

**T.C.
BAHÇEŞEHİR ÜNİVERSİTESİ**

**SEYAHAT SİGORTALARINDA
TAZMİNAT TALEP TAHMİNİ**

Yüksek Lisans Tezi

YUSUFCAN YALDIZ

İSTANBUL, 2020

T.C.
BAHÇEŞEHİR ÜNİVERSİTESİ

FEN BİLİMLERİ ENSTİTÜSÜ
MÜHENDİSLİK YÖNETİMİ

SEYAHAT SİGORTALARINDA
TAZMİNAT TALEP TAHMİNİ

Yüksek Lisans Tezi

YUSUFCAN YALDIZ

Tez Danışmanı: DR. ÖĞR. ÜYESİ ADNAN ÇORUM

İSTANBUL, 2020

T.C.
BAHÇEŞEHİR ÜNİVERSİTESİ

FEN BİLİMLERİ ENSTİTÜSÜ
MÜHENDİSLİK YÖNETİMİ

Tezin Adı: Seyahat sigortalarında poliçe sahiplerinin tazminat talebinde bulunup bulunmadığının Lojistik Regresyon ve Naive Bayes yöntemleriyle tahminlenmesi

Öğrencinin Adı Soyadı: Yusufcan YALDIZ
Tez Savunma Tarihi: 25.06.2020

Bu tezin Yüksek Lisans tezi olarak gerekli şartları yerine getirmiş olduğu Fen Bilimleri Enstitüsü tarafından onaylanmıştır.

Doç. Dr. Burak KÜNTAY
Enstitü Müdürü

Bu tezin Yüksek Lisans tezi olarak gerekli şartları yerine getirmiş olduğunu onaylarım.

Bu Tez tarafımızca okunmuş, nitelik ve içerik açısından bir Yüksek Lisans tezi olarak yeterli görülmüş ve kabul edilmiştir.

Jüri Üyeleri

İmzalar

Tez Danışmanı
Dr. Öğr. Üyesi Adnan ÇORUM

Üye
Dr. Öğr. Üyesi Ethem ÇANAKOĞLU

Üye
Doç. Dr. Özlem ŞENVAR

TEŞEKKÜR

Bana desteğini esirgemeyen sevgili annem ve babama teşekkür ediyorum. Yüksek lisans tezimin hazırlık sürecinde beni destekleyen sayın danışman hocam Dr. Öğr. Üyesi Adnan Çorum'a teşekkür ediyorum.

İstanbul, 2020

Yusufcan YALDIZ



ÖZET

SEYAHAT SİGORTALARINDA POLİÇE SAHİPLERİNİN TAZMİNAT TALEBİNDE BULUNUP BULUNMADIĞININ LOJİSTİK REGRESYON VE NAİVE BAYES YÖNTEMLERİYLE TAHMİNLENMESİ

Yusufcan YALDIZ

Mühendislik Yönetimi

Tez Danışmanı: Dr. Öğr. Üyesi Adnan ÇORUM

Haziran 2020, 60 Sayfa

Sigorta şirketleri için en önemli noktalardan biri sigorta poliçesi sahiplerinin yaşamış oldukları herhangi bir kaza veya olay sonrasında poliçenin hükümlerine dayanarak sigorta şirketine tazminat talebinde bulunması ve bu sürecin sonuçlandırılmasıdır. Sigorta tazminat talepleri uzun incelemeler sonucunda karara bağlanmaktadır. Bu karar süreci müşterinin bağlılığı ve memnuniyetine de etki etmektedir. Bu çalışma, seyahat sigortası tazminat talep süreçlerinin otomatize hale getirilebilmesi ve karar alma süreçlerine yardımcı olabilmesi amacıyla yapılmıştır. Bu çalışmada, sınırlı sayıda değişken ve dengesiz bir veri kümesi üzerinde Lojistik Regresyon ve Naive Bayes tahmin modelleri seyahat sigortası tazminat talebinin bulunup bulunmadığının tespiti için kullanılmıştır. Aynı zamanda Temel Bileşen Analizi kullanılarak boyut indirgeme yapılmıştır. Tüm çalışma Python programlama dili ve kütüphaneleri kullanılarak gerçekleştirilmiştir. Çalışmanın sonunda kullanılan modellerin bir karar destek mekanizması olarak kullanılabileceği görülmüştür.

Anahtar Kelimeler: Seyahat Sigortası, Tazminat Talebi, Lojistik Regresyon, Naive Bayes, Temel Bileşen Analizi

ABSTRACT

FORECASTING TRAVEL INSURANCE CLAIM BY USING LOGISTIC REGRESSION AND NAIVE BAYES FOR POLICY HOLDERS

Yusufcan YALDIZ

ENGINEERING MANAGEMENT

Thesis Supervisor: Assist. Prof. Dr. Adnan ÇORUM

June 2020, 60 Pages

The one of the most significant point for the insurance companies is that demanding an insurance claim by the policy holders which after they have been exposed to an incident or accident based on the policy conditions and resulted the whole process. Insurance claim processing and giving a decision about the claim takes for a long time. This process also effects on the satisfaction of the customers and their loyalty. The main goal of this thesis is that making travel insurance claims as having an automic process and help the decision making process. In the study, Logistic Regression and Naive Bayes prediction models have been used in order to predict the travel insurance claim status on a dataset which has a few variable and imbalanced dataset. Principal Component Analysis has been used as a dimension reduction technique as well. All of the study has been done in Python programming language and its libraries. As a consequence, it is seen that the models which used in the study can be used as a decision support mechanism.

Keywords: Travel Insurance, Insurance Claim, Logistic Regression, Naive Bayes, Principal Component Analysis

İÇİNDEKİLER

TABLolar.....	viii
ŞEKİLLER.....	ix
KISALTMALAR	xi
SEMBOLLER.....	xii
1. GİRİŞ.....	1
2. SİGORTA HAKKINDA GENEL KAVRAMLAR.....	3
2.1 SİGORTANIN TANIMI	3
2.2 SEYAHAT SİGORTASI.....	4
2.3 SİGORTA TAZMİNAT TALEPLERİ.....	5
2.3.1 Sigorta Dolandırıcılığı	7
3. MAKİNE ÖĞRENMESİ	9
3.1 DENETİMLİ ÖĞRENME	12
3.2 DENETİMSİZ ÖĞRENME.....	13
3.3 PEKİŞTİRMELİ ÖĞRENME.....	14
4. LİTERATÜR TARAMASI.....	15
4.1 DENGESİZ VERİ SETLERİ.....	15
4.2 SİGORTA TAZMİNAT TALEPLERİ.....	17
4.2.1 Tazminat Talepleriyle İlgili Genel Çalışmalar	17
4.2.2 Sigorta Tazminat Dolandırıcılığı Hakkında Çalışmalar	20
5. VERİ VE YÖNTEM	22
5.1 VERİ SETİNİN TANITIMI	22
5.2 VERİ ANALİZİ.....	23
5.2.1 Veri Görselleştirme ve Veri Temizleme.....	23
5.2.2 Korelasyon Analizi	32
5.2.2.1 Pearson Korelasyonu	33

5.2.2.2 Cramer's V.....	34
5.2.3 Minimum Maksimum Normalizasyonu.....	36
5.3 LOJİSTİK REGRESYON	37
5.4 NAİVE BAYES	38
5.5 TEMEL BİLEŞEN ANALİZİ.....	39
5.6 SMOTE.....	41
5.7 MODEL ÖLÇÜMLENDİRME	42
5.8 K KATLAMALI ÇAPRAZ DOĞRULAMA.....	45
5.9 ENCODING TEKNİKLERİ.....	45
5.9.1 Kukla Değişken.....	45
5.9.2 Hedef Ortalama	46
6. BULGULAR.....	48
6.1 LOJİSTİK REGRESYON SONUÇLARI	49
6.2 NAİVE BAYES SONUÇLARI	56
7. TARTIŞMA VE SONUÇ	59
KAYNAKÇA.....	61

TABLÖLAR

Tablo 5.1: Değişken açıklamaları.....	22
Tablo 5.2: Değişken istatistikleri.....	25
Tablo 5.3: Karmaşıklık matrisi.....	42
Tablo 5.4: Ölçüm kriterleri.....	43
Tablo 6.1: LR model 1 için karmaşıklık matrisi.....	49
Tablo 6.2: LR model 1 sonuçları.....	49
Tablo 6.3: LR model 2 için karmaşıklık matrisi.....	50
Tablo 6.4: LR model 2 sonuçları.....	51
Tablo 6.5: LR model 3 için karmaşıklık matrisi.....	51
Tablo 6.6: LR model 3 sonuçları.....	52
Tablo 6.7: LR model 4 için karmaşıklık matrisi.....	52
Tablo 6.8: LR model 4 sonuçları.....	53
Tablo 6.9: LR model 5 için karmaşıklık matrisi.....	53
Tablo 6.10: LR model 5 sonuçları.....	54
Tablo 6.11: LR model 6 için karmaşıklık matrisi.....	55
Tablo 6.12: LR model 6 sonuçları.....	55
Tablo 6.13: NB model 1 için karmaşıklık matrisi.....	56
Tablo 6.14: NB model 1 sonuçları.....	56
Tablo 6.15: NB model 2 için karmaşıklık matrisi.....	57
Tablo 6.16: NB model 2 sonuçları.....	57
Tablo 6.17: NB model 3 için karmaşıklık matrisi.....	58
Tablo 6.18: NB model 3 sonuçları.....	58

ŞEKİLLER

Şekil 2.1: Tazminat süreçleri genel akış.....	6
Şekil 3.1: Makine öğrenmesi ve sıradan programlama farkı.....	9
Şekil 3.2: Makine öğrenmesi ve kesişim alanları.....	10
Şekil 3.3: Makine öğrenmesi akışı.....	11
Şekil 3.4: Sınıflandırma problemleri.....	12
Şekil 3.5: Regresyon problemleri.....	13
Şekil 3.6: Kümeleme.....	14
Şekil 5.1: İşlenmemiş veri görünümü.....	23
Şekil 5.2: Boş veri bilgisi.....	24
Şekil 5.3: Boş veri ısı haritası.....	24
Şekil 5.4: Yaş kutu grafiği.....	26
Şekil 5.5: Süre kutu grafiği.....	26
Şekil 5.6: Veri sözlüğü.....	27
Şekil 5.7: Destinasyon dağılımı.....	27
Şekil 5.8: Tazminat talep dağılımı.....	28
Şekil 5.9: Acente kodu dağılımı.....	28
Şekil 5.10: Ürün dağılımı.....	29
Şekil 5.11: Tazminat talep ve net satış ilişkisi.....	29
Şekil 5.12: Destinasyon ve yaş ilişkisi.....	30
Şekil 5.13: Dağıtım kanalı, yaş ve tazminat talep ilişkisi.....	30
Şekil 5.14: Destinasyon, yaş ve tazminat talep ilişkisi.....	31
Şekil 5.15: Net satış dağılım grafiği.....	31
Şekil 5.16: Komisyon dağılım grafiği.....	32

Şekil 5.17: Süre dağılım grafiği.....	32
Şekil 5.18: Sayısal değişkenler arası ilişki için saçılım grafiği.....	33
Şekil 5.19: Sayısal değişkenler için korelasyon matrisi.....	34
Şekil 5.20: Cramer's v fonksiyonu.....	35
Şekil 5.21: Kategorik değişkenler için korelasyon matrisi.....	35
Şekil 5.22: PCA fonksiyonu.....	40
Şekil 5.23: Smote algoritması.....	41
Şekil 5.24: Dengesiz verilerde doğruluk ölçütü.....	43
Şekil 5.25: Duyarlılık-kesinlik ilişkisi.....	44
Şekil 5.26: K katlamalı çapraz doğrulama.....	45
Şekil 5.27: Hedef ortalama.....	47
Şekil 6.1: İçerik diyagramı.....	48
Şekil 6.2: Genel akış diyagramı.....	48
Şekil 6.3: 13 Temel bileşenin varyansları.....	50
Şekil 6.4: 9 Temel bileşenin varyansları.....	54

KISALTMALAR

BM	:	Birleşmiş Milletler
Ç.Kartil	:	Çeyrek Kartil
FN	:	False Negative
FP	:	False Positive
KNN	:	K Nearest Neighbour
LR	:	Lojistik Regresyon
NB	:	Naive Bayes
PCA	:	Principal Component Analysis
RF	:	Random Forest
SMOTE	:	Synthetic Minority Over-sampling Technique
SVM	:	Support Vector Machine
TBA	:	Temel Bileşen Analizi
TN	:	True Negative
TP	:	True Positive

SEMBOLLER

Beta : β

Çarpım Sembolü : Π

Doğal Logaritma : e

Ortalama : μ

Standart Sapma : σ

Toplam Sembolü : Σ



1. GİRİŞ

Dünyanın hemen hemen tüm sektörlerinde olduğu gibi sigortacılık alanında da iş süreçlerinin yapısı ve nihayetinde müşteri memnuniyeti önem arz etmektedir. Sigorta firmaları, satmış oldukları poliçeler aracılığıyla müşterilerle olan bağını yürütmekte ve müşteri ihtiyaçlarını karşılamaya çalışmaktadır.

Müşteriler, yaşadıkları bir kaza, problem veya olay sonrasında daha önceden satın almış oldukları sigorta poliçesinin koşulları çerçevesinde ilgili sigorta şirketine zararlarının karşılanabilmesi için tazminat taleplerinde bulunmaktadır. Sigorta şirketleri de müşterilerinin bu taleplerini inceler, karara bağlar ve tazminat sürecini müşteri veya sigorta şirketi lehine sonuçlandırır.

Tüm dünyada olduğu gibi sigortacılıkta da bilgi teknolojilerinin kullanıldığı alanlar yoğunlaşmaktadır. Günümüzde birçok sigorta türüne (seyahat, araç vb) ait poliçeler web sayfaları üzerinden 7 gün 24 saat boyunca kolayca satın alınabilir veya yenilenebilir durumdadır.

Büyük veri kavramının da yaygınlaşmasıyla birlikte veriden anlam ve değer elde etme olgusu kritik bir konu haline gelmiştir. Bu çalışmada, Python programlama dili kütüphaneleriyle bazı istatistiksel makine öğrenmesi teknikleri kullanılarak seyahat sigortası poliçesi yaptırmış olan müşterilerin sigorta şirketinden tazminat talep edip etmeyeceklerine yönelik bir tahminleme yapılmıştır. Dengesiz bir veri kümesi ve sınırlı sayıda değişken ile çalışılmıştır. Sigorta şirketlerinin tazminat taleplerini hızlı ve doğru sonuçlandırabilmesi, tazminat taleplerinin incelenmesindeki operasyonel süreçlerin minimuma indirilerek bir karar destek sistemi oluşturulabilmesi amacıyla bu tez çalışması yapılmıştır.

Tezin ikinci bölümünde genel sigorta kavramları ele alınarak sigortanın tanımından, seyahat sigortasından, tazminat süreçlerinden ve sigorta dolandırıcılığından bahsedilmiştir. Üçüncü bölümde makine öğrenmesinin genel yapısından ve öğrenme tekniklerinden bahsedilmiştir. Dördüncü bölüm olan literatür taramasında dengesiz veri

setlerinden, seyahat sigortası ve diğ er sigorta t rleriyle ilgili tazminat talep s re lerinden ve sigorta dolandırıcılığıyla ilgili daha  nce yapılan  alıřmalardan bahsedilmiřtir. Beřinci b l m olan Veri ve y ntemde  alıřılan verinin niteliđi, modeller, teknikler ve kriterler anlatılmıřtır. Altıncı b l m olan bulgular b l m nde Lojistik Regresyon ve Naive Bayes modelleriyle yapılan farklı řekillerdeki analizlerin sonu ları g sterilmiřtir. Yedinci ve son b l m tartıřma ve sonu  kısmında ise elde edilen  ıktıların arařtırma problemiyle olan bađlantısı vurgulanmıř, ileriye d n k yapılabilecek  alıřmalardan bahsedilmiřtir.



2. SİGORTA HAKKINDA GENEL KAVRAMLAR

2.1 SİGORTANIN TANIMI

Sigorta, belirli bir ücret karşılığında olabilecek kazalara veya olaylara karşı kişilerin kendini önceden teminat altına almasıdır. Bu güvence, kişiye verilen ve bir sözleşme niteliği taşıyan sigorta poliçesi belgesi üzerinden sağlanmaktadır (Yıldırım, 2013). Sigorta, kişi ya da kişilerin uğrayacağı ekonomik zararın devre dışı bırakılmasını amaçlamaktadır. Sigorta poliçesinde sigortayı yapan şirketle ilgili bilgilerin yanı sıra sigortalının kimlik bilgileri, sigortanın kapsamı, ne kadarlık bir teminata kadar karşılanacağı, süresi gibi birçok bilgi yer almaktadır (Türkiye Sigorta Birliği, 2013).

Olası problemlere karşı kendini güvencede hissetmek isteyen kişilerin poliçe satın alarak yapmış olduğu ödemeler bir yerde biriktirilmektedir. Bu bir havuz yapısı gibi düşünülebilir. Sigorta şirketleri, sigortalıların karşılaşmış oldukları olaya ilişkin zararı tespit ettikten sonra bu havuzda birikmiş miktarlarla sigortalıların zararlarını karşılamaktadır. Bu zararın boyutu genelde küçük meblağlar olmadığından müşteriler önceden yaptırmış oldukları poliçe sayesinde kendilerini ekonomik zararlardan korumaktadır (Günbatır, 2019). Sigorta aynı zamanda bir garanti olarak da adlandırılabilir. Örneğin, ürün satın aldıktan sonra ürüne ilişkin verilen garanti belgesi aslında bir sigortadır. Kimi zaman müşterilere belirli bir ücret karşılığında garantinin uzatıp uzatılmayacağına dair talepte bulunulur. Müşteri ek bir ücret karşılığında satın almış olduğu ürünün garantisini 1 ila 2 yıl arasında uzatabilir. Bu süreçte almış olduğu ürünün arızalanacağı kesin değildir. Ancak müşteri bu teminatı satın alarak ürünü ek bir süre için daha güvence altına alabilir. Bu durum da sigortanın tanımını açıklamak için verilebilecek örneklerden biridir.

Sigorta sektöründe şirketler, acenteler ve reasüranslar bulunmaktadır. Poliçelerdeki kuralları belirleyen, prim ödemelerinin takibini yapan sigorta şirketlerinin kendisidir. Şirketler, acenteleri yetkilendirerek poliçelerin müşterilere satışını ve böylece poliçelerin müşterilere ulaşmasını sağlarlar (Aydoğan ve Akpınar, 2016).

Sigortaların birçok türü bulunmaktadır (Türkiye Sigorta Birliği, 2013). Başlıca sigorta türlerine örnek olarak şunlar gösterilebilir:

- a. Seyahat Sigortası
- b. Hayat Sigortaları
- c. Araç Sigortası
- d. Yangın ve Deprem Sigortası

2.2 SEYAHAT SİGORTASI

Seyahat sigortası, seyahat öncesi, sırası veya sonrasında sigortalının başına gelebilecek çeşitli olaylara karşı poliçe sahibi kişinin zararının tazmin edilebilmesi amacıyla düzenlenmiş belge veya belgeler bütünü olarak tanımlanır. Hastalık, kaza, terör, bagaj kaybolması, çalınması veya gecikmesi gibi durumlarda seyahat sigortası müşterilerin zararlarının karşılanmasında kullanılır. Vize başvurularında başvuran kişiden talep edilen zorunlu belgelerden biridir. Bu sigortanın kapsamı sigorta şirketleri bazında farklılık gösterebilmektedir. Örnek olarak Anadolu Sigorta şirketi hastalık durumları ve nakil işlemleri, planlanan seyahatin iptal edilmesi, bagaj kayıpları, ilaç temini, seyahat belgesinin (pasaport, bilet, kimlik vb) kaybedilmesi, ulaşım şirketlerinin fazla bilet satışı politikasından kaynaklı olarak yaşanan gecikmeler gibi durumlarda yaşanacak mağduriyetleri kapsamına almıştır (Anadolu Sigorta, 2019). Axa Sigorta da yukarıda belirtilen benzer seyahat sigortası kapsamlarına sahiptir (Axa Sigorta, 2018). Unico Sigorta teminat sürelerinin önemli olduğundan bahsetmiş, gelecek seyahatler için uzun vadeli sigorta yaptırılmasını tavsiye etmiştir (Unico Sigorta, 2018).

Seyahat sigortası yaptırmak zorunluluktan ziyade bir güvence sağlamaktadır. Örneğin, yurtdışında seyahat sigortası yaptırmak zorunda olunmayan bir ülkeye seyahat ettikten sonra ilgili ülkede herhangi bir sağlık problemi meydana geldiğinde o ülkenin hastanelerine gittikten sonra yabancı bir vatandaş olduğunuz zaman çok yüksek ücretlerle karşılaşılabilir. Yaptırmış olduğunuz poliçe ile birlikte böyle bir risk karşısında kendinizi önceden garanti altına alabilirsiniz.

Seyahat sigortası yaptırırken diğ er  nemli olan hususlar ise sigortanın ge erli olduėu b lgeler ve tarihleridir. Birleřmiř Milletler (BM) tarafından Suriye, Irak, Afganistan gibi savař kořullarında bulunan  lkelerde poli eler ge erli deėildir. Ayrıca, poli enin tarih aralıėı seyahat ettiėiniz tarih aralıėı ile uyumlu olmak zorundadır (Fiba Emeklilik, 2012).

2.3 SİGORTA TAZMİNAT TALEPLERİ

Sigorta tazminat talebi, sigorta poli esi sahiplerinin yařamıř oldukları bir kaza veya olay sonrasında poli enin sınırları dahilinde olmak kaydıyla poli eyi d zenlemiř olan sigorta řirketine zararlarının karřılanması amacıyla yapmıř oldukları maddi taleptir.

Yapılan talep bir ihbar bildirimi sonrasında iřleme alınır. Daha sonra s reci takip etmesi i in sigorta řirketince bir yetkili atanmaktadır. Atanan yetkili gelen tazminat talebinin kapsamını, dosyanın i eriėini, ger ek olup olmadıėını incelemektedir. İnceleme tamamlandıktan sonra tazminat talep  demeleri ger ekleřtirilmektedir (Muslu, 2009). Tazminat talebinde bulunan sigortalıların talep gerek elerini belirten t m destekleyici dok manları sigorta řirketine ulařtırması gerekmektedir. Bu dok manlar polis kaydı, medikal dok man veya eėer seyahat iptal edilmiřse seyahatin  denmiř faturaları ve iptale neden olan duruma ait belgeler  rnek g sterilebilir. Bu s re te dok manların doėruluėu ve tazminat iddiasının durumu incelenmektedir.

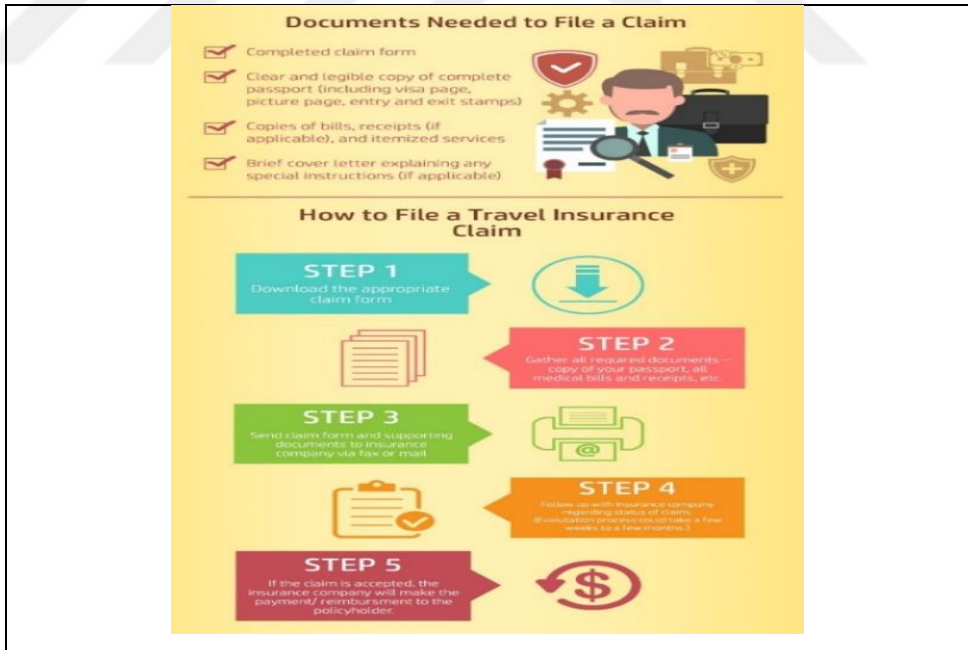
Gulf Sigorta poli e kapsamına giren kaza t r ne g re (bagaj kayıp, hastalık vb)  eřitli belgeler talep etmekte ve bununla ilgili bir formun kendilerine ulařtırılmasını beklemektedir (Gulf Sigorta, 2017). Mapfre Sigorta m řterilerden bir form talep etmektedir (Mapfre Sigorta, 2020). Benzer řekilde Sompo Sigorta m řterilerden bir tazminat talep formunu doldurması ve belgeleri g ndermesini beklemektedir (Sompo Sigorta, 2020). Taleplerin sonu landırılma s resi řirketten řirkete deėiřiklik g sterebilmektedir. Allianz řirketi seyahat sigortası tazminat talep t r ne g re istenen dok manlar ulařtıktan sonra 7 ile 10 iř g n  arasında karar verilebileceėini belirtmektedir (Allianz Travel Insurance, 2020).

Sigorta tazminat talep s reci boyunca titiz incelemeler yapılmakta ve sonu  sigortalıya bildirilmektedir. Bu s re  doėru takip edilmelidir.   nk , haklı olan bir m řterinin

tazminat talebini haksız gerekçelerle reddetmek ya da olması gerekenden düşük bir tazminat ödemek müşterinin tüketici mahkemelerinde açacağı davalar sonrasında şirketin itibarını ve maddi durumunu zedeleyebilir. Sigortacıların, tazminat taleplerini değerlendirip bu ödemeleri yaparken şeffaf bir şekilde değerlendirme yaparak müşteri memnuniyetini sağlamaları esastır. Bu süreçte, tespit edilen ihlallerle ilgili olarak sigorta şirketleri birtakım önlemler almaktadır. Buna tazminatın ödenmemesi, poliçenin geçersiz hale getirilmesi, yargı sürecinin başlatılması gibi durumlar dahildir. Axa Sigorta açıklamış olduğu istatistiklerde poliçe sahiplerinin yüzde 1.9'unun tazminat talebinde bulunduğunu, bu yüzde 1.9'luk kesimin yüzde 94.2'sinin tazminat taleplerinin şirket tarafından karşılandığını belirtmiştir. Bu ödemelerin de ortalamasının 855 İngiliz Sterlini (01.01.2020 itibarıyla 6737,4 TL'ye karşılık gelmektedir) olduğu belirtilmiştir (Axa Sigorta, 2019).

Şekil 2.1'de seyahat sigortası tazminat süreçlerinin genel akışı özetlenmiştir.

Şekil 2.1: Tazminat süreçleri genel akış



Kaynak: Visitor Insurance Reviews, 2015

Yapay zeka; Sigorta sektöründe en çok tazminat talebi süreçlerinde, poliçelerin onaylanmasında, risk planlamada, dolandırıcılık tespitlerinde, müşteri hizmetlerinde ve pazarlamada kullanılmaktadır. Örneğin, Accenture şirketinin sunmuş olduğu bir raporda

müşterilerin hizmet alımları sırasında ChatBot yapay zeka sistemlerinden faydalanarak bilgi aldıkları görülmüştür. Sigorta almış kişiler, uğradıkları herhangi bir kaza ya da sigorta kapsamına giren başka bir hususa ithafen zararlarının karşılanması için başlattıkları sürecin daha hızlı sonuçlanmasını istemektedirler. Bu süreçlerin hızlandırılması, müşteri memnuniyetini sağlayacaktır (Hall, 2017). Sigorta sektöründe veri madenciliği yaklaşımı, kar artışı, karar alma süreçleri, yeni müşterilerin kazanılması, dolandırıcılık aktivitelerinin önlenmesi, yeni ürün geliştirme, risk yönetiminin yapılması ve mevcut müşterilerin elde tutulması gibi amaçlarla da kullanılabilir (Kulkarni ve Devale, 2012).

2.3.1 Sigorta Dolandırıcılığı

Sigorta poliçesinin hazırlanmış olduğu tarihten bağımsız olarak hukuka aykırı şekilde hareket edilip sigorta şirketlerini bilerek yanlış yönlendirerek yapılan bir nitelikli dolandırıcılık faaliyetidir (Türkiye Sigorta Birliği, 2013). Belgelerde tahrifat yaparak, bilerek mala zarar vermek veya buna benzer yollarla sigorta firmasını yanıltıp firma üzerinden para kazanmak amaçlanır. Sigorta dolandırıcılıklarının fazla olması tazminat talep süreçlerinin gidişatını olumsuz etkilemektedir. Hem süreç yavaşlamakta hem de gerçekten ihtiyacı olduğu için tazminat talebinde bulunmuş sigortalılar mağdur olmaktadır.

Yalan beyan vererek sigorta şirketlerini kandırmak, kaza uydurmak gibi durumlar sigorta dolandırıcılığıdır. Bu tür dolandırıcılık faaliyetleri sigorta şirketlerinin zarar etmesine sebebiyet vermektedir (Yıldırım, 2013). Kısa bir zaman periyodunda yüksek meblağlı faturalar düzenlenmesi, tıbbi durumdan kaynaklı tazminat iddiasına rağmen tıbbi bir raporun bulunmaması, poliçe tarihleri geçtikten sonra tazminat iddiasında bulunulması (bazı durumlarda bu tolere edilebilmektedir) gibi durumlar tazminat dolandırıcılığı konusunda şüphe uyandırıcı faktörlerdir (Kırlıdoğ ve Asuk, 2012).

Sigorta sektörünün yapısı gereği bu sektör üzerinden işlenen suçların açığa çıkartılması kolay değildir. Bu sebeple bu sektörde gerçekleşecek olası bir infial durumunda gerek sigortalı gerekse de sigorta şirketleri büyük zararlar görmektedir. Sigorta şirketleri

üzerlerinde baskı yaratan mali problemleri karşılayabilmek adına bunu sigorta primlerine yansıtabilirler. Primlerin artması da aslında hiçbir suça karışmamış ve zamanında ödemelerini yapmış dürüst sigortalıların da durumdan dolaylı olarak etkilenmesi anlamına gelmektedir (Günbatar, 2019).



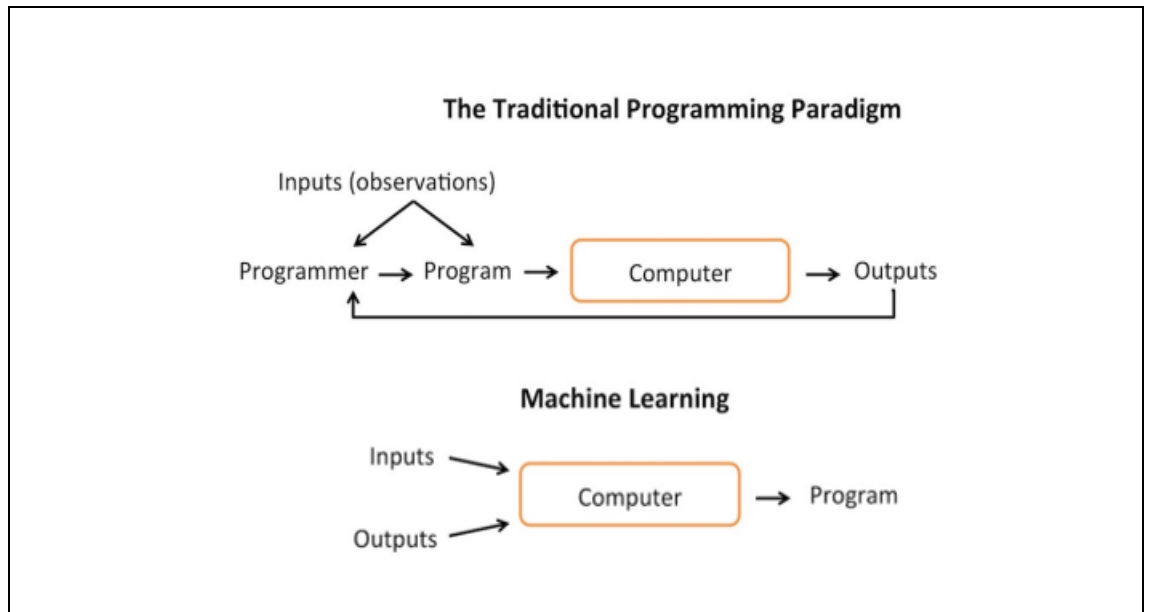
3. MAKİNE ÖĞRENMESİ

Geçmiş verileri kullanarak gelecek hakkında çıkarım veya tahmin yapmaya yarayan istatistiksel bilgisayar algoritmalarının genel adıdır (Orhan, 2017). Makine öğrenmesi verilerden öğrenerek hiç bilinmeyen tahminini yapmaktadır. Süreçlerin otomatize hale gelmesinde ve hızlandırılmasında ciddi fayda sağlamaktadır. Makine öğrenmesi ilk olarak Arthur Lee Samuel tarafından dile getirilmiş ve bilgisayarların öğrenebilme yeteneği kazanması olarak tanımlanmıştır (Raschka, 2018).

Makine öğrenmesi, şirketlerin kısa veya uzun vadede karar alma mekanizmalarına etki eder. Geleceğe yönelik öngörü ve tahminlerle birlikte şirketin çizeceği yol haritalarının şekillenmesine yardımcı olur. Verinin arka planında yer alan gizli bilgi önyüze çıkarılır ve şirketler bu veriyi anlamlandırarak karar alma süreçlerinde kullanırlar. Makine öğrenmesinde bilgisayar bilimi ve istatistik bir araya gelmektedir. Makine öğrenmesi, istatistik tabanlı algoritmaları kullandığından mutlaka yeterli verinin bulunması gerekir. Öğrenme ancak veriyle mümkün olabilir.

Şekil 3.1’de gösterildiği üzere makine öğrenmesi ve sıradan yazılım geliştirme arasında genel ölçekte bir farklılık bulunmaktadır.

Şekil 3.1: Makine öğrenmesi ve sıradan programlama farkı

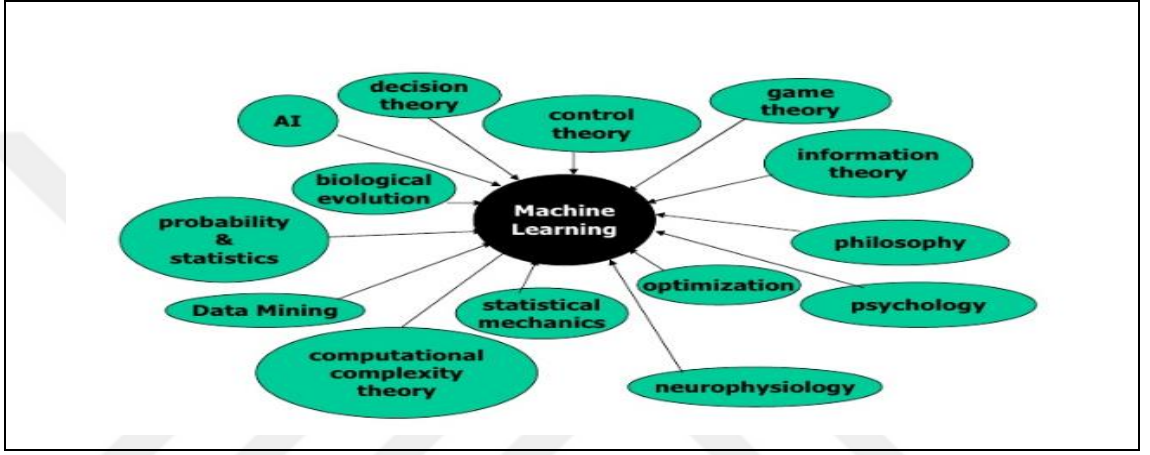


Kaynak: Raschka, 2018

Sıradan yaklaşımda yazılımcı tarafından geliştirilen bilgisayar programı ve girdilerimiz bir sonuç üretirken, makine öğrenmesi yaklaşımında ise girdilerimiz ve sonuçlarımız bilgisayara gönderilir ve çıktısında bir program oluşturulmaktadır.

Şekil 3.2’de gösterildiği gibi makine öğrenmesi alanı istatistik, optimizasyon, karar verme, psikoloji ve nöropsikoloji gibi birçok alanla kesişim halindedir.

Şekil 3.2: Makine öğrenmesi ve kesişim alanları



Kaynak: Intrator, 2010

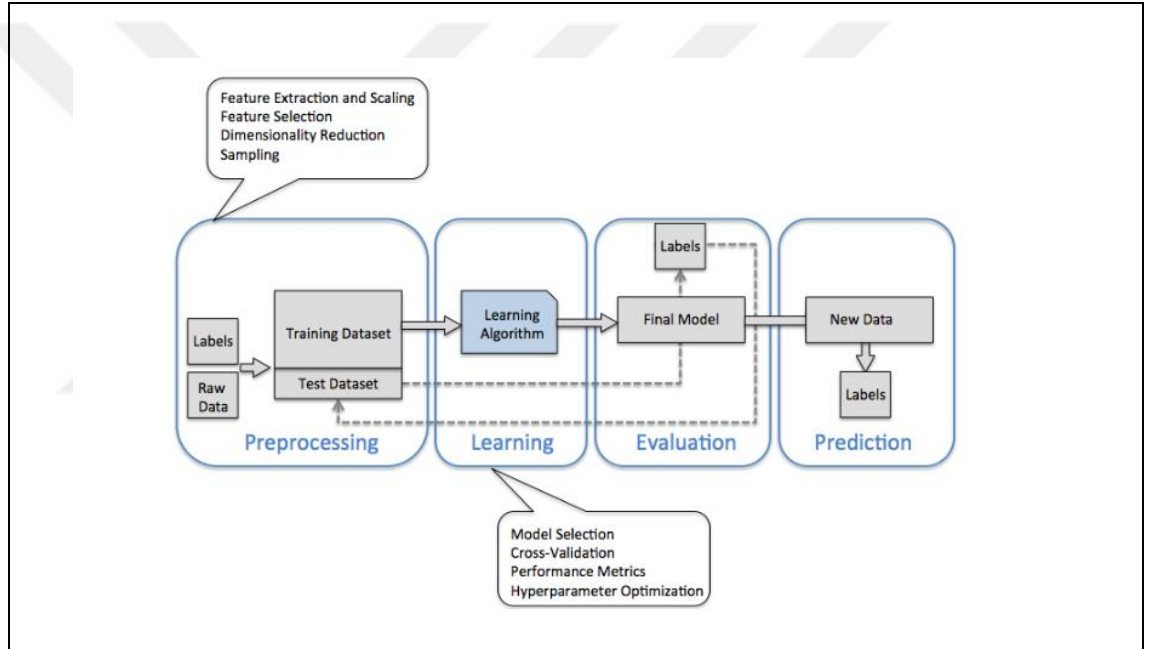
Makine öğrenmesi birçok disiplinin bir araya geldiği bir alandır. Örneğin, makine öğrenmesi uzmanı bir kişi ile bir tıp doktoru bir araya gelip bilgi paylaşımı yaparak ortaya katma değerli bir iş çıkarabilirler. İş bilgisi oldukça kıymetli bir bilgidir. Bankacılıkla, sigortacılıkla veya başka hangi alanla ilgili olursa olsun bir çalışma yapılmadan önce muhakkak iş bilgisi ve iş kurallarının neler olduğuna hakimiyet ihtiyacı vardır. Makine öğrenmesi yaklaşımları kullanırken belirlenecek değişkenler iş bilgisinin nasıl kullanıldığına göre şekillenmektedir.

Makine öğrenmesi, hayatın içinde birçok problemin çözümünde kullanılmaktadır. İstenmeyen maillerin tespiti, ses ve yüz tanıma, metin sınıflandırma, dolandırıcılık tespitleri, görüntü işleme gibi birçok alanda makine öğrenmesi bizzat vardır (Balcan, 2018). Makine öğrenmesi ve veri madenciliği algoritmaları churn analizi, segmentasyon, satış rakamlarına yönelik tahminler, sigorta poliçesi talepleri, öneri sistemleri, fraud işlemi tanıma gibi alanlarda kullanılabilir. Finans, bankacılık, sigorta ve birçok başka sektörde veri madenciliği uygulamaları mevcuttur (Ahmed, 2004).

Sigorta şirketlerinde veri madenciliği ve makine öğrenmesi çözümlerinin kullanılması gelir-gider dağılımının dengelenerek tasarruf yapılmasının sağlanması konusunda yardımcı olmaktadır. Veri madenciliğinin üst yönetimlerin karar alma süreçlerinde yer alması, finansal analiz ve stratejilerinin belirlenmesinde kullanılması sigorta şirketlerinin de makine öğrenmesinden yararlanmasını sağlamıştır. İleriye dönük poliçe satın alma, kayıp müşteri analizi, satışların artırılması, risk analizi gibi alanlarda da sigorta sektöründe veri madenciliği ve makine öğrenmesi kullanılmaktadır (Akpınar, 2018).

Şekil 3.3'te makine öğrenmesinin genel akışı ifade edilmiştir.

Şekil 3.3: Makine öğrenmesi akışı



Kaynak: Roman, 2018

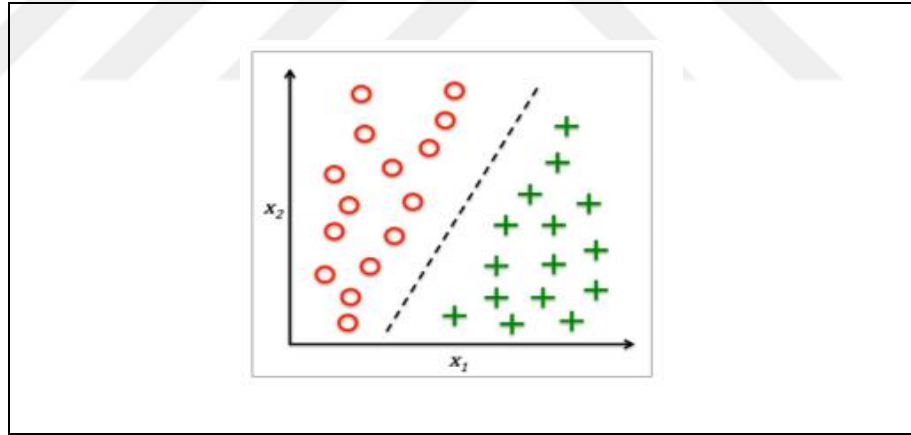
Veri ön işleme aşamasında veri eğitim ve test verisi olarak ayrılmaktadır. Eğitim verisi adı üzerinde istatistiksel modelin eğitilmesi gereken verilerden oluşur. Test verisi ise oluşturulan istatistiksel tahmin modelinin başarısının ölçüleceği veri grubunu oluşturur. Test verisine gerçek hayatta hiç karşılaşmadığımız verilermiş gibi davranılır. Öğrenme algoritması çalıştıktan sonra üretilen çıktı model tarafından bulunmuş olan sonucu ifade eder.

3.1 DENETİMLİ ÖĞRENME

Denetimli Öğrenme (Supervised Learning), verilen değişkenlerden yararlanarak çıktının ne olacağını tahmin eder. Bu bir fotoğrafta yer alan nesnenin ne olduğu, çalan bir şarkının ne olduğunu tahmin etmek bile olabilir. Sınıflandırma ve regresyon problemleri olarak incelenmektedir. Sınıflandırma problemlerinde çıktı kategorik bir veri tipindedir. Örneğin, doküman sınıflandırma, film kategorisi sınıflandırma, hastalıkların sınıflandırılması, spam mail olup olmama durumları, bir görüntü verisinde yüzün kime ait olduğu gibi problemler sınıflandırma problemlerinin kapsamına girmektedir. Regresyon problemlerinde ise genelde sürekli değişkenlere yönelik bir tahmin yapılır. Yani, çıkan değer genelde sayısalıdır. Örneğin, bir maaş verisinin tahminlenmesi, bir ürünün satış rakamlarının tahmin edilmesi, hava durumu tahminleri, döviz verilerinin gelecek yıllardaki tahminleri buna örnek gösterilebilir (Paul 2018; Balcan 2018).

Şekil 3.4'te sınıflandırma problemleri gösterilmektedir.

Şekil 3.4: Sınıflandırma problemleri

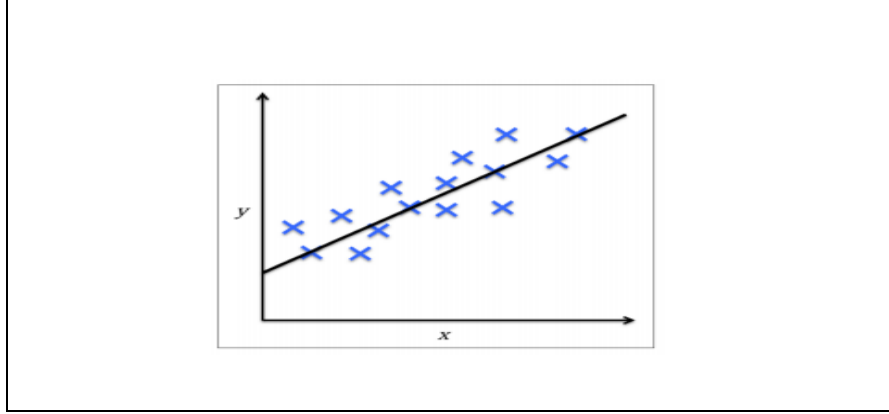


Kaynak: Paul, 2018

Kurulan istatistiksel modelin sonucu sınıflandırılarak ifade edilir. Örnekte X1 ve X2 sınıfı olarak gösterilmiştir. Evet/Hayır, Var/Yok ifadeleri gibi ayırıcı sonuçlar elde edilir.

Şekil 3.5'te doğrusal regresyon problemleri gösterilmektedir.

Şekil 3.5: Regresyon problemleri



Kaynak: Paul, 2018

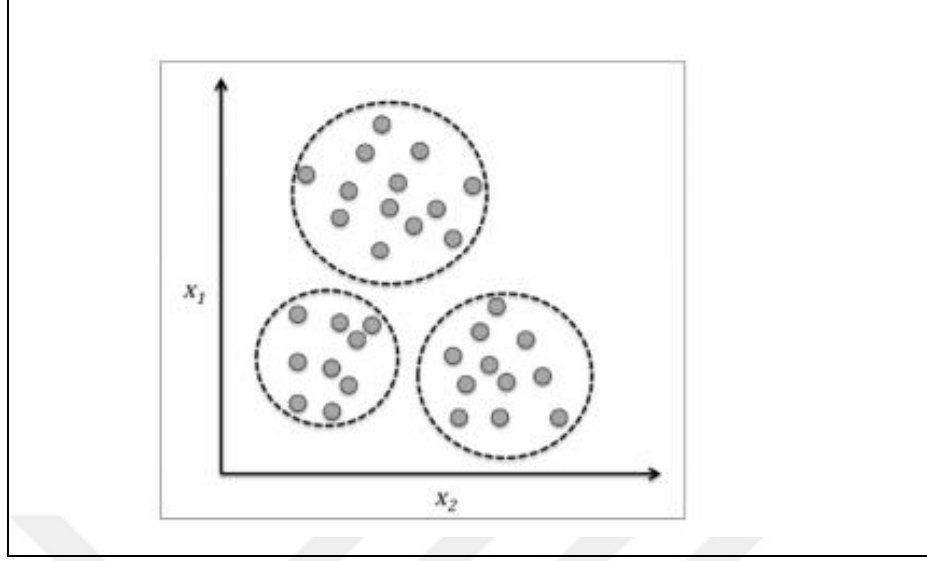
3.2 DENETİMSİZ ÖĞRENME

Denetimsiz Öğrenme (Unsupervised Learning) verinin karakterini ortaya çıkarma tekniğidir. Sonuca yönelik doğrudan bir tahmin üretmez. Veri seti içerisinde normal veri kümesinin olağan hareketlerinin dışında bir durumu analiz etmeye çalışır. Bu duruma anomali tespiti örnek verilebilir. Örneğin, bir bankacılık veri setinde yer alan dolandırıcılık aktivitesi bir anomalidir. Bunu tespit ederken Kümeleme (Clustering) tekniği kullanılmaktadır. Kümeleme tekniğiyle beraber veri setinde bulunan gruplar kendi içerisinde kategorilendirilir (Paul, 2018).

Denetimsiz Öğrenmede kümeleme tekniğiyle veri kendine benzer nitelikler taşıyan, özellikleri büyük oranda aynı eğilimler gösterenler kümeler içerisinde gruplandırılır. Kümeleme modelinde kullanılacak küme sayısı çoğu zaman önceden belirlenmektedir (Orhan, 2017).

Şekil 3.6'da Kümeleme tekniği gösterilmektedir. Üç ayrı kümenin oluşturulduğu gözlemlenmektedir.

Şekil 3.6: Kümeleme



Kaynak: Paul, 2018

Denetimsiz Öğrenme tekniklerinde veri setimizde bir çıktı yoktur. Denetimli Öğrenme tekniklerinden farklı olarak elimizde kesin bir veri sonucu bulunmadığından, birbirine en yakın verilerden kümelerin oluşturulması ve verinin buna göre gruplandırılarak istatistiksel modelin oluşturulması gerekmektedir.

3.3 PEKİŞTİRMELİ ÖĞRENME

Pekiştirmeli Öğrenme (Reinforcement Learning) yapay zeka kavramının bir alt birimi olarak düşünülmektedir. Kendi kendine öğrenebilen sistemlerdir. Atari oyunlarının bilgisayar sistemine karşı oynanması, satranç, Tic-Tac-Toe oyunu örnek olarak verilebilir. Kendisine karşı oynayan kişileri dahi yenebilen bu sistemler Pekiştirmeli Öğrenme algoritmalarıyla çalışmaktadır (Zemel vd., 2016).

4. LİTERATÜR TARAMASI

Tez kapsamında çalışılan veri setinin literatürde ayrı ve önemli bir problem olarak değerlendirilen dengesiz veri kümelerinden oluşması nedeniyle dengesiz veri kümeleriyle yapılan genel çalışmalar incelenmiştir. Sigorta tazminat talepleriyle ilgili olarak ise bu zamana kadar yapılan çalışmaların çoğunluğu tazminat taleplerinin dolandırıcılık niyetiyle yapılıp yapılmadığının araştırıldığı çalışmalardır. Tazminat süreçlerinin kendisiyle ilgili çalışmaların sayıca daha az olduğu görülmüştür. Tazminat talep süreçlerinin otomatize edilmesiyle ilgili çalışmaların yanı sıra araştırma konumuzu dolaylı olarak ilgilendirmesi sebebiyle sigorta dolandırıcılığı faaliyetleriyle ilgili yapılmış çalışmalara da yer verilmiştir. Seyahat sigortası özelinde veri madenciliği ve makine öğrenmesi yöntemleriyle yapılan pek fazla çalışmaya rastlanmamıştır. Bu sebeple yapılan literatür araştırması diğer sigorta türleriyle ilgili yapılan makine öğrenmesi ve veri madenciliği çalışmalarından oluşmaktadır.

4.1 DENGESİZ VERİ SETLERİ

Teknolojinin gelişmesi aynı zamanda büyük verinin varlığını da oluşturmuştur. Verilerin hacminin büyümesi aynı zamanda dengesiz veriye sebep olabildiğinden sınıflandırma problemlerinin çözümü zorlaştırmıştır. Sınıflandırma problemlerinin çözümünde karşılaşılan en büyük zorluklardan biri dengesiz veri kümeleridir (Longadge vd., 2013). Dengesiz veri kümeleri incelenen araştırma probleminde çıktılarından birinin diğerine göre daha nadir sayıda olmasından kaynaklıdır ve gerçek hayattaki uygulamalarda araştırma probleminin konusuna göre dengesiz verilerle sıkça karşılaşılabilir. Bu tür veri kümelerinde azınlık sayıdaki sınıf istatistiksel algoritmalar tarafından gürültülü bir veriymiş gibi algılanır. Bankacılık işlemlerindeki anomali tespitleri (dolandırıcılık faaliyetleri), müşterilerin ilgili firmayı kullanmama veya terk etme eğilimlerinin incelenmesi, hastalıklarla ilgili veriler, metin veya görüntü işleme gibi problemlerde dengesiz veri olabilir (Gosain ve Sardana 2017; Guo vd., 2008).

Dengesiz veri setlerinde önemli olan popülasyonu az olan sınıfın tespitinin yapılmasıdır. Bu tür veri kümelerinde istatistiksel algoritmalar popülasyonu fazla olan çıktıyı daha fazla tahmin etme eğilimindedir. Bu durum, sayıca az olan sınıfın model tarafından

yeterince tespit edilememesine sebep olur (Longadge vd., 2013). Dengesiz veri kümeleriyle çalışırken, istatistiksel algoritmalarla model oluşturmaya çalışıldığında modelin başarısının ölçülmesinde farklı kriterler de değerlendirilmelidir. Bu sebeple sadece doğruluk (accuracy) değeri bir ölçüt olarak kullanılmamalıdır (Ali vd., 2015; Longadge vd., 2013). Rahman ve Davis (2013) Kardiyovasküler hastalık tespitinin yapıldığı dengesiz bir veri seti üzerinde çalışmışlardır. Hasta olan kişilerin olmayan kişilere göre daha az sayıda olduğunu görmüşlerdir. Gerçekten hastalık taşıyan bir kişiye hastalık tanısının konulamaması büyük bir problemdir. Öte yandan, hastalığı olmayan bir kişiyi hasta gibi sınıflandırmak da bir problemdir, ancak getirdiği maliyet diğerine göre daha düşüktür. Makine öğrenmesi algoritmaları kullanarak kuracağımız modeller ve yapacağımız tahminlerde popülasyonu az olan sınıfın yeterince doğru tahmin edilmemesi ciddi sonuçlar doğurabilmektedir. Aynı çalışmada, dengesiz veri problemlerinin çözümünde örneklem azaltma ve artırma yöntemlerine değinilmiştir. SMOTE yöntemi kullanılarak popülasyonu az olan sınıfın yapay olarak sayısının artırılmasıyla veri seti dengeli hale getirilmiştir. Makine öğrenmesi algoritmalarının dengeli olmayan veri setlerinde yüksek popülasyona sahip sınıfı tespit etmekte başarılı, popülasyonu az olan sınıfı tespit etmekte başarısız olduğu anlatılmıştır. SMOTE'un verinin eğitilme süresini uzatması dışında başarılı bir yöntem kabul edilebileceğinden bahsedilmiştir. Elrahman ve Abraham (2013) dolandırıcılık tespit sistemleri (bankacılık, sigorta sektöründekiler gibi) dengesiz veri kümeleriyle çalışılabileceğini ve popülasyonu az olan sınıfların tespitinin yapılmasının önemli olduğunu belirtmişlerdir.

Amin vd. (2016) makine öğrenmesi algoritmalarının dengesiz veri problemi çözülmeden doğrudan uygulanmasının başarısız sonuçlar oluşturabileceğini belirtmişlerdir. Aynı çalışmada, müşteri terk eğilimine ilişkin bir veri seti incelenerek örneklem artırma metotlarından biri olan SMOTE'un kullanılmış olduğu görülmüştür. Klasikleşmiş örneklem artırma tekniklerinden ziyade, SMOTE tekniğinin kullanımının daha etkili olduğu belirtilmiştir. Zhang ve Wang (2011) Alman bir bankanın verilerinden yararlanarak müşteri kredi risklerini dengesiz bir veri seti üzerinden araştırmışlar ve SMOTE ile SMOTE'un geliştirilmiş bazı tekniklerini kullanarak dengesiz veri problemini çözmeye çalışmışlardır. Kartal ve Özen (2018) dengeli olmayan veri probleminin oluşumunu diyabet hastalığı tespiti için toplanan bir veri seti üzerinden

örneklendirmişlerdir. Problemin çözüm önerilerinde çalışmada kullanılabilecek veri sayılarının arttırılması veya ölçüm kriterlerinin farklılaştırılması gerektiğinden bahsetmişlerdir. Çalışmış oldukları veri setini yüzde 70 eğitim ve yüzde 30 test olarak ayırarak yapay sinir ağları modellerinin çeşitli örnekleme metotlarıyla veriyi dengelemişlerdir. SMOTE uygulayarak yüzde 84 doğruluk, yüzde 91 duyarlılık, yüzde 79 belirlilik elde etmişlerdir. Boyle (2019) dengesiz bir veri kümesiyle Lojistik Regresyon ve SMOTE kullanarak yüzde 86 duyarlılık yüzde 18 F1 skoru elde etmiştir.

4.2 SİGORTA TAZMİNAT TALEPLERİ

4.2.1 Tazminat Talepleriyle İlgili Genel Çalışmalar

Sigorta sektöründe istatistiksel modellerin kullanımı iş süreçlerinin gelişiminde önemlidir. Ancak modellerin başarısının, verinin kalitesine de bağlı olabileceği göz ardı edilmemelidir. İstatistiksel modeller sigorta şirketlerinin poliçe başlatma süreçleri, sigorta prim belirleme ve sigorta tazminat talep süreçlerinde kullanılabilir.

Yunos vd. (2016) motorlu taşıtlar için sigorta tazminatı talebi ve maliyetine yönelik analizler gerçekleştirmişlerdir. Geri Yayılımlı Yapay Sinir Ağları, Lineer Regresyon ve Bulanık Sistemler kullanmışlardır. Antwi ve Zhao (2012) sağlık sigortası yaptıran kişilerin tazminat taleplerini Lojistik Regresyon kullanarak araştırmışlardır. Koforidua Bölgesel Hastanesi'nden 1 Ocak 2011 – 31 Aralık 2011 tarihleri arası 4549 adet veri rastgele seçilmiştir. Cinsiyet, yaş, medeni durum, hastanın ikametinin hastaneye olan mesafesi ve hastanede kalış süresinin tazminat talebine etkiyen önemli faktörlerden olduğu görülmüştür. Sağlık durumu, ücretler ve gelir seviyesinin ise modelin tahminleme gücünde kritik bir rolü olmadığı saptanmıştır. Sigorta şirketlerinin bu konuda etkin bir risk yönetimi yaklaşımını belirlemeleri gerektiğinden bahsedilmiştir. Şirketler, hangi talebin dolandırıcılık niyetiyle olup olmadığını müşterilerin karakteristik özelliklerini analiz ederek bu konuda bir yaklaşım belirlemelidir. Çalışmanın en sonunda, 0-39 yaş aralığında çalışan gruptaki kişilerin tazminat talebi istemeye daha yatkın oldukları görülmüştür.

Kaşçelan vd. (2016) veri madenciliği tekniklerinden Kümeleme, Destek Vektör Regresyonu ve Lojistik Regresyon kullanarak araç sigortalarıyla ilgili tazminat talepleri, primleri ve risklerini araştırmışlardır. Lojistik Regresyon sonucuna göre yüzde 76.70 kesinlik değeri yüzde 56.33 tazminat talebi yapanlar için duyarlılık (recall) elde edildiği görülmüştür. Kumar vd. (2010) sigorta tazminat talep süreçlerinde sigortalıların yeniden değerlendirilme ihtiyacı olup olmayacağını veri madenciliği çalışması üzerinden analiz etmişlerdir. Çalışma Accenture şirketinde bir Amerikan sigorta şirketinin verileriyle gerçekleştirilmiştir. Sigortacılık alanında daha çok dolandırıcılık faaliyetlerinin araştırma konusu olduğu, ancak tazminat talep süreçlerinin (claim processing) kendisinin literatürde daha az incelendiğinden bahsetmişlerdir. Re-work'ten kaynaklı yaklaşık yüzde 30'lara varan operasyonel maliyet oluştuğunu gözlemlemişlerdir. Bu bağlamda, tazminat talep süreçlerinde kullanılmak üzere bir re-work engelleme aracı geliştirmişlerdir.

Pozzolo (2010) veri madenciliği tekniklerini bir sigorta tazminat talebi veri seti üzerinden karşılaştırmıştır. Karar Ağaçları, Regresyon, Temel Bileşen Analizi, Rastgele Orman gibi birçok istatistiksel modeli uygulayarak sonuçları birbirleriyle karşılaştırmıştır. Danso (2006) sigorta sektörüyle ilgili yoğun ve dağınık bir veri kümesinde istatistiksel modeller kullanmıştır. Lojistik Regresyonda yüzde 57 duyarlılık, yüzde 11 kesinlik, Naive Bayes'te yüzde 47 duyarlılık, yüzde 13 kesinlik elde edilmiştir. Duyarlılık ve kesinlik arasındaki ödünleşim vurgulanmıştır. Muslu (2009) sigorta alanında risk yönetimini 2000 adet hasar ihbar kaydı verisini rastgele seçerek veri madenciliği teknikleriyle incelemiştir. Uygulamış olduğu modeller sonrasında yüzde 81.59 doğruluk, yüzde 85.70 duyarlılık, yüzde 74.05 özgüllük değeri elde etmiştir.

Eason (2018) tıbbi maliyetlerin araştırıldığı bir veri setine sigorta tazminat talep durumu değişkeni ekleyerek LR modeliyle bir tahmin yapmış ve yüzde 86 doğruluk elde etmiştir. Hapon (2019) makine öğrenmesi modellerinin tazminat talep süreçlerinin otomatize edilmesinde kullanılabileceğini belirtmiştir. Shroff (2019) yapay zekanın sigorta sektöründe kullanılmasının gerekliliğinden bahsedilmiş ve makine öğrenmesiyle birlikte tazminat talep süreçlerinin hızlandırılabilceğinden söz etmiştir. Young (2017) analitik çözümlerin sigorta sektörünü nasıl değiştirdiği ve sigorta tazminat talep süreçlerinde analitiğin nasıl kullanılabileceği belirtilmiştir. Tazminat süreçlerinin otomatize

edilmesinden, karar destek sistemleri oluşturulmasından bahsedilmiştir. Malhotra ve Sharma (2018) tazminat talep süreçlerinin değerlendirilmesinde operasyonel verimin artırılması için makine öğrenmesinden faydalanılmasının önemli olduğu vurgulanmıştır. Bu süreçlerin otomatize edilerek müşteri memnuniyetinin sağlanmasının amaçlandığı belirtilmiştir. Bu durum şirketlerin kar marjlarını da arttıracaktır. D'Ambrosio (2019) seyahat sigortası tazminat talep süreçleriyle ilgili bir otomatize sistem örneğinden bahsetmiştir. Uçak yolcuları sigorta şirketlerinden uçuşlarının takip edilmesini isteyerek bir gecikme durumunda yolcunun planlarının etkileneceğinden otomatik olarak bir sigorta tazminat talebinin sistemselsel olarak oluşturulması ve ödemenin yolcunun hesabına doğrudan aktarılması şeklinde bir çözüm kullanılmıştır. Bu şekilde bir tespit yapılarak, yolcu henüz havaalanından dahi dönmemişken, hesabında tazminat bedelinin yatırılmış olduğunu görebilir. Bir sigortacı için tazminat süreçlerinin genelde klasik yaklaşımlar olduğu görülmektedir. Bürokrasi, kağıt evraklar, telefon aramaları gibi çok fazla manuel süreç mevcuttur. Yapay zeka ve makine öğrenmesi kullanılarak bu tür süreçler azaltılabilir veya tamamen kaldırılabilir. Seyahat sigorta poliçesi sahipleri en sık uçuş iptali kaynaklı olarak tazminat talep etmektedir. Eğer sigorta firması, uçuş kodunu sigortalı için sistemlerine kaydeder ve takip ederse tazminat talebi müşteri talep etmeden dahi oluşabilir ve ücret derhal müşteri hesabına yatırılabilir (Reese, 2019).

Millican vd. (2017) Brezilya'daki araç sigortaları için olan tazminat talepleri incelemişlerdir. Veri setinde 595213 adet veri ve 57 adet değişken bulunmaktadır. Ridge Regression, Least Squares Lasso Regression, LR, Naive Bayes, Rastgele Orman, Gradient Boosting ve Perceptron gibi algoritmalar tazminat talebinin sonucunun tahminlenmesinde kullanılmıştır. Veride 21694 adet tazminat talebi olduğu, 573518 tanesinin ise tazminat talebinde bulunmadığı görülmektedir. Bunun anlamı veri setinin dengesiz bir veri kümesi olduğuna işaretir. Temel Bileşen Analizi'nin veri setine çok fazla uygun olmadığı yönünde bir değerlendirme yapılmıştır. Hall (2017) istatistiksel modellerden yararlanarak bazı şirketlerin tazminat talep süreçlerini iyileştirdiğinden bahsetmiştir. Zürih Sigorta'nın tazminat talep süreçlerinin 40000 saatlerden 5 saniyelere indiği belirtilen bir yapay zeka çözüm sisteminden bahsedilmiştir. Limon Sigorta, yapay zeka algoritma çözümlerinden faydalanarak tazminat süreçlerinin 3 saniyelere indiğini belirtmiştir. Sigortalı kişinin tazminat talebinde bulunmasının ardından 18 adet fraud

algoritmasının müşterinin talebini incelediği, banka ile iletişime geçtiği, sigortalıyı bilgilendirdiği ve süreci kısa sürede bitirebildiğinden bahsedilmiştir. Huangfu (2015) araç sigortası poliçesi sahipleri için tazminat talep tahmin durumunu araştırmıştır. Lojistik Regresyon, Karar Ağaçları ve PCA kullanılmıştır. PCA kullanılmadan Karar Ağaçları'nın sınıflandırmada LR'a göre daha başarısız olduğu görülmüştür. En başarılı sonuç ise PCA kullanılarak Regresyon Ağaçları modelinde elde edilmiştir.

4.2.2 Sigorta Tazminat Dolandırıcılığı Hakkında Çalışmalar

Kırlıdoğ ve Asuk (2012) Türkiye'de sağlık sigortalarında dolandırıcılık faaliyetlerinin veri madenciliğiyle bulunmasına yönelik bir çalışmayı anomali tespiti metotları kullanarak yapmışlardır. SVM modelinin kullanıldığı çalışmada veri madenciliği yazılımı olası sonuçları yüzde 50 ve üzerinde olanları anomali olarak nitelendirmiştir. 808348 kayıttan 6595 tanesi yüzde 50 ve yüzde 67.3 arasında bir derecelendirmeye sahip olmuştur. Asıl amacı dolandırıcılık tespiti olan bu çalışma, veri madenciliği uygulamalarının bu tespitte başarılı olduğunu göstermektedir. Ghuse vd. (2017) sağlık sigortası tazminat talepleri için dolandırıcılık araştırması yapmışlardır. RF kullanılmış, Yapay Sinir Ağları'nın fazla ve kirli veriyle çalışma gücünün zorluğundan kaynaklı olarak başka çalışmalarda kullanılabileceğini anlatmışlardır. Brennan (2012) araç sigortaları için dolandırıcılık problemini araştırmıştır. Çalışmada kullanılan 15420 adet verinin 14497'si fraud niteliği olmayan, 923 tanesi ise fraud niteliği olan verilerdir. Bu durum ciddi bir dengesiz veri kümesi varlığına işaret etmektedir. İstatistiksel modeller ile tazminat talep durumlarına ilişkin tahmin yapılmıştır. Perera (2013) klasik makine öğrenmesi tekniklerinin yanı sıra ileri düzey yöntemler de uygulayarak bir dolandırıcılık problemini incelemiştir. Yıldırım (2013) Türkiye'deki sigorta dolandırıcılıklarıyla ilgili araştırma yapmıştır. Sigorta sektöründeki dolandırıcılığın sigortanın türünden bağımsız olarak tamamında gerçekleşebileceğinden bahsetmiştir. Sigorta Bilgi ve Gözetim Merkezi'ne 2012 yılında yansıyan istatistiklerden bahsedilen çalışmada en yüksek oranın sahte kaza tespit tutanağı olduğu görülmektedir. Sigorta dolandırıcılığının engellenme yöntemlerine de değinilmiştir. Günbatar (2019) Türkiye'de yapılmış araç sigortaları üzerinden dolandırıcılık vakalarını makine öğrenmesi yöntemleriyle incelemiştir. Çalışmada, Sigorta Sahteciliklerini Engelleme Bürosu'nun sağlamış olduğu veri seti üzerinde LR, Karar Ağaçları, Bayes Ağları, Yapay Sinir Ağları gibi modeller

kullanılmıştır. Moon vd. (2019) otomobil sektöründe poliçe sahiplerinin yapmış oldukları araç sigorta tazminat taleplerini incelemiştir. Fraud aktivitelerin sigorta şirketine yıllık maliyetinin 31 milyar doları bulduğu ve bu rakamın gün geçtikçe arttığı belirtilmiştir. 14927 fraud olarak değerlendirilmeyen veriye karşın 923 adet fraud veri bulunmaktadır. Bu dengesiz bir veri setiyle çalışıldığını göstermektedir. Naive Bayes ve LR da dahil olmak üzere Lasso ve Rastgele Orman gibi birçok farklı sınıflandırma algoritması ve K Katlamalı Çapraz Doğrulama ($k=10$ alınarak) kullanılmıştır. Mambo (2019) sağlık sigortalarında fraud araştırmasını K-NN, Naive Bayes ve Karar Ağacı modelleriyle incelemiştir.



5. VERİ VE YÖNTEM

Bu çalışmada, Open Data Commons lisansıyla Kaggle platformunda (kaggle.com/mhdzahier/travel-insurance) paylaşımına açık olan bir seyahat sigortası veri seti kullanılmıştır. Veriler Singapur’da toplanmış 3.parti bir sigorta şirketinin verileridir. Seyahat sigortası poliçesi sahiplerinin sigorta tazminat talebinde bulunup bulunmadığı incelendiğinden problem bir sınıflandırma problemi olarak tanımlanmaktadır. Çalışmaya ilişkin kaynak kodlar Python programlama dilinin Anaconda platformu üzerindeki Spyder adlı geliştirme ortamında hazırlanmıştır.

5.1 VERİ SETİNİN TANITIMI

Veri setinde 1 tanesi bağımlı, 10 tanesi bağımsız olmak üzere 11 adet değişken bulunmaktadır. Veri setinde toplamda 63326 adet kayıt bulunmaktadır.

Tablo 5.1’de değişkenlerin açıklamaları, türleri ve içeriğiyle ilgili bilgiler yer almaktadır.

Tablo 5.1: Değişken açıklamaları

Değişken Adı	Türü	Durumu
Tazminat Talep	Kategorik	Bağımlı Değişken
Acenta Kodu	Kategorik	Bağımsız Değişken
Acenta Tipi	Kategorik	Bağımsız Değişken
Dağıtım Kanalı	Kategorik	Bağımsız Değişken
Ürün Adı	Kategorik	Bağımsız Değişken
Süre	Nümerik	Bağımsız Değişken
Destinasyon	Kategorik	Bağımsız Değişken
Net Satış	Nümerik	Bağımsız Değişken
Komisyon	Nümerik	Bağımsız Değişken
Cinsiyet	Kategorik	Bağımsız Değişken
Yaş	Nümerik	Bağımsız Değişken

- Tazminat Talep: Sigorta tazminatı talebinde bulunulma durumunu evet veya hayır olarak gösterir.
- Acenta Kodu: Sigorta poliçesini düzenleyen acentanın kodunu gösterir.

- c. Acenta Tipi: Sigorta poliçesini düzenleyen acentanın tipini ifade eder.
- d. Ürün Adı: Seyahat sigortasıyla ilgili bir sigortacılık ürününü ifade eder.
- e. Süre: Seyahat edilen süreyi ifade eder.
- f. Destinasyon: Sigortalının seyahat ettiği yeri göstermektedir.
- g. Net Satış: Acentanın net satış miktarını gösterir.
- h. Komisyon: Acentanın tahsil ettiği komisyonu gösterir.
- i. Cinsiyet: Sigortalının cinsiyetini gösterir.
- j. Yaş: Sigortalının yaşını gösterir.

5.2 VERİ ANALİZİ

5.2.1 Veri Görselleştirme ve Veri Temizleme

Veri biliminin en çok zaman alan ve en önemli safhalarından biri veri analizi safhasıdır. Bu aşamada verinin doğru anlaşılabilmesi, yeterince analiz edilememesi, gürültülü veya aykırı gözlemlerin uygun yöntemler kullanarak elimine edilememesi istatistiksel modellerin kalitesini etkilemektedir.

Şekil 5.1’de verinin hiçbir ön işlemeden geçirilmeden önceki içeriği gösterilmiştir.

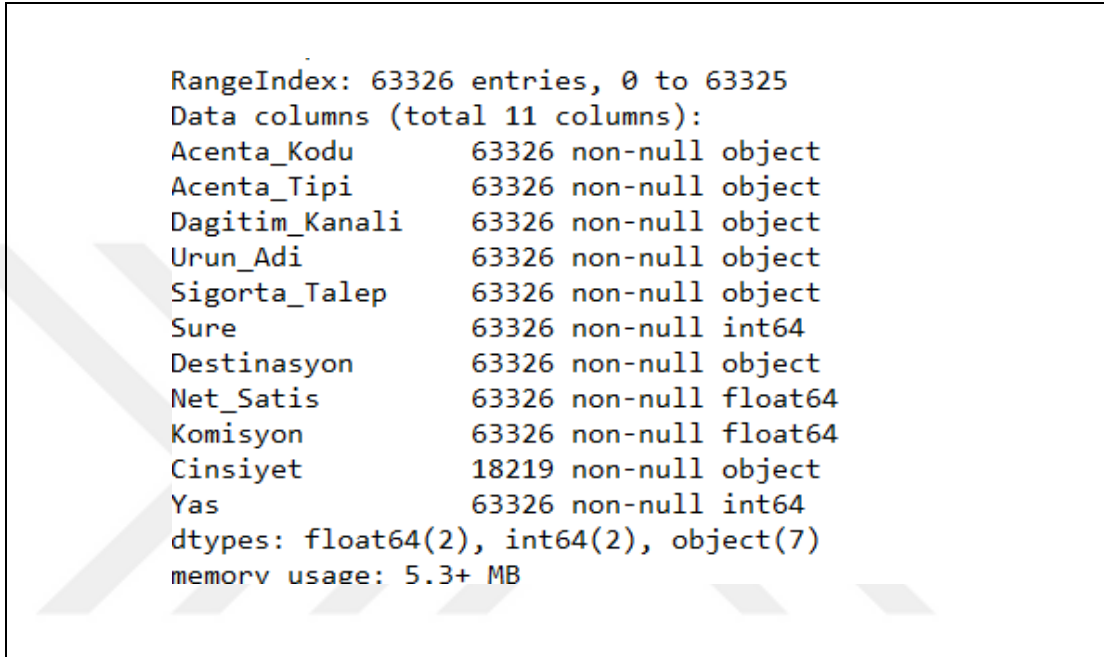
Şekil 5.1: İşlenmemiş veri görünümü

sigorta_frame - DataFrame											
Index	Acenta_Kodu	Acenta_Tipi	Dagitim_Kanali	Urun_Adi	Sigorta_Talep	Sure	Destinasyon	Net_Satis	Komisyon	Cinsiyet	Yas
0	CBH	Travel Age...	Offline	Comprehens...	No	186	MALAYSIA	-29	9.57	F	81
1	CBH	Travel Age...	Offline	Comprehens...	No	186	MALAYSIA	-29	9.57	F	71
2	CWT	Travel Age...	Online	Rental Veh...	No	65	AUSTRALIA	-49.5	29.7	nan	32
3	CWT	Travel Age...	Online	Rental Veh...	No	60	AUSTRALIA	-39.6	23.76	nan	32
4	CWT	Travel Age...	Online	Rental Veh...	No	79	ITALY	-19.8	11.88	nan	41
5	JZI	Airlines	Online	Value Plan	No	66	UNITED STA...	-121	42.35	F	44
6	CWT	Travel Age...	Online	Rental Veh...	No	47	UNITED STA...	-39.6	23.76	nan	32
7	CWT	Travel Age...	Online	Rental Veh...	No	63	AUSTRALIA	-108.9	65.34	nan	29
8	CWT	Travel Age...	Online	Rental Veh...	No	57	THAILAND	-19.8	11.88	nan	44
9	CWT	Travel Age...	Online	Rental Veh...	No	186	AUSTRALIA	-99	59.4	nan	37
10	JZI	Airlines	Online	Basic Plan	No	33	KOREA, DEML	-26	9.1	nan	118
11	JZI	Airlines	Online	Basic Plan	No	1	MALAYSIA	-18	6.3	M	47
12	KML	Travel Age...	Online	Premier PL...	No	53	NORWAY	-130	49.4	F	48
13	CWT	Travel Age...	Online	Rental Veh...	No	5	MALAYSIA	-39.6	23.76	nan	64
14	EPX	Travel Age...	Online	2 way Comp...	No	39	VIET NAM	-25	0	nan	36
15	CWT	Travel Age...	Online	Rental Veh...	No	6	DENMARK	-19.8	11.88	nan	53
16	CWT	Travel Age...	Online	Rental Veh...	No	48	DENMARK	-79.2	47.52	nan	43
17	CWT	Travel Age...	Online	Rental Veh...	No	11	UNITED STA...	-29.7	17.82	nan	58
18	JZI	Airlines	Online	Basic Plan	No	3	MALAYSIA	-18	6.3	M	47
19	CWT	Travel Age...	Online	Rental Veh...	No	14	THAILAND	-69.3	41.58	nan	37
20	CWT	Travel Age...	Online	Rental Veh...	No	136	NORWAY	-108.9	65.34	nan	32

Python'un info adlı fonksiyonu kullanılarak veri setinin her bir değişkeninde boş veri analizi yapılmaktadır. Cinsiyet kolonu hariç, hiçbir kolonda boş veri bulunmadığı tespit edilmiştir. Cinsiyet kolonunda ise 45107 adet boş kayıt bulunmaktadır.

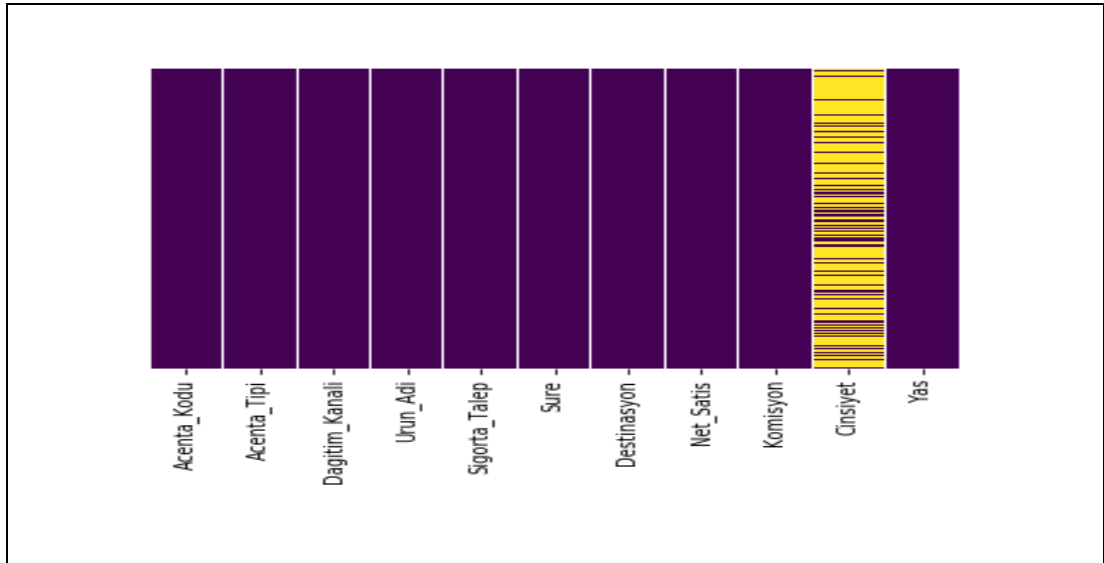
Şekil 5.2'de boş veri analizine ilişkin bilgiler gösterilmiştir.

Şekil 5.2: Boş veri bilgisi



Şekil 5.3'te Seaborn kütüphanesinin boş olan değerlere göre çizdiği ısı haritasına göre cinsiyet kolonunda boş değer sayısının çok fazla olduğu görülmektedir.

Şekil 5.3: Boş veri ısı haritası



Cinsiyet değişkeni çok fazla boş değer içermesi sebebiyle modelden tamamen çıkarılmıştır.

Tablo 5.2’de Python’un describe fonksiyonu kullanılarak float64 ve int64 tipindeki değişkenlere ait sayısal analizler gösterilmiştir.

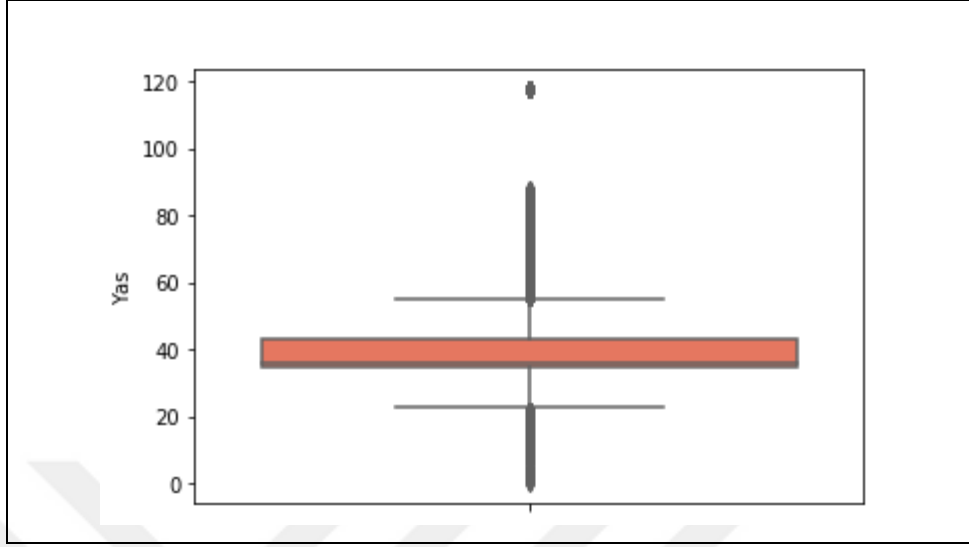
Tablo 5.2: Değişken istatistikleri

	Net Satış	Komisyon	Süre	Yaş
Ortalama	40.702018	9.809992	49.317074	39.969981
Std.Sapma	48.845637	19.804388	101.791566	14.017010
En Küçük	-389	0	-2	0
1.Ç.Kartil	18	0	9	35
Medyan	26.53	0	22	36
3.Ç.Kartil	48	11.55	53	43
En Büyük	810	283.50	4881	118

Örneğin, net satış değişkeninin ortalaması 40.702018, komisyon değişkeninin ortalaması ise 9.809992’dir. Aynı zamanda standart sapma, medyan, minimum, maksimum değerlerin yanı sıra yüzde 25’lik, yüzde 50’lik ve yüzde 75’lik çeyrek kartil değerleri de gösterilmektedir. Süre değişkeninin standart sapmasının yüksek olduğu, minimum olarak negatif değerler barındırdığı, maksimum değeri 4881 olduğu görülmektedir. Yaş değişkeninde ise maksimum 118 yaş bilgisi olduğu görülmektedir. Bu bilgiler veri setinin aykırı gözlem içerdiğine işaret etmektedir. Verilerin farklı ölçeklerde ve aralıklarda bulunmuş olması veri setinin sayısal değişkenlerinde bir normalizasyona ihtiyaç duyulduğunu göstermektedir. Bu normalizasyon verinin eğitim ve test verisi olarak ayrıldığı aşamada yapılacaktır.

Şekil 5.4'te yaş değişkenine ait kutu grafiği (box plot) gösterilmektedir.

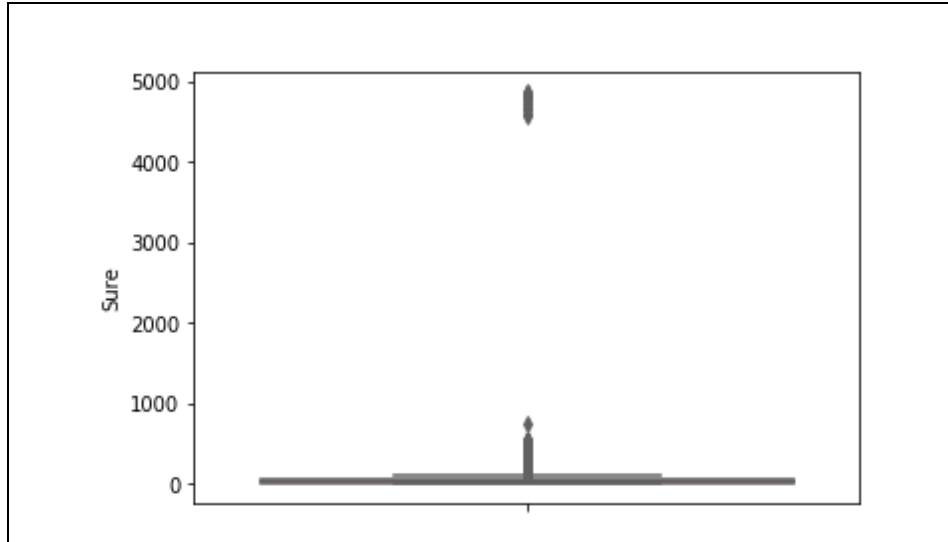
Şekil 5.4: Yaş kutu grafiği



Yaş değişkeninde aykırı gözlemlerin temizlenmesi için yaş bilgisi 118'den düşük olan kayıtların yaşlarının aritmetik ortalaması hesaplanmış (38.7383) ve yaş alanında 118 olarak gözükten değerlere bu değer konulmuştur.

Şekil 5.5'te süre değişkenine ait kutu grafiği gösterilmektedir.

Şekil 5.5: Süre kutu grafiği



Süre değişkeninde negatif değerlerin temizlenmesi maksadıyla süresi 0'a eşit ve 0'dan büyük olan kayıtların sürelerinin aritmetik ortalaması hesaplanmış (49.3210) ve süre alanındaki negatif değerlere bu değer konulmuştur.

Şekil 5.6’da gösterildiği üzere Destinasyon değişkeni için bir veri sözlüğü oluşturulmuştur.

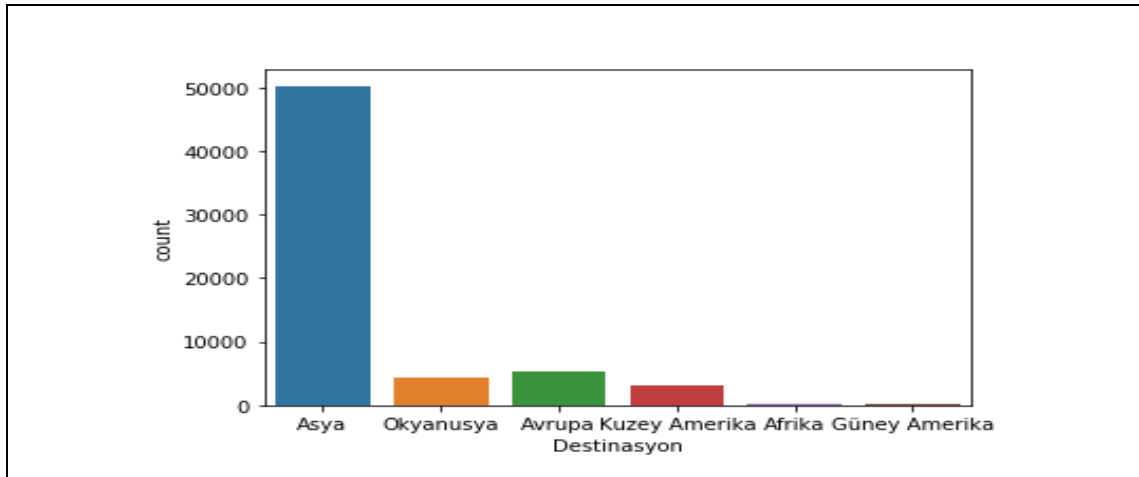
Şekil 5.6: Veri sözlüğü

```
dictg = {'AUSTRIA': 'Avrupa',  
        'ALBANIA': 'Avrupa',  
        'BELARUS': 'Avrupa',  
        'BELGIUM': 'Avrupa',  
        'BOSNIA AND HERZEGOVINA': 'Avrupa',  
        'BULGARIA': 'Avrupa',  
        'CROATIA': 'Avrupa',  
        'CYPRUS': 'Avrupa',  
        'CZECH REPUBLIC': 'Avrupa',  
        'DENMARK': 'Avrupa',  
        'ESTONIA': 'Avrupa',  
        'FAROE ISLANDS': 'Avrupa',  
        'FINLAND': 'Avrupa',  
        'FRANCE': 'Avrupa',  
        'GERMANY': 'Avrupa',  
        'GREECE': 'Avrupa',  
        'HUNGARY': 'Avrupa',  
        'ICELAND': 'Avrupa',  
        'IRELAND': 'Avrupa',  
        'ITALY': 'Avrupa',  
        'LATVIA': 'Avrupa',  
        'LITHUANIA': 'Avrupa',  
        'LUXEMBOURG': 'Avrupa',  
        'MACEDONIA, THE FORMER YUGOSLAV REPUBLIC OF': 'Avrupa',  
        'MALTA': 'Avrupa',  
        'MOLDOVA, REPUBLIC OF': 'Avrupa',  
        'NETHERLANDS': 'Avrupa',  
        'NORWAY': 'Avrupa',
```

Veri sözlüğü destinasyon alanında çok fazla tekil değer bulunması sebebiyle oluşturulmuştur. Bu sayının fazlalığı istatistiksel modellerin tahmin gücünü zayıflatabilir. Bu sebeple veri seti içerisinde bulunan ülkeler için bulundukları kıtaları belirten bir veri sözlüğü oluşturulmuş ve destinasyon alanı bu sözlüğe göre doldurulmuştur. Buna göre; Asya, Avrupa, Afrika, Kuzey Amerika, Güney Amerika ve Okyanusya’da bulunan ülkelerin veri setinin destinasyon alanında yer almış oldukları görülmüştür.

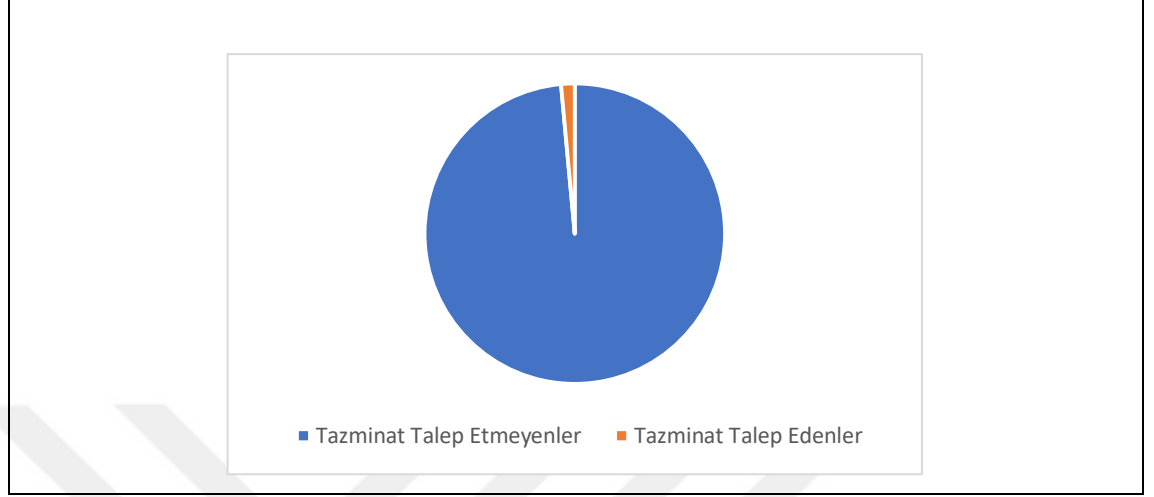
Şekil 5.7’ye göre en fazla Asya kıtasındaki ülkelere seyahat gerçekleştirilmiştir.

Şekil 5.7: Destinasyon dağılımı



Şekil 5.8’de görülebileceği üzere sigorta tazminat talebinde bulunmayan sigortalı sayısı 62399, tazminat talebinde bulunan sayı ise 927’dir.

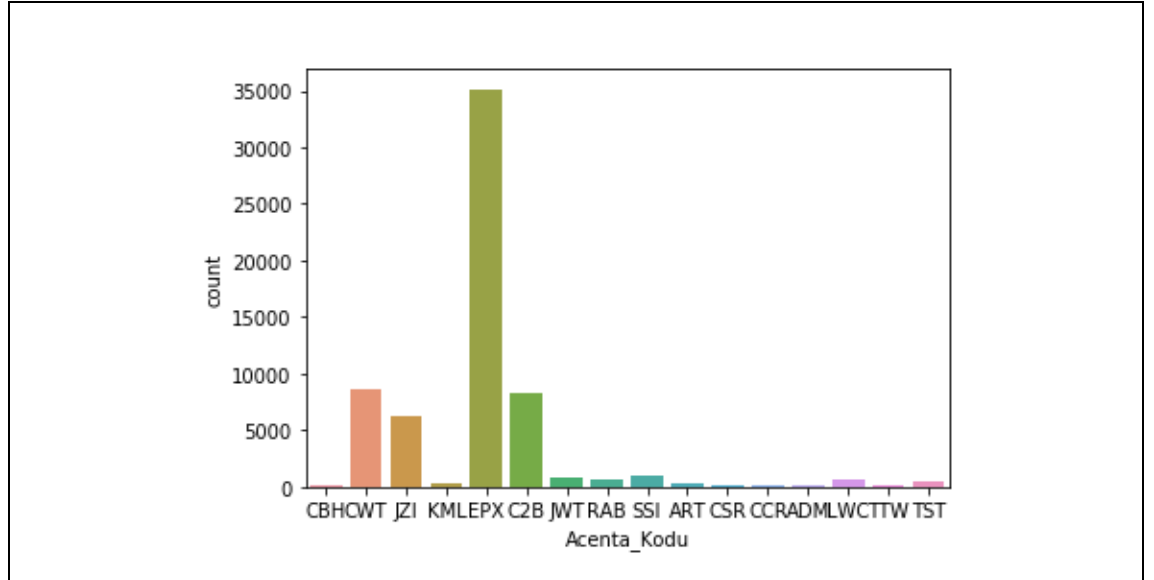
Şekil 5.8: Tazminat talep dağılımı



Tazminat talebinde bulunanların oranı yaklaşık yüzde 1.46’ya eşittir. Bu durum dengesiz veri kümesi problemini göstermektedir.

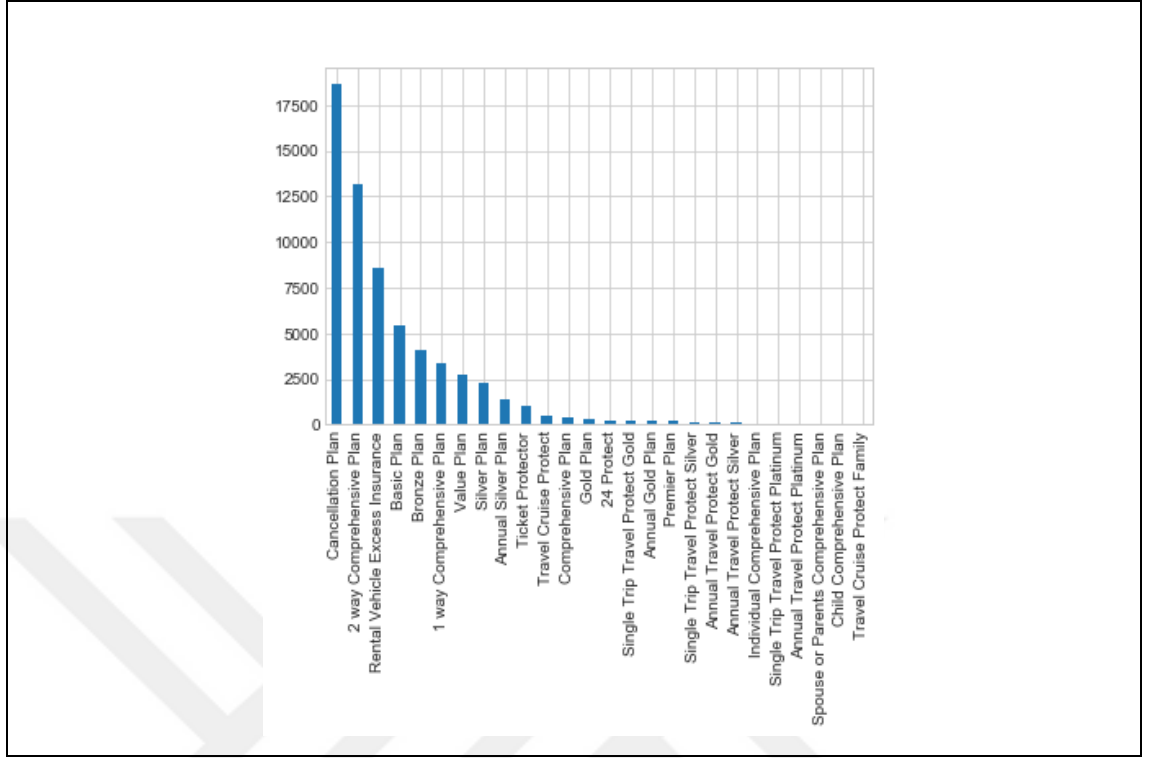
Şekil 5.9’da seyahat sigortası acentalarının sayısal olarak dağılımı gösterilmektedir.

Şekil 5.9: Acente kodu dağılımı



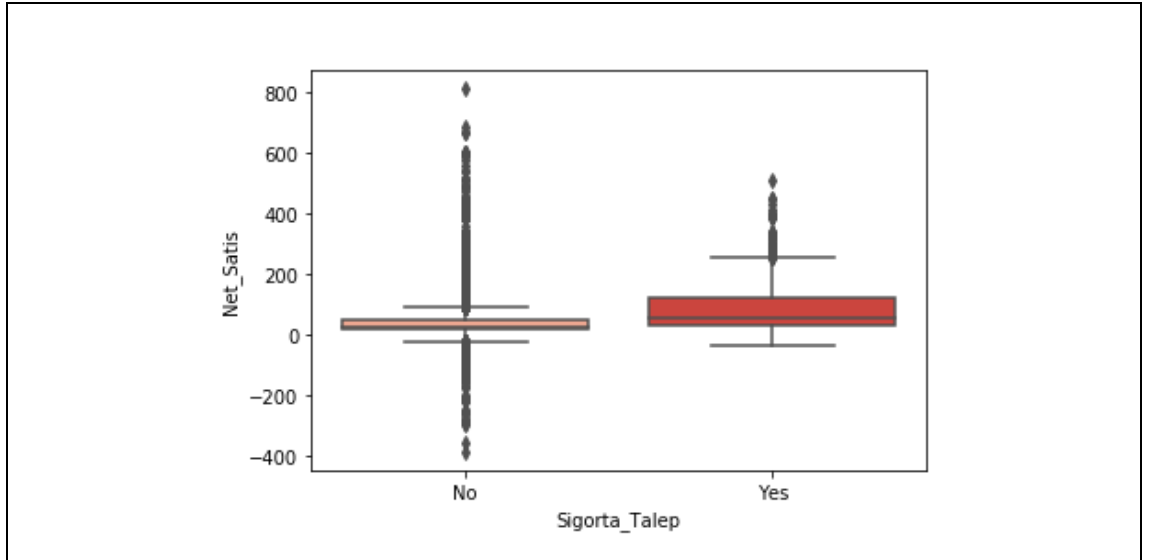
Şekil 5.10’da ürün dağılımları gösterilmektedir. En fazla iptal planı ürününün olduğu görülmektedir.

Şekil 5.10: Ürün dağılımı



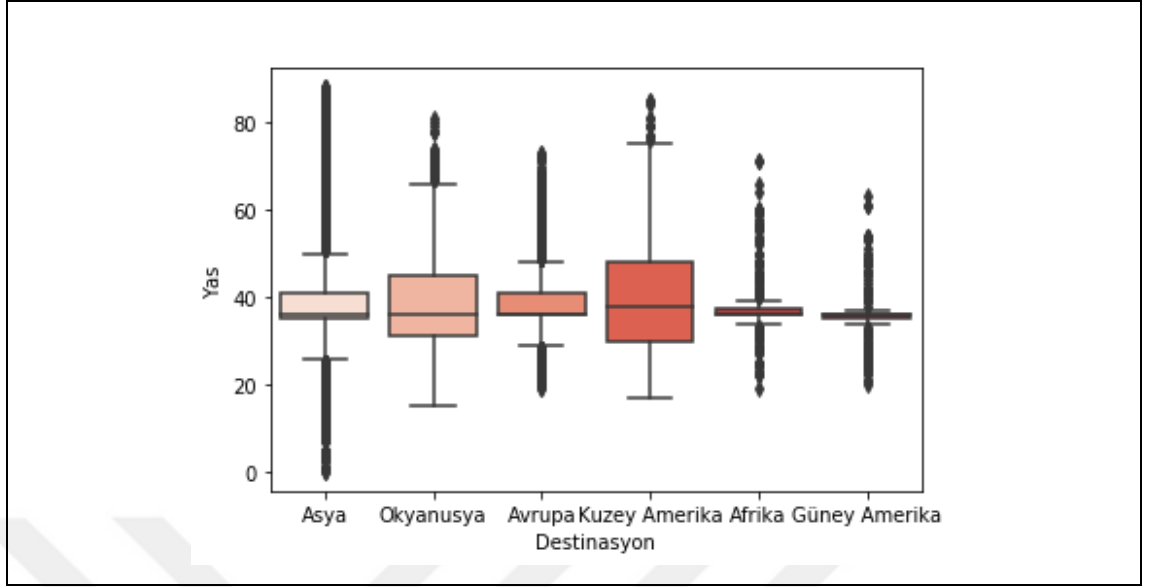
Şekil 5.11’de sigorta tazminat talebinde bulunan ve bulunmayan kişilerin net satış değişkeniyle ilişkili kutu grafiği gösterilmektedir.

Şekil 5.11: Tazminat talep ve net satış ilişkisi



Şekil 5.12’de destinasyon ve yaş arasındaki kutu grafiği gösterilmektedir.

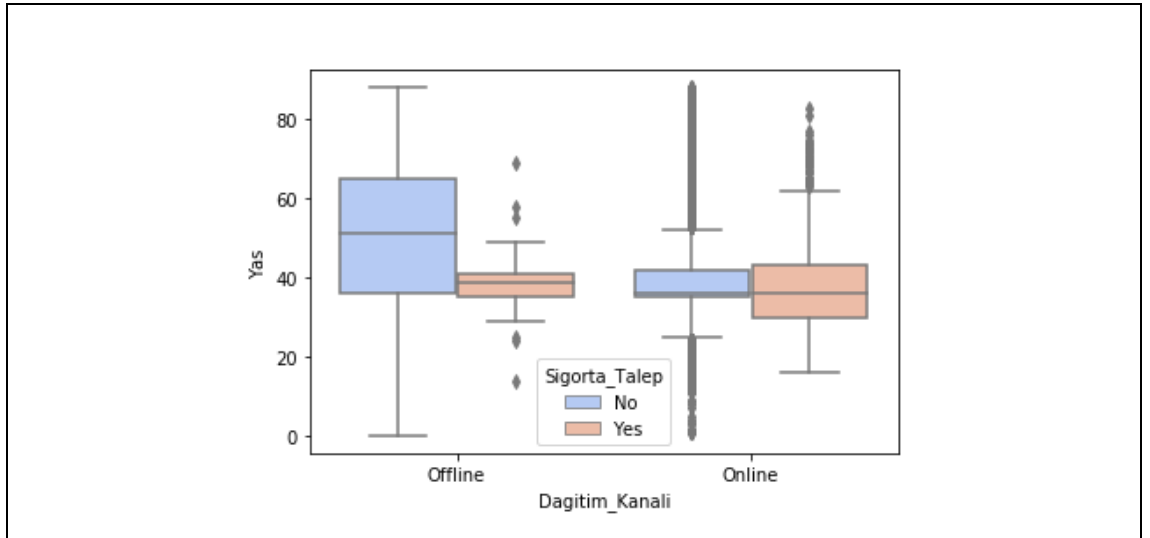
Şekil 5.12: Destinasyon ve yaş ilişkisi



Kuzey Amerika kıtasında popülasyonun genelde 3.çeyrek kartil civarında toplandığı görülmektedir. Asya kıtasında aykırı gözlem yaş değeri sayısı oldukça yüksektir.

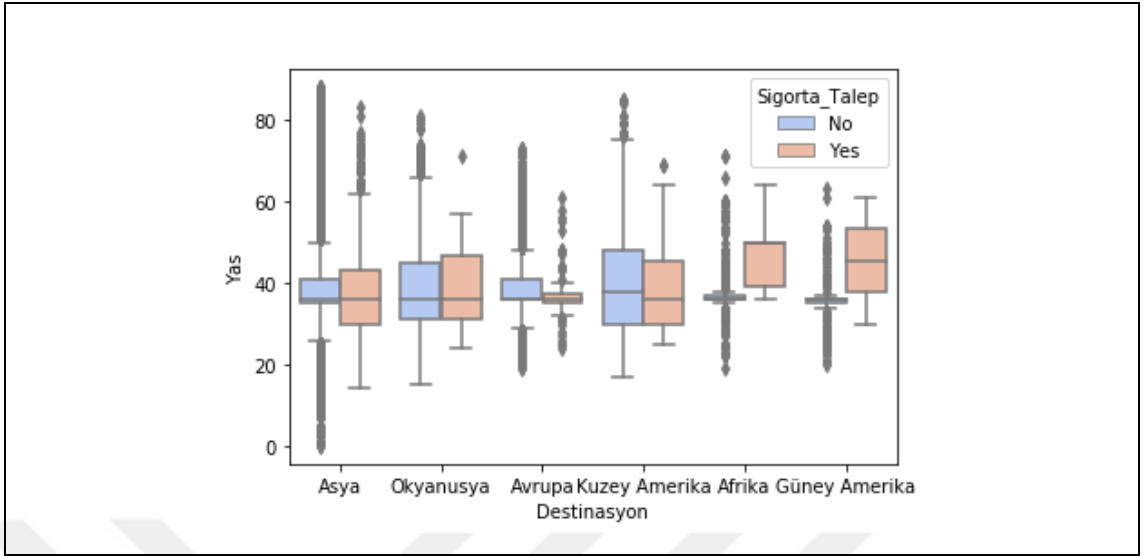
Şekil 5.13'te dağıtım kanalı, yaş ve sigorta tazminat talep durumu arasındaki ilişki gösterilmiştir.

Şekil 5.13: Dağıtım kanalı, yaş ve tazminat talep ilişkisi



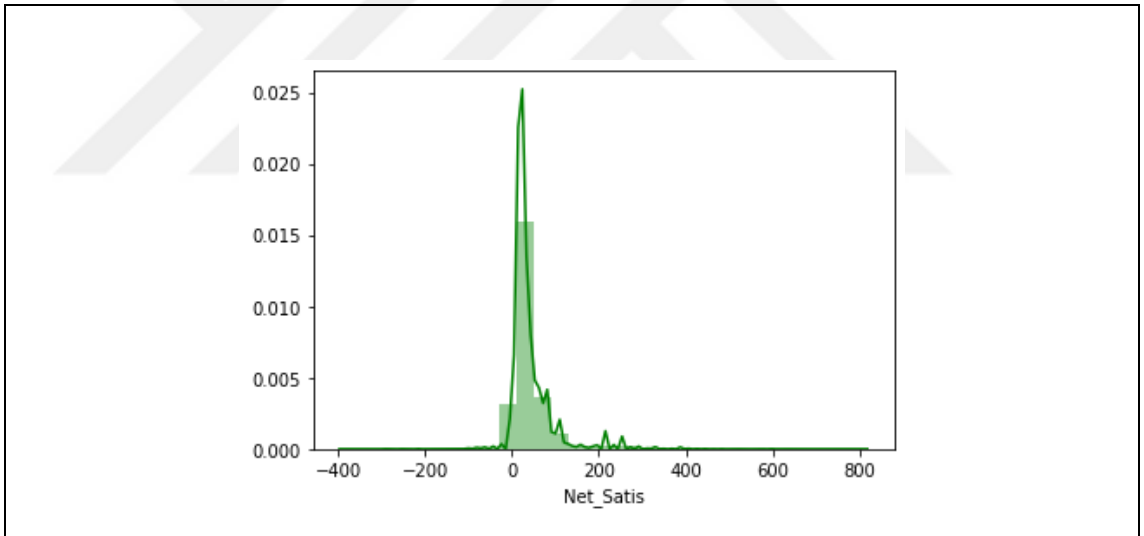
Şekil 5.14'te destinasyon, yaş ve sigorta tazminat talep durumu arasındaki ilişki gösterilmiştir.

Şekil 5.14: Destinasyon, yaş ve tazminat talep ilişkisi



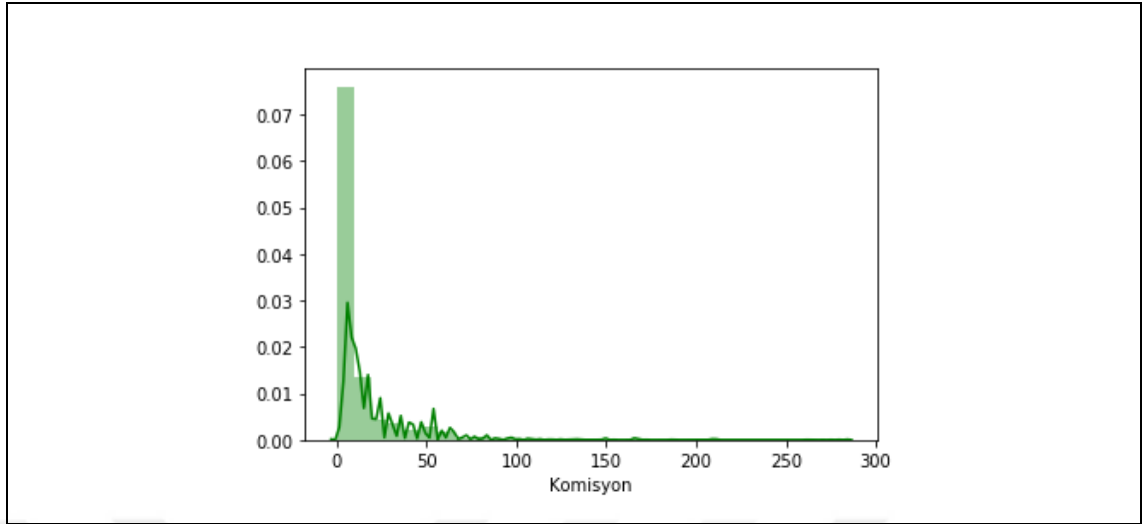
Şekil 5.15'te net satış verisine ilişkin bir dağılım eğrisi gösterilmektedir.

Şekil 5.15: Net satış dağılım grafiği



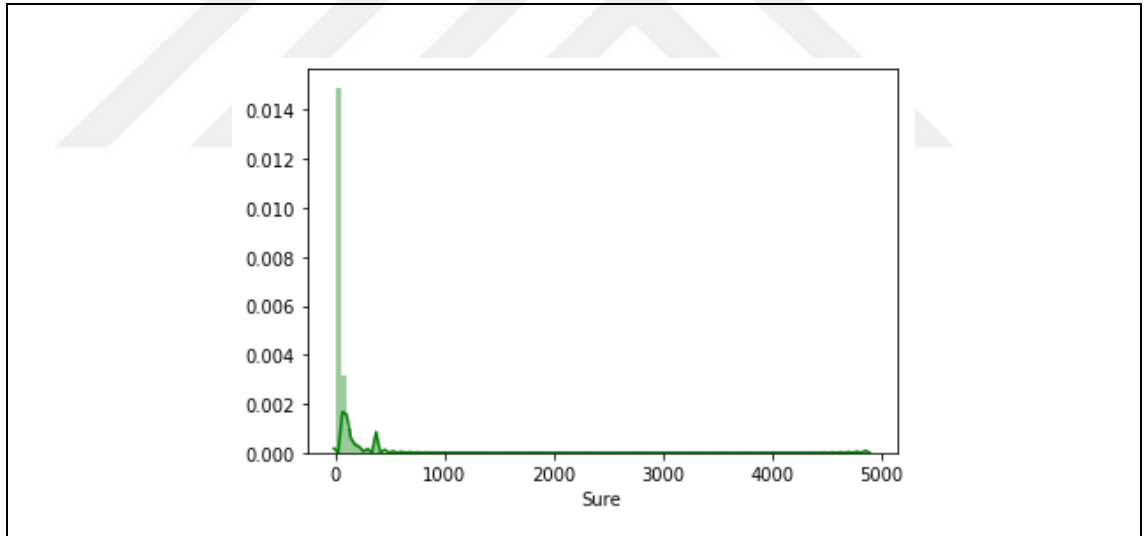
Şekil 5.16'da komisyon değişkenine ilişkin bir dağılım eğrisi gösterilmektedir.

Şekil 5.16: Komisyon dağılım grafiği



Şekil 5.17’de süre değişkenine ait bir dağılım eğrisi gösterilmektedir.

Şekil 5.17: Süre dağılım grafiği



5.2.2 Korelasyon Analizi

Değişkenler arasındaki ilişkiler belirlenirken Korelasyon Analizi yapılmaktadır. Korelasyon neden-sonuç ilişkisine işaret edebilir veya etmeyebilir. Bu konuda herhangi bir kesinlik yoktur.

5.2.2.1 Pearson Korelasyonu

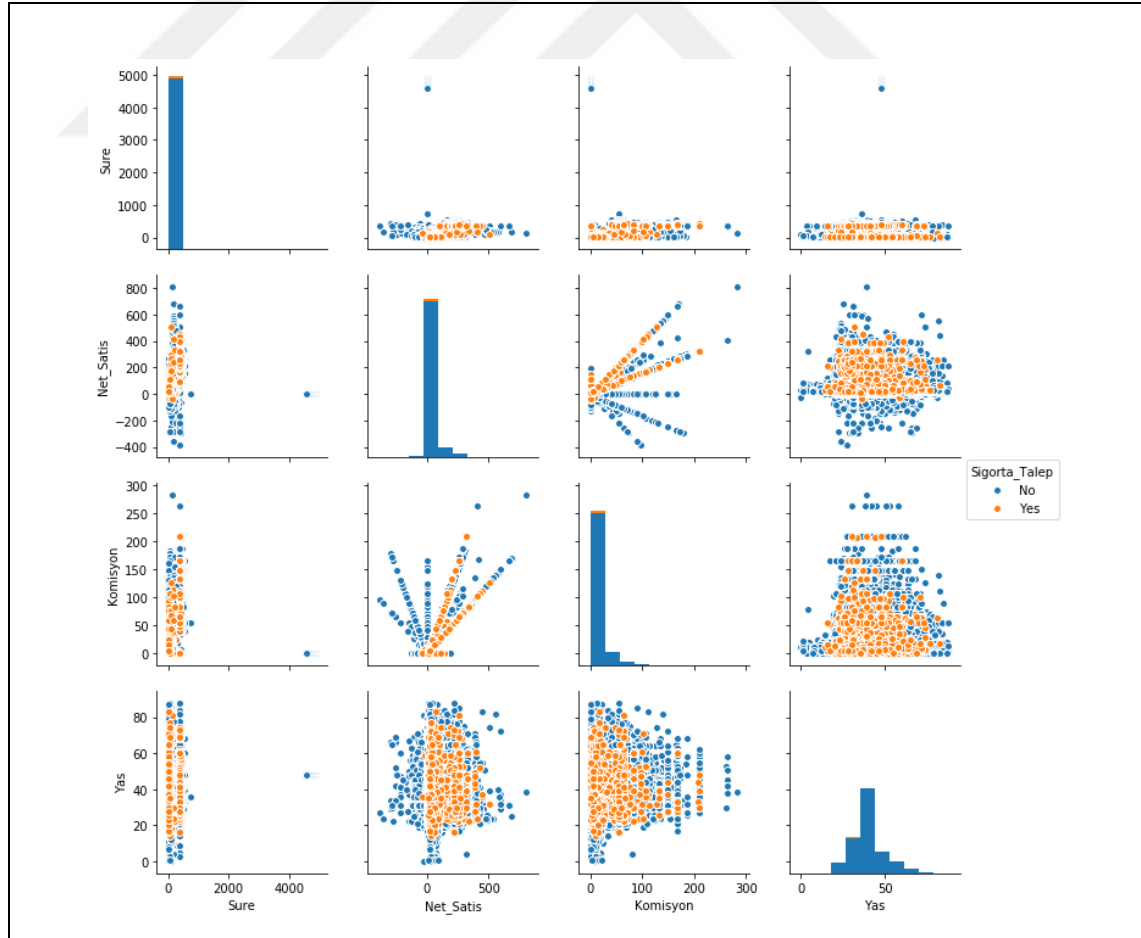
Sayısal değişkenler arasındaki ilişkinin belirlenmesinde Pearson Korelasyonu kullanılmaktadır. Pearson Korelasyonu için katsayı en küçük -1, en büyük ise +1 olabilir. -1 negatif yönlü bir ilişkinin, +1 pozitif yönlü ilişkinin varlığına işaret ederken, 0 ise hiçbir ilişkinin olmadığını göstermektedir.

Denklem 5.1’de Pearson Korelasyonu’nun hesaplanması gösterilmiştir.

$$r_{XY} = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2} \sqrt{\sum_{i=1}^n (Y_i - \bar{Y})^2}} \quad (5.1)$$

Şekil 5.18’de veri setinde yer alan sayısal değişkenlerin arasındaki ilişkiyi belirten ait saçılım grafiği (scatter plot) gösterilmektedir.

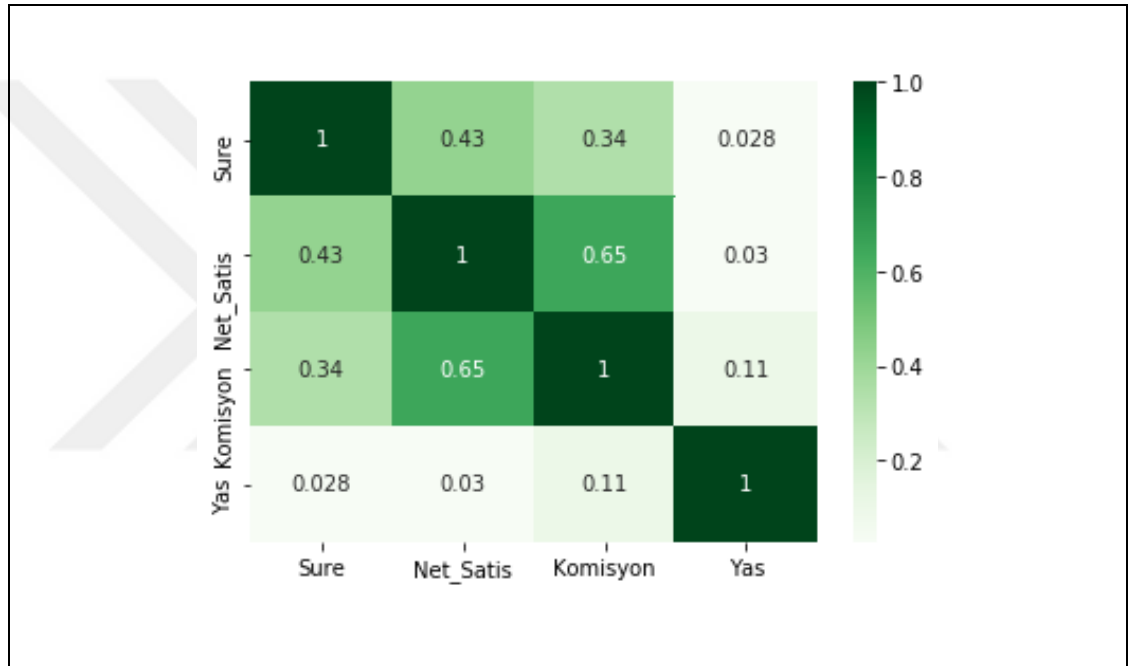
Şekil 5.18: Sayısal değişkenler arası ilişki için saçılım grafiği



Bu saçılım grafiği aynı zamanda hedef değişkenimiz olan sigorta tazminat talep durumuyla da ilişkilendirilmiştir. Turuncu renk olanlar tazminat talebinde bulunanları, mavi renkli olanlar ise talepte bulunmayanları göstermektedir. Tazminat talep durumu değişkenini hesaba katmaksızın, grafikten değişkenlerin birbirleriyle olan ilişkileri de tespit edilebilmektedir.

Şekil 5.19'da Python'un SeaBorn kütüphanesinde bulunan corr fonksiyonu ile sayısal değişkenlere ait hesaplanmış korelasyon matrisi gösterilmektedir.

Şekil 5.19: Sayısal değişkenler için korelasyon matrisi



Örneğin, komisyon ve net satışın arasında yüzde 65 korelasyon, net satış ve süre arasında yüzde 43'lük bir korelasyon bulunmaktadır. Komisyon ve yaş arasında yüzde 11 korelasyon bulunmaktadır.

5.2.2.2 Cramer's V

Bir veri setinde yer alan kategorik değişkenler arasındaki ilişki belirlenirken kullanılan tekniklerden biri Cramer's V'dir. Cramer's V için en küçük değer 0, en büyük ise +1 olabilir. Negatif bir ilişki türü yoktur. +1 pozitif yönlü ilişkinin varlığına işaret ederken,

0 ise hiçbir ilişkinin olmadığını göstermektedir (people.vcu.edu). Cramer's V aynı zamanda Cramer's Phi olarak da bilinmektedir.

Denklem 5.2'de Cramer's V'nin hesaplanması ifade edilmiştir.

$$\text{Cramér } V = \sqrt{\frac{\chi^2/n}{\min(r-1, c-1)}} \quad (5.2)$$

Şekil 5.20'de Cramer's V'nin Python'da yazılan fonksiyonu gösterilmektedir.

Şekil 5.20: Cramer's V fonksiyonu

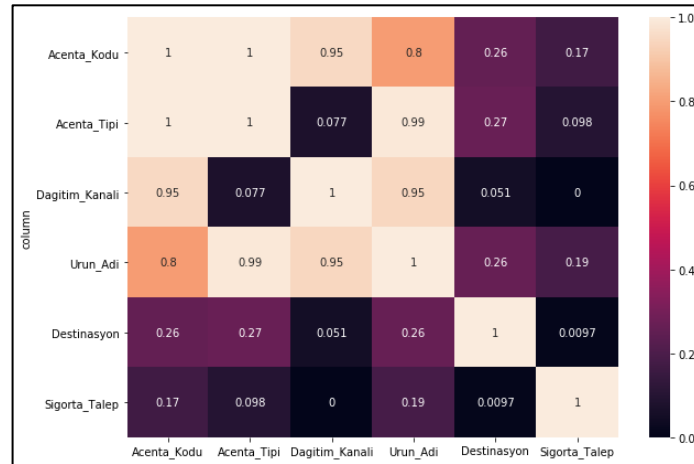
```
def cramers_v(x, y):
    confusion_matrix = pd.crosstab(x,y)
    chi2 = ss.chi2_contingency(confusion_matrix)[0]
    n = confusion_matrix.sum().sum()
    phi2 = chi2/n
    r,k = confusion_matrix.shape
    phi2corr = max(0, phi2-((k-1)*(r-1))/(n-1))
    rcorr = r-((r-1)**2)/(n-1)
    kcorr = k-((k-1)**2)/(n-1)
    return np.sqrt(phi2corr/min((kcorr-1),(rcorr-1)))
```

Kaynak: Zychlinski, 2018

Yazılmış olan bu cramers_v fonksiyonu çalıştırıldığında Cramer's V katsayıları kategorik değişkenler için bulunmaktadır.

Şekil 5.21'de kategorik değişkenler arası korelasyon değerleri gösterilmektedir.

Şekil 5.21: Kategorik değişkenler için korelasyon matrisi



Acenta bilgisi (Acenta Kodu) ile sigorta tazminat talebi değişkeni arasında yüzde 17, ürün adı ve sigorta tazminat talebi arasında yüzde 19 korelasyon olduğu görülmektedir.

Dağıtım kanalı ve sigorta tazminat talebi arasında ise hiçbir ilişkinin olmadığı dolayısıyla korelasyonun yüzde 0 olduğu anlaşılmaktadır. Cinsiyet değişkeni, çok fazla boş değer içermesi sebebiyle modelden çıkartıldığı için Cramer's V katsayısı hesaplamasında kullanılmamıştır. Destinasyon değişkeni de tekil değerlerden kurtarılarak hazırlanmış olan veri sözlüğüne göre matrise dahil edilmiştir.

5.2.3 Minimum Maksimum Normalizasyonu

Veri setindeki sayısal değişkenler farklı ölçütlerde olduğunda istatistiksel modellerin performansları düşmektedir. Bunu önlemek maksadıyla sayısal değişkenlerin normal dağılıma yaklaştırılması gerekmektedir (Raschka, 2014).

Makine öğrenmesi algoritması, farklı birim veya skalalarda bulunan verilerin aynı model içerisinde tahminlenmesinde iyi sonuçlar vermeyebilir. Bunun çözülebilmesi için feature scaling (özellik ölçekleme) adı verilen bir yöntem uygulayarak, farklı değişken aralıklarını normalize etmek veya standartlaştırmak gerekmektedir. Bu işlem sonrasında veri setinde yer alan tüm sayısal değişkenler aynı skala içerisinde yer alacaktır. Bu durum, istatistiksel algoritmalarla kurulan modellerin performanslarına doğrudan pozitif etki edecektir (Nayak vd., 2014).

Minimum Maksimum Normalizasyonu veri setinde sayısal değişkenin bulunduğu ilgili kolondaki en küçük ve en büyük değeri bulduktan sonra denklem 5.3'te gösterilen hesaplama göre kolondaki tüm değerleri 0 ve 1 aralığında (0 ve 1 dahil) olacak şekilde değiştirmektedir.

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}} \quad (5.3)$$

Normalizasyon işlemi veri setinin eğitim ve test olarak ayrıldığı safhada uygulanmalıdır.

5.3 LOJİSTİK REGRESYON

Sınıflandırma problemlerinin çözümünde kullanılan algoritmalarından biridir. Bağımlı değişkenin (y) ayrık değişken olduğu durumlarda bu değişkenin çıktısının ne olacağını tahmin ederken kullanılmaktadır. Evet/Hayır, Var/Yok gibi sonuçların tespitinde LR modeli kullanılabilir. Kısacası, bağımlı değişkenimiz kategorik olmalıdır (Ching, 2020). Lineer Regresyon tabanlı ancak olasılık teoremine dayalı bir algoritmadır. Sonuçların başarılı olabilmesi için regresyonda çoklu ilişki problemi (multicollinearity) olmaması gerekmektedir.

z sayısı $(-\infty, +\infty)$ aralığında sürekli bir değişken olmak üzere Sigmoid (Lojistik) fonksiyonu denklem 5.4'te gösterilmektedir.

$$\theta(z) = \frac{e^z}{1 + e^z} = \frac{1}{1 + e^{-z}} \quad (5.4)$$

Sigmoid fonksiyonu; z sayısı çok büyük değerler aldığında yaklaşık olarak 1, z sayısı çok küçük veya negatif değerler aldığında yaklaşık olarak 0, z sayısı 0 değerini aldığında ise 0.5 değerine eşit olmaktadır.

Denklem 5.5'te basit lineer regresyon denklemi gösterilmektedir.

$$\hat{y} = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p \quad (5.5)$$

Basit lineer regresyondan elde edilen çıktı yerine yazıldığında denklem 5.6 elde edilmektedir.

$$\theta(z = \hat{y}) = \frac{1}{1 + e^{-\hat{y}}} = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \dots + \beta_p x_p)}} \quad (5.6)$$

Bu denklem Lojistik Regresyonu ifade etmektedir. LR, maliyet fonksiyonu itibariyle lineer regresyondan ayrılmaktadır.

Denklem 5.7'de odds oranı gösterilmektedir.

$$\text{Odds} = \frac{\theta}{1 - \theta} \quad (5.7)$$

Pay kısmı gerçekleşme olasılığını, payda kısmı ise gerçekleşmeme olasılığını ifade eder.

Denklem 5.8’de odds fonksiyonu için her iki tarafın doğal logaritmasının alınması sonrası oluşan Logit fonksiyonu yer almaktadır.

$$\text{Logit} = \log(\text{odds}) = \log\left(\frac{\theta}{1 - \theta}\right) \quad (5.8)$$

LR modeli, Logit fonksiyonu ve bağımsız değişkenler arasında lineer bir ilişki olduğunu varsaymaktadır.

Python’un Scikit-Learn kütüphanesinde bulunan LogisticRegression fonksiyonu kullanılarak model kurulumu gerçekleştirilmiştir.

5.4 NAIVE BAYES

Sınıflandırma problemlerinin çözümünde kullanılan algoritmalarından biridir. Bayes olasılığı tabanlı bir algoritmadır. Gerçek dünyada çok sayıda problemin çözümünde (istenmeyen mail tespiti, metin madenciliği, öneri sistemleri gibi) kullanılmaktadır.

Naive Bayes (NB) yaklaşımı tüm bağımlı değişkenlerin birbirleri arasında herhangi bir etkisinin olmadığını ve çıktıya eşit ölçüde etki ettiğini varsaymaktadır (Gupta, 2017). Bu durum NB algoritmasının hızlı çalışmasında önemli rol oynamaktadır (Roman, 2019).

Denklem 5.9’da A ve B olmak üzere toplam 2 adet olay için Bayes teoreminin formülü gösterilmektedir.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \quad (5.9)$$

- $P(A|B)$: B olayı varken A olayının olma ihtimalini ifade eder.
- $P(B|A)$: A olayı varken B olayının olma ihtimalini ifade eder.
- $P(A)$: A olayının olma ihtimalini ifade eder.
- $P(B)$: B olayının olma ihtimalini ifade eder.

Denklem 5.10’da Bayes teoreminin birden çok olay olduğu durumlardaki hesaplanışı gösterilmektedir.

$$P(Y = k | X) = \frac{P(X|Y = k) * P(Y = k)}{P(X)} \quad (5.10)$$

Denklem 5.11’de NB algoritmasının matematiksel ifadesi gösterilmiştir. Olaylar birbirlerinden bağımsız kabul edildiği için çarpım şeklinde gösterilmektedir.

$$P(Y = k | X_1 \dots X_n) = \frac{P(X_1|Y = k) * P(X_2|Y = k) \dots * P(X_n|Y = k) * P(Y = k)}{P(X_1) * P(X_2) \dots * P(X_n)} \quad (5.11)$$

Denklem 5.12’de NB’in daha sade hali gösterilmektedir.

$$P(Y = k | X_1 \dots X_n) = P(Y = k) * \prod_{i=1}^n P(X_i | Y = k) \quad (5.12)$$

Denklem 5.13’de Gaussian NB’e ait formül gösterilmektedir.

$$P(x_i | y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp \left(-\frac{(x_i - \mu_y)^2}{2\sigma_y^2} \right) \quad (5.13)$$

Bu değişkenlerin sürekli değişken olduğu durumlarda kullanılmaktadır. Bu çalışmada da Gaussian NB kullanılmıştır.

Python’un Scikit-Learn kütüphanesinde bulunan GaussianNB fonksiyonu kullanılarak model kurulumu gerçekleştirilmiştir.

5.5 TEMEL BİLEŞEN ANALİZİ

Veri setinin boyut indirgeme (dimension reduction) tekniğiyle basitleştirilmesine Temel Bileşen Analizi (TBA) denir (Collins, 2010). Veri keşfinde, görselleştirmesinde, örüntü tanımadada, görüntü işlemede, yüz tanıma sistemlerinde ve zaman serilerinin çözümlerinde kullanılabilir (Song vd., 2010). TBA’de orijinal veri setine lineer dönüşüm uygulanarak belirlenen her bir temel bileşenin varyans hesaplaması yapılmaktadır (Korjus, 2014). TBA sanal bileşenler oluşturup veri setindeki değişkenlerin sayısının azaltılarak model başarısının artırılmasını amaçlamaktadır.

Alakasız veya modele çok fazla etkisi olmayan değişkenlerin kullanımı model performansını etkilemektedir. İlk temel bileşen, varyansı (değişkenliği) en yüksek olan temel bileşendir. Boyut indirgeme yapılırken kritik nokta istatistiksel öğrenme algoritmasını eğitme aşamasında bilgi kaybının önüne geçilmesi noktasıdır. Bunu yaparken korelatif olmayan sanal değişkenler yani bileşenler yaratılmış olur (Jolliffe ve Cadima, 2015).

Denklem 5.14’de kovaryans matrisinin hesaplanması gösterilmektedir.

$$cov(X,Y) = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{x})(Y_i - \bar{y}) \quad (5.14)$$

Tüm veri setinin kovaryans matrisi bulunduktan sonra matrisin özdeğer ve özvektörler hesaplanmalıdır.

Denklem 5.15’de özdeğerlerin bulunacağı determinant hesaplaması yer almaktadır.

$$\det(A - \lambda I) = 0 \quad (5.15)$$

A bir kare matrisi (veri setini), λ özdeğerleri ve I birim matrisi ifade etmek üzere determinant hesaplaması yapılmalıdır. Özdeğerler bulunduktan sonra özvektörlere ulaşılır. Lineer dönüşüm uygulandıktan sonra yönü değişmeyen vektörler özvektörlerdir. En büyük özdeğere sahip özvektörler seçilir ve dxk boyutundaki W matrisine ulaşılarak temel bileşenler bulunur.

Şekil 5.22’de TBA’nin Python üzerinden bir örnek kod bloğu yer almaktadır.

Şekil 5.22: PCA fonksiyonu

```
pca = PCA(n_components=7, random_state=42)
x_train = pca.fit_transform(x_train)
x_test = pca.transform(x_test)
explained_variance = pca.explained_variance_ratio_
```

7 adet temel bileşen seçilmiştir. Bu şekilde yapıldığında ilk 7 temel bileşenin varyansı hesaplanacak ve modeldeki değişken sayısı 7 olarak düşünülecektir. Varyans toplamı yükseldikçe temel bileşenin modeldeki belirsizliği açıklayabilme gücü artmaktadır.

Python'un Scikit-Learn kütüphanesinde bulunan PCA fonksiyonu kullanılarak boyut indirgeme gerçekleştirilmiştir.

5.6 SMOTE

SMOTE (Synthetic Minority Over-sampling Technique) dengesiz veri problemlerinin çözümünde kullanılan önemli bir örneklem arttırma tekniğidir. Bağımlı değişkende sayıca az olan sınıfı tespit edip bu sınıfın sayısını arttırarak sayıca fazla olan sınıfla eşitler ve dengesiz veri problemini çözer. Bunu yaparken en yakın komşuluk algoritmasını kullanır (Chawla vd., 2002).

Şekil 5.23'te SMOTE örneklem arttırma tekniğinin açık algoritması verilmiştir.

Şekil 5.23: Smote algoritması

```

Algorithm SMOTE(T, N, k)
Input: Number of minority class samples T; Amount of SMOTE N%;
       Number of nearest neighbors k
Output: (N/100) * T synthetic minority class samples
1. (* If N is less than 100%, randomize the minority class samples as
   only a random percent of them will be SMOTEd. *)
2. if N < 100
3.   then Randomize the T minority class samples
4.     T = (N/100) * T
5.     N = 100
6.   endif
7. N = (int)(N/100) (* The amount of SMOTE is assumed to be in
   integral multiples of 100. *)
8. k = Number of nearest neighbors
9. numattrs = Number of attributes
10. Sample[ ]: array for original minority class samples
11. newindex: keeps a count of number of synthetic samples generated,
   initialized to 0
12. Synthetic[ ]: array for synthetic samples
   (* Compute k nearest neighbors for each minority class sample only. *)
13. for i ← 1 to T
14.   Compute k nearest neighbors for i, and save the indices in
   the nnarray
15.   Populate(N, i, nnarray)
16. endfor

   Populate(N, i, nnarray) (* Function to generate the synthetic sam-
   ples. *)
17. while N ≠ 0
18.   Choose a random number between 1 and k, call it nn. This
   step chooses one of the k nearest neighbors of i.
19.   for attr ← 1 to numattrs
20.     Compute: dif = Sample[nnarray[nn]][attr] - Sample[i][attr]
21.     Compute: gap = random number between 0 and 1
22.     Synthetic[newindex][attr] = Sample[i][attr] + gap *
     dif
23.   endfor
24.   newindex++
25.   N = N - 1
26. endwhile
27. return (* End of Populate. *)
End of Pseudo Code.

```

Kaynak: Chawla vd., 2002

X_i sayıca az olan sınıfın değişkeni kabul edilirse SMOTE öncelikli olarak bu sınıfın k en yakın komşusunu bulmaktadır. Sonrasında rastgele bir komşu seçilmekte ve x_i ile komşuyu birbirine bağlayan yapay örneklemeler yaratılmaktadır (Chawla vd., 2002). Örneklem azaltma bilgi kaybına sebep olacağından örneklem artırma tekniği olan SMOTE tercih edilmiştir. Python'un Imbalanced-Learn kütüphanesinde bulunan smote fonksiyonu kullanılarak örneklem artırma yapılmıştır.

5.7 MODEL ÖLÇÜMLENDİRME

Karmaşıklık matrisinde yer alan değerlere göre model ölçülenmesi için bazı kriterler hesaplanmaktadır. Bağımlı değişkenin gerçekten pozitif ve negatif olduğu durumlara karşılık istatistiksel modelin yaptığı tahminler bu matriste yer almaktadır.

Tablo 5.3'te karmaşıklık matrisi gösterilmektedir.

Tablo 5.3: Karmaşıklık matrisi

Bağımlı Değişken	Negatif Tahmin Edilen (0)	Pozitif Tahmin Edilen (1)
Gerçekten Negatif (0)	TN (Doğru Negatif)	FP (Yalancı Pozitif)
Gerçekten Pozitif (1)	FN (Yalancı Negatif)	TP (Doğru Pozitif)

Tablo 5.4'te model ölçüm kriterleri, tanımları ve formülleri gösterilmektedir.

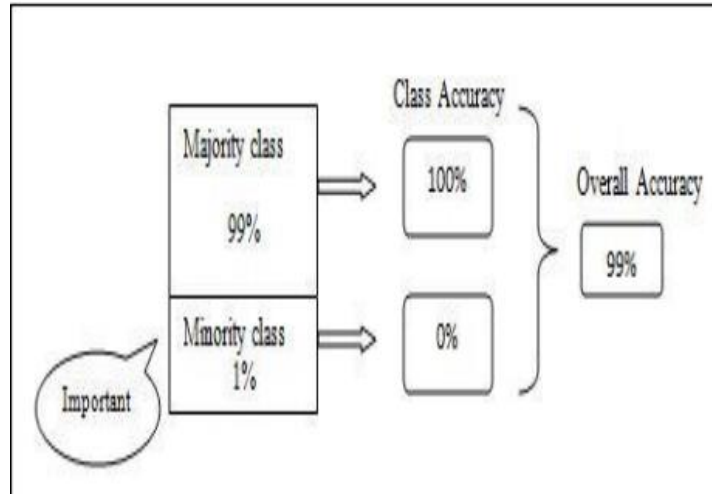
Tablo 5.4: Ölçüm kriterleri

Ölçüm Kriteri	Açıklama	Formül
Doğruluk (Accuracy)	Sınıflandırma modelinin genel başarısıdır.	$TP+TN / TP+TN+FP+FN$
Duyarlılık (Recall)	Pozitif olanları doğru tespit edebilme ölçütüdür.	$TP / TP+FN$
Kesinlik (Precision)	Pozitif olanların ne kadar doğru olduğunu belirleyen ölçüttür.	$TP / TP + FP$
F1 Skor (F1-Score)	Duyarlılık ve kesinliğin harmonik ortalamasıdır.	$2*(Kesinlik*Duyarlılık/Kesinlik+Duyarlılık)$

Modeller değerlendirilirken özellikle dengesiz veri kümelerinde sadece doğruluk ölçütünün baz alınması yanlıştır.

Şekil 5.24'te dengesiz veri setlerinde bulunan problemlerden biri gösterilmektedir.

Şekil 5.24: Dengesiz verilerde doğruluk ölçütü



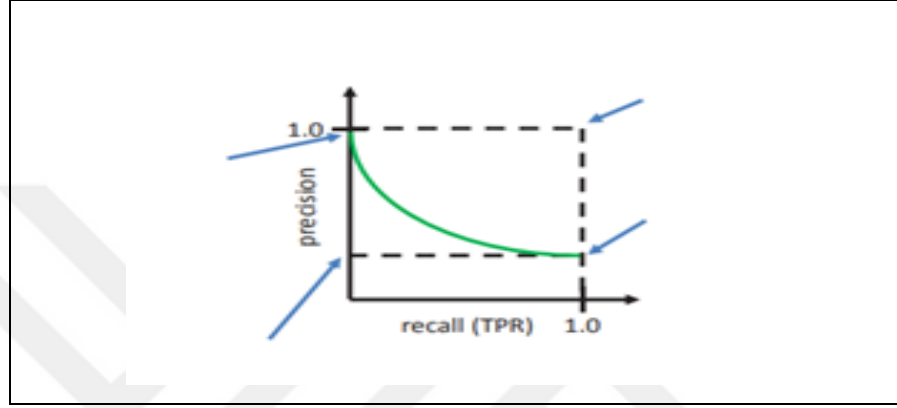
Kaynak: Shakeel vd., 2017

Popülasyonu fazla olan sınıftan kaynaklı olarak yüksek bir doğruluk değeri elde edilir, ancak bu durum modelin azınlık sınıfının tespit edilmesinde aslında başarısız olduğunun

bir göstergesi niteliğindedir. Bu sebeple, örneklem dengelemesi yapıldıktan sonra (örneğin SMOTE ile) duyarlılık, kesinlik ve F1 skorları da değerlendirilmelidir (Guo vd., 2008). Duyarlılık ve kesinlik arasında ters bir orantı (trade-off) vardır. Yani bu iki ölçütten genelde birinin seçilerek ödünleşim yapılması gerekir (Kallada, 2016).

Şekil 5.25'te duyarlılık – kesinlik ödünleşimi gösterilmektedir.

Şekil 5.25: Duyarlılık-kesinlik ilişkisi



Kaynak: Wu, 2017

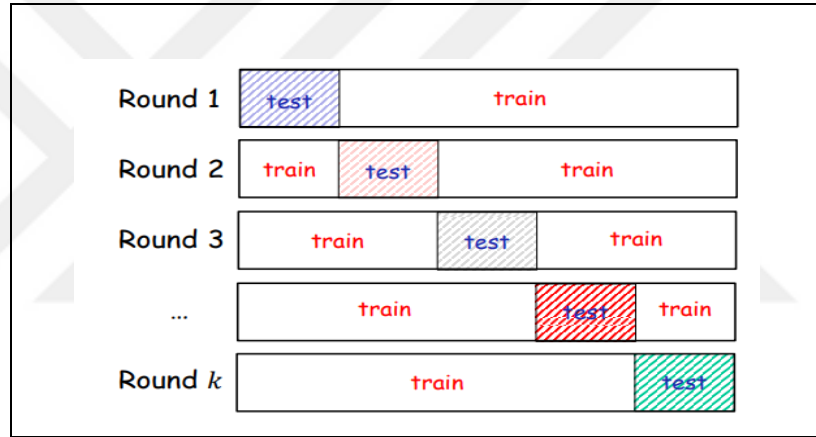
İkisinden hangisinin seçileceği durumlar araştırma probleminin konusuyla ilgilidir. Gerçekten kanser olan bir hastanın istatistiksel model tarafından hasta olmadığı şeklinde sınıflandırılması hastanın ölümüne sebep olacağından bu büyük bir sorundur. Bu sorunun yaşanmaması için yüksek duyarlılık değerleri öncelikli olarak değerlendirilir. Aynı şekilde, dolandırıcılık hareketi olan bir kredi kartı işleminin olağan bir işlem olarak değerlendirilmesi müşterinin memnuniyetsizliğine ve para kaybına sebep olacağından ciddi bir durumdur. Yalancı pozitiflik durumunun getirdiği maliyetin yüksek olduğu durumlarda ise kesinlik ölçütü tercih edilebilir (Shung, 2018). Sigorta tazminat taleplerinde kimin tazminat talep edeceğinin belirlenememesi ciddi bir müşteri memnuniyetsizliğidir. Aynı zamanda, müşterinin finansal veya manevi kayıplarıyla sonuçlanabilir. Tazminat talep eden birinin etmemiş olarak sınıflandırılması, aslında tazminat talebi yapmamış birinin talep yapmış gibi sınıflandırılmasına oranla daha yüksek bir maliyet içermektedir. Bu sebeple duyarlılık değerlerinin yüksekliği daha fazla önem arz etmektedir.

5.8 K KATLAMALI ÇAPRAZ DOĞRULAMA

K-Katlamalı Çapraz Doğrulama (K-Fold Cross Validation) ile veri seti seçilen k tane eşit gruba bölünmektedir. K değeri genelde 5 ya da 10 olarak belirlenmektedir (Neale 2019; Osuna 2005). Bölünen gruptan 1 tanesi test seti geride kalan $k-1$ tanesi ise eğitim verisi olarak kabul edilir. Bu işlem tüm test setleri k kez tekrarlanır. Model k değeri kaç kez seçilmişse o kadar eğitilmiş olur. Eğitim sürecinin tamamlanmasının ardından test setleri için her bir metriğin ortalaması alınmaktadır. Çok hızlı bir yöntem olmadığı için çok büyük veri setlerinde kullanılması uygun olmayabilir (Şirin, 2017).

Şekil 5.26'da K Katlamalı Çapraz Doğrulamanın akış süreci gösterilmektedir.

Şekil 5.26: K katlamalı çapraz doğrulama



Kaynak: Balcan, 2018

Her bir adımın sonunda elde edilen performans kriterlerinin en son aşamada ortalaması alınarak model başarısının artırılması amaçlanmaktadır.

Python'un Scikit-Learn kütüphanesinde bulunan `cross_val` fonksiyonu kullanılarak çapraz doğrulama yapılmıştır.

5.9 ENCODING TEKNİKLERİ

5.9.1 Kukla Değişken

Kategorik değişkenlerdeki değerlerin ayrıştırılmasında kullanılan bir tekniktir. Örneğin, cinsiyet verisinin yer aldığı bir veri setinde erkek cinsiyeti 1, kadın cinsiyeti 0 olarak

kodlanabilir (Schnatter, 2012). Bu şekilde ikili (binary) bir deęişken oluşturularak cinsiyet deęeri anlaşılır. Bu alıřmada, sigorta tazminat talep edenler 1 (evet), tazminat talep etmeyenler 0 (hayır) olarak kukla deęişken (dummy variable) yöntemiyle ayrıştırılabilmektedir. Ancak, bir deęişkende ikiden fazla seviye deęer olduęunda her bir deęer ayrı bir deęişkene dönüřtürölür. Örnek olarak, veri setinde destinasyon deęişkeninde 6 adet kıta bulunmaktadır. Her bir kıta ayrı bir deęişkene dönüřtürölür ve o sütunda 1 yazanlar o kıtaya seyahat etmiř kiřilerdir. Deęişkenlerde yer alan deęerler arasında bir seviye farkı bulunmadıęı için Label Encoding teknięi tercih edilmemiřtir.

5.9.2 Hedef Ortalama

Hedef ortalama (target mean), kategorik deęişkenlerin baęımlı deęişkene göre nümerik hale dönüřtürölebildięi tekniktir. Hedef deęişken kategorikse olasılık hesaplaması ile baęımsız kategorik deęişkenler de nümerik hale dönüřtürölebilmektedir (Halford, 2018). Sigorta tazminat talep durumu kategorik bir deęişkendir. Veri setinde yer alan baęımlı kategorik deęişkenlerin deęerleri hedef deęişkene göre olasılıkları hesaplanarak nümerik hale dönüřtürölerek istatistiksel modelin tamamen rakamlardan oluřan bir veri setiyle tahmin yapması mümkün hale getirilebilmektedir.

řekil 5.27’de hedef ortalama encoding teknięine göre kategorik deęişkenlerde (Acenta Kodu, Tipi, Daęıtım Kanalı gibi) meydana gelen deęişiklikler gösterilmektedir.

Şekil 5.27: Hedef ortalama

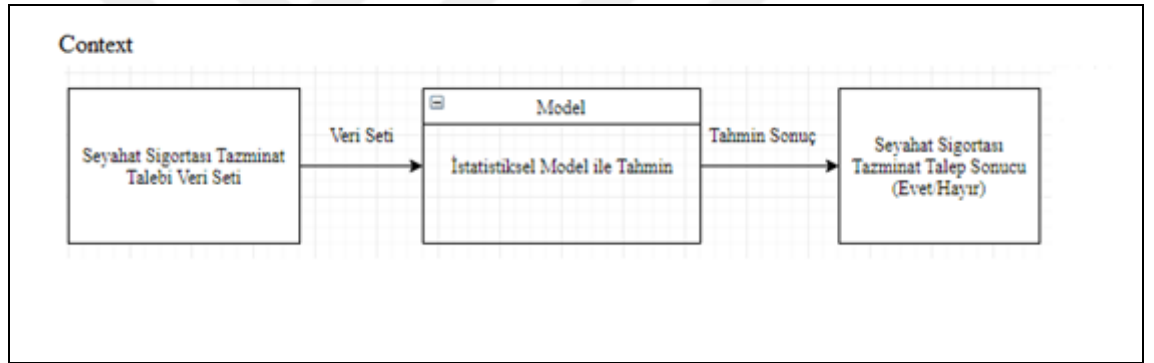
Index	Acenta_Kodu	Acenta_Tipi	Dagitim_Kanali	Urun_Adi	Sure	Destinasyon	Net_Satis	Komisyon	Yas	Sigorta_Talep
0	0.0134453	0.00737273	0.0152037	0.0141439	186	0.0153251	-29	9.57	81	0
1	0.0134453	0.00737273	0.0152037	0.0141439	186	0.0153251	-29	9.57	71	0
2	0.0101792	0.00737273	0.0146258	0.0101792	65	0.010138	-49.5	29.7	32	0
3	0.0101792	0.00737273	0.0146258	0.0101792	60	0.010138	-39.6	23.76	32	0
4	0.0101792	0.00737273	0.0146258	0.0101792	79	0.0124575	-19.8	11.88	41	0
5	0.0053389	0.03353	0.0146258	0.0077584	66	0.0141828	-121	42.35	44	0
6	0.0101792	0.00737273	0.0146258	0.0101792	47	0.0141828	-39.6	23.76	32	0
7	0.0101792	0.00737273	0.0146258	0.0101792	63	0.010138	-108.9	65.34	29	0
8	0.0101792	0.00737273	0.0146258	0.0101792	57	0.0153251	-19.8	11.88	44	0
9	0.0101792	0.00737273	0.0146258	0.0101792	186	0.010138	-99	59.4	37	0
10	0.0053389	0.03353	0.0146258	0.00474806	33	0.0153251	-26	9.1	38.7384	0
11	0.0053389	0.03353	0.0146258	0.00474806	1	0.0153251	-18	6.3	47	0
12	0.0179069	0.00737273	0.0146258	0.0149627	53	0.0124575	-130	49.4	48	0
13	0.0101792	0.00737273	0.0146258	0.0101792	5	0.0153251	-39.6	23.76	64	0
14	0.00562951	0.00737273	0.0146258	0.0108777	39	0.0153251	-25	0	36	0
15	0.0101792	0.00737273	0.0146258	0.0101792	6	0.0124575	-19.8	11.88	53	0
16	0.0101792	0.00737273	0.0146258	0.0101792	48	0.0124575	-79.2	47.52	43	0
17	0.0101792	0.00737273	0.0146258	0.0101792	11	0.0141828	-29.7	17.82	58	0
18	0.0053389	0.03353	0.0146258	0.00474806	3	0.0153251	-18	6.3	47	0
19	0.0101792	0.00737273	0.0146258	0.0101792	14	0.0153251	-69.3	41.58	37	0
20	0.0101792	0.00737273	0.0146258	0.0101792	136	0.0124575	-108.9	65.34	32	0

6. BULGULAR

Veri temizliği ve ilgili encoding tekniğinin uygulanması sonrasında veri seti eğitim ve test olmak üzere iki parçaya ayrılmaktadır. Eğitim verisi, istatistiksel modelin eğitilmesinde kullanılmaktadır. Test veri seti ise eğitilen modelin başarısının sınıandığı veri grubudur. Model eğitilerek test veri grubu üzerinde tahminler yapılmaktadır. Tüm veri setinin yüzde 80'i eğitim yüzde 20'si test olarak ayrılmıştır. Literatürde genelde 80-20 veya 70-30 ayrımı tercih edilmektedir. SMOTE uygulandıktan sonra eğitim verisinde yer alan evet ve hayır sayıları 49913 olarak eşitlenmiştir. Karmaşıklık matrisleri çapraz doğrulama sonrası sonuçları ifade etmemektedir.

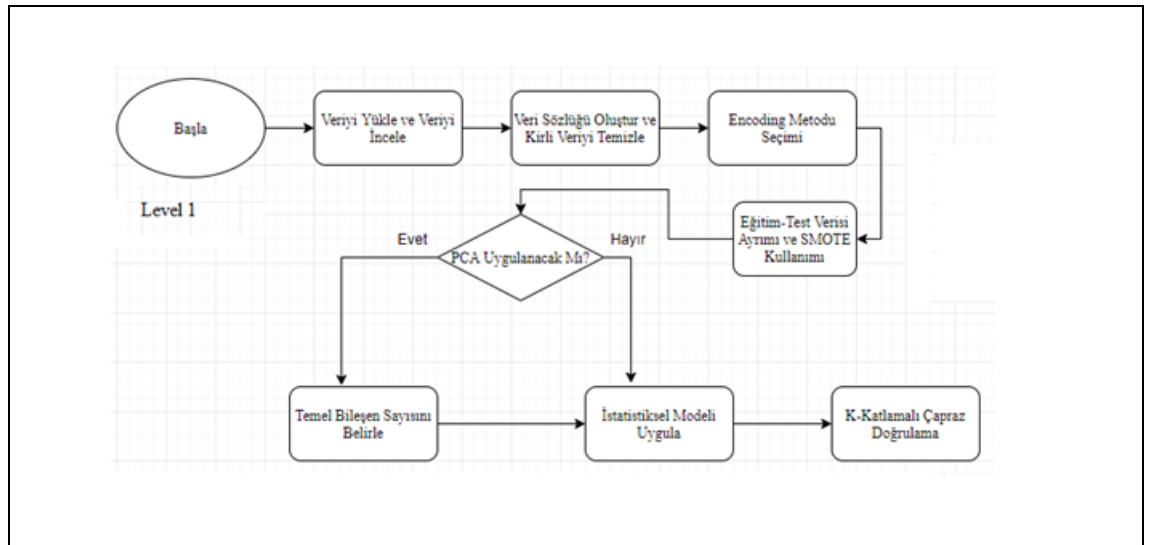
Şekil 6.1'de içerik (context) diyagram gösterilmektedir.

Şekil 6.1: İçerik diyagramı



Şekil 6.2'de genel akış diyagramı yer almaktadır.

Şekil 6.2: Genel akış diyagramı



6.1 LOJİSTİK REGRESYON SONUÇLARI

Tablo 6.1’de model no 1 için karmaşıklık matrisi gösterilmektedir.

Tablo 6.1: LR model 1 için karmaşıklık matrisi

Gerçek Değer	Tazminat Talep Etmeyen (Tahmin)	Tazminat Talep Eden (Tahmin)
Tazminat Talep Etmeyen	9991	2495
Tazminat Talep Eden	49	131

Tablo 6.2’de TBA uygulanmayan ve kukla değişken encoding yöntemi (50+ sayıda değişken elde edilerek) kullanılarak yapılan LR modeli sonuçları yer almaktadır.

Tablo 6.2: LR model 1 sonuçları

Model No:1 TBA: Kullanılmadı Encoding: Kukla Değişken (50+ sayıda)	
Ölçümler	Çapraz Doğrulama (k=10)
Doğruluk: 0.80	Doğruluk: 0.75
Duyarlılık: 0.73	Duyarlılık: 0.70
Kesinlik: 0.05	Kesinlik: 0.78
F1: 0.09	F1: 0.74

K Katlamalı Çapraz Doğrulama öncesinde 0.80 doğruluk ve 0.73 duyarlılık elde edilmiştir. Çapraz Doğrulama sonrası modelin kesinliğinde ve F1 skorunda iyileşme gözlemlenmiştir. Çapraz Doğrulama sonrası doğruluk ve duyarlılık ölçütlerinde düşüş yaşansa da modelin genel performansının iyileştiği görülmektedir.

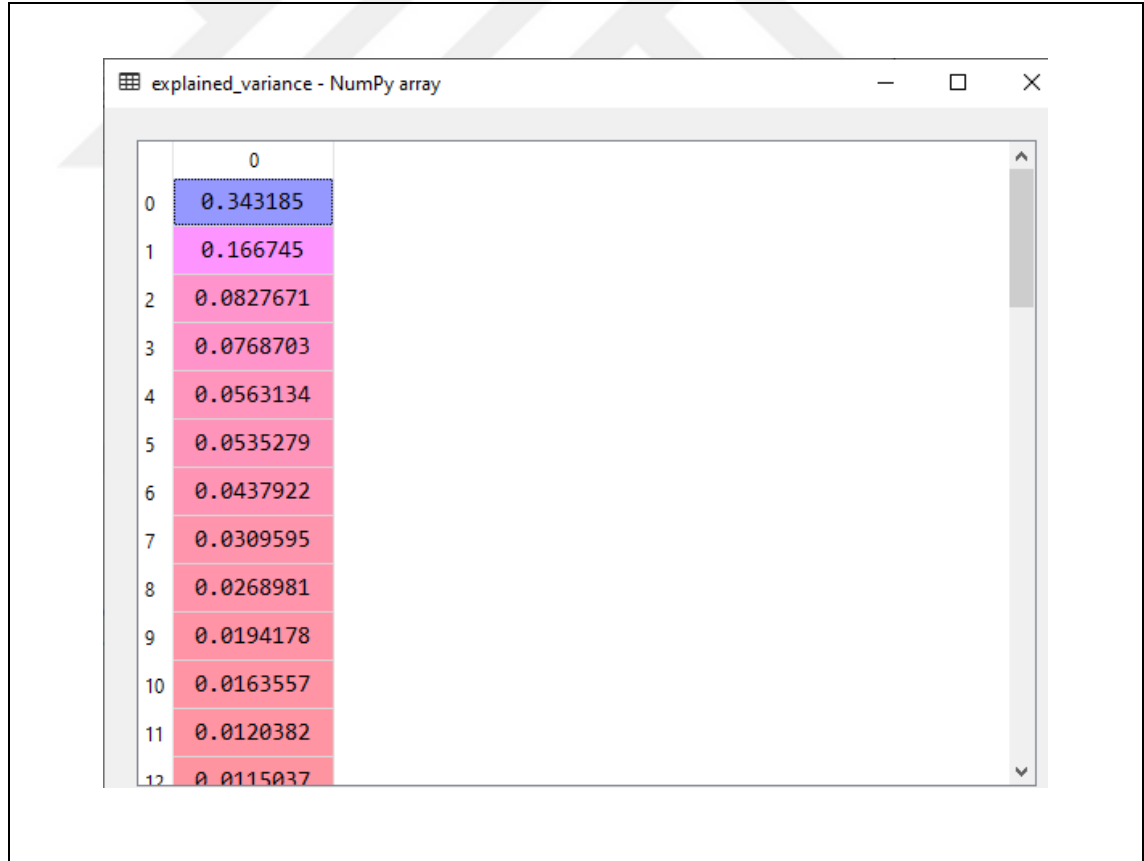
Tablo 6.3’de model 2 için karmaşıklık matrisi gösterilmektedir.

Tablo 6.3: LR model 2 için karmaşıklık matrisi

Gerçek Değer	Tazminat Talep Etmeyen (Tahmin)	Tazminat Talep Eden (Tahmin)
Tazminat Talep Etmeyen	10278	2208
Tazminat Talep Eden	55	125

Şekil 6.3’te 13 temel bileşenin oluşturduğu sanal değişkenlerin varyansları gösterilmiştir.

Şekil 6.3: 13 Temel bileşenin varyansları



Tablo 6.4’de LR modeli TBA ile boyut azaltma yöntemiyle birlikte uygulanmıştır.

Tablo 6.4: LR model 2 sonuçları

Model No:2 TBA: n:13 varyans:92 Encoding: Kukla Değişken (50+ sayıda)	
Ölçümler	Çapraz Doğrulama (k=10)
Doğruluk: 0.82	Doğruluk: 0.75
Duyarlılık: 0.69	Duyarlılık: 0.67
Kesinlik: 0.05	Kesinlik: 0.79
F1: 0.10	F1: 0.73

Kukla değişken tekniğiyle 50’den fazla olan değişken sayısı 13’e indirilmiştir. 13 temel bileşen sürecin belirsizliğini yaklaşık yüzde 92 oranında açıklamaktadır. Oluşturulan modelin doğruluğu 0.82, duyarlılığı 0.69 olarak elde edilmiştir. Çapraz Doğrulama sonrası modelin kesinliğinde ve F1 skorunda iyileşme gözlemlenmiştir. Çapraz Doğrulama sonrası doğruluk ve duyarlılık ölçütlerinde düşüş yaşansa da modelin genel performansının iyileştiği görülmektedir.

Tablo 6.5’de model 3 için karmaşıklık matrisi gösterilmektedir.

Tablo 6.5: LR model 3 için karmaşıklık matrisi

Gerçek Değer	Tazminat Talep Etmeyen (Tahmin)	Tazminat Talep Eden (Tahmin)
Tazminat Talep Etmeyen	9938	2548
Tazminat Talep Eden	54	126

Tablo 6.6’da LR modeli TBA ile boyut azaltma yöntemiyle birlikte uygulanmıştır.

Tablo 6.6: LR model 3 sonuçları

Model No:3 TBA: n:7 varyans:81 Encoding: Kukla Değişken (50+ sayıda)	
Ölçümler	Çapraz Doğrulama (k=10)
Doğruluk: 0.79	Doğruluk: 0.74
Duyarlılık: 0.70	Duyarlılık: 0.68
Kesinlik: 0.05	Kesinlik: 0.77
F1: 0.09	F1: 0.72

Kukla değişken tekniğiyle 50’den fazla olan değişken sayısı 7’ye indirilmiştir. 7 temel bileşen sürecin belirsizliğini yaklaşık yüzde 81 oranında açıklamaktadır (Şekil 6.3’te 0 ile 6 arası indekslenmiş temel bileşenlerin varyansları toplandığında yaklaşık yüzde 81 olduğu görülmektedir). Oluşturulan modelin doğruluğu 0.79, duyarlılığı 0.70 olarak elde edilmiştir. Çapraz Doğrulama sonrası modelin kesinliğinde ve F1 skorunda iyileşme gözlemlenmiştir. Çapraz Doğrulama sonrası doğruluk ve duyarlılık ölçütlerinde düşüş yaşansa da modelin genel performansının iyileştiği görülmektedir.

Tablo 6.7’de model 4 için karmaşıklık matrisi gösterilmektedir.

Tablo 6.7: LR model 4 için karmaşıklık matrisi

Gerçek Değer	Tazminat Talep Etmeyen (Tahmin)	Tazminat Talep Eden (Tahmin)
Tazminat Talep Etmeyen	10485	2001
Tazminat Talep Eden	61	119

Tablo 6.8’de TBA uygulanmayan ve hedef ortalama encoding yöntemi (9 bağımsız değişken) kullanılarak yapılan LR modeli sonuçları yer almaktadır.

Tablo 6.8: LR model 4 sonuçları

Model No:4 TBA: Kullanılmadı Encoding: Hedef Ortalama (9 değişken)	
Ölçümler	Çapraz Doğrulama (k=10)
Doğruluk: 0.84	Doğruluk: 0.75
Duyarlılık: 0.66	Duyarlılık: 0.65
Kesinlik: 0.06	Kesinlik: 0.81
F1: 0.10	F1: 0.72

K Katlamalı Çapraz Doğrulama öncesinde 0.84 doğruluk ve 0.66 duyarlılık elde edilmiştir. Çapraz Doğrulama sonrası modelin kesinliğinde ve F1 skorunda iyileşme gözlemlenmiştir. Çapraz Doğrulama sonrası doğruluk ve duyarlılık ölçütlerinde düşüş yaşansa da modelin genel performansının iyileştiği görülmektedir.

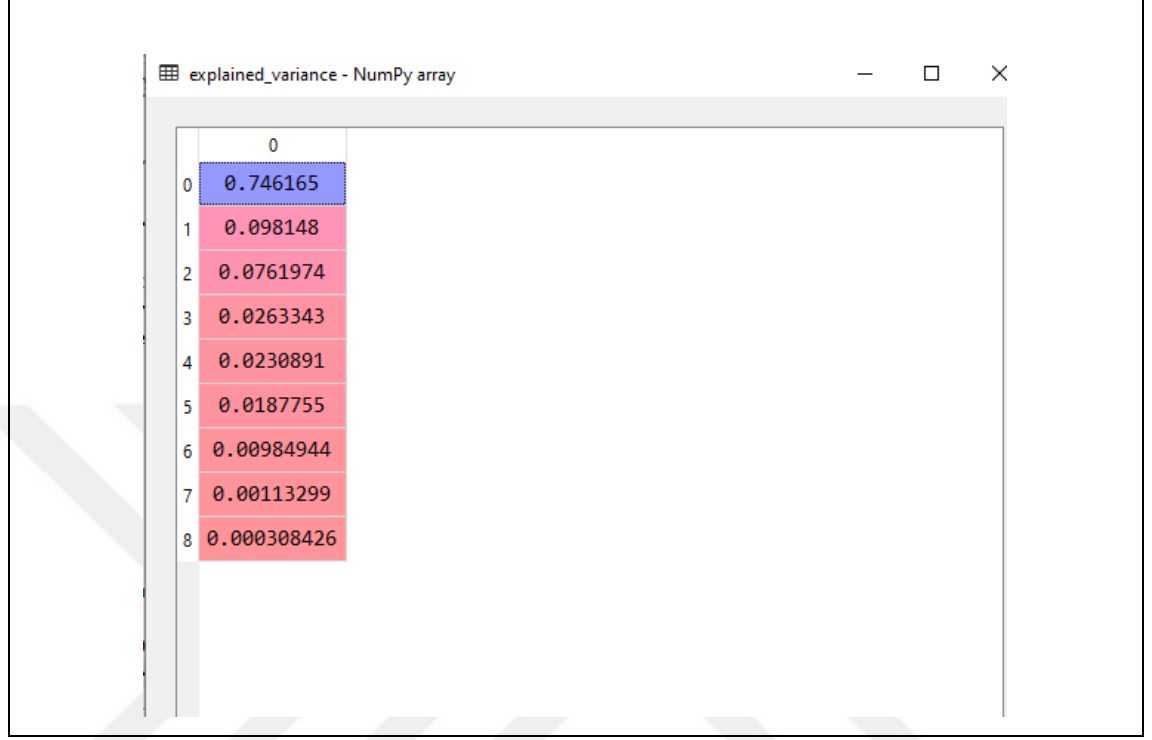
Tablo 6.9’da model 5 için karmaşıklık matrisi gösterilmektedir.

Tablo 6.9: LR model 5 için karmaşıklık matrisi

Gerçek Değer	Tazminat Talep Etmeyen (Tahmin)	Tazminat Talep Eden (Tahmin)
Tazminat Talep Etmeyen	10558	1928
Tazminat Talep Eden	62	118

Şekil 6.4'te 9 temel bileşenin oluşturduğu sanal değişkenlerin varyansları gösterilmiştir.

Şekil 6.4: 9 Temel bileşenin varyansları



Tablo 6.10'da LR modeli TBA ile boyut azaltma yöntemiyle birlikte uygulanmıştır.

Tablo 6.10: LR model 5 sonuçları

Model No:5 TBA: n:6 varyans:97 Encoding: Hedef Ortalama (9 değişken)	
Ölçümler	Çapraz Doğrulama (k=10)
Doğruluk: 0.84	Doğruluk: 0.74
Duyarlılık: 0.66	Duyarlılık: 0.64
Kesinlik: 0.06	Kesinlik: 0.81
F1: 0.11	F1: 0.72

Hedef ortalama encoding yöntemiyle 9 bağımsız değişken sayısı 6 değişkene indirilmiştir. 6 temel bileşen sürecin belirsizliğini yaklaşık yüzde 97 oranında açıklamaktadır (Şekil 6.4'te 0 ile 5 arası indekslenmiş temel bileşenlerin varyansları toplandığında yaklaşık yüzde 97 olduğu görülmektedir). Oluşturulan modelin doğruluğu 0.84, duyarlılığı 0.66 olarak elde edilmiştir. Çapraz Doğrulama sonrası modelin

kesinliğinde ve F1 skorunda iyileşme gözlemlenmiştir. Çapraz Doğrulama sonrası doğruluk ve duyarlılık ölçütlerinde düşüş yaşansa da modelin genel performansının iyileştiği görülmektedir.

Tablo 6.11’de model 6 için karmaşıklık matrisi gösterilmektedir.

Tablo 6.11: LR model 6 için karmaşıklık matrisi

Gerçek Değer	Tazminat Talep Etmeyen (Tahmin)	Tazminat Talep Eden (Tahmin)
Tazminat Talep Etmeyen	10711	1775
Tazminat Talep Eden	64	116

Tablo 6.12’de LR modeli TBA ile boyut azaltma yöntemiyle birlikte uygulanmıştır.

Tablo 6.12: LR model 6 sonuçları

Model No:6 TBA: n:3 varyans:91 Encoding: Hedef Ortalama (9 değişken)	
Ölçümler	Çapraz Doğrulama (k=10)
Doğruluk: 0.85	Doğruluk: 0.74
Duyarlılık: 0.64	Duyarlılık: 0.62
Kesinlik: 0.06	Kesinlik: 0.82
F1: 0.11	F1: 0.71

Hedef ortalama encoding yöntemiyle 9 bağımsız değişken sayısı 3 değişkene indirilmiştir. 3 temel bileşen sürecin belirsizliğini yaklaşık yüzde 91 oranında açıklamaktadır (Şekil 6.4’te 0 ile 2 arası indekslenmiş temel bileşenlerin varyansları toplandığında yaklaşık yüzde 91 olduğu görülmektedir). Oluşturulan modelin doğruluğu 0.85, duyarlılığı 0.64 olarak elde edilmiştir. Çapraz Doğrulama sonrası modelin kesinliğinde ve F1 skorunda iyileşme gözlemlenmiştir. Çapraz Doğrulama sonrası

doğruluk ve duyarlılık ölçütlerinde düşüş yaşansa da modelin genel performansının iyileştiği görülmektedir.

6.2 NAIVE BAYES SONUÇLARI

Kullanılan veri setinde hem kategorik hem sürekli değişkenler bulunmaktadır. Kategorik olan değişkenler kukla değişkene dönüştürülüp 1 ve 0'lar ile ifade edilerek sayısal ayrık değişkenlere dönüştürülebilmektedir. Ancak NB algoritmasının ikili değişkenlere karşı performanslı çalışmaması sebebiyle NB kullanılırken kukla değişken encoding tekniğine yer verilmemiş, sürekli değişkenlerle çalışabilmek adına sadece hedef ortalama tekniği kullanılarak modeller oluşturulmuştur. Bu yüzden GaussianNB kullanılmıştır.

Tablo 6.13'de NB model 1 için karmaşıklık matrisi gösterilmektedir.

Tablo 6.13: NB model 1 için karmaşıklık matrisi

Gerçek Değer	Tazminat Talep Etmeyen (Tahmin)	Tazminat Talep Eden (Tahmin)
Tazminat Talep Etmeyen	10669	1817
Tazminat Talep Eden	67	113

Tablo 6.14'de TBA uygulanmayan ve hedef ortalama encoding yöntemi kullanılarak yapılan NB modeli sonuçları yer almaktadır.

Tablo 6.14: NB model 1 sonuçları

Model No:1 TBA: Kullanılmadı Encoding: Hedef Ortalama (9 değişken)	
Ölçümler	Çapraz Doğrulama (k=10)
Doğruluk: 0.85	Doğruluk: 0.74
Duyarlılık: 0.63	Duyarlılık: 0.62
Kesinlik: 0.06	Kesinlik: 0.81
F1: 0.11	F1: 0.70

K Katlamalı Çapraz Doğrulama öncesinde 0.85 doğruluk ve 0.63 duyarlılık elde edilmiştir. Çapraz Doğrulama sonrası modelin kesinliğinde ve F1 skorunda iyileşme gözlemlenmiştir. Çapraz Doğrulama sonrası doğruluk ve duyarlılık ölçütlerinde düşüş yaşansa da modelin genel performansının iyileştiği görülmektedir.

Tablo 6.15’de NB model 2 için karmaşıklık matrisi gösterilmektedir.

Tablo 6.15: NB model 2 için karmaşıklık matrisi

Gerçek Değer	Tazminat Talep Etmeyen (Tahmin)	Tazminat Talep Eden (Tahmin)
Tazminat Talep Etmeyen	10426	2060
Tazminat Talep Eden	65	115

Tablo 6.16’da NB modeli TBA ile boyut azaltma yöntemiyle birlikte uygulanmıştır.

Tablo 6.16: NB model 2 sonuçları

Model No:2 TBA: n:6 varyans:97 Encoding: Hedef Ortalama (9 değişken)	
Ölçümler	Çapraz Doğrulama (k=10)
Doğruluk: 0.83	Doğruluk: 0.73
Duyarlılık: 0.64	Duyarlılık: 0.62
Kesinlik: 0.05	Kesinlik: 0.79
F1: 0.10	F1: 0.69

Hedef ortalama encoding yöntemiyle 9 bağımsız değişken sayısı 6 değişkene indirilmiştir. 6 temel bileşen sürecin belirsizliğini yaklaşık yüzde 97 oranında açıklamaktadır (Şekil 6.4’te 0 ile 5 arası indekslenmiş temel bileşenlerin varyansları toplandığında yaklaşık yüzde 97 olduğu görülmektedir). Oluşturulan modelin doğruluğu 0.83, duyarlılığı 0.64 olarak elde edilmiştir. Çapraz Doğrulama sonrası modelin kesinliğinde ve F1 skorunda iyileşme gözlemlenmiştir.

Çapraz Doğrulama sonrası doğruluk ve duyarlılık ölçütlerinde düşüş yaşansa da modelin genel performansının iyileştiği görülmektedir.

Tablo 6.17’de NB model 3 için karmaşıklık matrisi gösterilmektedir.

Tablo 6.17: NB model 3 için karmaşıklık matrisi

Gerçek Değer	Tazminat Talep Etmeyen (Tahmin)	Tazminat Talep Eden (Tahmin)
Tazminat Talep Etmeyen	10879	1607
Tazminat Talep Eden	74	106

Tablo 6.18’de NB modeli TBA ile boyut azaltma yöntemiyle birlikte uygulanmıştır.

Tablo 6.18: NB model 3 sonuçları

Model No:3 TBA: n:3 varyans:91 Encoding: Hedef Ortalama (9 değişken)	
Ölçümler	Çapraz Doğrulama (k=10)
Doğruluk: 0.87	Doğruluk: 0.73
Duyarlılık: 0.59	Duyarlılık: 0.58
Kesinlik: 0.06	Kesinlik: 0.82
F1: 0.11	F1: 0.68

Hedef ortalama encoding yöntemiyle 9 bağımsız değişken sayısı 3 değişkene indirilmiştir. 3 temel bileşen sürecin belirsizliğini yaklaşık yüzde 91 oranında açıklamaktadır (Şekil 6.4’te 0 ile 2 arası indekslenmiş temel bileşenlerin varyansları toplandığında yaklaşık yüzde 91 olduğu görülmektedir). Oluşturulan modelin doğruluğu 0.87, duyarlılığı 0.59 olarak elde edilmiştir. Çapraz Doğrulama sonrası modelin kesinliğinde ve F1 skorunda iyileşme gözlemlenmiştir. Çapraz Doğrulama sonrası doğruluk ve duyarlılık ölçütlerinde düşüş yaşansa da modelin genel performansının iyileştiği görülmektedir.

7. TARTIŞMA VE SONUÇ

Bu tez çalışmasında sigorta poliçesi sahiplerinin sigorta şirketlerinden tazminat taleplerinde bulunup bulunmadığına yönelik bir araştırma seyahat sigortası ürünü özelinde açık bir veri seti kullanılarak gerçekleştirilmiştir. Oluşturulan istatistiksel modellerle amaç, şirketlerin önceden aksiyon almalarını sağlayarak hızlı, süreci otomatize edebilen, operasyonel yapısı az ve tazminat süreçlerinde müşteri memnuniyetini sağlayabilecek bir karar destek sistemi oluşturmaktır. Bu zamana kadar yapılan çalışmalar daha çok sigorta dolandırıcılığının araştırılmasında yoğunlaşmıştır. Tazminat talep süreçlerinin kendisinin araştırıldığı çalışmalar daha az sayıdadır. Çalışılan veri setinin niteliği gereği problem bir sigorta dolandırıcılığı problemi olarak ele alınmamıştır.

Bu tez çalışmasına konu olan araştırma problemi denetimli öğrenme teknikleriyle çözümlenmiştir. Lojistik Regresyon ve Naive Bayes modelleri kullanılarak tazminat talep tahminleri gerçekleştirilmiştir. Modelleri kullanırken kukla değişken ve hedef ortalama encoding teknikleri ayrı ayrı senaryolarda kullanılmıştır. Aynı zamanda boyut indirgeme tekniği olan Temel Bileşen Analizi de her iki model için de ayrıca kullanılmıştır. Veri bilimi literatürünün önemli problemlerinden biri olan dengesiz veri kümesi sorunu SMOTE yöntemiyle çözümlenmiştir.

Duyarlılık ve kesinlik ölçütleri arasında ters orantı olması bulunması bizi iki ölçütten birini seçmeye yönlendirmektedir. Bu bağlamda çapraz doğrulama öncesi modellerde genel olarak duyarlılık ölçütünün daha iyi olduğu görülmektedir. Duyarlılık değerinin iyi olması tazminat talebinde bulunmuş kişilerin tespitinde önemlidir. Böylece şirket tarafından gerçekten talepte bulunanlar tespit edilerek asıl amaca ulaşılabilir. Modellerin dezavantajlı noktalarının çapraz doğrulama aşaması öncesinde kesinlik değerini iyi tahmin edememesi olduğu görülmektedir. Bunun sebepleri olarak kullanılan modellerin kendisi veya veri setinin yapısı gösterilebilir. Kesinlik değerinin ilk başta düşük olmasının sebebi tazminat talep etmediği halde talep etti şeklinde yapılan sınıflandırma sayısının biraz daha fazla olmasıdır. Bu da talepte bulunmamış kişilerle de sigortacı tarafından iletişim kurulmasını gerektirebilir.

Her ne kadar bu durum süreçleri biraz etkileyecek olsa da tazminat talep ettiği halde etmedi şeklinde sınıflandırılan bir senaryoya nazaran (duyarlılığı düşük olan) maliyeti daha düşüktür. Müşteri memnuniyetsizliğinin oluşmasının önüne geçilmelidir.

Riskin minimize edilmesi adına modellerin performansları çapraz doğrulama yapılarak artırılmış, bunun da modellerde kesinlik ve F1 değerlerini oldukça pozitif etkilediği görülmüştür. Lojistik Regresyon'un Naive Bayes'e göre daha başarılı sonuçlar verdiği görülmektedir. Nihayetinde, kurulan istatistiksel modellerin üst yönetimlerin stratejik karar alma süreçlerinde yardımcı olabileceği anlaşılmaktadır.

İlerleyen zamanlarda yapılacak çalışmalarda dengesiz veri setlerinin çözülmesinde AdaBoost gibi farklı teknikler kullanılabilir. Mümkünse daha fazla veri ve daha fazla değişken toplanarak ilerlenebilir. Ayrıca, Yapay Sinir Ağları ya da Rastgele Orman gibi ağaç bazlı algoritma modelleri kullanılarak tahmin yapılabilir. Problem tazminat talep durumu bağımlı değişkeninin kaldırılmasıyla bir denetimsiz öğrenme problemine dönüştürülebilir ve kümeleme algoritmaları kullanılarak anomali tespiti yapılabilir.

Kurulmuş olan mevcut sistem kimlerin tazminat talep edip etmediğine yönelik bir tahminle sınırlı olduğundan, bu sistemin arkasına bir dolandırıcılık tespit sistemi entegre edilebilirse tazminat taleplerinin sonuçlarının daha doğru değerlendirilmesi mümkün olacaktır. Bunun için dolandırıcılık aktivitesinin tespitini sağlayacak daha net değişkenlere ihtiyaç vardır. Bu sağlanabilirse sigorta tazminat talep süreci ve dolandırıcılık tespit modeli eş zamanlı olarak çalıştırılır ve çok daha etkin sonuçlar elde edilir. Örnek olarak, Kuzey Amerika'ya yapılan seyahatlerde bir anomali görülebilirse sigorta şirketi o bölgeye yapılan seyahatler için farklı aksiyonlar alarak ilerleyebilir.

KAYNAKÇA

Sürekli Yayınlar

- Ahmed, S. R., 2004. Applications of Data Mining in Retail Business. Las Vegas, NV, USA, USA, International Conference on Information Technology: Coding and Computing, 2004. Proceedings. ITCC 2004.
- Akpınar, Ö., 2018. Sigorta Sektöründe Veri Madenciliği ve Kullanım Alanları. Dumlupınar Üniversitesi Sosyal Bilimler Dergisi, Issue 57, pp. 103-119.
- Ali, A., Shamsuddin, S. M. & Ralescu, A., 2015. Classification with class imbalance problem: A Review. Int. J. Advance Soft Compu. Appl, 7(3).
- Amin, A. ve diğerleri, 2016. Comparing Oversampling Techniques to Handle the Class Imbalance Problem: A Customer Churn Prediction Case Study. IEEE Access, Cilt 4, pp. 7940 - 7957.
- Antwi, S. & Zhao, X., 2012. A logistic regression model for Ghana National Health Insurance claims. International Journal of Business and Social Research (IJBSR), 2(7), pp. 139-147.
- Aydoğan, N. & Akpınar, Ö., 2016. Sigorta Sektöründe Kara Para Aklama: FATF Raporları ve Aklama Tipolojileri. Marmara Üniversitesi İktisadi ve İdari Bilimler Dergisi, 38(1), pp. 31-55.
- Brennan, P., 2012. A comprehensive survey of methods for overcoming the class imbalance problem in fraud detection. Yüksek Lisans Tezi.
- Chawla, N. V., Bowyer, K. W., Hall, L. O. & Kegelmeyer, W. P., 2002. SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research, Cilt 16, p. 321–357.
- Danso, S. O., 2006. An Exploration of Classification Prediction Techniques in Data Mining: The insurance domain. Yüksek Lisans Tezi.
- Elrahman, S. & Abraham, A., 2013. A Review of Class Imbalance Problem. Journal of Network and Innovative Computing, Cilt 1, pp. 332-340.
- Ghuse, N., Pawar, P. & Potgantwar, A., 2017. An Improved Approach For Fraud Detection In Health Insurance Using Data Mining Techniques. Int. J. Sc. Res. in Network Security and Communication, 5(5), pp. 27-32.

- Gosain, A. & Sardana, S., 2017. Handling class imbalance problem using oversampling techniques: A review. Udupi, India, 2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI).
- Guo, X. ve diğeri, 2008. On the Class Imbalance Problem. Jinan, China, 2008 Fourth International Conference on Natural Computation.
- Günbatar, E., 2019. Türkiye'de Otomobil Sigortası Sahtekârlıklarının Makine Öğrenmesi Yöntemleri ile Tespit Edilmesi. Yüksek Lisans Tezi.
- Hall, S., 2017. How Artificial Intelligence Is Changing The Insurance Industry. CIPR Newsletter, Cilt 22, pp. 2-7.
- Huangfu, D., 2015. Data Mining for Car Insurance Claims Prediction. Yüksek Lisans Tezi.
- Jolliffe, I. & Cadima, J., 2015. Principal component analysis:a review and recent developments. Phil. Trans. R. Soc. A, Cilt 374.
- Kaščelan, V., Kaščelan, L. & Burić, M. N., 2016. A nonparametric data mining approach for risk prediction in car insurance: a case study from the Montenegrin market. Economic Research-Ekonomska Istraživanja, 29(1), pp. 545-558.
- Kırlıdoğ, M. & Asuk, C., 2012. A fraud detection approach with data mining in health insurance. Procedia - Social and Behavioral Sciences, Cilt 62, pp. 989-994.
- Kulkarni, R. & Devale, A., 2012. Applications of Data Mining Techniques in Life Insurance. International Journal of Data Mining & Knowledge Management Process (IJDMP), 2(4), pp. 31-40.
- Kumar, M., Ghani, R. & Mei, Z., 2010. Data mining to predict and prevent errors in health insurance claims processing. Washington, DC, USA, The 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 65-74.
- Longadge, R., Dongre, S. & Latesh, M., 2013. Class Imbalance Problem in Data Mining: Review. International Journal of Computer Science and Network (IJCSN), 2(1).
- Mambo, S. R., 2019. Use of Data Mining to Detect Fraud Health Insurance Claims. Yüksek Lisans Tezi.
- Millican, M., Zhang, L. & Kimball, D., 2017. Predicting Insurance Claims in Brazil. basım yeri bilinmiyor:CS 229 Final Report.

- Moon, H., Pu, Y. & Ceglia, C., 2019. A Predictive Modeling for Detecting Fraudulent Automobile Insurance Claims. *Theoretical Economics Letters*, Cilt 9, pp. 1886-1900.
- Muslu, D., 2009. Sigortacılık Sektöründe Risk Analizi: Veri Madenciliği Uygulaması. Yüksek Lisans Tezi.
- Nayak, S. C., Misra, B. B. & Behera, H., 2014. Impact of Data Normalization on Stock Index Forecasting. *International Journal of Computer Information Systems and Industrial Management Applications*, Cilt 6, pp. 257-269.
- Perera, B. K., 2013. A Class Imbalance Learning Approach to Fraud Detection in Online Advertising. Yüksek Lisans Tezi.
- Pozzolo, A. D., 2010. Comparison of Data Mining Techniques for Insurance Claim Prediction. Yüksek Lisans Tezi.
- Rahman, M. & Davis, D., 2013. Addressing the Class Imbalance Problem in Medical Datasets. *International Journal of Machine Learning and Computing*, 3(2), pp. 225-228.
- Shakeel, F., Sabhitha, A. & Sharma, S., 2017. Exploratory Review on Class Imbalance Problem: An Overview. Delhi, India, 2017 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT).
- Song, F., Guo, Z. & Mei, D., 2010. Feature selection using principal component analysis. Yichang, China, 2010 International Conference on System Science, Engineering Design and Manufacturing Informatization.
- Yıldırım, İ., 2013. Türk Sigortacılık Sektörünün Yumuşak Karnı: Sigorta Suistimalleri Sorunu. *Sosyal ve Beşeri Bilimler Dergisi*, 5(1), pp. 331-340.
- Yunos, Z. M. ve diğerleri, 2016. Comparative Performance For Predictive Modelling In Motor Insurance Claims. *Jurnal Teknologi*, Cilt 78, pp. 29-34.
- Zhang, L. & Wang, W., 2011. A Re-sampling Method for Class Imbalance Learning with Credit Data. Nanjing, Jiangsu, China, 2011 International Conference of Information Technology, Computer Engineering and Management Sciences, pp. 393-397.

Diğer Yayınlar

Allianz Travel Insurance, 2020. Required Documentation Checklist. [Çevrimiçi]

Available at: <https://www.allianztravelinsurance.com/claims/required-documents>
[Erişildi: 23 Şubat 2020].

Anadolu Sigorta, 2019. Yurtdışı Seyahat Sigortası: Anadolu Sigorta. [Çevrimiçi]

Available at: <https://www.anadolusigorta.com.tr/tr/seyahat-ve-ferdi-kaza/yurt-disi-seyahat-sigortasi> [Erişildi: 23 Şubat 2020].

Anon., tarih yok Nominal Association: Phi and Cramer's V. [Çevrimiçi]

Available at:

<http://www.people.vcu.edu/~pdattalo/702SuppRead/MeasAssoc/NominalAssoc.html> [Erişildi: 1 Mart 2020].

Anon., tarih yok Open Database License (ODbL) v1.0. [Çevrimiçi]

Available at: <https://opendatacommons.org/licenses/odbl/1.0/>
[Erişildi: 2 Haziran 2019].

Axa Sigorta, 2018. Yurtdışı Seyahat Sağlık Sigortası. [Çevrimiçi]

Available at: <https://www.axasigorta.com.tr/yurt-disi-seyahat-saglik-sigortasi>
[Erişildi: 23 Şubat 2020].

Axa Sigorta, 2019. Making a travel insurance claim. [Çevrimiçi]

Available at: <https://www.axa.co.uk/travel-insurance/claims/>
[Erişildi: 23 Şubat 2020].

Balcan, M. F., 2018. k-Fold Cross-Validation. [Çevrimiçi]

Available at: <http://www.cs.cmu.edu/~10715-f18/lectures/k-fold.pdf>
[Erişildi: 19 Kasım 2019].

Balcan, M. F., 2018. Machine Learning 10-401. [Çevrimiçi]

Available at: <http://www.cs.cmu.edu/~ninamf/courses/401sp18/lectures/intro-ml-401-01-17.pdf>
[Erişildi: 10 Haziran 2019].

Boyle, T., 2019. Dealing with Imbalanced Data. A guide to effectively handling imbalanced datasets in Python. [Çevrimiçi]

Available at: <https://towardsdatascience.com/methods-for-dealing-with->

imbalanced-data-5b761be45a18

[Erişildi: 8 Eylül 2019].

Ching, C., 2020. Logistic regression theory for practitioners. [Çevrimiçi]

Available at: <https://towardsdatascience.com/the-data-scientists-field-guide-to-logistic-regression-part-1-intuition-97084b11bd68> [Erişildi: 1 Mart 2020].

Collins, R., 2010. Principal Components Analysis. [Çevrimiçi]

Available at:

<http://www.cse.psu.edu/~rtc12/CSE586Spring2010/lectures/pcaLectureShort.pdf>
[Erişildi: 7 Aralık 2019].

D'Ambrosio, R., 2019. Technology Has Only Begun to Impact Travel Insurance.

[Çevrimiçi]

Available at: <https://www.travelmarketreport.com/articles/Technology-Has-Only-Begun-to-Impact-Travel-Insurance> [Erişildi: 29 Aralık 2019].

E., 2018. Tasting Python Machine Learning : Insurance Claim Prediction. [Çevrimiçi]

Available at: <https://medium.com/@easonlai888/tasting-python-machine-learning-insurance-claim-prediction-8775cd0aa926> [Erişildi: 1 Mart 2020].

Fiba Emeklilik, 2012. Seyahat Sağlık Sigortası. [Çevrimiçi]

Available at: <https://www.fibaemeklilik.com.tr/saglik-sigortasi/seyahat-saglik-sigortasi> [Erişildi: 23 Şubat 2020].

Gulf Sigorta, 2017. What are the things that must be done in case of Travel claim?.

[Çevrimiçi]

Available at: <https://www.gulfsigorta.com.tr/pdf/incoming-seyahat-bolgo-beyan-formu-engpdf-7700.pdf> [Erişildi: 23 Şubat 2020].

Gupta, P., 2017. Naive Bayes in Machine Learning. [Çevrimiçi]

Available at: <https://towardsdatascience.com/naive-bayes-in-machine-learning-f49cc8f831b4> [Erişildi: 16 Ağustos 2019].

Halford, M., 2018. Target encoding done the right way. [Çevrimiçi]

Available at: <https://maxhalford.github.io/blog/target-encoding/>
[Erişildi: 19 Ağustos 2019].

Hapon, M., 2019. 6 Ways Machine Learning is Changing Insurance. [Çevrimiçi]

Available at: <https://www.netguru.com/blog/machine-learning-insurance-insurtech-fintech-underwriting> [Erişildi: 29 Aralık 2019].

Intrator, N., 2010. Machine Learning: Foundations. [Çevrimiçi]

Available at: <https://www.slideshare.net/butest/machine-learning-foundations-course-number-0368403401-3858837> [Erişildi: 10 Haziran 2019].

Kallada, M., 2016. No Free Lunch Theorem, Better Model Comparison, Ensemble Learning. [Çevrimiçi]

Available at: <https://web.cs.dal.ca/~kallada/stat2450/lectures/Lecture13.pdf> [Erişildi: 24 Ağustos 2019].

Kartal, E. & Özen, Z., 2018. Dengesiz Veri Setlerinde Sınıflandırma. [Çevrimiçi]

Available at: <http://www.elektrik.gen.tr/wp-content/uploads/2018/01/8.-Dengesiz-Veri-Setlerinde-S%C4%B1n%C4%B1fland%C4%B1rma.pdf> [Erişildi: 14 Mart 2020].

Korjus, K., 2014. PRINCIPAL COMPONENT ANALYSIS. [Çevrimiçi]

Available at: https://courses.cs.ut.ee/MTAT.03.227/2014_spring/uploads/Main/lecture-notes-9.pdf [Erişildi: 7 Aralık 2019].

Malhotra, R. & Sharma, S., 2018. MACHINE LEARNING IN INSURANCE. [Çevrimiçi]

Available at: https://www.accenture.com/_acnmedia/pdf-84/accenture-machine-learning-insurance.pdf [Erişildi: 29 Aralık 2019].

Mapfre Sigorta, 2020. Make a claim on your travel insurance. [Çevrimiçi]

Available at: <https://www.mapfreinsurance.com/en/claims/travel/> [Erişildi: 23 Şubat 2020].

Nasrudin, Z., 2019. [Çevrimiçi]

Available at: <https://www.kaggle.com/mhdzahier/travel-insurance> [Erişildi: 2 Haziran 2019].

Neale, C., 2019. Cross Validation: A Beginner's Guide. [Çevrimiçi]

Available at: <https://towardsdatascience.com/cross-validation-a-beginners-guide-5b8ca04962cd> [Erişildi: 16 Kasım 2019].

Orhan, U., 2017. [Çevrimiçi]

Available at: <https://bmb.cu.edu.tr/uorhan/DersNotu/ML-01.pdf> [Erişildi: 13 Haziran 2019].

- Osuna, R. G., 2005. LECTURE 13: Cross-validation. [Çevrimiçi]
Available at: https://www.cs.tau.ac.il/~nin/Courses/NC05/pr_113.pdf
[Erişildi: 16 Kasım 2019].
- Paul, M., 2018. What is Machine Learning? INFO-4604, Applied Machine Learning
University of Colorado Boulder. [Çevrimiçi]
Available at: https://cmci.colorado.edu/classes/INFO-4604/files/slides-1_intro.pdf
[Erişildi: 15 Haziran 2019].
- Raschka, S., 2014. About Feature Scaling and Normalization and the effect of
standardization for machine learning algorithms. [Çevrimiçi]
Available at:
https://sebastianraschka.com/Articles/2014_about_feature_scaling.html
[Erişildi: 3 Aralık 2019].
- Raschka, S., 2018. STAT 479: Machine Learning Lecture Notes. [Çevrimiçi]
Available at: https://sebastianraschka.com/pdf/lecture-notes/stat479fs18/01_ml-overview_notes.pdf
[Erişildi: 10 Haziran 2019].
- Reese, D., 2019. AI Trends in Insurance for 2020 and Beyond. [Çevrimiçi]
Available at: <https://becominghuman.ai/ai-trends-in-insurance-for-2020-and-beyond-ac3b461f6042> [Erişildi: 29 Aralık 2019].
- Roman, V., 2018. Machine Learning Introduction: A Comprehensive Guide.
[Çevrimiçi]
Available at: <https://towardsdatascience.com/machine-learning-introduction-a-comprehensive-guide-af6712cf68a3> [Erişildi: 10 Haziran 2019].
- Roman, V., 2019. Naive Bayes Algorithm: Intuition and Implementation in a Spam
Detector. [Çevrimiçi]
Available at: <https://towardsdatascience.com/naive-bayes-intuition-and-implementation-ac328f9c9718> [Erişildi: 16 Ağustos 2019].
- Schnatter, S., 2012. Dummy Variables. [Çevrimiçi]
Available at:
[http://statmath.wu.ac.at/~fruehwirth/Oekonometrie_I/Folien Econometrics I teil6.pdf](http://statmath.wu.ac.at/~fruehwirth/Oekonometrie_I/Folien_Econometrics_I_teil6.pdf) [Erişildi: 12 Ağustos 2019].

Shroff, R., 2019. How Are Insurance Companies Implementing Artificial Intelligence (AI)?. [Çevrimiçi]

Available at: <https://towardsdatascience.com/how-are-insurance-companies-implementing-artificial-intelligence-ai-aaf845fce6a7>

[Erişildi: 29 Aralık 2019].

Shung, K. P., 2018. Accuracy, Precision, Recall or F1?. [Çevrimiçi]

Available at: <https://towardsdatascience.com/accuracy-precision-recall-or-f1-331fb37c5cb9> [Erişildi: 24 Ağustos 2019].

Sompo Sigorta, 2020. SEYAHAT SAĞLIK SİGORTASI TAZMİNAT TALEP FORMU. [Çevrimiçi]

Available at: <https://www.somposigorta.com.tr/uploads/file/67dbbecb-123b-4a7c-90a6-e3cde9d1fdde.pdf> [Erişildi: 23 Şubat 2020].

Şirin, E., 2017. Bir Bakışta K-Fold Cross Validation. [Çevrimiçi]

Available at: <https://www.veribilimiokulu.com/bir-bakista-k-fold-cross-validation/> [Erişildi: 16 Kasım 2019].

Türkiye Sigorta Birliği, 2013. Sigorta Branşları: Türkiye Sigorta Birliği. [Çevrimiçi]

Available at: <https://www.tsb.org.tr/sigorta-branslari.aspx?pageID=622> [Erişildi: 23 Şubat 2020].

Türkiye Sigorta Birliği, 2013. Sigorta Tanımları: Türkiye Sigorta Birliği. [Çevrimiçi]

Available at: <https://www.tsb.org.tr/sigorta-tanimlari.aspx?pageID=648> [Erişildi: 23 Şubat 2020].

Unico Sigorta, 2018. Yurtdışı Seyahat Sigortası Nedir?. [Çevrimiçi]

Available at: <https://www.unicosigorta.com.tr/blog/yurtdisi-seyahat-sigortasi-nedir> [Erişildi: 23 Şubat 2020].

Visitor Insurance Reviews, 2015. The Visitors Insurance Claims Process. [Çevrimiçi]

Available at: <https://www.visitorsinsurancereviews.com/visitors-insurance-claims-process> [Erişildi: 10 Mart 2020].

Wu, J., 2017. Advanced Evaluation, Imbalanced Data. [Çevrimiçi]

Available at:

[https://www.cs.hmc.edu/~yjw/teaching/cs158/lectures/10_EvaluationImbalance.p](https://www.cs.hmc.edu/~yjw/teaching/cs158/lectures/10_EvaluationImbalance.pdf)
[df](https://www.cs.hmc.edu/~yjw/teaching/cs158/lectures/10_EvaluationImbalance.pdf) [Erişildi: 24 Ağustos 2019].

- Young, J., 2017. AI, Machine Learning, and Predictive Analytics – How Technology is Changing the Insurance Landscape. [Çevrimiçi]
Available at: <https://us.hitachi-solutions.com/blog/predictive-analytics-insurance/>
[Erişildi: 29 Aralık 2019].
- Zemel, R., Urtasun, R. & Fidler, S., 2016. CSC 411: Lecture 19: Reinforcement Learning. [Çevrimiçi]
Available at:
https://www.cs.toronto.edu/~urtasun/courses/CSC411_Fall16/19_rl.pdf
[Erişildi: 15 Haziran 2019].
- Zychlinski, S., 2018. The Search for Categorical Correlation. [Çevrimiçi]
Available at: <https://towardsdatascience.com/the-search-for-categorical-correlation-a1cf7f1888c9>
[Erişildi: 1 Mart 2020].