin sonuçları Tablo 4.17’de gösterildiği gibi elde edilmiştir.

Tablo 4.17 Parametreleri Optimize Edilmiş Sınıflandırıcıların Karşılaştırılması

Model	Doğruluk	F1	FNR	FPR	NPV	Kesinlik	ROC	Hatırlama	Özgüllük	Süre
LightGBM	0,882	0,668	0,433	0,038	0,883	0,813	0,887	0,567	0,962	0,871
CatBoost	0,880	0,674	0,407	0,049	0,889	0,782	0,878	0,593	0,951	114,89
XGBoost	0,879	0,647	0,471	0,032	0,875	0,831	0,882	0,529	0,969	1,054
RO	0,878	0,652	0,454	0,038	0,879	0,807	0,879	0,547	0,962	5,760
DT	0,868	0,613	0,468	0,060	0,873	0,723	0,836	0,532	0,940	0,240
YSA	0,867	0,579	0,506	0,062	0,864	0,700	0,853	0,494	0,938	19,706
DVM	0,843	0,521	0,599	0,041	0,845	0,742	0,822	0,401	0,959	46,897
KNN	0,839	0,501	0,622	0,038	0,841	0,743	0,831	0,378	0,962	1,321
Adaboost	0,835	0,542	0,544	0,066	0,854	0,668	0,855	0,456	0,934	8,199
NB	0,791	0,484	0,567	0,105	0,844	0,548	0,796	0,433	0,895	0,167
LR	0,785	0,252	0,837	0,038	0,797	0,555	0,709	0,163	0,962	1,978
ZeroR	0,773	-	-	-	-	-	-	-	-	-

Parametreleri optimize edilmiş sınıflandırma modellerinin karşılaştırma sonuçlarına göre doğruluk değeri en yüksek olan modeller LightGBM, CatBoost, XGBoost, RO ve Karar Ağaçları algoritmalarıdır. F1-skor, NPV, kesinlik, AUC, hassasiyet ve özgüllük kriterlerine göre de bu modeller en iyi değerlere sahiptir.

Verilerin yeniden örnekleme, dengesiz bir veri kümesiyle başa çıkmak için en sık tercih edilen yaklaşımlardan biridir. Verilerin yeniden örnekleme, düşük örnekleme (under sampling) ve yüksek örnekleme (over sampling) olmak iki üzere iki tür yöntem bulunmaktadır. Düşük örneklemede, bazı önemli verileri içeren örneklerin veri setinden

çıkarılmasından dolayı, çoğu durumda düşük örnekleme tekniklerine göre aşırı örnekleme tercih edilmektedir. Sağlık turizmi veri setinin aşırı dengesiz bir yapıda olmamasına rağmen ADASYN, ROSE, SMOTE-ENN teknikleri ile veri seti dengeli bir hale getirilmiş ve sınıflandırıcı performansları incelenmiştir.

ADASYN (ADaptive SYNthetic) yaklaşımı ile azınlık sınıf örneklerinin yoğunluğuna göre çalışmaktadır. Yeni örneklerin üretilmesi azınlık sınıfı örneklerinin yoğunluğuyla ters orantılıdır. Azınlık sınıfı örnek yoğunluğunun düşük olduğu veya hiç olmadığı özellik alanı bölgesinde daha fazla örnek, yüksek yoğunluklu alanda ise daha az örnek üretilmektedir. ADASYN örnekleme yaklaşımı ile dengeli hale getirilmiş veri seti üzerinde sınıflandırıcılara ait sonuçlar Tablo 4.18’de verilmiştir.

Tablo 4.18 ADASYN ile Sınıflandırıcıların Karşılaştırılması

Model	Doğruluk	F1	FNR	FPR	NPV	Kesinlik	ROC	Hatırlama	Özgüllük	Süre
RO	0,925	0,852	0,049	0,083	0,985	0,771	0,978	0,951	0,917	1,184
DT	0,914	0,833	0,055	0,095	0,982	0,745	0,925	0,945	0,906	0,064
XGBoost	0,908	0,817	0,093	0,092	0,971	0,743	0,939	0,907	0,908	1,005
LightGBM	0,878	0,747	0,204	0,099	0,938	0,703	0,923	0,797	0,901	0,460
CatBoost	0,864	0,721	0,224	0,111	0,931	0,673	0,917	0,776	0,889	7,920
YSA	0,815	0,657	0,218	0,176	0,928	0,566	0,887	0,782	0,825	7,920
KNN	0,805	0,663	0,154	0,207	0,946	0,545	0,907	0,846	0,793	0,204
Adaboost	0,785	0,607	0,267	0,199	0,911	0,519	0,844	0,733	0,801	0,689
DVM	0,785	0,623	0,215	0,216	0,926	0,516	0,849	0,785	0,785	19,056
ZeroR	0,773	-	-	-	-	-	-	-	-	-
LR	0,679	0,506	0,276	0,334	0,892	0,389	0,743	0,724	0,666	0,286
NB	0,531	0,480	0,044	0,594	0,970	0,321	0,801	0,956	0,406	0,011

ADASYN uygulanmış veri seti üzerinde sınıflandırma modellerinin karşılaştırma sonuçlarına göre başarımlar en yüksek olan modeller RO, DT, XGBoost, LightGBM ve CatBoost algoritmalarıdır.

ROSE (Random Over Sampling Examples), örneklemeyle dayalı yapay veri üreten, dengesiz sınıfların varlığında ikili sınıflandırma problemleri ile başa çıkacak işlevler sağlayan önemli bir yaklaşımdır. ROSE ile seçilen örnek üzerinde merkezlenen bir olasılık dağılımına göre gerçek olanlara mümkün olduğunca benzeyen sentetik veri noktaları oluşturularak nadir sınıfın aşırı örnekleme amaçlanmaktadır. Yapay dengeli örnekler, yumuşatılmış önyükeme yaklaşımına göre üretilmektedir. ROSE ile dengeli hale getirilmiş veri seti üzerinde sınıflandırıcılara ait sonuçlar Tablo 4.19’da verilmiştir.

Tablo 4.19 ROSE ile Sınıflandırıcıların Karşılaştırılması

Model	Doğruluk	F1	FNR	FPR	NPV	Kesinlik	ROC	Hatırlama	Özgüllük	Süre
RO	0,925	0,852	0,047	0,084	0,985	0,770	0,987	0,954	0,917	1,408
XGBoost	0,908	0,821	0,064	0,101	0,980	0,732	0,940	0,936	0,900	0,231
DT	0,907	0,820	0,064	0,101	0,980	0,730	0,917	0,936	0,899	0,069
LightGBM	0,883	0,781	0,081	0,127	0,973	0,680	0,933	0,919	0,873	0,307
Catboost	0,870	0,761	0,084	0,144	0,972	0,651	0,925	0,916	0,856	3,380
YSA	0,810	0,668	0,154	0,201	0,947	0,552	0,897	0,846	0,799	10,175
KNN	0,797	0,656	0,148	0,219	0,947	0,533	0,912	0,852	0,781	0,814
DVM	0,778	0,632	0,160	0,240	0,942	0,506	0,860	0,840	0,760	36,434
ZeroR	0,773	-	-	-	-	-	-	-	-	-
Adaboost	0,769	0,614	0,189	0,244	0,932	0,494	0,858	0,811	0,756	0,426
LR	0,667	0,486	0,305	0,342	0,880	0,373	0,732	0,695	0,658	0,595
NB	0,533	0,481	0,047	0,590	0,968	0,321	0,802	0,954	0,410	0,206

ROSE ile dengeli bir hale getirilmiş veri seti üzerinde sınıflandırma modellerinin karşılaştırma sonuçlarına göre doğruluk değeri en yüksek olan modeller RO, XGBoost, DT, LightGBM ve CatBoost algoritmalarıdır. SMOTE (Synthetic Minority Oversampling Technique), azınlık sınıfı için sentetik örnekler oluşturarak dengesiz veri kümelerinin üstesinden gelmek üzere özel olarak tasarlanmış bir aşırı örnekleme tekniğidir. SMOTE, önyargıyı azaltarak ve azınlık sınıfının önemli özelliklerini yakalayarak, daha doğru tahminlere ve daha iyi model performansına katkıda bulunmaktadır. SMOTE, rastgele aşırı örneklemenin ortaya çıkardığı aşırı uyum sorununun üstesinden gelmeye yardımcı olmaktadır. Birlikte bulunan pozitif örnekler arasında interpolasyon yardımıyla yeni örnekler oluşturmak için özellik alanına odaklanmaktadır. SMOTE örnekleme yaklaşımına ait sonuçlar Tablo 4.20’de verilmiştir.

Tablo 4.20 SMOTE ile Sınıflandırıcıların Karşılaştırılması

Model	Doğruluk	F1	FNR	FPR	NPV	Kesinlik	ROC	Hatırlama	Özgüllük	Süre
RO	0,928	0,855	0,055	0,078	0,983	0,781	0,980	0,945	0,923	1,206
DT	0,923	0,848	0,055	0,084	0,983	0,768	0,931	0,945	0,917	0,083
XGBoost	0,908	0,819	0,087	0,093	0,973	0,742	0,940	0,913	0,907	0,257
LightGBM	0,878	0,750	0,189	0,103	0,942	0,698	0,923	0,811	0,897	0,308
CatBoost	0,864	0,722	0,224	0,110	0,931	0,674	0,920	0,776	0,890	5,768
YSA	0,822	0,658	0,241	0,160	0,922	0,581	0,897	0,759	0,840	14,072
KNN	0,817	0,678	0,148	0,193	0,949	0,564	0,917	0,852	0,807	0,184
Adaboost	0,791	0,612	0,270	0,192	0,911	0,527	0,840	0,730	0,808	0,459
DVM	0,789	0,624	0,227	0,207	0,923	0,523	0,852	0,773	0,793	16,552
ZeroR	0,773	-	-	-	-	-	-	-	-	-
LR	0,683	0,499	0,302	0,322	0,884	0,388	0,735	0,698	0,678	0,220
NB	0,539	0,482	0,055	0,580	0,963	0,323	0,801	0,945	0,420	0,026

SMOTE ile dengeli bir hale getirilmiş veri seti üzerinde sınıflandırma modellerinin karşılaştırma sonuçlarına göre doğruluk değeri en yüksek olan modeller RO, DT, XGBoost, LightGBM ve CatBoost algoritmalarıdır.

SMOTE-ENN (SMOTE-Edited Nearest Neighbor), Batista vd. (2004) tarafından geliştirilen, azınlık sınıfı için sentetik örnekler üretmeye yönelik SMOTE yeteneğini ve gözlemin sınıfı ile K-en yakın komşu çoğunluğu arasında farklı sınıfa sahip olduğu tanımlanan her iki sınıftan bazı gözlemleri silmeye yönelik ENN yeteneğini birleştirmektedir. SMOTE-ENN örnekleme yaklaşımına ait sonuçlar Tablo 4.21’de verilmiştir.

Tablo 4.21 SMOTE-ENN ile Sınıflandırıcıların Karşılaştırılması

Model	Doğruluk	F1	FNR	FPR	NPV	Kesinlik	ROC	Hatırlama	Özgüllük	Süre
RO	0,909	0,812	0,134	0,078	0,959	0,764	0,959	0,866	0,922	1,126
XGBoost	0,895	0,784	0,157	0,090	0,952	0,732	0,920	0,843	0,910	0,237
DT	0,891	0,782	0,140	0,100	0,957	0,717	0,880	0,861	0,900	0,058
LightGBM	0,872	0,734	0,224	0,100	0,932	0,695	0,906	0,776	0,900	0,270
CatBoost	0,861	0,708	0,256	0,105	0,923	0,676	0,902	0,744	0,895	5,662
YSA	0,828	0,666	0,244	0,151	0,922	0,595	0,890	0,756	0,849	9,428
KNN	0,815	0,663	0,198	0,181	0,934	0,564	0,889	0,802	0,819	0,179
Adaboost	0,798	0,611	0,302	0,172	0,903	0,543	0,841	0,698	0,828	0,450
DVM	0,796	0,626	0,247	0,192	0,918	0,535	0,842	0,753	0,808	15,478
ZeroR	0,773	-	-	-	-	-	-	-	-	-
LR	0,686	0,488	0,340	0,307	0,874	0,387	0,730	0,660	0,693	0,272
NB	0,551	0,483	0,076	0,559	0,952	0,327	0,797	0,924	0,441	0,014

SMOTE-ENN ile dengeli bir hale getirilmiş veri seti üzerinde sınıflandırma modellerinin karşılaştırma sonuçlarına göre doğruluk değeri en yüksek olan modeller RO, XGBoost, DT, LightGBM ve CatBoost algoritmalarıdır.

En iyi performans gösteren beş sınıflandırıcıya ait temel, parametreleri optimize edilmiş, ADASYN, ROSE, SMOTE ve SMOTE-ENN teknikleri uygulanarak doğruluk, F1-skor, kesinlik, ROC, hassasiyet, özgüllük ve çalışma hızı değerleri Tablo 4.22’de verilmiştir. Araştırma problemi için en iyi performansı gösteren sınıflandırıcı modellerin, topluluk öğrenmesi algoritmaları olduğu görülmektedir. Dengeli hale getirilmiş veri seti üzerine uygulanan teknikler içerisinde en iyi doğruluk (0,928) ve F1-skor (0,855) değerlerine sahip SMOTE tekniği uygulanmış Rastgele Orman algoritması olmuştur. Rastgele Orman algoritmasının ADASYN, ROSE ve SMOTE-ENN teknikleri için de çok yakın sonuçlar ürettiği için tercih edilebilir. Temel ve parametreleri optimize edilmiş Rastgele Orman algoritmaları yüksek çalışma süresi, düşük doğruluk değerlerinden dolayı daha az tercih edilmiştir.

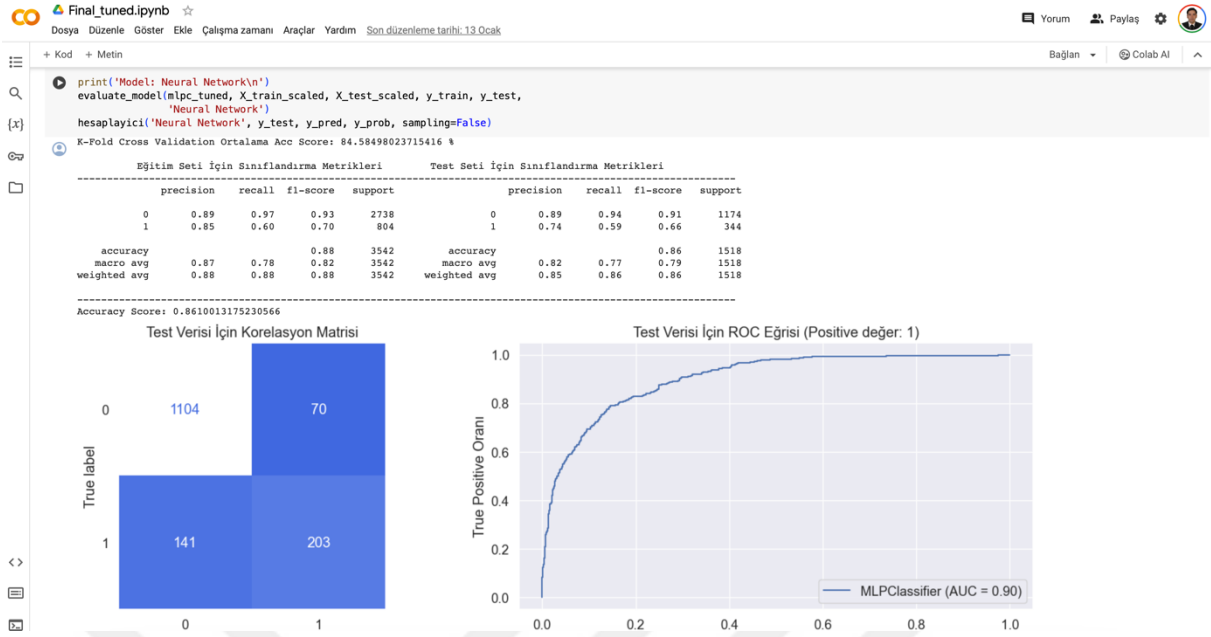
Tablo 4.22 En İyi Performansa Sahip Sınıflandırıcılara ait Özet Karşılaştırma

Model	Teknik	Doğruluk	F1-Skor	Kesinlik	ROC	Hassasiyet	Özgüllük	Hız
CatBoost	Temel	0,873	0,670	0,813	0,888	0,570	0,962	6,834
	Optimize	0,880	0,674	0,782	0,878	0,593	0,951	114,893
	ADASYN	0,864	0,721	0,673	0,917	0,776	0,889	7,920
	ROSE	0,870	0,761	0,651	0,925	0,916	0,856	3,380
	SMOTE	0,864	0,722	0,674	0,920	0,776	0,890	5,768
	SMOTE-ENN	0,861	0,708	0,676	0,902	0,744	0,895	5,662
XGBoost	Temel	0,858	0,653	0,735	0,883	0,587	0,938	0,867
	Optimize	0,879	0,647	0,831	0,882	0,529	0,969	1,054
	ADASYN	0,908	0,817	0,743	0,939	0,907	0,908	1,005
	ROSE	0,908	0,821	0,732	0,940	0,936	0,900	0,231
	SMOTE	0,908	0,819	0,742	0,940	0,913	0,907	0,257
	SMOTE-ENN	0,895	0,784	0,732	0,920	0,843	0,910	0,237
LightGBM	Temel	0,862	0,650	0,762	0,886	0,567	0,948	0,797
	Optimize	0,882	0,668	0,813	0,887	0,567	0,962	0,871
	ADASYN	0,878	0,747	0,703	0,923	0,797	0,901	0,460
	ROSE	0,883	0,781	0,680	0,933	0,919	0,873	0,307
	SMOTE	0,878	0,750	0,698	0,923	0,811	0,897	0,308
	SMOTE-ENN	0,872	0,734	0,695	0,906	0,776	0,900	0,270
RO	Temel	0,866	0,639	0,822	0,876	0,523	0,967	4,749
	Optimize	0,878	0,652	0,807	0,879	0,547	0,962	5,760
	ADASYN	0,925	0,852	0,771	0,978	0,951	0,917	1,184
	ROSE	0,925	0,852	0,770	0,987	0,954	0,917	1,408
	SMOTE	0,928	0,855	0,781	0,980	0,945	0,923	1,206
	SMOTE-ENN	0,909	0,812	0,764	0,959	0,866	0,922	1,126
DT	Temel	0,823	0,578	0,628	0,796	0,535	0,907	0,162
	Optimize	0,868	0,613	0,723	0,836	0,532	0,940	0,240
	ADASYN	0,914	0,833	0,745	0,925	0,945	0,906	0,064
	ROSE	0,907	0,820	0,730	0,917	0,936	0,899	0,069
	SMOTE	0,923	0,848	0,768	0,931	0,945	0,917	0,083
	SMOTE-ENN	0,891	0,782	0,717	0,880	0,861	0,900	0,058

4.6.6. Uygulama

Araştırmada veri setini analize hazır hale getirebilmek için hedef niteliğin oluşturulması, veri temizleme, veri ayırıklaştırma, aykırı değerlerin ve tekrarlı verilerin temizlenmesi, değişken seçimi uygulanmıştır. Veri ön işleme adımları Google Cloud BigQuery ve Python kullanılarak yapıldıktan sonra analizler için Google Colab ortamına geçilmiştir. Kişisel bilgisayarların RAM, CPU ve GPU açısından yetersiz olması büyük veri setleri ile çalışırken performans sorunlarına yol açtığı görülmüştür. Google colab bulut hizmeti ücretsiz sunulan GPU desteği ile veri ön işleme ve modellemenin araştırmacıların işini çok kolaylaştırdığı görülmüştür.

Sınıflandırma tahmin problemi için hazırlanan sınıflandırıcılar, karşılaştırma metriklerine göre Tablo 4.22’de listelenmiştir. Tahmin modellerinin eğitim ve tahmin süreçleri, Google Colab üzerinde yer alan .ipynb uzantılı dosyalar kullanılarak Şekil 4.22’de gösterildiği gibi gerçekleştirilmiştir.



Şekil 4.22 Google Colab Çalışma Ortamı

Sınıflandırıcı tahmin modelleri, gerçek hayat problemlerine ve canlı sistemlere aktarılabilirdiği sürece işe değer katmakta ve anlamı olmaktadır. Tahmin modellerinin dışa aktarılabilmesi için kullanılabilecek Pickle, Joblib, PMML (Predictive Model Markup Language) formatları bulunmaktadır. Bu formatlar makine öğrenimi modellerinin yapılandırılmalarını, parametrelerini, model mimarisini ve parametre ağırlıklarını kolaylıkla saklamayı sağlamaktadır. Araştırma kapsamında Pickle ve PMML formatları tercih edilmiştir. Bu formatların dışa aktarımı Şekil 4.23'te gösterilmiştir.

```
# Modelin dışa aktarımı için tercih ettiğimiz pickle kütüphanesini içe aktarıyoruz.
import pickle

# XGBoostClassifier() modeli bu aşamada 'xgb' olarak varsayılmıştır.
# Dışa aktarılacak modelin ismini belirleyelim.
model_name = "xgb.pkl"

# Modeli pickle kütüphanesi üzerinden kaydetme süreci.
with open(model_name, "wb") as model_file:
    pickle.dump(xgb, model_file)
```

Şekil 4.23 Modelin Pickle ile Dışarı Aktarılması

Tahmin modellerinin canlı sisteme aktarılabilmesi için model “.pkl” formatında dışarıya aktarılmıştır. Dışarıya aktarılan eğitilmiş tahmin modellerinin web tabanlı karar destek sistemi içerisine aktarılabilmesi için Flask teknolojisi tercih edilmiştir.

Rezervasyon Kartı

1 Tedavi Bilgisi

2 Rezervasyon Bilgisi

Uzmanlık Alanı	Bariatric Surgery	Varış Tarihi	10/05/2024
Girişim Türü	Tüp Mide Ameliyatı	Çıkış Tarihi	10/26/2024
Tedavi Ücreti	1140	Rezervasyon Oluşturulma Tarihi	06/05/2024

< Geri

İleri >

%96.81 ihtimalle iptal edilecek

Şekil 4.24 Sağlık Turizmi Rezervasyon Formu Örneği

Şekil 4.24'te gösterilen rezervasyon formu hazırlanmış ve gerekli bağımsız değişkenler, hazırlanan form yardımıyla alınmıştır. Rezervasyon formunu sadeleştirebilmek amacıyla “varis_suresi”, “gece_sayisi” değişkenleri forma eklenmeyip arka planda hesaplanmaktadır.

```
filename = './models/xgb.pkl'
model = pickle.load(open(filename, 'rb'))
@app.route('/submit', methods=['POST'])
def submit_form():
    try:
        data = request.json
        data_df = pd.DataFrame.from_dict(data, orient='index').T
        pred = model.predict_proba(data_df)
        pred = round(pred[0][1]*100,4)
        return str(pred)
    except Exception as e:
        return jsonify({'error': str(e)}), 400
```

Şekil 4.25 Tahmin Modelinin Canlı Sisteme Yüklenmesi

Rezervasyon formunda yer alan alanlar doldurulup gönderildikten sonra JSON formatlı veri pandasDataFrame'e dönüştürülerek Şekil 4.25'te gösterildiği gibi tahmine hazır hale getirilmektedir. Formdan alınan veri predic_proba metoduyla tahmin edilmekte ve elde edilen tahmin değerinin 0. indeksli 1. elemanı olan TP (True Positive) değerini ifade etmektedir.

TARTIŞMA VE SONUÇ

Sağlık turizmi sektöründe, rezervasyonunu iptal etmeyi planlayan sağlık turistlerini tahmin etmek, sağlık turizmi paydaşlarının gelir yönetimi ve kaynak yönetimi açısından büyük önem taşımaktadır. Rezervasyon iptal tahminlerini ve iptale neden olan bağımsız değişkenleri tespit etmek, işletmeler açısından çok değerlidir. Bilişim teknolojileri alanındaki gelişmeler sayesinde, düşük maliyetler ile daha yüksek hesaplama gücü elde edilmiştir. Sağlık turizmi paydaşlarının rezervasyon iptal tahmin ihtiyaçlarını karşılamak için bilişim teknolojilerinin sunmuş olduğu imkanları ve makine öğrenmesi yöntemlerini birleştirerek, sağlık turizmi rezervasyon iptallerini başarılı bir şekilde tahmin etme ve karar destek sistemlerine entegre etme mümkün hale gelmiştir.

Teknoloji ve veri bilimi alanından meydana gelen gelişmeler, bilimsel çalışmaların vazgeçilmez unsuru olan literatür tarama metodolojisinde de değişikliklere yol açmıştır. Araştırma sayısının günden güne artarak büyük veri oluşturduğu günümüzde, literatür taramalarının klasik yöntemler ile yapılması çok zaman ve enerji kaybına yol açmaktadır. Bu araştırma kapsamında; WoS ve Scopus veritabanlarında detaylı arama ifadeleri kullanılarak, çalışma verilerini yığın bir şekilde bilgisayarda depolayan ve metin madenciliği teknikleri ile sistematik literatür tarama metodolojisi uygulanmıştır. Araştırmada tercih edilen literatür tarama yaklaşımının ilgi çekmesi üzerine, çalışmalarında sistematik literatür tarama metodolojisini kullanmak isteyen araştırmacıların kullanımına açılacaktır. Hackers & Painters kitabının yazarı Paul Graham'ın 2004 senesinde belirttiği gibi, yazılımlar artık masaüstü ortamda tercih edilmemekte, bulut sunucu ortamında çalıştırılmaktadır. Doktora araştırmasında kullanılan literatür tarama yazılımı da lokal bilgisayarlarda çalışan ve çok sayıda R paketine bağlı olan bir yazılımdır. Sistematik literatür tarama yazılımının daha gelişmiş özelliklere sahip olan online sürümünün geliştirmelerine Python programlama dili ile başlanmış olup, yazılım tamamlandıktan sonra (<https://github.com/erokan>) adresinde paylaşılacaktır.

Sistematik literatür taraması metodolojisinde belirtildiği üzere; konaklama, transfer, tur, vb. diğer hizmet sektörlerinde, rezervasyon iptal tahmini üzerine çalışmalar yer almaktadır. Turizm sektörü özelinde, gelişmiş makine öğrenimi sınıflandırma algoritmalarını kullanarak otel rezervasyon iptallerinin tahminini yapan çok sayıda araştırma bulunmasına rağmen, sağlık turizmi alanında rezervasyon iptallerinin tahmini üzerine bir çalışma bulunmamaktadır. Bu araştırma, sağlık turizmi alanında bulunan bu boşluğu doldurmayı hedeflemiştir.

Araştırma genelinde, veri madenciliği süreçlerini etkin bir şekilde yönetebilmek için, sektörde en çok tercih edilen standartlaştırılmış bir metodoloji olan CRISP-DM tercih edilmiştir. Problemin tanımlanması, veriyi anlama, veriyi hazırlama, modelleme, değerlendirme ve uygulama adımlarına ait tüm görev ve çıktılar detaylı bir şekilde ikinci bölümde açıklanmış ve dördüncü bölümde tüm adımlar adım adım uygulanmıştır. Doktora tezinin önemli katkılarından bir diğeri, CRISP-DM metodolojisinin bir çalışmada baştan sonra kadar, tüm adımları ile nasıl işletileceğinin uygulamalı olarak gösterilmesidir.

Rezervasyon iptallerini tahmin edebilen modeller geliştirebilmek için, veri madenciliği metodolojileri ve teknikleri kullanılmaktadır. Literatür çalışmalarında, rezervasyon iptal tahminlerini başarı ile yapan çalışmalar birinci bölümde listelenmiştir. Bu çalışmalara ek olarak, Uluslararası Veri Bilimi ve Uygulamaları Konferansı- ICONDATA 2018’da sunumu yapmış olduğumuz “Otel Rezervasyon İptallerinin Makine Öğrenmesi Teknikleriyle Tahmin Edilmesi” bildiri, makale olarak yayımlanmıştır (Boz vd., 2018). Doktora araştırmasının temelini oluşturan bu çalışmada, CRISP-DM metodolojisi takip edilerek, makine öğrenmesi teknikleri ile, beş farklı otele ait toplam 38.826 kayıttan oluşan otel rezervasyon verisi kullanılarak, otellerin gelecekteki rezervasyonlarının iptal durumları tahmin edilmeye çalışılmıştır. Çalışmada, sınıflandırma algoritmalarından Rastgele Orman, Destek Vektör Makineleri, K-En Yakın Komşu ve C4.5 Karar Ağacı algoritmaları kullanılarak dört farklı model R programlama dili oluşturulmuş ve modellerin performans karşılaştırmaları yapılmıştır. En iyi sonuç %73 doğruluk oranı ile C4.5 Karar Ağacı Algoritmasından elde edilmiştir. Bu çalışma esnasında elde edilen deneyim, doktora araştırmasının temelini oluşturmuştur.

Sağlık turizminde rezervasyon iptallerinin tahmini çalışmalarında, hangi tahmin modellerin iptal tahmininde daha başarılı olduğu, hangi bağımsız değişkenlerin rezervasyon iptallerini daha iyi açıkladığını tespit etmek, araştırmanın en önemli diğer çıktılarıdır. Turizm verilerine ek olarak sağlık verilerinin de veri setinde yer alması, veri seti hazırlama ve tahmin sürecini zorlu ve karmaşık bir hale getirmesine rağmen, çalışmanın özgün yönünü oluşturma bakımından değerlidir. Araştırmada, kişisel veri içermeyen, anonim hale getirilmiş 5.060 adet sağlık turisti rezervasyon verisi, on iki farklı makine öğrenmesi yöntemi (ZeroR, Naive Bayes, K-En Yakın Komşu, Destek Vektör Makinaları, Lojistik Regresyon, Karar Ağaçları, Yapay Sinir Ağları, Rastgele Orman, XGBoost, CatBoost, AdaBoost ve LightGBM) ve on farklı model değerlendirme metriği (doğruluk, kesinlik, hassasiyet, F1-skor, özgüllük, FPR, FNR, NPV, ROC ve hız) kullanılarak sağlık turizmi rezervasyon iptal tahmin modelleri geliştirilmiştir.

Sağlık Turizmi Yönetim Sistemi yazılımının farklı tablolarından bir araya getirilmiş 39 değişken bulunmaktadır. Tahmin modeli oluşturma esnasında, veri setinde daha çok bağımsız değişkenin bulunması her zaman daha yüksek başarıyı vermemektedir. Daha kolay anlaşılabilen sonuçlar ve daha iyi bir tahmin modeli oluşturabilmek için veri setinde değişken azaltma işlemi gerçekleştirmek gerekmektedir. Sánchez-Medina ve C-Sánchez, 2020 otel rezervasyon iptal tahmini çalışmalarında sadece 13 adet değişken kullanarak %98 tahmin başarıyı yakalanabileceğini göstermiştir. Ayrıca, dört resort otelden alınan veri kümesi ve veri bilimi kullanarak bir sınıflandırma problemi olarak rezervasyon iptali tahminini ele alan ve rezervasyon iptallerini %90'ın üzerinde doğrulukla tahmin edici modeller oluşturma mümkün olduğunu gösterilmiştir. (Antonio vd., 2017b).

Sağlık turizmi rezervasyon iptallerinin etkin bir şekilde tahmini için 12 tahmin modeli ve 10 karşılaştırma kriterini içeren, altı farklı tahmin test tasarım ortamı oluşturulmuştur. Birinci test ortamı: “Dengesiz veri seti üzerinde, parametre optimizasyonu yapılmamış tahmin modelleri”, ikinci test ortamı: “Dengesiz veri seti üzerinde, parametre optimizasyonu yapılmış tahmin modelleri”, üçüncü test ortamı: “ROSE tekniği ile dengeli hale getirilmiş veri seti üzerinde tahmin modelleri”, dördüncü test ortamı: “ADASYN tekniği ile dengeli hale getirilmiş veri seti üzerinde tahmin modelleri”, beşinci test ortamı: “SMOTE tekniği ile dengeli hale getirilmiş veri seti üzerinde tahmin modelleri” ve son olarak altıncı test ortamı: “SMOTE-ENN tekniği ile dengeli hale getirilmiş veri seti üzerinde tahmin modelleri” olmak üzere tüm sonuçlar değerlendirme bölümünde raporlanmıştır.

Araştırma problemi için en iyi performansı gösteren sınıflandırıcı modellerin, topluluk öğrenmesi algoritmaları olduğu görülmektedir. SMOTE tekniği ile dengeli hale getirilmiş veri seti üzerine uygulanan teknikler içerisinde en iyi doğruluk (0,928) ve F1-skor (0,855) değerlerine sahip Rastgele Orman Algoritması olmuştur. Rastgele Orman Algoritmasının ADASYN, ROSE ve SMOTE-ENN teknikleri dengeli hale getirilmiş veri seti için de çok yakın sonuçlar ürettiği görülmüştür. “varis_suresi”, “konaklama_ucreti”, “tedavi_satis_ucreti”, “tekrarlanan_misafir” ve “tur_durumu” bağımsız değişkenleri, tahmin modeli sonucuna en çok etki eden nitelikler olarak ön plana çıkmaktadır.

Bir rezervasyon sisteminde, sistem tarafından iptal edilmesi muhtemel olan rezervasyonların tanımlanması çok kıymetli bir özelliktir. Bu özellik ile işletmeler, daha fazla sayıda müşteriyle, daha verimli bir şekilde iletişim kurma imkanına sahip olurlar. İptal olasılığı yüksek olan rezervasyonların sistematik olarak tahmin edilmesi neticesinde, oteller iptalleri önlemek için harekete geçebilir ve iptallerin yüzde 37 oranında düşmesini sağlayabilirler (Antonio vd., 2019b).

“Veri madenciliği teknikleri ve makine öğrenimi modelleriyle sağlık turizmi rezervasyon verileri kullanarak, rezervasyon iptallerini tahmin etmek mümkündür.” ve “Rezervasyon iptallerini başarılı bir şekilde tahmin eden tahmin modeli çevrimiçi karar destek sistemi içerisine yerleştirilebilir.” araştırma soruları başarılı bir şekilde doğrulanmıştır. Araştırma kapsamında en iyi tahmin model sonucunu elde edebilmek için en iyi parametre araştırmaları yapılmış ve sistemin uygulamaya nasıl konulacağına ait çalışmalar başarı ile tamamlanmıştır.

Tüm araştırmalarda olduğu gibi bu araştırmanın da sınırlılıkları bulunmakta birlikte, bu durum daha ileri araştırmalar için bir fırsat sunmaktadır. Rezervasyon sistemin gerçek dünya verilerinin geri bildirimi ile tahmin modelini otomatik olarak geliştirmesi konusu, gelecekteki araştırmacılar tarafından ele alınabilir. Üzerinde işlem yapılan rezervasyonlar, işlem yapılmayan rezervasyonlara göre daha az sıklıkta iptal edildiğinden, hangi eylemin uygulandığını gösteren bir nitelik, model performansına destek sağlayabilir. İptali önlemek için her rezervasyon için gerçekleştirilen eylemlerin kaydedilmesi (promosyon verme, üst pakete terfi ettirme veya indirim yapma), tahmin edilen rezervasyonların her biri için hangi eylemlerin gerçekleştirilmesi gerektiğini önerebilen başka bir makine öğrenimi modeli hazırlanabilir.

KAYNAKÇA

- Abiteboul, S., Hull, R. ve Vianu, V. (1995). *Foundations of Databases*. Addison-Wesley, US.
- Abraham, A. (2005). *Handbook of Measuring System Design: Artificial Neural Networks*. Wiley.
- Adil, M., Ansari, M. F., Alahmadi, A., Wu, J. Z. ve Chakraborty, R. K. (2021). “Solving the Problem of Class Imbalance in the Prediction of Hotel Cancellations: A Hybridized Machine Learning Approach”. *Processes*, 9(10): 1-17.
- Akdeve, E., Barca, M. ve Gedik, İ. (2013). “Türkiye Sağlık Turizm Sektörünün Analizi ve Strateji Önerileri”. *İşletme Araştırmaları Dergisi*, 5(3): 64-92.
- Akkaya, G. C., Demireli, E. ve Yakut, Ü. H. (2009). “İşletmelerde Finansal Başarısızlık Tahminlemesi: Yapay Sinir Ağları ile İMKB Üzerine Bir Uygulama”. *Eskişehir Osmangazi Üniversitesi Sosyal Bilimler Dergisi*, 10(2): 187-216.
- Akpınar, H. (2017). *DATA: Veri Madenciliği-Veri Analizi*. (2. bs.). Papatya Yayıncılık, İstanbul.
- Aktepe, C. (2021). “Sağlık Turizminde Yeni Fırsatlar ve Türkiye’de Yerleşik Sağlık İşletmelerinin Pazarlama Çabaları”. *İşletme Araştırmaları Dergisi*, 5(1): 170-188.
- Albon, C. (2018). *Machine Learning with Python Cookbook: Practical Solutions from Preprocessing to Deep Learning*. O’Reilly Media, CA.
- Allee, V. (1997). “12 Principles of Knowledge Management”. *Training & Development*, 51: 71-74.
- Almotiri, S., Alosaimi, N. ve Abdullah, B. (2021). “Using API with Logistic Regression Model to Predict Hotel Reservation Cancellation by Detecting the Cancellation Factors”. *(IJACSA) International Journal of Advanced Computer Science and Applications*, 12(6): 750-759.
- Altunışık, R., Coşkun, R., Bayraktaroğlu, S. ve Yıldırım, E. (2012). *Sosyal Bilimlerde Araştırma Yöntemleri SPSS Uygulamaları*. Sakarya Kitabevi, Sakarya.
- Altunkaynak, B. (2019). *Veri Madenciliği Yöntemleri ve R Uygulamaları*. Seçkin Yayıncılık, Ankara.
- Anderson, D. R., Sweeney, D. J. ve Williams, T. A. (2010). *Statistics for Business and Economics*. (11th ed.). South-Western College Pub, CA.
- Andriawan, Z. A., Purnama, S. R., Darmawan, A. S., Ricko, Wibowo, A., Sugiharto, A. ve Wijayanto, F. (2020). “Prediction of Hotel Booking Cancellation using CRISP-DM”.

4th International Conference on Informatics and Computational Sciences (ICICoS).
10-11 November 2020, Semarang.

- Antonio, N. (2019). *Hotel Revenue Management: Using Data Science to Predict Booking Cancellations*. Doctoral Thesis. Instituto Universitario de Lisboa, Lisboa.
- Antonio, N., Almeida, A. M. De ve Nunes, L. (2016). "Using Data Science to Predict Hotel Booking Cancellations". *Handbook of Research on Holistic Optimization Techniques in the Hospitality, Tourism, and Travel Industry*, 141-167, IGI Global.
- Antonio, N., Almeida, A. M. De ve Nunes, L. (2017a). "Predicting Hotel Bookings Cancellation with a Machine Learning Classification Model". *16th IEEE International Conference on Machine Learning and Applications (ICMLA)*. 18-21 December 2017, Cancun, 1049-1054.
- Antonio, N., Almeida, A. M. De ve Nunes, L. (2017b). "Predicting Hotel Booking Cancellations to Decrease Uncertainty and Increase Revenue". *Tourism & Management Studies*, 13(2): 25-39.
- Antonio, N., Almeida, A. M. De ve Nunes, L. (2019a). "Big Data in Hotel Revenue Management: Exploring Cancellation Drivers to Gain Insights into Booking Cancellation Behavior". *Cornell Hospitality Quarterly*, 60(4): 298-319.
- Antonio, N., Almeida, A. M. De ve Nunes, L. (2019b). "An Automated Machine Learning Based Decision Support System to Predict Hotel Booking Cancellations". *Data Science Journal*, 18(1): 1-20.
- Arı, A. ve Önder, H. (2013). "Farklı Veri Yapılarında Kullanılabilecek Regresyon Yöntemleri". *Anadolu Tarım Bilimleri Dergisi*, 28(3): 168-174.
- Arvanitakis, K., Karydis, I., Kermanidis, K. ve Avlonitis, M. (2019). "A Machine Learning Approach for Asperities' Location Identification". *Evolving Systems*, 10(1): 41-50.
- Aslanova, K. (2013). "Türkiye'de Sağlık Turizmi ve Sağlık Turizmi Hukuku". *Avrasya Uluslararası Araştırmalar Dergisi*, 3(3): 129-145.
- Ateş, E. C. (2021). *Önleyici Kolluk Faaliyetleri Kapsamında Sosyal Medyada Veri Madenciliği: Suçları Önlemek için Twitter Analytic'i Kullanma*. Doktora Tezi. Ankara Üniversitesi Sağlık Bilimleri Enstitüsü, Ankara.
- Aydın, S. (2007). *Veri Madenciliği ve Anadolu Üniversitesi Uzaktan Eğitim Sisteminde Bir Uygulama*. Doktora Tezi. Anadolu Üniversitesi Sosyal Bilimler Enstitüsü, Eskişehir.
- Balaban, M. ve Kartal, E. (2015). *Veri Madenciliği ve Makine Öğrenmesi Temel Algoritmaları ve R Dili ile Uygulamaları*. Çağlayan Kitabevi, İstanbul.

- Batista, G., Prati, R. ve Monard, M.-C. (2004). "A Study of the Behavior of Several Methods for Balancing Machine Learning Training Data". *SIGKDD Explorations*, 6: 20-29.
- Bell, D., ve Mgbemena, C. (2018). "Data-driven Agent-based Exploration of Customer Behavior". *Simulation*, 94(3): 195-212.
- Berry, M. J. A. ve Linoff, G. S. (2004). *Data Mining Techniques for Marketing, Sales, and Customer Relationship Management*. (2nd ed.). Wiley, Indianapolis.
- Bhatia, P. (2019). *Data Mining and Data Warehousing Principles and Practical Techniques*. Cambridge University Press.
- Bhatt, G. D. (2001). "Knowledge Management in Organizations: Examining the Interaction between Technologies, Techniques, and People". *Journal of Knowledge Management*, 5(1): 68-75.
- Bilgiç, H. H., Yağlı, H., Koç, A. ve Yapıcı, A. (2016). "Deneyisel Bir Organik Rankine Çevriminde Yapay Sinir Ağları (YSA) Yardımıyla Güç Tahmini". *Selçuk Üniversitesi Mühendislik, Bilim ve Teknoloji Dergisi*, 4(1): 7-17.
- Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer, Singapore.
- Blei, D. (2012). "Probabilistic Topic Models". *Communications of the ACM*, 55(4): 77-84.
- Blei, D. M. ve Lafferty, J. D. (2009). *Text Mining: Topic Models*. Chapman and Hall/CRC.
- Blei, D. M., Ng, A. Y. ve Jordan, M. I. (2003). "Latent Dirichlet Allocation". *The Journal of Machine Learning Research*, 3: 993-1022.
- Bookman, M. Z. ve Bookman, K. R. (2007). *Medical Tourism in Developing Countries*. Palgrave Macmillan, London.
- Boschetti, A. ve Massaron, L. (2016). *Python Data Science Essentials*. (2nd ed.). Wiley Publishing, US.
- Boz, M., Canbazoglu, E., Özen, Z. ve Gülseçen, S. (2018). "Otel Rezervasyon İptallerinin Makine Öğrenmesi Yöntemleri ile Tahmin Edilmesi". *Veri Bilimi*, 1(1): 7-14.
- Bramer, M. (2020). *Principles of Data Mining*. Springer, London.
- Breiman, L. (1996). "Bagging Predictors". *Machine Learning*, 24(2): 123-140.
- Breiman, L. (2001). "Random forests". *Machine learning*, 45: 5-32.
- Budak, H. ve Erpolat, S. (2012). "Kredi Riski Tahmininde Yapay Sinir Ağları ve Lojistik Regresyon Analizi Karşılaştırılması". *Academic Journal of Information Technology*, 3(9): 23-30.
- Burma, Z. A. (2009). *Veri Tabanı Yönetim Sistemleri ve SQL/PL-SQL/T-SQL*. Seçkin Yayıncılık, Ankara.

- Can, Ş. (2022). “R” Programlama Dilinde Tahmin Edici Veri Madenciliği Algoritmalarının Modellenmesi ve Performanslarının Karşılaştırılması. Doktora Tezi. Manisa Celal Bayar Üniversitesi Sosyal Bilimler Enstitüsü, Manisa.
- Canapero, M. ve Bertinato, L., “Health Tourism-The Growth Phenomenon”. <https://www.asianhhm.com/healthcare-management/health-tourism> (erişim tarihi: 28.01.2024).
- Canty, M. J. (2014). *Image Analysis, Classification and Change Detection in Remote Sensing: With Algorithms for ENVI/IDL and Python*. (3rd ed.). CRC Press, Boca Raton.
- Cemaloğlu, N. ve Duykuluoğlu, A. (2020). *Sosyal Bilimlerde Veri Madenciliği*. Pegem Akademi Yayıncılık, Ankara.
- Chadha A., (2021). “Defining Knowledge, Information, Data”. <https://www.phpkb.com/kb/article/defining-knowledge-information-data-239.html> (erişim tarihi: 27.11.2023).
- Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C. ve Wirth, R. (2000). “CRISP-DM 1.0: Step-by-step Data Mining Guide.” *SPSS inc*, 9(13): 1-73.
- Chatterjee, S. ve Hadi, A. S. (2012). *Regression Analysis by Example*. (5th ed.). John Wiley & Sons, New Jersey.
- Chen, T. ve Guestrin, C. (2016). “XGBoost: A Scalable Tree Boosting System”. *KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785-794.
- Chengsheng, T., Huacheng, L., & Bing, X. (2017). “AdaBoost Typical Algorithm and Its Application Research”. *MATEC Web of Conferences* 139: 222.
- Chhetry, S., “Handling Imbalanced Dataset In Machine Learning”. <https://shivamkc01.medium.com/handling-imbalanced-dataset-in-machine-learning-9ac075787e07> (erişim tarihi: 16.05.2023).
- Cios, K., Pedrycz, W., Swiniarski, R. ve Kurgan, L. (2007). *Data Mining: A Knowledge Discovery Approach*. Springer, Boston.
- Cirillo, C., Bastin, F. ve Hetrakul, P. (2018). “Dynamic Discrete Choice Model for Railway Ticket Cancellation and Exchange”. *Transportation Research Part E: Logistics and Transportation Review*, 110: 137-146.
- Clarke, B., Fokoue, E. ve Zhang, H. H. (2009). *Principles and Theory for Data Mining and Machine Learning*. Springer, New York.
- Cotha, N. K. P. ve Sokolova, M. (2015). “Multi-label Learning in Classification of Patients’ Quasi-identifiers”. *Progress in Artificial Intelligence*, 4(3): 37-48.

- Cover, T. ve Hart, P. (1967). "Nearest Neighbor Pattern Classification". *IEEE Transactions on Information Theory*, 13(1): 21-27.
- Cristianini, N. ve Shawe-Taylor, J. (2000). *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*. Cambridge University Press.
- C.-Sánchez, E., Sánchez-Medina, A.J., Romero-Domínguez, L. (2022). "Forecasting Hotel-booking Cancellations Using Personal Name Records: An Artificial Intelligence Approach". J.L. Reis, E.P. López, L. Moutinho, J.P.M.d. Santos (Ed.), *Marketing and Smart Technologies. Smart Innovation, Systems and Technologies*. Springer, Singapore.
- Cunningham, P. ve Delany, S. J. (2020). "k-Nearest Neighbour Classifiers - A Tutorial". *ACM Computing Surveys (CSUR)*, 54(6): 1-25.
- Çavuşoğlu, T. (2021). *Yapay Sinir Ağları ile Türkiye Ekonomisindeki Doğrudan Yabancı Sermaye Yatırımları*. Yüksek Lisans Tezi. Karabük Üniversitesi Lisansüstü Eğitim Enstitüsü, Karabük.
- Dangeti, P. (2017). *Statistics for Machine Learning*. Packt Publishing, Birmingham.
- Davenport, T. H. ve Prusak, L. (2000). *İş Dünyasında Bilgi Yönetimi: Kuruluşlar Elleriindeki Bilgiyi Nasıl Yönetirler?*. (Çev. G. Günay). Rota Yayın Yapım Tanıtım, İstanbul.
- Demir, E. ve Dinçer, S. E. (2020). "Üretim Sektöründe Veri Madenciliği Uygulamaları: Literatür Taraması". *Anadolu Üniversitesi İşletme Fakültesi Dergisi*, 2(1): 1-12.
- Deshpande, S. P., & Thakare, V. M. (2010). "Data Mining System and Applications: A Review". *International Journal of Distributed and Parallel Systems*, 1(1): 32-44.
- Duman, O. (2016, Haziran). *Batı Anadolu Bölgesindeki Depremlerin Yapay Sinir Ağı Yöntemiyle Tahmini*. Yüksek Lisans Tezi. Gümüşhane Üniversitesi Sosyal Bilimler Enstitüsü, Gümüşhane.
- Edwards, D., Griffin, T. ve Hayllar, B. (2008). "Urban Tourism Research: Developing an Agenda". *Annals of Tourism Research*, 35(4): 1032-1052.
- Ersöz, F. (2019). *Veri Madenciliği Teknikleri ve Uygulamaları*. Seçkin Yayıncılık, Ankara.
- Field, A. (2018). *Discovering Statistics Using IBM SPSS Statistics*. (5th ed.). SAGE.
- Filiz, E. ve Öz, E. (2017). "Classification of BIST-100 Index' Changes via Machine Learning Methods". *Marmara Üniversitesi İktisadi ve İdari Bilimler Dergisi*, 39(1): 117-129.
- Finney, J. M. (1972). "Indirect Effects in Path Analysis". *Sociological Methods & Research*, 1(2): 175-186.

- Freund, Y. ve Schapire, R. E. (1997). "A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting". *Journal of Computer and System Sciences*, 55(1): 119-139.
- Gaganis, C., Pasiouras, F., Spathis, C. ve Zopounidis, C. (2007). "A Comparison of Nearest Neighbors, Discriminant and Logit Models for Auditing Decisions". *Intelligent Systems in Accounting, Finance & Management: International Journal*, 15(1-2): 23-40.
- Gayar, N. F. El, Saleh, M., Atiya, A., El-Shishiny, H., Zakhary, A. A. Y. F. ve Habib, H. A. A. M. (2011). "An Integrated Framework for Advanced Hotel Revenue Management". *International Journal of Contemporary Hospitality Management*, 23(1): 84-98.
- GeekforGeeks.org, "Güçlendirme Algoritmaları Arasındaki Farklar". <https://www.geeksforgeeks.org/gradientboosting-vs-adaboost-vs-xgboost-vs-catboost-vs-lightgbm> (erişim tarihi: 26.12.2023).
- Geron, A. (2019). *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*. (2nd ed.). O'Reilly Media.
- Google Machine Learning, "Imbalanced Data". <https://developers.google.com/machine-learning/data-prep/construct/sampling-splitting/imbalanced-data> (erişim tarihi: 12.05.2023).
- Gökçen, H. (2007). *Yönetim Bilgi Sistemleri*. Palme Yayıncılık, Ankara.
- Gökmen, Ş. (2014). *Müşteri İlişkileri Yönetiminde Bir Araç Olarak Veri Madenciliği ve Parekende Sektöründe Bir Uygulama*. Yüksek Lisans Tezi. Yıldız Teknik Üniversitesi Sosyal Bilimler Enstitüsü, İstanbul.
- Gürsoy, U. T. Ş. (2009). *Veri Madenciliği ve Bilgi Keşfi*. Pegem Akademi Yayıncılık, Ankara.
- Halperin, M., Blackwelder, W. C. ve Verter, J. I. (1971). "Estimation of the Multivariate Logistic Risk Function: A Comparison of the Discriminant Function and Maximum Likelihood Approaches". *Journal of Chronic Diseases*, 24(2): 125-158.
- Hamzaçebi, C. (2011). *Yapay Sinir Ağları: Tahmin Amaçlı Kullanımı, MATLAB ve NeuroSolutions Uygulamalı*. Ekin Yayınevi, Bursa.
- Han, J., Kamber, M. ve Pei, J. (2012). *Data Mining Concepts and Techniques*. (3rd ed.). Morgan Kaufmann Publishers, Waltham.
- Hand, D. J., Smyth, P. ve Mannila, H. (2001). *Principles of Data Mining*. MIT Press, Cambridge.

- Hardin, J.M., Chhieng, D.C. (2007). "Data Mining and Clinical Decision Support Systems". E.S. Berner (Ed.), *Clinical Decision Support Systems: Theory and Practice*. (2nd ed.). Springer, New York.
- Haykin, S. S. (2009). *Neural Networks and Learning Machines*. Prentice Hall.
- Hazim, L. R. (2018). *Four Classification Methods Naïve Bayesian, Support Vector Machine, K-Nearest Neighbors and Random Forest Are Tested For Credit Card Fraud Detection*. Master's Thesis. Altinbas University Master of Information Technology, Istanbul.
- Hosmer, D. W., Lemeshow, S. ve Sturdivant, R. X. (2013). *Applied Logistic Regression*. (3rd ed.). John Wiley & Sons, New Jersey.
- Huang, T. M., Kecman, V. ve Kopriva, I. (2006). *Kernel Based Algorithms for Mining Huge Data Sets (Vol. 1)*. Springer-Verlag, Heidelberg.
- Hussain, S., Calvo, R. ve Chen, F. (2013). "Automatic Cognitive Load Detection from Face, Physiology, Task Performance and Fusion During Affective Interference". *Interacting with Computers*, 26(3): 256-268.
- Ikonomakis, E., Kotsiantis, S. ve Tampakas, V. (2005). "Text Classification Using Machine Learning Techniques". *WSEAS Transactions on Computers*, 4(8): 966-974.
- Intezari, A., Taskin, N. ve Pauleen, D. J. (2017). "Looking Beyond Knowledge Sharing: An Integrative Approach to Knowledge Management Culture". *Journal of Knowledge Management*, 21(2): 492-515.
- Jabbar, M. A., Deekshatulu, B. L. ve Chandra, P. (2013). "Classification of Heart Disease Using K- Nearest Neighbor and Genetic Algorithm". *Procedia Technology*, 10: 85-94.
- Jackson, J. (2002). "Data Mining; A Conceptual Overview". *Communications of the Association for Information Systems*, 8(1): 19.
- Joint Commission International, "Search for JCI-Accredited Organizations". <https://www.jointcommissioninternational.org/who-we-are/accredited-organizations> (erişim tarihi: 10.03.2024).
- John, G. H. ve Langley, P. (1995). "Estimating Continuous Distributions in Bayesian Classifiers". *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, 338-345.
- Kantar, G. ve Işık, E. (2014). "Türkiye’de Sağlık Turizmi". *Sağlık Akademisyenleri Dergisi*, 1(1): 15-20.
- Kantardzic, M. (2011). *Data Mining: Concepts, Models, Methods, and Algorithms*. (2nd ed.). Wiley, New Jersey.

- Kapanoğlu, M. (2016). *Optimizasyon Yönelimli Karar Destek Sistemleri*. Anadolu Üniversitesi Açıköğretim Fakültesi, Eskişehir.
- Kaplan, M. (2016). “Kültürel Bir İnşa Süreci olarak Sağlık: Kavramsal ve Tarihsel Bir Bakış”. *Dört Öge*, (10): 11-18.
- Kaya, A. (2018). “Bilgi, Veri ve Enformasyon Arasındaki İlişki Nedir?” <https://techworm.com/bilgi-veri-ve-enformasyon-arasindaki-iliski-nedir> (erişim tarihi: 7.12.2023).
- Kaya, M., & Özel, S. A. (2014). “Açık Kaynak Kodlu Veri Madenciliği Yazılımlarının Karşılaştırılması”. *Akademik Bilişim’14 - XVI. Akademik Bilişim Konferansı Bildirileri*, 47-53.
- KDNuggets, “Python Leads the 11 Top Data Science, Machine Learning Platforms: Trends and Analysis”. <https://www.kdnuggets.com/2019/05/poll-top-data-science-machine-learning-platforms.html> (erişim tarihi: 22.09.2023).
- Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., ... & Liu, T. Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, 30, 3149-3157.
- Keen, P. G. W. ve Morton, M. S. S. (1978). *Decision Support Systems : An Organizational Perspective*. Addison-Wesley Publication.
- Kefe, İ. (2013). “6102 Sayılı Türk Ticaret Kanunu’nun Muhasebe Bilgi Sistemi Uygulamalarına Getirdiği Düzenlemeler”. *Elektronik Mesleki Gelişim ve Araştırmalar Dergisi*, 1(1), 20-33.
- Kelleher, J. D., ve Tierney, B. (2018). *Data Science*. MIT Press.
- Khamis, H. S., Cheruiyot, K. W. ve Kimani, S. (2014). “Application of k-Nearest Neighbour Classification in Medical Data Mining”. *International Journal of Information and Communication Technology Research*, 4(4): 121-128.
- Khan, M., Ding, Q. ve Perrizo, W. (2001). “K-Nearest Neighbor Classification on Spatial Data Streams Using P-Trees”. *In Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 517-528.
- Kimes, S. ve Wirtz, J. (2003). “Has Revenue Management become Acceptable? Findings from an International Study on the Perceived Fairness of Rate Fences”. *Journal of Service Research*, 6(2): 125-135.
- Kotsiantis, S. (2007). “Supervised Machine Learning: A Review of Classification Techniques”. *Informatica*, 31: 249-268.
- Köse, İ. (2018). *Veri Madenciliği Teori Uygulama ve Felsefesi*. Papatya Yayıncılık, İstanbul.

- Kumdereli, Ü. C. (2012). *Tıp Bilişimi ve Veri Madenciliği Uygulamaları: EEG Sinyallerindeki Epileptiform Aktiviteye Veri Madenciliği Yöntemlerinin Uygulanması*. Yüksek Lisans Tezi. Trakya Üniversitesi Fen Bilimleri Enstitüsü, Edirne.
- Land, S. ve Fischer, S. (2012). *RapidMiner 5: RapidMiner in Academic Use*. Rapid-I GmbH.
- Lane, H., Hapke, H. ve Howard, C. (2019). *Natural Language Processing in Action: Understanding, Analyzing, and Generating Text with Python*. Manning Publications, New York.
- Lantz, B. (2013). *Machine Learning with R*. Packt Publishing, Birmingham.
- Larose, D. T. (2005). *Discovering Knowledge in Data: An Introduction to Data Mining*. Wiley, New Jersey.
- Laudon, K. C. ve Laudon, J. P. (2014). *Management Information Systems: Managing the Digital Firm*. Pearson Educación, Mexico.
- Laurinen, P. (2006). *A Top-Down Approach for Creating and Implementing Data Mining Solutions*. University of Oulu Faculty of Technology, Finland.
- Lavesson, N., Boldt, M., Davidsson, P. ve Jacobsson, A. (2011). "Learning to Detect Spyware Using End User License Agreements". *Knowledge and Information Systems*, 26: 285-307.
- Ledolter, J. (2013). *Data Mining and Business Analytics with R*. Wiley, New Jersey.
- Lemke, C., Riedel, S. ve Gabrys, B. (2009). "Dynamic Combination of Forecasts Generated by Diversification Procedures Applied to Forecasting of Airline Cancellations". *IEEE/IAFE Conference on Computational Intelligence for Financial Engineering, Proceedings (CIFEr)* 85-91.
- Lemke, C., Riedel, S. ve Gabrys, B. (2013). "Evolving Forecast Combination Structures for Airline Revenue Management". *Journal of Revenue and Pricing Management*, 12(3): 221-234.
- Liaw, A., ve Wiener, M. (2002). "Classification and Regression by RandomForest". *R news*, 2(3): 18-22.
- Ma, J., Yu, Z., Qu, Y., Xu, J. ve Cao, Y. (2019). "Application of the XGBoost Machine Learning Method in PM2.5 Prediction: A Case Study of Shanghai". *Aerosol and Air Quality Research*, 20(1): 128-138.
- MacKay, D. J. C. (2003). *Information Theory, Inference and Learning Algorithms*. Cambridge University Press.
- Mårtensson, M. (2000). "A Critical Review of Knowledge Management as a Management Tool". *Journal of Knowledge Management*, 4(3): 204-216.

- Martin, S. ve Scott, J. T. (2000). "The Nature of Innovation Market Failure and the Design of Public Support for Private Innovation". *Research Policy*, 29(4): 437-447.
- McCulloch, W. S. ve Pitts, W. (1943). "A Logical Calculus of the Ideas Immanent in Nervous Activity". *The Bulletin of Mathematical Biophysics*, 5(4): 115-133.
- Mehrotra, R., ve Ruttley, J. (2006). *Revenue Management*. American Hotel and Lodging Association.
- Melgani, F. ve Bruzzone, L. (2004). "Classification of Hyperspectral Remote Sensing Images with Support Vector Machines". *IEEE Transactions on Geoscience and Remote Sensing*, 42(8): 1778-1790.
- Mert, M. (2016). *Yatay Kesit Veri Analizi Bilgisayar Uygulamaları*. Detay Yayıncılık.
- Mertler, C. ve Reinhart, R. (2017). *Advanced and Multivariate Statistical Methods: Practical Application and Interpretation*. (6th ed.). Routledge.
- Metzger, A., Franklin, R. ve Engel, Y. (2012). "Predictive Monitoring of Heterogeneous Service-Oriented Business Networks: The Transport and Logistics Case". Annual SRII Global Conference, 313-322.
- Mitchell, T. M. (1999). "Machine Learning and Data Mining". *Communications of the ACM*, 42(11): 30-36.
- Morales, D. R. ve Wang, J. (2010). "Forecasting Cancellation Rates for Services Booking Revenue Management Using Data Mining". *European Journal of Operational Research*, 202(2): 554-562.
- Murgia, A., Ortu, M., Tourani, P., Adams, B. ve Demeyer, S. (2018). "An Exploratory Qualitative and Quantitative Analysis of Emotions in Issue Report Comments of Open Dource Systems". *Empirical Software Engineering*, 23(1): 521-564.
- Murphy, K. P. (2012). *Machine Learning : A Probabilistic Perspective*. MIT Press, London.
- Nelder, J. A. ve Wedderburn, R. W. M. (1972). "Generalized Linear Models". *Journal of the Royal Statistical Society Series A: Statistics in Society*, 135(3): 370-384.
- Nikita, M. ve Chaney, N. (2020). "Tuning of the Latent Dirichlet Allocation Models Parameters". <https://cran.r-project.org/web/packages/ldatuning/ldatuning.pdf> (erişim tarihi: 23.12.2023).
- Oğuzlar, A., ve Kızılkaya, Y. M. (2019). *Metin Madenciliğinde Duygu Analizi*. Dora Yayıncılık, Bursa.
- Olson, D. L. ve Delen, D. (2008). *Advanced Data Mining Techniques*. Springer, US.

- Osowski, S., Tran, L. ve Markiewicz, T. (2004). "Support Vector Machine-Based Expert System for Reliable Heartbeat Recognition". *IEEE Transactions on biomedical engineering*, 51(4): 582-589.
- Öğücü, M. O. (2006). *Yapay Sinir Ağları ile Sistem Tanıma*. Yüksek Lisans Tezi. İstanbul Teknik Üniversitesi Fen Bilimleri Enstitüsü, İstanbul.
- Önder, H. ve Cebeci, Z. (2002). "Lojistik Regresyonlarda Değişken Seçimi". *Çukurova Üniversitesi Ziraat Fakültesi Dergisi*, 17: 105–114.
- Özdemir, K. (2004). *Paket Programlar ile İstatistiksel Veri Analizi*. (5. bs.). Kaan Kitabevi, Eskişehir.
- Özer, Ö., ve Songur, C. (2012). "Türkiye'nin Dünya Sağlık Turizmindeki Yeri ve Ekonomik Boyutu". *Mehmet Akif Ersoy Üniversitesi Sosyal Bilimler Enstitüsü Dergisi*, 4(7): 69-81.
- Özkan, M. ve Boran, L. (2014). "Veri Madenciliğinin Finansal Kararlarda Kullanımı". *Çankırı Karatekin Üniversitesi İktisadi ve İdari Bilimler Fakülte Dergisi*, 4(1): 59-82.
- Özkan, Y. (2016). *Veri Madenciliği Yöntemleri*. (4 bs.). Papatya Bilim.
- Özkurt, H. (2007). "Sağlık Turizmi Tahvilleri". *Maliye Dergisi*, 152(1): 122-142.
- Öztemel, E. (2006). *Yapay Sinir Ağları*. (2. bs.). Papatya Yayıncılık, İstanbul
- Öztürk, M. ve Bayat, M. (2011). "Uluslararası Turizm Hareketlerinde Sağlık Turizminin Rolü ve Kalite Çalışmalarının Önemi Bir Literatür Çalışması". *Kahramanmaraş Sütçü İmam Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi*, 1(2): 135-156.
- Patel, B. N., Prajapati, S. G. ve Lakhtaria, K. I. (2012). "Efficient Classification of Data Using Decision Tree". *Bonfring International Journal of Data Mining*, 2(1): 6.
- Pektaş, A. O. (2013). *SPSS ile Veri Madenciliği*. Dikeyksen Yayın Dağıtım, İstanbul.
- Pérez, F. A. A., Ortiz, G. E. R., Muñiz, E. R., Sacarello, F. J. O., Kang, J. E. ve Rodriguez-Roman, D. (2020). "Predicting Trip Cancellations and No-Shows in Paratransit Operations". *Transportation Research Record*, 2674(8): 774-784.
- Piramuthu, S. (2004). "Evaluating Feature Selection Methods for Learning in Data Mining Applications". *European Journal of Operational Research*, 156(2): 483-494.
- Powers, D. A. ve Xie, Y. (1999). *Statistical Methods for Categorical Data Analysis*. Emerald Group Publishing, Bingley.
- Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V. ve Gulin, A. (2018). "CatBoost: Unbiased Boosting with Categorical Features". *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, 6639-6649.
- Pujari, A. K. (2008). *Data Mining Techniques*. Universities Press.

- Pulugurtha, S. S. ve Nambisan, S. S. (2003). "A Decision-Support Tool for Airline Yield Management Using Genetic Algorithms". *Computer-Aided Civil and Infrastructure Engineering*, 18(3): 214-223.
- Puri, C., Akhil, Gaurav, Kush ve Kumar, N. (2017). "Opinion Ensembling for Improving Economic Growth through Tourism". *Procedia Computer Science*, 122: 237-244.
- Qi, C., Wang, Y., Tian, W. ve Wang, Q. (2016). "Multiple Kernel Boosting Framework Based on Information Measure for Classification". *Chaos, Solitons & Fractals*, 89: 175-186.
- Rakesh, M. V., Kumar, S. P., Yogitha ve Aishwarya., R. (2022). "Hotel Booking Cancellation Prediction using ML algorithms". *Proceedings of the 2nd International Conference on Artificial Intelligence and Smart Energy (ICAIS 2022)*, 466-471.
- Rashidi, M., Ghodrat, M., Samali, B., Kendall, B., & Zhang, C. (2017). "Remedial Modelling of Steel Bridges Through Application of Analytical Hierarchy Process (AHP)". *Applied Sciences*, 7(2): 168.
- Refaat, M. (2007). *Data Preparation for Data Mining Using SAS*. Elsevier, CA.
- Ripley, B. (2002). *Statistical Data Mining*. Springer-Verlag, New York.
- Sánchez, E. C., Sánchez-Medina, A. J., & Pellejero, M. (2020). "Identifying Critical Hotel Cancellations Using Artificial Intelligence". *Tourism Management Perspectives*, 35, 100718.
- Sánchez-Medina, A. J., & Eleazar, C. (2020). "Using machine learning and big data for efficient forecasting of hotel booking cancellations". *International Journal of Hospitality Management*, 89, 102546.
- Sarkar, D. (2019). *Text Analytics with Python: A Practitioner's Guide to Natural Language Processing*. (2nd ed.). APress, Bangalore.
- Satu, M. S., Ahammed, K. ve Abedin, M. Z. (2020). "Performance Analysis of Machine Learning Techniques to Predict Hotel booking Cancellations in Hospitality Industry". *ICCIT 23rd International Conference on Computer and Information Technology*, 19-21 December, 2020.
- Savaş, S., Topaloğlu, N., & Yılmaz, M. (2012). "Veri Madenciliği ve Türkiye'deki Uygulama Örnekleri". *İstanbul Ticaret Üniversitesi Fen Bilimleri Dergisi*, 11(21): 1-23.
- Seifert, J. W. (2004). "Data Mining and the Search for Security: Challenges for Connecting the Dots and Databases". *Government Information Quarterly*, 21(4): 461-480.

- Sharif Azadeh, S., Labib, R. ve Savard, G. (2013). "Railway Demand Forecasting in Revenue Management Using Neural Networks". *International Journal of Revenue Management*, 7(1): 18-36.
- Shehadeh, A., Alshboul, O., Al Mamlook, R. E. ve Hamedat, O. (2021). "Machine Learning Models for Predicting the Residual Value of Heavy Construction Equipment: An Evaluation of Modified Decision Tree, LightGBM, and XGBoost Regression". *Automation in Construction*, 129, 103827.
- Shiksha Online, "Difference Between Bagging and Boosting". <https://www.shiksha.com/online-courses/articles/bagging-and-boosting/> (erişim tarihi: 10.11.2023).
- Shmueli, G., Bruce, P. C. ve Patel, N. R. (2016). *Data Mining for Business Analytics: Concepts, Techniques, and Applications with XLMiner*. Wiley, New Jersey.
- Siering, M., Deokar, A. V ve Janze, C. (2018). "Disentangling Consumer Recommendations: Explaining and Predicting Airline Recommendations Based on Online Reviews". *Decision Support Systems*, 107: 52-63.
- Silahtaroglu, G. (2013). *Veri Madenciliği Kavram ve Algoritmalar*. (2. bs.). Papatya Yayıncılık Eğitim, İstanbul.
- Silverman, B. W. ve Jones, M. C. (1989). "E. Fix and J. L. Hodges (1951): An Important Contribution to Nonparametric Discriminant Analysis and Density Estimation: Commentary on Fix and Hodges (1951)". *International Statistical Review*, 57(3): 233-238.
- Smith, S., Parsa, H. G., Bujisic, M. ve Van Der Rest, J. P. (2015). "Hotel Cancellation Policies, Distributive and Procedural Fairness, and Consumer Patronage: A Study of the Lodging Industry". *Journal of Travel & Tourism Marketing*, 32(7): 886-906.
- Steinberg, D., ve Colla, P. (1995). *CART: Tree-Structured Non-Parametric Data Analysis*. Salford Systems, San Diego, CA.
- Stock, J. H. ve Watson Mark W. (2007). *Introduction to Econometrics*. Pearson.
- Suzuki, K. (2011). *Artificial Neural Networks - Methodological Advances And Biomedical Applications*. Intech.
- Şahin, S. K. (2021). "Sağlık Turizminin Geliştirilmesinde Sağlık Bakanlıklarının Koordinatör Kuruluşları: Türkiye, Malezya ve Güney Kore Karşılaştırması". *Gümüşhane Üniversitesi Sağlık Bilimleri Dergisi*, 10(4): 983-992.
- Şeker, Ş. E. (2018). "CRISP-DM: Endüstriler Arası Standart İşleme - Veri Madenciliği için (Cross Industry Standard Processing - Data Mining)", *YBS Ansiklopedi*, 5(2).

- Şengül, H. ve Bulut, A. (2019). “Dünyada ve Türkiye’de Sağlık Turizmi”. *Yönetim, Ekonomi ve Pazarlama Araştırmaları Dergisi*, 3(1): 45-62.
- Şerbetli, A. (2008). *Makro Ekonomik Göstergelerin Banka Karlılığı ve Sermaye / Aktif Getirisi Üzerindeki Etkisinin Yapay Sinir Ağları ile Test Edilmesi*. Yüksek Lisans Tezi. Marmara Üniversitesi Sosyal Bilimler Enstitüsü, İstanbul.
- Talluri, K. T. ve Van Ryzin, G. J. (2005). “The Theory and Practice of Revenue Management”. *International Series in Operations Research & Management Science*. Springer, US.
- Tengilimoğlu, D. (2021). “Sağlık Turizmi ve Devlet Teşvikleri”. *Journal of Life Economics*, 8(1): 1-10.
- Tolon, M. ve Güneri Tosunoğlu, N. (2008). “Tüketici Tatmini Verilerinin Analizi: Yapay Sinir Ağları ve Regresyon Analizi Karşılaştırması”. *Gazi Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi*, 10(2): 247-259.
- Topal, M. H. (2006). *Uluslararası Kuruluşların Teşviklere Bakışı ve Türk Teşvik Sisteminin Bu Çerçeve Analizi*. Yüksek Lisans Tezi. Karadeniz Teknik Üniversitesi Sosyal Bilimler Enstitüsü, Trabzon.
- Truett, J., Cornfield, J. ve Kannel, W. (1967). “A Multivariate Analysis of the Risk of Coronary Heart Disease in Framingham”. *Journal of Chronic Diseases*, 20(7): 511-524.
- Two Crows Corporation. (1999). *Introduction to Data Mining and Knowledge Discovery*. (3rd ed.). Two Crows Corp.
- Ulutürk, S. (2015). “Sağlık Ekonomisi, Sağlık Statüsü, Sağlığın Ölçülmesinde Kullanılan Ölçütler ve Önemi: Türkiye Örneği”. *Finans Politik ve Ekonomik Yorumlar*, (603): 47-63.
- Üzümcü, T. P. ve Çimen, S. (2019). “Kocaeli Sağlık Turizmine Yönelik Nitel Bir Çalışma”. *Uluslararası Turizm, Ekonomi ve İşletme Bilimleri Dergisi (IJTEBS)*, 3(2): 97-116.
- Vajirakachorn, T. ve Chongwatpol, J. (2017). “Application of Business Intelligence in the Tourism Industry: A Case Study of a Local Food Festival in Thailand”. *Tourism Management Perspectives*, 23: 75-86.
- Vercellis, Carlo. (2008). *Business Intelligence: Data Mining and Optimization for Decision Making*. Wiley, West Sussex.
- Vidhya, V., Jeyaram, G. ve Ishwarya, K. R. (2016). *Database Management Systems*. Alpha Science International Ltd., Oxford.

- Wang, T. (2021). "An Intelligent Passenger Flow Prediction Method for Pricing Strategy and Hotel Operations". *Complexity*, 2021: 1-12.
- Weiss, G. ve Davison, B. (2010). *Data Mining*. Wiley.
- Wikimedia Commons, "Single Layer Neural Network". <https://commons.wikimedia.org/wiki/File:SingleLayerNeuralNetwork.png> (erişim tarihi: 23.12.2023).
- Williams, G. (2011). *Data Mining with Rattle and R*. Springer, New York.
- Wirth, R. ve Hipp, J. (1998). "CRISP-DM: Towards a Standard Process Model for Data Mining". *Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining*. (1): 29-39.
- Yalçın, Ö. (2016). Veri Madenciliği Yöntemleri. Papatya Yayıncılık Eğitim, İstanbul.
- Yılmaz, H. ve Yılmaz, N. (2022). "Dünyada ve Türkiye’de Dijital Sağlık Turizmi Uygulamaları". *Sağlık Akademisyenleri Dergisi*, 9(1): 64-72.
- Yılmaz, M. (2009). "Enformasyon ve Bilgi Kavramları Bağlamında Enformasyon Yönetimi ve Bilgi Yönetimi". *Ankara Üniversitesi Dil ve Tarih-Coğrafya Fakültesi Dergisi*, 49(1): 95-118.
- Yurdakul, S. (2015). *Veri madenciliği ile lise öğrenci performanslarının değerlendirilmesi*. (Yüksek Lisans Tezi). Kırıkkale Üniversitesi Fen Bilimleri Enstitüsü, Kırıkkale.
- Zaccone, G. (2016). *Getting Started with Tensorflow*. Packt Publishing, Birmingham.
- Zakhary, A., Atiya, A. F., El-Shishiny, H. ve Gayar, N. E. (2011). "Forecasting Hotel Arrivals and Occupancy Using Monte Carlo Simulation". *Journal of Revenue and Pricing Management*", 10(4): 344-366.
- Zaki, M. J. (2014). *Data Mining and Analysis: Foundations and Algorithms*. Cambridge University Press, New York.
- Zaki, M. J., & Meira Jr, W. (2020). *Data Mining and Machine Learning: Fundamental Concepts and Algorithms*. (2nd ed.). Cambridge University Press, Cambridge.
- Zhou, H. (2020). *Learn Data Mining Through Excel*. Apress, Berkeley, CA.
- Zurada, J. (1992). *Introduction to Artificial Neural systems*. West Publishing.

ÖZGEÇMİŞ

Adı ve Soyadı	Erokan CANBAZOĞLU
EĞİTİM DURUMU	
Mezun Olduğu Lise	Şehitler Fen Lisesi / Polatlı Lisesi
Lisans Diploması	Marmara Üniversitesi, Teknik Eğitim Fakültesi Fakültesi, Bilgisayar ve Kontrol Öğretmenliği (İng.) Bölümü
Yüksek Lisans Diploması	Bahçeşehir Üniversitesi, Fen Bilimleri Enstitüsü, Bilgisayar Mühendisliği Ana Bilim Dalı
Yüksek Lisans Tez Konusu	The Role of Ipad in Dentistry Patient Education and Consultations
Yabancı Dil	İngilizce
BİLİMSEL FAALİYETLER	
Katıldığı Bilimsel Kongre/Sempozyum ve Bilimsel Toplantılar	
Ercan U., Irmak S., Çevik K.K., Canbazoğlu E., "Yapay Sinir Ağları Kullanılarak Konutlarda Elektrik Tüketimi Düzeylerinin Tahmin Edilmesi", SOSYO EKONOMI, no.46, pp.1-1, 2020 (Link)	
Boz M., Canbazoğlu E., Özen Z., Gülseçen S., "Otel Rezervasyon İptallerinin Makine Öğrenmesi Yöntemleri ile Tahmin Edilmesi", Veri Bilimi, cilt.1, ss.7-14, 2018 (Link)	
Canbazoğlu E., Ercan U., İpekçi Çetin E., "Ticari Taksi Araç Yenilemelerinde Ahs Ve Moora Yöntemlerine Dayalı Karar Destek Mobil Uygulaması", Uluslararası İktisadi ve İdari İncelemeler Dergisi, ss.117-134, 2018 (Link)	
Canbazoğlu E., Salman Y.B., Yildirim M., Merdenyan B., Ince I., "Developing a mobile application to better inform patients and enable effective consultation in implant dentistry.", Computational and structural biotechnology journal, vol.14, pp.252-61, 2016	
Yilmaz H., Koçer A., Canbazoğlu E., Güngör A., "Tarımsal Atıklardan Üretilen Peletlerin Parça Yoğunluğu Ve Dayanıklılık Direnci Arasındaki İlişkinin İncelenmesi", I. Uluslararası Türk Dünyası Mühendislik ve Fen Bilimleri Konferansı, Antalya, TÜRKİYE, pp.1187-1192	
Canbazoğlu E., İpekçi Çetin E., "Ticari Taksi Yenilemelerinde AHP ve MOORA Yöntemlerine Dayalı Karar Destek Mobil Uygulaması", 18. Uluslararası Ekonometri, Yöneylem Araştırması ve İstatistik Sempozyumu, TRABZON, TÜRKİYE, 5-7 Ekim 2017, pp.234-234	
Koçer A., Canbazoğlu E., Güngör A., "Karbon Ayak İzi Açısından Turizm Sektörü: Otel Örneği", I. Uluslararası Türk Dünyası Mühendislik ve Fen Bilimleri Konferansı, Antalya, TÜRKİYE, pp.130-136	
Canbazoğlu E., Ercan U., "Türkiye Hane Halklarının Balık Ve Kırmızı Et Tüketimlerinin Multinomial LOGIT Regresyon İle Analizi", 1st International Conference on Food and Agricultural Economics (ICFAEC) 2017,	
Canbazoğlu E., Koçer A., Aksoy E., Güngör A., "Mobil Uygulama İle Etkili Sunum Değerlendirme", I. Uluslararası Türk Dünyası Mühendislik ve Fen Bilimleri Konferansı, Antalya, TÜRKİYE, pp.60-69	
Canbazoğlu E., Aksoy E., San B.T., İpekçi Çetin E., "Katı Atık Depolama Alanı Seçiminde CBS Uygulamaları İçin Kriter Ağırlıklandırma Yazılımı", 18. Uluslararası Ekonometri, Yöneylem Araştırması ve İstatistik Sempozyumu, TRABZON,	

- Ercan U., Canbazoglu E., "**Hanehalkı Kırmızı Et Tüketimine Etki Eden Sosyoekonomik Ve Demografik Faktörlerin CHAID Analizi İle Belirlenmesi**", 1st International Conference on Food and Agricultural Economics (ICFAEC) 2017,
- Canbazoglu E., Koçer A., Yaka İ.F., Güngör A., "**Learning The Greenhouse Temperature And Wind Rates Via Mobile Phones**", 20th International Research/Expert Conference - Trends in the Development of Machinery and Associated Technology - TMT 2016, Venedik, ITALYA, 24 Eylül - 1 Ekim 2016, pp.193-196
- Canbazoglu E., Koçer A., "**Phonegap ile Çoklu Mobil Ortamlara Uygulama Geliştirme: Web Servis Örneği**", 18. Akademik Bilişim Konferansı, AYDIN, TÜRKİYE, 3-5 Şubat 2016, ss.00-00
- Canbazoglu E., Salman Y.B., Cheng H., "**Development of A Patient Education and Consultation Mobile Application for Implant Dentistry**", The 1st Asian Conference on Ergonomics and Design 2014: New Challenge in Asian Ergonomics and Design, ACED2014, Jeju Island, GUNEE KORE, 21-24 Mayıs 2014, pp.000-000

İDARİ GÖREVLER

Akdeniz Üniversitesi TBMYO Bilgisayar Programcılığı Bölüm Başkan Yardımcısı

İŞ DENEYİMİ

Görev Unvanı	Görev Yeri
Genel Müdür / Proje Yöneticisi /Uzman Yazılımcı	PHD Teknoloji, Antalya
Öğretim Görevlisi	Akdeniz Üniversitesi, Teknik Bilimler MYO, Bilgisayar Programcılığı, Antalya
Bilişim Teknolojileri Eğitmeni	Şişli ATL, EML, TL Bilişim Teknolojileri Alanı, İstanbul