İSTANBUL TEKNİK ÜNİVERSİTESİ ★ FEN BİLİMLERİ ENSTİTÜSÜ

STRATEJİK YÖNETİM PERSPEKTİFİNDEN SİGORTACILIK SEKTÖRÜNDE MAKİNE ÖĞRENMESİ ALGORİTMALARI İLE ANOMALİ TESPİTİ

YÜKSEK LİSANS TEZİ

Ayşe Nurbanu ŞAHAN

İşletme Mühendisliği Anabilim Dalı

İşletme Mühendisliği Programı

TEMMUZ 2020

İSTANBUL TEKNİK ÜNİVERSİTESİ ★ FEN BİLİMLERİ ENSTİTÜSÜ

STRATEJİK YÖNETİM PERSPEKTİFİNDEN SİGORTACILIK SEKTÖRÜNDE MAKİNE ÖĞRENMESİ ALGORİTMALARI İLE ANOMALİ TESPİTİ

YÜKSEK LİSANS TEZİ

Ayşe Nurbanu ŞAHAN
(507161005)

İşletme Mühendisliği Anabilim Dalı

İşletme Mühendisliği Programı

Tez Danışmanı: Doç. Dr. Tolga KAYA

TEMMUZ 2020

İTÜ, Fen Bilimleri Enstitüsü'nün 507161005 numaralı Yüksek Lisans Öğrencisi Ayşe Nurbanu Şahan, ilgili yönetmeliklerin belirlediği gerekli tüm şartları yerine getirdikten sonra hazırladığı "STRATEJİK YÖNETİM PERSPEKTİFİNDEN SİGORTACILIK SEKTÖRÜNDE MAKİNE ÖĞRENMESİ ALGORİTMALARI İLE ANOMALİ TESPİTİ" başlıklı tezini aşağıda imzaları olan jüri önünde başarı ile sunmuştur.

Tez Danışmanı :	**Doç. Dr. Tolga KAYA** İstanbul Teknik Üniversitesi	.....
Jüri Üyeleri :	**Doç. Dr. Başar ÖZTAYŞİ** İstanbul Teknik Üniversitesi	.....
	Dr.Öğr.Üye. Barış SOYBİLEN İstanbul Bilgi Üniversitesi	.....

Teslim Tarihi : 15 Haziran 2020
Savunma Tarihi : 13 Temmuz 2020

Hayatım boyunca verdikleri tüm desteklerden dolayı aileme ve dostuma içten teşekkür ediyorum.

ÖNSÖZ

Bu çalışmada poliçelendirme aşamasında anomali tespitine yönelik makine öğrenmesi teknikleri kullanılarak modeller geliştirilmiştir. Geliştirilen modeller içinden, sigorta şirketlerinin anomali potansiyeli taşıyan poliçe taleplerini en yüksek oranda tespit etmelerini sağlayan model seçilmiştir. Yürütmüş olduğum çalışmanın çok az sayıda akademik çalışmanın yer aldığı mülkiyet sigortasında sahtekârlık tespitine yönelik olması sebebi ile literatüre katkı sağladığı düşüncesindeyim.

Çalışma süreci boyunca beni destekleyen ve yönlendiren tez danışmanım ve değerli hocam Sayın Tolga Kaya'ya yardımları için sonsuz teşekkürlerimi sunarım. Çalışma boyunca sağlamış oldukları teknik altyapı ve sektör spesifik bilgiler için Ümit Şile ve ekibine teşekkür ederim. Hayatımın her aşamasında olduğu gibi bu çalışmayı yürütürken de beni destekleyen değerli aileme en içten teşekkürlerimi iletiyorum.

Haziran 2020

Ayşe Nurbanu ŞAHAN
(Endüstri Mühendisi)

İÇİNDEKİLER

Sayfa

ÖNSÖZ ……vii
İÇİNDEKİLER.....x
KISALTMALAR.....xii
SEMBOLLER.....xiv
ÇİZELGE LİSTESİ.....xvi
ŞEKİL LİSTESİ.....xviii
ÖZET ……xx
SUMMARY.....xxii
1. GİRİŞ ……1
1.1 Tezin Amacı 3
2. LİTERATÜR.....5
2.1 Sigortacılık 5
2.1.1 Sigorta branşları 7
2.1.2 Sigortacılık kavramları.....7
2.2 Stratejik Yönetim.....9
2.3 Literatür Araştırması 12
3. METODOLOJİ 18
3.1 Makine Öğrenmesi 18
3.1.1 Makine öğrenmesinin türleri 18
3.1.1.1 Denetimli öğrenme.....19
3.1.1.2 Denetimsiz öğrenme.....21
3.1.1.3 Pekiştirmeli öğrenme.....31
3.1.1.4 Makine öğrenmesi modellerini geliştirme yolları 31
3.2 Anomali Tespiti (Anomaly Detectıon).....34
3.3 K-Ortalamalar (K-Means) 36
3.4 Ağaç Temelli Modeller (Tree-Based Models) 40
3.4.1 Karar ağaçları (Decision trees).....40
3.4.2 Bagging 40
3.4.3 Rastgele Orman (Random Forest).....41
3.4.4 Boosting 42
3.4.4.1 AdaBoost.....42
3.4.4.2 Gradient Boosting Machine (GBM).....42
3.5 Model Sonuçlarının Yorumlanması ve Seçimi.....53
3.6 CRISP-DM Metodolojisi.....56
3.6.1 İşi Anlama 57
3.6.2 Veriyi Anlama.....57
3.6.3 Veriyi Hazırlama.....58
3.6.4 Modelleme.....58
3.6.5 Değerlendirme.....58

3.6.6 Uygulama 58
4. UYGULAMA 61
4.1 İşi Anlama 61
4.1.1 Sigorta sektöründe kullanılan göstergeler 61
4.2 Veriyi Anlama 64
4.2.1 Verilerin toplanması 64
4.2.2 Verilerin tanımlanması 65
4.2.3 Verilerin araştırılması 66
4.2.3.1 Hasar frekanslarının incelenmesi 66
4.2.3.2 Üretim frekanslarının incelenmesi 70
4.2.3.3 Dönüşüm oranının incelenmesi 74
4.2.3.4 Hasar prim oranının incelenmesi 74
4.2.3.5 Geçen sürenin incelenmesi 76
4.2.4 Veri kalitesinin doğrulanması 77
4.3 Veriyi Hazırlama 78
4.3.1 Verilerin seçimi 78
4.3.2 Verilerin temizlenmesi 79
4.3.3 Verilerin entegrasyonu 80
4.3.4 Verilerin oluşturulması (öznitelik çıkarımı) 80
4.3.5 Verilerin biçimlendirilmesi 84
5. MODELLEME 85
5.1 Kümeleme 85
5.1.1 K-Ortalamalar 85
5.1.1.1 Küme performansının ölçümü 85
5.1.1.2 Kullanılan değişkenler 85
5.1.1.3 Dirsek grafiği 87
5.1.1.4 Küme tanımları 88
5.2 Sınıflandırma 89
5.2.1 Kullanılan araçlar 89
5.2.2 Sınıflandırma modelleri 89
5.2.2.1 Birinci aşama: baz model değişkenleri üzerinden model performansı 90
5.2.2.2 İkinci aşama: türetilen değişkenler üzerinden model performansı 91
5.2.2.3 Üçüncü aşama: türetilen değişkenler üzerinden model performansı 92
5.2.2.4 Dördüncü aşama: hiper parametre optimizasyonu 92
5.2.2.5 Beşinci aşama: yeniden örnekleme 94
6. SONUÇLAR 99
7. GELECEK ÇALIŞMALAR 101
KAYNAKLAR 103
ÖZGEÇMİŞ 107

KISALTMALAR

HPO	**:** Hasar Prim Oranı
GOSS	**:** Gradient-based One- Side Sampling
EFB	**:** Exclusive Feature Bundling
CRISP DM	**:** Cross-industry standard process for data mining
SMOTE	**:** Synthetic Minority Over-sampling Technique
SİSEB	**:** Sigorta Sahteciliklerini Engelleme Bürosu (SİSEB)
DT	**:** Decision Trees
SGD	**:** Stochastic Gradient Descent
RF	**:** Random Forests
GBT	**:** Gradient Boosted Trees
AUROC	**:** Mean area under the receiver operating characteristic
PCC	**:** Mean percentage correctly classified

SEMBOLLER

$\boldsymbol{C_k}$	**:** Gözlemlerin ait oldukları kümeler
$\mathbf{W(C_k)}$	**:** $\mathbf{C_k}$ kümesi için küme içi değişkenlik
$\bar{\mathbf{x}}_{\mathbf{kj}}$	**:** $\mathbf{C_k}$ kümesindeki j özelliğinin ortalamasını temsil etmektedir.
K	**:** Küme sayısı

ÇİZELGE LİSTESİ

Sayfa

Çizelge 2.1 : 25 yönetim aracı. **11**
Çizelge 3.1 : Rastgele Ormanlar hiper parametreleri. **41**
Çizelge 3.2 : GBM hiper parametreleri. **43**
Çizelge 3.3 : XGBoost hiper parametreleri. **46**
Çizelge 3.4 : LightGBM hiper parametreleri. **49**
Çizelge 3.5 : CatBoost hiper parametreleri. **53**
Çizelge 3.6 : Hata matrisi. **54**
Çizelge 3.7 : CRISP-DM aşamaları içerisinde takip edilen adımlar. **60**
Çizelge 4.1 : Stratejik yönelikler bazında göstergeler. **62**
Çizelge 4.2 : Kurumsal karne yaklaşımı ile belirlenen göstergeler. **63**
Çizelge 4.3 : Veri setinde yer alan gözlem ve değişken sayısı. **64**
Çizelge 4.4 : Veri setlerinde yer alan değişkenler. **65**
Çizelge 4.5 : Hasar prim oranı en yüksek olan iller ve ilçeler. **75**
Çizelge 4.6 : Aynı gün içerisinde hasar gerçekleşmesinin en yoğun olduğu iller. **76**
Çizelge 4.7 : Aynı gün hasarın gerçekleşmesi durumunun acente, il ve branş bazında dağılımı. **77**
Çizelge 4.8 : Düzenlenen veri setinde yer alan gözlem ve değişken sayısı. **78**
Çizelge 4.9 : Düzenlenen veri setinde yer alan gözlem sayısı. **80**
Çizelge 5.1 : Kümeleme değişkenleri. **86**
Çizelge 5.2 : Ortalama Silloutte değerleri. **87**
Çizelge 5.3 : Oluşturural kümelerin değişken bazında ortalama değerleri. **88**
Çizelge 5.4 : Kullanılan araçlar. **89**
Çizelge 5.5 : Birinci aşama model performansları. **91**
Çizelge 5.6 : İkinci aşama model performansları. **91**
Çizelge 5.7 : Üçüncü aşama model performansları **92**
Çizelge 5.8 : Hiper parametre optimizasyonu için belirlenen değerler. **93**
Çizelge 5.9 : Dördüncü aşama model performansları **93**
Çizelge 5.10 : Yeniden örnekleme sonrası model performansları. **95**

ŞEKİL LİSTESİ

Sayfa

Şekil 3.1 : Çapraz Endüstri Standart Süreç Modeli (CRISP-DM)..... **57**
Şekil 4.1 : 2018 yılında en çok hasar oluşan 5 il (a) illerde gerçekleşen toplam hasar, (b) ilk iki çeyrek dağılımı, (c) ilk çeyrek dağılımı, (d) ikinci çeyrek dağılımı. **67**
Şekil 4.2 : 2017 yılında en çok hasar oluşan 5 il (a) illerde gerçekleşen toplam hasar, (b) ilk iki çeyrek dağılımı, (c) ilk çeyrek dağılımı, (d) ikinci çeyrek dağılımı. **68**
Şekil 4.3 : 2018 yılında en çok hasar oluşan 5 acente (a) acentelerde gerçekleşen toplam hasar, (b) ilk iki çeyrek dağılımı, (c) ilk çeyrek dağılımı, (d) ikinci çeyrek dağılımı. **69**
Şekil 4.4 : 2017 yılında en çok hasar oluşan 5 acente (a) acentelerde gerçekleşen toplam hasar, (b) ilk iki çeyrek dağılımı, (c) ilk çeyrek dağılımı, (d) ikinci çeyrek dağılımı. **70**
Şekil 4.5 : 2018 yılında en çok üretim yapılan 5 il (a) illerde gerçekleşen toplam üretim, (b) ilk iki çeyrek dağılımı, (c) ilk çeyrek dağılımı, (d) ikinci çeyrek dağılımı..... **71**
Şekil 4.6 : 2017 yılında en çok üretim yapılan 5 il (a) illerde gerçekleşen toplam üretim, (b) ilk iki çeyrek dağılımı, (c) ilk çeyrek dağılımı, (d) ikinci çeyrek dağılımı..... **71**
Şekil 4.7 : 2018 yılında en çok üretim yapılan 5 acente (a) acentelerde gerçekleşen toplam üretim, (b) ilk iki çeyrek dağılımı, (c) ilk çeyrek dağılımı, (d) ikinci çeyrek dağılımı. **72**
Şekil 4.8 : 2017 yılında en çok üretim yapılan 5 acente (a) acentelerde gerçekleşen toplam üretim, (b) ilk iki çeyrek dağılımı, (c) ilk çeyrek dağılımı, (d) ikinci çeyrek dağılımı. **73**
Şekil 4.9 : Üretilen poliçelerin hasara dönüşüm oranı en fazla olan 5 il (a) 2018, (b) 2017..... **74**
Şekil 4.10 : Hasar prim oranı toplamları en fazla olan 5 il..... **75**
Şekil 4.11 : Hasar prim oranı toplamları en fazla olan 5 ilçe. **75**
Şekil 4.12 : Aynı gün içerisinde gerçekleşme frekansı..... **76**
Şekil 5.1 : Dirsek grafiği..... **87**
Şekil 5.2 : XGBoost tarafından belirlenen değişken önemleri. **97**

STRATEJİK YÖNETİM PERSPEKTİFİNDEN SİGORTACILIK SEKTÖRÜNDE MAKİNE ÖĞRENMESİ ALGORİTMALARI İLE ANOMALİ TESPİTİ

ÖZET

Sigortacık sektöründe yaşanan sahtekârlık faaliyetleri, sigorta şirketlerine büyük kayıplar yaşatmaktadır. Otomobil ve mülkiyet sigortası, sigortacılık sektöründe sunulan sözleşmeli ana hizmetlerin toplam hasar maliyetlerinin %76'sını oluşturmaktadır. Literatür incelendiğinde, mülkiyet sigortasında sahtekârlık tespitinin büyük ölçüde ihmal edildiği ve ağırlıklı olarak araç sigortasında sahtekârlık tespiti üzerine çalışmalar yürütüldüğü görülmektedir.

Büyük sigorta verilerinde sahtekârlık tespitine yönelik yürütülen çalışmalarda veri madenciliği tekniklerinden yararlanılmaktadır. Sahtekârlık kayıtları bulmada kullanılan en yaygın tekniklerden biri anomali tespitidir. Bu teknik, alışılmış paternlerden sapan aykırı değerleri veya anomalileri tespit etmeyi amaçlamaktadır

Bu tezin amacı makine öğrenmesi algoritmalarını kullanarak mülkiyet sigortası alanında yüksek hasara sahip poliçelerin, oluşturulma aşamasında önceden tespitinin mümkün olup olmadığını, anomali tespiti yöntemleri ile incelemektir.

Çalışma kapsamında sigorta şirketinden temin edilen mülkiyet sigortasına ait etiketsiz veriler incelenmiştir. Etiketli verilerin bulunmaması sebebi ile anomali tespiti aşamasında oluşturulan anomali senaryoları doğrultusunda denetimsiz öğrenme tekniklerinden yararlanılmış olup, veri etiketleri elde edilmiştir. Literatürde yer alan çalışmalarda hasar prim oranı değerlerinin sahtekârlık tespitinde etkisi olduğu görülmektedir. Tez kapsamında anomali senaryosu oluşturulurken yüksek hasar prim oranlarına sahip il, ilçeler ve anomali durumların tespitini destekleyen diğer değişkenler dikkate alınmıştır. Belirlenen bu anomali senaryosunda yer alan değişkenlerde farklılaşma gösteren gözlemleri bir araya getiren bir küme arayışı ile anomali tespitinin nihayetlendirilmesi hedeflenmiştir. Bu yaklaşım ile elde edilen küme elemanları anomali olarak işaretlenmiştir.

Elde edilen etiketler ile, poliçe oluşturulma aşamasında anormal durumların sınıflandırılmasında kullanılacak tahmin modellerinin oluşturulmasına yönelik denetimli öğrenme tekniklerine başvurulmuştur. Poliçe oluşturulma aşamasında yoğunluklu olarak kategorik değişkenler özelinde bilgi girişi olmaktadır. Bu nedenle çalışılacak sınıflandırma algoritmaları belirlenirken kategorik değişken yoğunluklu veri setlerinde iyi performans verdiği bilinen ağaç temelli algoritmaların performansları incelenmiştir. Anomali durumların tespitinde model performanslarını geliştirmeye yönelik yeniden örnekleme yöntemleri kullanılmıştır.

Çalışma sistematiği oluşturulurken ve takip edilmesi gereken adımlar belirlenirken Çapraz Endüstri Standart Süreç Modeli (CRISP-DM) yaklaşımı benimsenmiş olup R programlama dili kullanılmıştır.

Çalışmada yüksek hasar prim oranlarının anomali olma durumunda ayrıştırıcı olduğu gözlemlenmiştir. Çalışma sonucunda, tahmin modelleri arasında en iyi performans gösteren algoritmanın, undersampling yeniden örnekleme yaklaşımı ile 60:40 oranında dengelenmiş veri setinde 0,92 F1 skora sahip olduğu tespit edilmiştir.

AN APPLICATION OF MACHINE LEARNING TO ANOMALY DETECTION IN INSURANCE INDUSTRY USING STRATEGIC MANAGEMENT APPROACH

SUMMARY

The main purpose of the insurance approach is to share the losses of the individuals that cannot be covered individually with other insured people in a similar position. Individuals who are in a similar position come together within a company or institution and the management of this process is carried out by the relevant institutions. The institutions undertaking this task are insurance companies. The relevant parties are individuals who request insurance from insurance companies; the insured, and the company providing the insurance; the insurer (underwriter).

The insurance process involves decision making on both the insured's and the insurer's part to reach a policy agreement. First, the insured makes an application to the insurance company or credit agency. The application is evaluated by the insurance company and the insurance company makes the decision to approve or reject the application. A range of premium amounts are offered for the policies approved by the insurance company. The premium amount for the chosen policy is paid by the insured and the insurance contract (policy) is created. In case the insured has an accident, the damage caused by the accident is paid for by the insurance company according to the policy.

More than a thousand companies operate in the insurance industry worldwide. Each year, these companies collect over a trillion dollars in premiums. 76% of the total damage costs of the main contracted services provided in the insurance industry are constituted by automobile and property insurance.

Fraud activities in the insurance industry cause great losses to insurance companies. Insurance fraud is when businesses or individuals make false claims in order to gain compensation from their insurance companies. In the literature, it is seen that fraud detection in property insurance is largely neglected while studies on fraud detection are predominantly carried out in automobile insurance.

Data mining techniques are used in studies carried out to detect fraud in large sets of insurance data. Anomaly detection is one of the most common techniques used in finding fraud records. This technique aims to detect anomalies or outliers deviating from conventional patterns.

It can be said that the skewed data is one of the problems encountered in the studies carried out to detect fraud. The main source of this is that abnormal situations and normal situations do not spread evenly. It is seen that in most of the studies examined, the percentage of observations identified as fraud remained below 30% of the total data set. Some methods are used to artificially rebalance the class distribution against the problem of unbalanced distribution of classes. Some of these methods can be re-sampling methods such as sampling from the negative class,

sampling from the positive class or creating synthetic positive samples as a pre-processing step before the training phase.

Another problem that can be encountered in detecting fraud is that the fraudulent/ anomaly labels are not included in the data sets. In the studies carried out in order to detect fraud in unlabeled data, specialists are employed for labeling the fraudulent situations, different data groups with similar qualities are combined with master data or clustering approaches are used.

Insurance companies generally carry out their studies for fraud detection through the method of retrospective classification of incoming claims. There are also studies that adopt the opposite of this approach. These studies were carried out to make the fraud detection during the underwriting process.

In order to clearly understand the business dimension of this study on insurance data; interviews were held with experts working in the insurance industry. Sector reports, annual reports, and strategic plans of companies were examined. In addition, studies on the field in the literature have been scanned. As a result of these interviews and studies, comprehensive information was obtained to determine which parameters should be focused on during analysis. Especially preferred indicators to be followed in reaching the sector targets have been determined and these indicators have been used to create the data set that will be run on the model in the later stages of this study.

Cross-industry standard process for data mining (CRISP-DM) approach has been adopted while establishing the working system and determining the steps to be followed. From a data science project management approach, CRISP-DM divides the life cycle of a data mining project into six stages. These steps are as follows; business understanding, data understanding, data preparation, modeling, evaluation, and deployment. This methodology, especially for beginners, guides the process model, how the work should progress, helps in structuring the project, and makes suggestions for each task involved in the process. It prevents overlooking/omitting any of the steps for each task through a determined checklist.

The purpose of this thesis is to examine whether high loss ratios in the field of property insurance can be detected beforehand during the underwriting step with anomaly detection methods by using machine learning algorithms.

Within the scope of the study, unlabeled data of fire insurance obtained from insurance companies were examined. Due to the lack of labeled data during the anomaly detection stage, data labels were obtained by using unsupervised learning techniques in line with the anomaly scenarios created. While creating anomaly scenarios within the scope of the thesis, provinces and districts with high loss ratios and other extracted features to support the detection of anomaly situations were taken into account. It is aimed to finalize the anomaly detection by seeking a cluster that brings together observations that differ in the variables in this anomaly scenario. The cluster elements obtained with this approach are labeled as anomalies.

Supervised learning techniques were used to create prediction models for the classification of anomaly situations during the underwriting phase with the obtained labels. During the underwriting phase, there is extensive information related to categorical variables; so while determining the classification algorithms to be studied, the performances of tree-based algorithms were examined, which are known for having note-worthy performance in categorical variable dense data sets.

Resampling methods were used to improve model performance in detecting anomaly situations.

The cluster representing anomaly states were determined by the K-means cluster approach.

Decision Tree, Bagging Tree, Random Forest, Xgboost, Lightgbm, Catboost are the tree-based algorithms whose performances were evaluated for the model, which is aimed to be developed for the detection of anomaly situations during the creation of policies. F1 scores were taken into consideration when comparing model performances.

The study was carried out to capture the best model performance with the selected models and was designed in 5 stages.

Firstly, models were run on the variables used during the underwriting phase. Here, in order to enable the model to respond more quickly, feature engineering was performed in multi-factor categorical variables such as province and district. With the defined 'same city' variable, whether the insured city and the risk city are the same or not were determined and included in the model. The same study was done for the district.

In the second stage, the variables derived to detect excessive increases in production and losses are included in the model. While extracting the features following this increase, the total production quantity and loss quantity of one month before the relevant date were taken into consideration and the information from one week ago was taken into consideration. In this way, it was aimed to follow the sudden increase in production or loss. However, with these new features, the model performances did not reach the desired levels.

In the third stage, daily loss ratios were calculated for some categorical variables to improve model performances (e.g. agency-based daily loss ratios). With these new variables, the conversion rates to loss of the policies produced within one day were followed. Still model performances did not reach the desired levels.

In the fourth step, cross-validated hyper parameter optimization was performed to improve model performances. Considering the sensitivity values of the hyperparameter optimization, it was determined that the Lightgbm algorithm showed the highest performance in detecting true anomaly situations as an anomaly. When comparing F1 scores, Catboost was the most suitable model for the study carried out. However, detection percentages remained at approximately 30%.

Due to the inability to reach the desired performance levels with the first four steps of the study, the final step was initiated. A resampling approach was used in the final step taking into account that the unbalanced data, which are encountered in many studies in the literature, are balanced by resampling and that the model performances are increased. The under-sampling method is adopted from the resampling approaches, and data performances at different balance levels are arranged and studied again. A significant improvement was observed in the results obtained using this method.

When model outputs are analyzed with the same comparison parameters, considering the F1 score values when the dataset is balanced by 50:50, the highest performing algorithms are Random Forest, Catboost and Decision Tree. When the data set is balanced at 60:40, the Xgboost, Random Forest and Catboost algorithms were the

highest performing. At 70:30 it was again determined that Random Forest, Xgboost and Catboost algorithms were the highest performing. Finally, Random Forest, Xgboost and Bagging algorithms came to the forefront with the 80:20 ratio.

When the performances of the determined algorithms were compared according to the F1 scores obtained, considering all the ratios selected to balance the data set, the highest performance was observed in the XGBoost algorithm at a rate of 60:40.

The most important variables determined by the XGBoost algorithm are taken into consideration as the factors that the insurance companies should consider while examining the anomaly of the requests received during the underwriting phase and developing strategies for these situations. The issues determined to be focused on by insurance companies are listed as follows; from which region the policy request comes from, whether there is a remarkable increase in the number of loss occurrences in the agency that constitutes the policy, the rate of conversion of daily loss ratios in the district where the policy request comes from, and under which tariff the policy was created.

In the study, it was observed that high loss ratio is a distinguishing factor for anomalies. As a result of the study, it has been determined that the algorithm that performs best among the prediction models has an F1 score of 0.92 in the data set balanced with an under-sampling approach at a ratio of 60:40.

R programming language was been used while completing this study.

Decision Tree, Bagging Tree, Random Forest, Xgboost, Lightgbm, Catboost are the tree-based algorithms that were used in this study, however; artificial neural networks and deep learning methods are also recommended for future studies. While within the scope of this study, K-means was preferred at the clustering stage, an intensity based clustering approach can also be preferred in future studies. In this study, classifications were made by determining a single threshold value in ensemble methods while using classification algorithms. However, by using a variety of threshold values, it is possible to examine which threshold is the most appropriate for the problem. This is another area that can be examined in future studies.

1. GİRİŞ

Dünya çapınca sigortacılık sektöründe binden fazla şirket faaliyet göstermektedir. Her yıl bu şirketler bir trilyon doların üstünde prim toplamaktadırlar. İşletmelerin ya da bireylerin tazminat kazanmak amacı ile yanlış sigorta tazminat taleplerinde bulunması durumu sigorta sahtekârlığı olarak bilinmektedir. Toplam sigorta sahtekârlığı tutarının kırk milyar doların üstünde olduğu tahmin edilmektedir (Roy, n.d.). Dolayısı ile sigortacılık sektöründe sahtekârlık tespiti, şirketler için zorlu bir süreç olmasına rağmen büyük önem taşımaktadır.

Türkiye'de Sigorta Sahteciliklerini Engelleme Bürosu (SİSEB) tarafından suiistimal tespitinde kullanılan analitik model yardımıyla 2016 yılında, 5,03 milyon TL tutarında suiistimal gerçekleştiren organize suistimal networkü tespit edilmiş ve ilgili dosyalara yönelik hukuksal süreç başlatılmıştır (*TSB Faaliyet Raporu*, 2016). Sürekli değişiklik gösterek suiistimal paternlerini tespit edebilmek adına suiistimal tespit modelleri düzenli olarak güncellenmektedir. 2017 yılı haziran ayında güncellenen SİSEB analitik modeli, güncelleme sonrası tespit oranlarında artış göstermiştir. 2017 yılı için tespit edilen suiistimal miktarı 42,70 milyon TL'dir (*TSB Faaliyet Raporu*, 2017).

SİSEB, sigorta şirketlerinin paylaşabileceği sigorta sahtekârlığı durumu örneklerini; 'mahkemede karara bağlanmış suistimal durumları, savcılık soruşturmasına konu hasar başvuruları durumları, şirketçe ortaya çıkarılan suistimal durumlarında sigortalıdan feragatname alındığı durumlar, sahte belgeli hasar başvuruları, gerçek dışı beyana dayalı şüpheli hasar başvuruları, hasardan sonra poliçe düzenlenmesi durumu, sahte hasar şüphesi ile reddedilen ve sigortalının şirketi dava ettiği durumlar, alkollü araç kullanımına ilişkin durumlar, sürücü Bilgilerine İlişkin durumlar, şirketlerin sahtekârlık şüphesiyle ek araştırma gereksinimi duyduğu durumlar' şeklinde gruplamıştır (*Sigorta Suistimalleri Bilgi Paylaşım Sistemi*).
Sigortacılığın farklı alanlarında yaşanan sahtekârlık faaliyetlerinin sigorta şirketlerine büyük kayıplar yaşattığına dair çeşitli örneklere rastlanılmaktadır.

Sağlık sigortası yüksek harcama seviyelerine sahip bir alan olması sebebi ile dolandırıcılık faaliyetleri için potansiyel bir hedef olarak tanımlanmaktadır. FBI tarafından dolandırıcılığın tüm tıbbi maliyetlerin % 3-10'unu oluşturduğu tahmin edilmekte olup, dolandırıcılık tespiti ile ABD sağlık programları ve faydalanıcıları için önemli tasarrufların sağlanacağı belirtilmiştir (Bauder, R. A., ve Khoshgoftaar, 2017).

Sigorta şirketlerinin maliyetleri arasında araç sigortası önemli bir paya sahiptir. Araç sigortasında sahtekârlık tespiti sigorta şirketlerinin maliyetlerini düşürmede kritik bir öneme sahiptir (Wang ve Xu, n.d.).

Sigortacılık sektöründe otomobil ve mülkiyet sigortası, sunulan sözleşmeli ana hizmetlerin toplam hasar maliyetlerinin %76'sını (Palacio, 2019) oluşturmaktadır. Literatür incelendiğinde, mülkiyet sigortasında sahtekârlık tespitinin büyük ölçüde ihmal edildiği ve ağırlıklı olarak araç sigortasında sahtekârlık tespiti üzerine çalışmalar yürütüldüğü görülmektedir. Bunun temel sebebinin, tanıkların daha az olması ya da tanık olacak kişilerin hasar ihbarında bulunan kişiler ile aynı mülkü paylaşıyor olmalarından kaynaklandığı düşünülmektedir (Palacio, 2019). Ayrıca sigorta şirketlerinin rekabetçi bilgilerini açık etmemek ve bu bilgiyi gizli tutmak amacı ile nadiren gerçek sahtekârlık verilerini paylaşmaları bunun bir diğer sebebi olarak görülmektedir (Palacio, 2019). Yürütülen çalışma çok az sayıda akademik çalışmanın yer aldığı mülkiyet sigortasında sahtekârlık tespitine yönelik olması sebebi ile önem arz etmektedir.

Sahtekârlık tespitine yönelik yürütülen çalışmalarda karşılaşılan problemlerin başında anormal durumlar ile normal durumların dengeli dağılmayışı kaynaklı verinin çarpık olmasının geldiği söylenebilir. İncelenen çalışmaların çoğunda sahtekâr olarak tespit edilen gözlemlerin yüzdesinin, toplam veri setinin %30'unun altında kaldığı görülmektedir (C. Phua ve diğ., 2010). Sınıfların dengesiz dağılıyor olması problemine karşı sınıf dağılımını yapay olarak yeniden dengelemek için; training aşaması öncesinde ön işleme adımı olarak, negatif sınıftan örnekleme yapma, pozitif sınıftan örnekleme yapma (Chawla ve diğ., 2002) ya da sentetik pozitif örnekler oluşturma (Synthetic Minority Over-sampling Technique/ SMOTE) (Chawla ve diğ., 2002; Nagrecha, n.d.) gibi yeniden örnekleme yollarına gidilebilmektedir.

Sahtekârlık tespitinde karşılaşılabilecek bir diğer problem ise veri setlerinde sahtekâr/ anomali etiketlerinin yer almayışıdır. Etiketsiz verilerde sahtekârlık tespitine yönelik yürütülecek çalışmalarda, sahtekâr durumların etiketlenmesi noktasında uzman bireyler ile çalışılmakta (Domingues, 2015), benzer nitelikte farklı veri grupları ile ana veri birleştirilmekte (Bauder ve diğ., 2018) ve kümeleme (Moradi ve Hosseinkhani, 2017; Shan ve diğ., 2008; Williams ve Huang, 1997) yaklaşımlarından yararlanılmaktadır.

Sigorta şirketleri sahtekârlık tespitine yönelik çalışmalarını genelde, gelen tazminat taleplerinin geriye dönük sınıflandırılması yöntemi ile gerçekleştirmektedir. Bu yaklaşımın tersini benimseyen, sahtekârlık tepsitinin poliçeler oluşturulurken yapılabilmesine yönelik yapılmış çalışmalar mevcuttur (Nagrecha ve diğ., 2018).

1.1 Tezin Amacı

Bu çalışmanın amacı, yüksek hasara sahip poliçelerin, oluşturulma aşamasında önceden tespitinin mümkün olup olmadığını, anomali tespiti yöntemleri ile incelemektir. Bu doğrultuda, etiket içermeyen veri kümesinde ilk aşamada denetimsiz öğrenme teknikleri ile etiketleri elde etmek, devamında elde edilen etiketleri poliçe oluşturma aşamasında tahmin etmektir. Buna yönelik takip edilen adımlar aşağıda belirtildiği gidbidir;

- Anomali senaryosunun oluşturulması, senaryo ile ilişkili değişkenlerin veri setlerinden türetilmesi
- Denetimsiz öğrenme teknikleri ile anomali gözlemlerin tespit edilmesi
- Poliçe aşamasında anomalilerin tespit edilmesine yönelik denetimli öğrenme algoritmalarının uygulanması
- Algoritmaların anomali tespit performanslarının geliştirilmesine yönelik yeni değişkenlerin türetilmesi, analize dahil edilmesi
- Algoritmaların anomali tespit performanslarının geliştirilmesine yönelik hiper parametre optimizasyonu yaklaşımının belirlenmesi
- Algoritmaların anomali tespit performanslarının geliştirilmesine yönelik yeniden örnekleme yapılması
- Yüksek accuracy değerlerine ulaşmak yerine yüksek sensitivity, AUC ve F1 score değerlerine ulaşılmasının hedeflenmesi.

2. LİTERATÜR

2.1 Sigortacılık

> Sigorta, aynı türden tehlikeyle karşı karşıya olan kişilerin, belirli bir miktar para ödemesi yoluyla toplanan tutarın, sadece o tehlikenin gerçekleşmesi sonucu fiilen zarara uğrayanların zararını karşılamada kullanıldığı, bir risk transfer sistemidir. Bu sistem sayesinde kişiler, karşı karşıya bulundukları tehlikelerin neden olabileceği, parayla ölçülebilen zararlarını, nispeten küçük miktarlarda ödemiş oldukları primler yoluyla paylaşmaktadırlar. (*Sigorta Tanımları*)

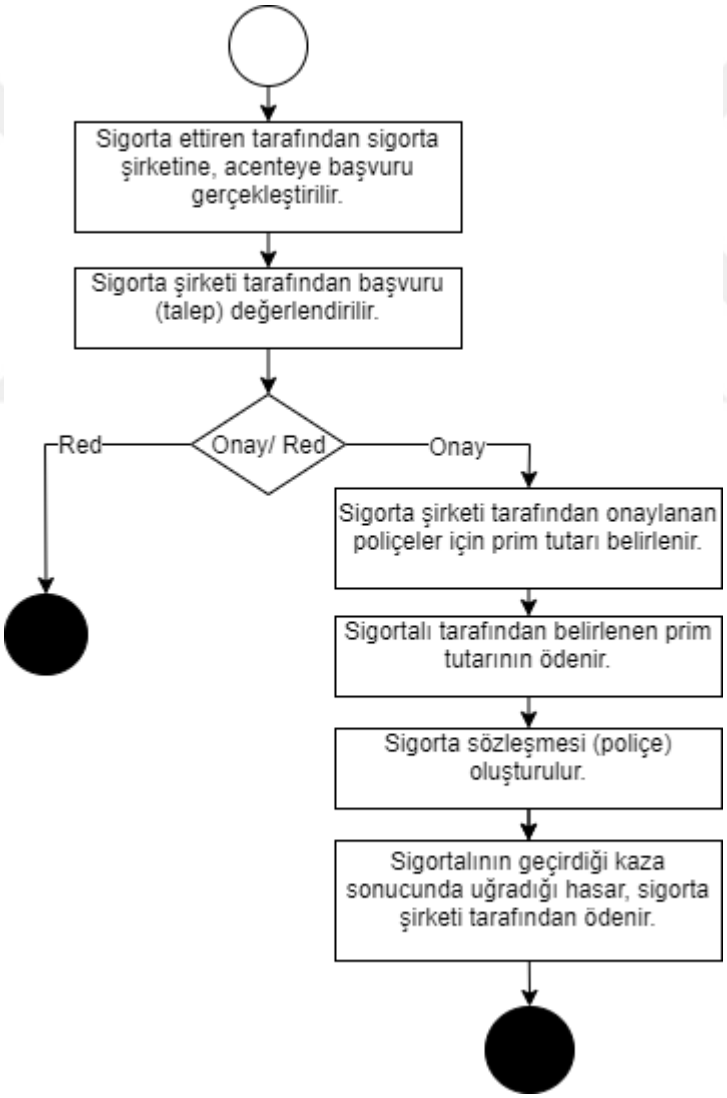
Sigorta yaklaşımında temel amaç kişilerin bireysel olarak karşılaması mümkün olmayan miktarlardaki zararlarını benzer konumda bulunan diğer bireyler ile paylaşmaktır. Benzer konumda olan bireyler bir şirket/ kurum bünyesinde bir araya gelmektedirler ve bu sürecin yönetimi ilgili kurumlar tarafından gerçekleştirilmektedir. Bu görevi üstlenen kurumlar sigorta şirketleridir. Sigorta şirketlerinden sigorta talebinde bulunan bireyler; sigorta ettiren (sigortalı-genelde aynı kişi), sigortayı sağlayan şirket ise; sigortacıdır.

Tehlike ile karşılaşması muhtemel taraf olan sigorta ettiren ve sigorta ettirene zararının tazmininin teminatını veren, yasal olarak yetkilendirilmiş sigortacı, sigorta kanunu ve mevzuatına göre sigorta sözleşmesinin iki tarafını oluşturmaktadır.

Genel anlamda sigorta yaptırma süreci incelenecek olursa; sigorta ettiren, sigorta şirketine talepte bulunmaktadır. İlgili başvurular direk sigorta şirketine gerçekleştirilebileceği gibi ayrıca acenteler üzerinden de gerçekleştirilebilmektedir. Gelen talep doğrultusunda sigorta şirketi bireylere teklifler sunmaktadır. Karşılıklı mutabakat sağlanan teklif üzerinden poliçelendirmeye (sigorta sözleşmesi) gidilmektedir. Poliçeler yıllık olarak kesilmektedir. Poliçelerdeki her bir revizyon zeyl olarak adlandırılmaktadır. Poliçelendirme işleminin nihayetlenmesi için, ilgili sigortacının sigortalıyı korumaya yönelik verdiği taahhütlere karşılık gelen poliçe için belirlemiş olduğu prim tutarının, sigortalı tarafından ödenmesi gerekmektedir. Sigortalı (poliçe sahibi), poliçelendirme işlemi tamamlandıktan sonra poliçesi

kapsamında yer verilen teminatlardan birinin kapsamına giren bir tehlike durumu ile karşılaştığında hasarının tazminini talep etmede yasal olarak yetkili konuma gelmektedir.

İlgili poliçe kapsamında hasara uğrayan bireyler sigorta şirketlerinden tazminat talebinde bulunmaktadır. Gerekli değerlendirme aşamaları tamamlandıktan sonra zarara uğrayan bireylere sigorta şirketleri tarafından ödeme gerçekleştirilmektedir. Tazminat taleplerinin ödemesi kısa sürede gerçekleşebileceği gibi eksper incelemeleri, raporlama ve eğer varsa dava süreçleri kaynaklı poliçe yılını aşan sürelerde sonuçlanma durumları ile de karşılaşılmaktadır. Sigorta yaptırma süreci Şekil 2.1'de özetlenmiştir.



Şekil 2.1 : Sigorta yaptırma süreci.

2.1.1 Sigorta branşları

2007/1 sayılı yayınlanan tebliğde Türkiye'de uygulanacak Sigorta Branşlarına yer verilmiştir. Hayat dışı grubunda yer alan sigorta branşlar sırası ile;

- 'Sigorta Branşları: Kaza (meslek hastalıkları dahil), Hastalık/ Sağlık, Kara araçları, Raylı araçlar, Hava araçları, Su araçları (deniz, göl ve nehir araçları), Nakliyat (ticari mal, bagaj ve tüm diğer mallar), Yangın ve doğal afetler (yangın, patlama, deprem, sel, deprem ve sel dışındaki doğal afetler, nükleer enerji, toprak kayması), Genel zararlar (Dolu veya dondan kaynaklanan zararlar, 8'de sayılan haller dışında hırsızlık ve diğer tüm olaylardan kaynaklanan zararlar), Kara araçları sorumluluk, Hava araçları sorumluluk, Su araçları sorumluluk (deniz, göl ve nehir araçları), Genel sorumluluk, Kredi, Emniyeti suiistimal, Finansal Kayıplar, Hukuksal Koruma, Destek,
- Birden Fazla Branş İçin Verilen Ruhsat Tanımları; Kaza ve Hastalık/Sağlık Sigortası, Kara Taşımacılığı Sigortası, Nakliyat ve Ulaşım Sigortası, Havacılık Sigortası, Yangın Sigortası ve Diğer Zararlar, Sorumluluk Sigortaları, Kredi ve Emniyeti Suistimal Sigortaları' şeklindedir.

2.1.2 Sigortacılık kavramları

Hasar: 'Herhangi bir mala, eşyaya verilen fiziki hasarı veya kişinin gerek bedeni gerekse manevi zararını ifade etmek için kullanılır.'

Zeyil: 'Poliçe ve tecditler üzerinde herhangi bir değişiklik, ilave, iptal gibi durumlar olduğunda düzenlenen belgelerdir.'

Muallak Hasar (Outstanding Loss): 'Meydana gelmiş ve sigorta şirketi tarafından bilinen bir hasara ilişkin olarak, ileride ödenmesi muhtemel hasar ödemesidir ve bu muhtemel ödeme için ayrılan miktara da muallak hasar rezervi denir.'

Hasar Prim Oranı (Loss Ratio): 'Gerçekleşen hasarların kazanılmış prime olan oranıdır. Hasar prim oranının hesabı şu şekilde yapılmaktadır:

(Ödenen Hasarlar + Muallak Hasarlar Karşılığı - Devreden Muallak Hasarlar Karşılığı) / (Yazılan Prim - Kazanılmamış Primler Karşılığı + Devreden

Kazanılmamış Primler Karşılığı - Devam Eden Riskler Karşılığı + Devreden Devam Eden Riskler Karşılığı)'

İş Yılı (Underwriting Year): Poliçe yılıdır. Bir yıllık işlem gören poliçelerin düzenlendiği yılı refere eder. Poliçe için belirlenen primin sigortalı tarafından sigortacıya takip eden yıl içerisinde ödenmesi gerekir. Hasar oluşması durumunda, hasar ödemesi için aradan geçen süreden bağımsız ödenen hasar bedeli poliçenin düzenlendiği yıla göre muhasebe kayıtlarına girilir. Dolaysısı ile muhasebeleştirme noktasında önem taşımaktadır. (*Sigorta Tanımları*, n.d.).

2.2 Stratejik Yönetim

‘Stratejik Yönetim, özel, kamu ve kar amacı gütmeyen gönüllü sektörlerde faaliyet gösteren tüm organizasyonlarda geleceğe yönelik amaç ve hedeflerin belirlenmesine ve bu hedeflere ulaşılabilmesi için yapılması gerekli işlemlerin tespit edilmesine imkan sağlayan bir yönetim tekniğidir’ (Aktan, 2008).

Şirketler, geleceğe yönelik ideal ve ortak bakışlarını, geliştirdikleri stratejiler ile yansıtırlar. Şirketin içerisinde bulunduğu mevcut durum ve çevre şartları sonucunda ortaya çıkan ihtiyaçlar, ulaşmak istediği vizyona erişmesinde birer ara basamak olarak da nitelendirilebilecek amaçların ve bu amaçları gerçekleştirmeye yönelik hedeflerin ortaya konmasını sağlar.

Andrews’a (1981) göre, işletme açısından strateji “İşletmenin hangi işi yaptığını veya yapmak istediğini, ne tür bir işletme olduğunu veya olmak istediğini tanımlayan amaç, hedef ve görevlerin tümü ve bunları gerçekleştirmek için gerekli yöntemlere verilen addır”. Strateji işletmenin tüm paydaşlarının ortak yönde hareket etmelerini sağlar. Stratejisi olmayan işletme, önünü göremeyen ve yönü olmayan bir işletmedir.

Giderek zorlaşan rekabet koşulları karşısında ayakta kalabilmek adına şirketler farklılaşmaya ihtiyaç duymuş, stratejik düşünme, stratejik planlama ve stratejik karar alma konularına verilen önem artmaya başlamışlardır (Aktan, 2008).

Stratejik yönetim araçlarından biri olan stratejik planlarda, işletmeye ait amaçlara, stratejik plan dönemi içerisinde amaçların gerçekleştirilme düzeyini resmeden hedeflere ve hedeflere yönelik proje ve faaliyetlere yer verilir. Hedeflere yönelik belirlenen proje ve faaliyetler, bütçe ve performans dikkate alınarak şekillendirilir.

Stratejik planlar ile belirlenen amaç, hedef, performans göstergeleri ve stratejilerin tanımlarına sırası ile yer verilmiştir.

Şirketin vizyon çalışmaları ile belirlediği genel çerçevede ulaşmayı hedeflediği kavramsal çıktılar, elde edilmek istenen sonuçlar şirketin stratejik amaç ifadelerini belirlemektedir. Belirlenen amaç ifadelerinin şirketin misyonu ve vizyonu ile uyumlu olması beklenir ve şirketi belirlenmiş olan vizyonuna taşıması istenir.

Belirlenen amaçların gerçekleştirilmesine yönelik, belirli bir zaman diliminde gerçekleştirilmesi istenen gelişim, stratejik hedef ifadeleri ile belirlenmektedir. Hedeflerin kurumun misyonu, vizyonu, değerleri ve amaçları ile tutarlı olması ve zaman ve miktar ile ifade edilebilir olması beklenmektedir.

Belirlenen hedeflere yönelik, ilgili hedefin ne kadarının gerçekleştirildiğini takip etmede kullanılmak üzere performans göstergeleri belirlenmektedir. Girdi, çıktı, verimlilik, sonuç ve kalite göstergeleri olmak üzere toplamda beş tip performans göstergesi bulunmaktadır. Performans göstergelerinin ölçülebilirliği önem arz ettiğinden zaman ve miktar ile ifade edilebilir olmaları beklenmektedir. Sonuç ve çıktı göstergeleri amaç ve hedefler odağında ne düzeyle başarı elde edildiğinin ölçülmesinde kullanılmaktadır (*Kamu İdareleri İçin Stratejik Planlama Kılavuzu*, 2018). Hedef performans göstergeleri belirlenirken, geçmiş dönem performansları dikkate alınarak, ilgili göstergeye çok yüksek ya da yok düşük değerlerin atanmasının önüne geçilmesi gerekir.

Belirlenen hedeflere ulaşmada izlenecek yöntemler, belirlenen politikalar ve benimsenen yaklaşımlar stratejiler ile ortaya konmaktadır. Stratejiler şirketlerin, diğer şirketlerden farklılaşmasını sağlayan noktalardır.

Stratejik yönetim araçlarından bir diğeri, işletmeleri farklı boyutlara indirgeyerek, o boyutlar düzleminde işletmenin vizyonunu ve belirlemiş olduğu stratejileri destekleyecek performans göstergelerinin belirlenmesinde kullanılmak üzere tasarlanmış kurumsal karne (balanced scorecard) yaklaşımıdır.

Kaplan ve Norton tarafından 1996 yılında ortaya konulan bu yöntemde, finansal boyut, müşteri boyutu, işletme içi süreçler boyutu ve öğrenme-gelişme boyutu olmak üzere dört boyut bulunmaktadır. Kurumsal karne ile yöneticilerin bu boyutları dikkate alarak işletmeyi farklı açılardan değerlendirmesi amaçlanmıştır (Kaplan ve Norton, 1996).

Yöntemin isminde yer alan dengeli kavramı arka planda işletmeyi indirgemiş olduğu boyutların kendi aralarında sağladığı dengeye işaret etmektedir. Burada belirlenen performans göstergeleri ile elde edilen değerler üzerinden işletme performansını değerlendirirken ve geleceğe yönelik adımlar belirlenirken sadece finansal değil,

finansal olmayan ölçüleri de göz önünde bulunduran dengeli bir sistem vurgulanmak istenmiştir (Kaplan ve Norton, 2003).

İşletmelerin bu kapsamda kullandıkları çeşitli yönetim araçları vardır. Bain ve Company dünya çapında, yöneticilerin stratejik yönetim perspektifinde hangi araçların kullanımını önceliklendirdiklerini ve kullanılan araçların vermiş olduğu sonuçların yöneticiler nezdinde memnuniyet düzeylerini, uyguladığı anketler ile araştırmaktadır. 2015 yılı raporunda, uluslararası ölçekte, stratejik planlama kullanım oranına göre 4. sırada yer alırken, kurumsal karne 6. sırada yer almıştır. Ancak sürpriz bir şekilde o dönem için diğer araçlarla kıyaslandığında görece daha yeni bir araç olan büyük veri analizi, kullanım oranında her ne kadar geri sıralarda kalsada, memnuniyet noktasında ilk sırayı almıştır (Rigby ve Bilodeau, 2017). 2017 yılı raporunda ise şirketlerde en popüler 25 yönetim aracına yer verilmiştir. Ankette yer verilen 25 yönetim aracı Çizelge 2.1 belirtilmiştir.

Çizelge 2.1 : 25 yönetim aracı.

1. İleri analitik/ Advanced Analytics	13. Dijital Dönüşüm
2. Çevik Yönetim/ Agile Management	14. Çalışan Katılımı Sistemi/
3. Dengeli Kurumsal Karne/ BSC	Employee Engagement Systems
4. Benchmarking	15. Nesnelerin İnterneti/ IoT
5. İş Süreçlerinin Yeniden Yapılandırılması	16. Birleşme ve Satın Almalar
6. Değişim Yönetimi Programları/	17. Misyon ve Vizyon İfadeleri
7. Karmaşıklığın Azaltılması/	18. Organizasyonel Zaman Yönetimi
Complexity Reduction	19. Fiyat Optimizasyonu Modelleri
8. Temel Yetkinlikler/ Core Competencies	20. Senaryo ve Acil Durum Planlaması
9. Müşteri Deneyimi Analizi	21. Stratejik İşbirliği
10. Müşteri İlişkileri Yönetimi/ CRM	22. Stratejik Planlama
11. Müşteri Memnuniyeti Sistemleri	23. Tedarik Zinciri Yönetimi
12. Müşteri Segmentasyonu	24. Toplam Kalite Yönetimi
	25. Sıfır Tabanlı Bütçe

Anket sonuçları incelendiğinde en çok tercih edilen yönetim aracının %48 kullanım oranı ile stratejik planlama olduğu tespit edilmiştir. En çok tercih edilen ilk dört araç geçmiş dönem ile kıyaslamalı incelendiğinde çalışan katılımı anketlerinin sıralamadan düştüğü ve onun yerine %42 kullanım oranı ile ileri analitiğin geldiği

görülmektedir. Kurumsal karne %29 kullanım oranı ile 14. sıraya gerilemiştir (Rigby ve Bilodeau, 2017).

Yönetim araçlarındaki dinamik değişimde göz önüne alınarak özellikle performans göstergeleri odağında ileri analitik uygulamaları çalışmanın, verimli sonuçlar doğuracağı düşünülmektedir. Dolayısı ile yürütülen çalışma kapsamında işi anlama aşamasında ve kurulacak sınıflandırma modellerinde yer verilecek değişkenler belirlenirken ve türetilirken, sigorta sektörü raporları ve stratejik planları incelenmiş, buralarda yer veriler göstergeler odağında model değişkenlerinin belirlenmesi yoluna gidilmiştir.

2.3 Literatür Araştırması

Büyük sigorta verilerinde sahtekârlık tespitine yönelik yürütülen çalışmalarda veri madenciliği tekniklerinden yararlanılmaktadır. Sahtekârlık kayıtları bulmada kullanılan en yaygın tekniklerden biri anomali tespitidir. Bu teknik, alışılmış paternlerden sapan aykırı değerleri veya anomalileri tespit etmeyi amaçlamaktadır (Kirlidog ve Asuk, 2012).

Sigortacılık sektöründe otomobil ve mülkiyet sigortası, sunulan sözleşmeli ana hizmetlerin toplam hasar maliyetlerinin %76'sını (Palacio, 2019) oluşturmaktadır. Mülkiyet sigortası yangın, hırsızlık ve bazı hava hasarları gibi olaylar sonucunda doğabilecek mülk risklerine karşı koruma sağlamaktadır. Mülkiyet sigortası kapsamında giren özel sigorta biçimleri yangın sigortası, taşkın sigortası, deprem sigortası ve ev sigortası şeklinde tanımlanmaktadır (*Property insurance*). Otomobil sigortası ise otomobil, kamyon, motosiklet ve diğer karayolu taşıtları için koruma sağlamaktadır (*Automobile insurance*). Literatür incelendiğinde, mülkiyet sigortasında sahtekârlık tespitinin büyük ölçüde ihmal edildiği ve ağırlıklı olarak araç sigortasında sahtekârlık tespiti üzerine çalışmalar (Clifton Phua ve diğ., 2004) yürütüldüğü görülmektedir. Bunun temel sebebinin, tanıkların daha az olması ya da tanık olacak kişilerin hasar ihbarında bulunan kişiler ile aynı mülkü paylaşıyor olmalarından kaynaklandığı düşünülmektedir (Palacio, 2019). Ayrıca sigorta şirketlerinin rekabetçi bilgilerini açık etmemek ve bu bilgiyi gizli tutmak amacı ile nadiren gerçek sahtekârlık verilerini paylaşmaları bunun bir diğer sebebi olarak görülmektedir (Palacio, 2019).

Bu kapsamda literatür araştırmasının odağı geniş tutulmuş olup, sigorta sektöründeki tüm branşlar için yürütülen anomali ve sahtekârlık tespiti çalışmalarının yanı sıra farklı konular üzerinede incelemelerde bulunulmuştur. Sigortacılık sektöründe sahtekârlık tespitine yönelik çalışmalarda denetimsiz öğrenme ve denetimli öğrenme yaklaşımlarının, farklı uygulama örneklerine rastlanmaktadır.

1997 and 2013 yılları arasında sigorta sektöründe sahtekârlık tespitine yönelik yapılan çalışmaların %57'sinin sınıflandırma yöntemlerini kullandığı görülmektedir. Bu döneme yönelik yapılan araştırmada en sık kullanılan sınıflandırma tekniklerinin Decision tree, Naïve Bayes ve Support Vector Machine olduğu tespit edilmiştir. Sınıflandırma algoritmalarının yanı sıra kümeleme, regresyon, uç değer analizi, tahminleme ve görselleştirme kullanılan diğer teknikler olarak karşımıza çıkmaktadır (Shan ve diğ., 2008).

Denetimsiz öğrenme üzerine yapılan sahtekârlık ve anormallik tespiti çalışmalarında, kümeleme (Williams ve Huang, 1997), uç değerlerin tespiti ve ilişkilendirme kurallarının (Shan ve diğ., 2008) kullanıldığı görülmektedir (Joudaki ve diğ., 2015).

Williams ve Huang (1997) çalışmalarında, küme sınıflarının tespitinde k-means, kuralların tanımlanmasında C4.5 decision tree ve kuralların değerlendirilmesinde alan bilgisi, istatistiksel özetler ve görselleştirme araçları olmak üzere; üç adımlı bir proses uygulamışlardır (Williams ve Huang, 1997).

İncelenen bir diğer çalışmada, bir sigorta şirketinin sağlık sigortası müşterilerine uygun hizmetler sunabilmesine yönelik farklı müşteri gruplarını oluşturmak ve bu grupların karakteristiklerini belirlemek amacıyla kümeleme algoritmalarından yararlanılmıştır. Çalışmada tercih edilen kümeleme yöntemlerinden biri k-means'dir (Moradi ve Hosseinkhani, 2017).

Sahtekârlık tespitine yönelik farklı uygulamalar dikkat çekmektedir. Şüpheli mülkiyet sigortası taleplerinin tespit edilmesine yönelik yürütülen bir çalışmada genetik programlama kullanılarak bulanık kuralların geliştirilmesi amaçlanmıştır (Bentley, 2000).

Sahtekârlık tespitine yönelik yürütülen çalışmalarda karşılaşılan temel problem sınıfların dengesiz dağılımıdır. İncelenen çalışmaların çoğunda sahtekârlık olarak tespit edilen gözlemlerin yüzdesinin, toplam veri setinin %30'unun altında kaldığı görülmektedir (C. Phua ve diğ., 2010). Sınıfların dengesiz dağılıyor olması problemine karşı sınıf dağılımını yapay olarak yeniden dengelemek için; training aşaması öncesinde ön işleme adımı olarak, negatif sınıftan örnekleme yapma, pozitif sınıftan örnekleme yapma (Chawla ve diğ., 2002) ya da sentetik pozitif örnekler oluşturma (SMOTE) (Chawla ve diğ., 2002; Nagrecha, n.d.) gibi yeniden örnekleme yollarına gidilebilmektedir. Bu yaklaşımın avantajı mevcutta bulunan sınıflandırma algoritmalarından herhangi biri için geçerli olmasıdır. Dengesiz veri setlerinde vermiş oldukları ikna edici performans sonuçları sebebiyle bu yaklaşımlar çok dikkat çekmektedir (Galar ve diğ., 2011; Khoshgoftaar, T. M. ve diğ., 2010).

Phua ve diğ. 2004 yılında yapmış olduğu çalışmada az sayıda bulunan gözlemleri oversampling ile yeniden örneklendirerek elde ettiği veri setinde backpropagation neural networks, naive Bayes, ve C4.5 sınıflandırma algoritmalarını kurmuştur.

Bauder ve diğ. (2017) sahtekâr ve sahtekâr olmayan sınıfların dağılımını dengelemeye yönelik az sayıda bulunan sınıf için oversampling, yüksek sayıda bulunan sınıf için undersampling tekniklerini uygulamış ayrıca sahtekârlık olarak etiketlenmiş gözlemlere dokunmadan normal gözlemlerden yerine koymadan rasgele örnekleme (randomly sample without replacement) yaklaşımını benimseyerek sınıfların dağılımını 80:20 olacak şekilde yeniden düzenlemiştir.

Hassan ve Abraham (2016) çalışmasında sigortacılık sektöründe dengesiz veri setlerinde sınıflandırma modellerinin performansını arttırmaya yönelik, çoğunlukta bulunan sınıf için, bölümlere ayrılmış undersampling yöntemini önermiştir. Elde edilen sonuçlar, sınıflandırıcıların, orijinal dengesiz verilerde gösterdikleri performansların çok üstündedir. Önerilen yaklaşımın ensemble tekniklerde kullanılması tavsiye edilmiştir.

Nagrecha ve diğ. (2018) çalışmalarında, araç sigortalarını kapsayan veri setinde, dengeli olmayan veri dağılımına yönelik random oversampling, random undersampling yeniden örnekleme yöntemlerini kullanmıştır. Geleneksel olarak yorumlanabilir modellere (DT, Stochastic Gradient Descent (SGD), Random Forests

(RF) and Gradient Boosted Trees (GBT)) yer verilen çalışmada, yeniden örneklemenin tahmin performanslarında pozitif bir etkisi olduğu görülmüştür (Nagrecha ve diğ., 2018).

Sahtekârlık tespitine yönelik yürütülen çalışmalarda, model performansları kıyaslanırken accuracy değerleri üzerinden kıyaslamasının doğru olmayacağı belirtilmektedir. Araç sigortasında sahtekârlık tespitinde yapılan bir diğer çalışmada kullanılan standart ikili sınıflandırma algoritmalarının (logistic regression, C4.5 decision tree, k-nearest neighbor, Bayesian learning multilayer perceptron neural network, least-squares support vector machine, naive Bayes, and tree-augmented naive Bayes classification) sonuçları kıyaslanırkern PCC (mean percentage correctly classified) ve AUROC (mean area under the receiver operating characteristic) kullanmıştır (Viaene ve diğ., 2002). Ayrıca area under the precision–recall curve ve f1 score un kullanıldığı örneklerde bulunmaktadır (Nagrecha ve diğ., 2018).

Sahtekârlık eğilimi gözlemlenirken yüksek hasar prim oranına sahip gözlemler dikkate alınmaktadır. Tüm üretilen poliçeler, poliçelerin maliyetleri (tazminat talebi tutarı) ve faydaları (ödenen prim) dikkate alınan diğer değişkenlerdir (Nagrecha, n.d.). Otomotiv sektöründe karı düşük ve sahtekârlık olması muhtemel sürücüler belirlenirken hasar prim oranı (loss ratio) değerlerinden yararlanılmaktadır (Nagrecha ve diğ., 2018). Doğal afetlerin sigortacılar üzerindeki etkilerinin incelendiği bir çalışmada; sermaye oranı ve primlerin karlılık üzerinde olumlu etkisi olduğunu tespit edilirken, kayıp oranı, beklenmedik sıklık ve gişe rekorları kıran sigortaların kârlılık üzerinde olumsuz bir etkisi olduğunu vurgulamıştır (Benali ve Feki, 2017). Bu değişkenlerde, sahtekârlık tespitinde dikkate alınabilir.

Sigorta şirketleri sahtekârlık tespitine yönelik çalışmalarını genelde, gelen tazminat taleplerinin geriye dönük sınıflandırılması yöntemi ile gerçekleştirmektedir. Bu yaklaşımın tersini benimseyen, sahtekârlık tepsitinin poliçeler oluşturulurken yapılabilmesine yönelik yapılmış çalışmalar mevcuttur. Nagrecha ve arkadaşları (2018), çalışmasında önerdiği FraudBuster ile üretim veri setinde sahtekârlığın tespit edilmesine yönelik veri güdümlü bir yaklaşım önerilmiştir. Ayrıca poliçenin oluşturulması aşamasında, sigortalama kararının doğruluğu üzerinde insan faktörünün etkisini düşürmeyi hedefleyen, otomatikleştirilmiş sigortalama kontrolü yaklaşımını öneren çalışmalarda bulunmaktadır (Mkrtychev ve Enik, 2019).

3. METODOLOJİ

3.1 Makine Öğrenmesi

Makine öğrenmesi üç ana araştırma odağı etrafında şekillenmiştir (Michalski ve Tecuci, 1994). Bunlar sırası ile;

- Görev odaklı çalışmalar- Önceden belirlenmiş bir görev grubunda performansı artırmak için öğrenme sistemlerinin geliştirilmesi ve analizi (mühendislik yaklaşımı olarak da bilinir)
- Bilişsel simülasyon- İnsan öğrenme süreçlerinin araştırılması ve bilgisayar simülasyonu
- Teorik analiz- Muhtemel öğrenme yöntemleri ve bağımsız algoritmaların alanının teorik olarak keşfi

şeklindedir.

3.1.1 Makine öğrenmesinin türleri

Makine öğrenmesi genellikle iki ana grup ile tanımlanır.

Bunlardan ilki kestirimci (predictive) ya da denetimli öğrenme (supervised learning) yaklaşımıdır. Amaç, etiketli girdi-çıktı çiftleri üzerinden x girdisi ve y çıktısı üzerine bir çıkarımda bulunmaktır. y kategorik değişken olduğunda problem sınıflama yada örüntü tanıma (pattern recognition) şeklinde tanımlanır. y nümerik, gerçek değer aldığında problem regresyon problemi olarak tanımlanır.

Bunlardan ikincisi ise tanımlayıcı (descriptive) ya da denetimsiz öğrenme (unsupervised learning) yaklaşımıdır. Burada sadece girdi değişkeni verilir ve ilginç örüntülerin bulunması amaçlanır. Bazen bilgi keşfi (knowledge discovery) olarak da tanımlanmaktadır.

Makine öğrenmesinin üçüncü türü takviyeli/ pekiştirmeli öğrenme (reinforcement learning) olarak bilinmekledir. Bu yaklaşım ara sıra ödül veya ceza verildiğinde nasıl hareket edileceğini veya davranacağını öğrenmek için kullanışlıdır. Karar teorisi takviyeli/ pekiştirmeli öğrenmenin temelini oluşturur (Murphy, 2012).

3.1.1.1 Denetimli öğrenme

Modelini eğitim verisinde yer alan girdiler ve bu girdilere ait etiketli çıktıları kullanarak kuran ve test verisi üzerinde kurduğu bu modeli doğrulayan/sınayan öğrenme yöntemine denetimli öğrenme denmektedir (Onan, 2015). Denetimli öğrenme yaklaşımı ile yürütülecek çalışmalarda etiketli veriler bulunur, bu etiketli veri kümesi ve uygun algoritmalar kullanılarak analiz yapılır. Denetimli öğrenme regresyon ve sınıflama olmak üzere iki grupta incelenmektedir.

Regresyon yöntemi

Çıktının bir sayı olduğu uygulamalara regresyon denir. Bağımlı değişkenin, bağımsız bir veya birden fazla değişken tarafından modellenmesi şeklinde tanımlanabilir. Bağımsız değişkenlerin katsayıları, bağımlı değişkene etki düzeylerini belirtir (Kılınç, 2018). Regresyon yöntemi sayısal bağımlı değişkenlerin tahmin edildiği hipotezler için uygun bir yöntemdir.

Doğrusal Regresyon Yöntemi: Denetimli öğrenmenin çok basit bir çeşididir. Sayısal cevapların tahmininde çok kullanılan bir araçtır. Bir sabit değişken ve birden çok bağımsız değişkenin parametreler (katsayılar) ile çarpımından oluşmaktadır. Basit doğrusal regresyon sadece bir adet bağımsız değişkenden oluşurken, çoklu doğrusal regresyonda birden fazla bağımsız değişken yer alabilmektedir. Kullanılan çoğu yöntem doğrusal regresyonun geliştirilmiş versiyonları şeklinde kabul edilebilir (James ve diğ., 2013).

Doğrusal Olmayan Regresyon Yöntemi: Doğrusal regresyon tahmin gücü noktasında belirgin kısıtları olan bir yaklaşımdır. Bunun nedeni doğrusallık yaklaşımının, çoğu zaman sadece tahmin/ yaklaşık bir değer bulma eğiliminde olmasıdır. Doğrusal olmayan regresyon, bağımsız değişkenler ile bağımlı değişkenler arasında doğrusal bir ilişki olmadığı durumlarda kullanılmaktadır. Yapay Sinir Ağları (Neural Networks), K-En Yakın Komşular (K-Nearest Neighors), Destek Vektör Makineleri (Support Vector Machines) ve ağaç tabanlı modeller doğrusal olmayan regresyon modellerine örnektirler.

Sınıflandırma Yöntemi

Regresyon modellerinde bağımlı değişkenin nicel olduğu belirtilmiştir. Ancak bazı durumlarda bağımlı değişken nicel değil niteliksel olabilmektedir. Bu duruma örnek olarak müşterinin kullandığı operatörü kullanmaya devam etmesi ya da etmemesi (churn) verilebilir. Genellikle nitel değişkenlerden kategorik değişken olarak bahsedilmekte olup, kategorik (nitel) cevapların tahmin edildiği yöntemlere sınıflandırma yöntemleri denmektedir.

Sınıflandırma algoritmaları, ilgili gözlemin ait olduğu sınıfın bilindiği veri kümesi ile eğitilip, yeni gelen bir gözlemin ait olduğunu sınıfın tahmin edilmesi odağında çalışan yöntemlerdir. Ait olunan sınıfı temsil eden veri kümesinde bulunan sınıf etiketlerinin ikili olduğu yöntemler, ikili sınıflandırma olarak tanımlanırken, ikiden fazla olduğu yöntemlere ise çoklu sınıflandırma adı verilmektedir.

Sınıflandırma yöntemleri doğrusal ve doğrusal olmayan sınıflandırma şeklinde iki kategoride incelenebilir.

Doğrusal Sınıflandırma Yöntemi: Veri kümesinde birbirine benzer özellikler gösteren örneklerin birlikte gruplanması amacıyla verilerin iki veya çok boyutlu düzlem üzerindeki dağılımları üzerinden ayrıştırılmalarını sağlayan yöntemlere sınıflandırma yöntemleri denmektedir. Eğer bu verilerin/ örneklerin dağıldıkları düzlem üzerinde sınıflandırmalarında birbirine uzaklıkları doğrusal ise o zaman doğrusal sınıflandırma teknikleri bu gözlemler için anlamlı sonuçlar verecektir.

Doğrusal olarak ayrıştırılabilen veriler iki boyutlu düzlem üzerinde düz çizgi ile gruplarına ayrıştırılabilirler. Bu yöntem iki boyutlu uzayda doğrusal çizgilerle ifade edilirken, çok boyutlu uzayda düzlemlere (hiper düzlemler (hyper surface) dönüşmektedir (Kılınç, 2018).

Gördüğümüz çoğu sınıflandırma algoritması, sınıfları ayırmak için doğrusal fonksiyonları kullanır. Lojistik Regresyon (Logistic Regression), Doğrusal Dikriminant Analizi (LDA) ve Destek Vektör Makineleri (SVM) doğrusal sınıflandırma algoritmalarına verilebilecek bazı örneklerdir.

Doğrusal Olmayan Sınıflandırma Yöntemi: Veriler doğrusal olarak ayrılmıyorlarsa, doğrusal bir sınıflandırma yöntemi iki sınıfı tam olarak ayırt edemez. Doğru veya doğrusal bir düzlem ile birbirinden ayrılamayan veri dağılımlarını sınıflandırmak için doğrusal olmayan sınıflandırma yöntemleri kullanılmaktadır (Kılınç, 2018). Örneğin

doğrusal olmayan sınıflandırma algoritmalarında biri olan KNN'de karar sınırının (decision boundry) şekline dair herhangi bir varsayım mevcut değildir. Dolayısı ile linear olmayan karar sınırlarında lojistik regresyon ve doğrusal disktiminant analizine kıyasla daha iyi performans gösterir.

3.1.1.2 Denetimsiz öğrenme

Denetimli öğrenmenin aksine denetimsiz öğrenmede, modelde kullanılacak veri kümesindeki örnekler için herhangi bir etiket tanımlaması bulunmamaktadır. Ayrıca bu öğrenme yaklaşımında toplamda kaç adet sınıfın (etiketin) olduğu bilinmemektedir.

Denetimsiz öğrenme temelde n gözlemde ölçülmüş belirli özelliklere (X1, X2, …, Xp) sahip veri setine yönelik bir dizi istatistiksel yöntemi kapsamaktadır. Bağımlı değişken (cevap değişkeni/Y) veri kümesinde yer almadığı için bu öğrenme yaklaşımı tahmin ile ilgilenmemekte, bunun yerine X1, X2.. Xp'deki ölçümler hakkında ilginç şeyler keşfetmek istenmektedir. Burada temel amaç birbiri ile çok benzerlik gösteren birey yada nesneleri aynı kümeye yerleştirmek ve bu kapsamda bir gözlemin diğer gözlemlerle ilgili özelliklerini değerlendirip, birbirleriyle nasıl benzer ya da birbirlerinden nasıl farklı olduklarını belirlemektir.

Literatürde en yaygın kullanımı olan denetimsiz öğrenme yöntemi kümelemedir. Kümeleme analizi çok kullanışlı ve faydalı bir istatistiki veri indirgeme tekniğidir. Kümeleme analizi ilk olarak Tryon tarafından kullanılmıştır (Frank ve Kotthoff, 2019). Bağımlı ve bağımsız değişken şeklinde bir ayrım bu yöntemde bulunmamaktadır. Veriyi sınıflara veya kümelere bölme işlemidir. Birbirine en çok benzeyen gözlemler aynı küme içerisinde yer alacak şekilde analizin gerçekleştirilmesi amaçlanmaktadır. Birbirleriyle karşılaştırıldığında aynı küme içerisindeki gözlemler benzer özellikler taşırken diğer sınıflardaki gözlemler farklı özellik taşırlar (James ve diğ., 2013). Kümeleme faydalı/kullanılabilir değişken kümeleri oluşturmak ve değişkenler arasındaki gizli kalmış özellikleri tanımlamada yardımcı bir yöntemdir.

Kümeleme yöntemleri uzaklık-yakınlık, yoğunluk ve komşuluk gibi hesaplamalar ile veri kümesinde yer alan gözlemler arasında benzerlik yakalamaya çalışmaktadır. Kümeleme yöntemlerinin çoğu gruplama işlemini gerçekleştirirken, örnekler

arasındaki uzaklık-yakınlık hesaplamaları ile elde ettiği uzaklık ölçütü değerlerini kullanmaktadır (Kılınç, 2018).

Uzaklık temelli farklılık ölçütlerinden (distance-based dissimilarity measures) bazıları; Öklid, Manhattan, Minkowski, Canberra ve Chebyshev'dir (Cichosz, 2015). Öklit uzaklığı, iki nokta/ örnek arasındaki uzaklığın ölçülmesinde en çok kullanılan yöntemdir.

Kümelemenin belli başlı özellikleri sırası ile şu şekilde özetlenebilir;

- Küme sayısı bilinmez.
- Kümelere ilişkin herhangi bir bilgi mevcut olmayabilir.
- Kümelemeden elde edilen sonuçlar dinamiktir.
- Aynı kümede yer alan örnekler arasında benzerlik fazladır.
- Farklı kümede yer alan örnekler arasında benzerlik azdır.
- Kümeleme yöntemi veri içinde gizlenmiş örüntüleri bulabilmelidir.

Kümeleme işleminde kullanılan farklı veri gruplama yaklaşımları bulunmaktadır (Aggarwal ve Reddy, 2014). Bunlar sırası ile;

1. Olasılığa Dayalı Yöntemler
2. Uzaklık Temelli Yöntemler
    a. Düz (Bölümleme Tabanlı)
    b. Hiyerarşik
3. Yoğunluk ve Izgara Tabanlı Yöntemler
    a. Yoğunluk Tabanlı
    b. Izgara Tabanlı
4. Spectral Yöntemler

şeklindedir.

Bölümleme tabanlı kümeleme yöntemleri

Bölümleme tabanlı kümeleme algoritmaları belirli bir amaç fonksiyonunu optimize ederek ve bölümlerin kalitesini yinelemeli olarak arttırarak, veriyi temsil eden grupları keşfetmeyi amaçlamaktadır. Bu algoritmalar belirlenen bir merkeze dayalı kümeler oluşturulmaya çalışır. Genellikle algoritma her bir kümeyi temsil eden prototip noktaları seçen belirli parametreler içermektedir. Bu özelliği sebebi ile

prototip tabanlı kümeleme şeklinde de tanımlanmaktadır. Bu yöntemlerde; belirlenen k adet kümeye, veri kümesinde yer alan gözlemler yerleştirilir. K-means, k-Medians ve k-Medoid bu kategoride yer alan algoritmalara örnektir (Aggarwal ve Reddy, 2014).

Hiyerarşik kümeleme yöntemleri

Hiyerarşik kümeleme algoritmaları dendrogram adı verilen ikili ağaç tabanlı veri yapısını oluşturarak kümeleme sorununa yaklaşır. Dendrogram oluşturulduktan sonra, kümeleme algoritmasını yeniden çalıştırmadan, aynı veri kümesi için farklı kümeleme çözümleri elde etmeye yönelik ağaç farklı düzeylerde bölünerek otomatik olarak doğru sayıda kümenin seçimi gerçekleştirilebilir. Hiyerarşik kümeleme, aşağıdan yukarıya ve yukarıdan aşağıya kümeleme olmak üzere iki farklı şekilde gerçekleştirilebilir.

Hiyerarşik kümeleme yöntemlerinin anlaşılması ve gerçekleştirmesi kolaydır. Bölümleme tabanlı kümeleme algoritmalarının aksine küme sayısının (k değerinin) baştan belirlenmesine gerek yoktur. Ancak bu yöntemler veri kümesi içerisinde dışa düşen değerlerin ve gürültünün olduğu ya da veri gruplarının şeklinin küresel olmadığı durumlarda iyi sonuç vermeyebilir (Aggarwal ve Reddy, 2014). Hiyerarşik kümeleme yöntemi iki alt gruba ayrılır:

- Birleştirici Kümeleme (Agglomerative Clustering): İlk olarak, belirli bir yakınlık ölçüsü kullanılarak bir benzerlik matrisi oluşturulur. Tüm örnekler birer küme olarak varsayılır. En yakın kümeler her seviyede birleştirilir ve benzerlik matrisi buna göre güncellenir. Bu birleştirme işlemi tek bir küme kalana kadar sürdürülür.

- Ayrıştırıcı Kümeleme (Divisive Clustering): Ayrıştırıcı hiyerarşik kümeleme yöntemi, tek tek yapraklara kadar hiyerarşinin oluşturulmasına gerek duyulmadığı durumlarda birleştirici hiyerarşik kümeleme yöntemlerinden daha avantajlı olmaktadır. Bu yöntem yukarıdan aşağı bir yaklaşımdır. İlk olarak tek bir kümeyle başlanır. Devamında bütün küme içerisinde birbirine en az benzeyen iki örnek bulunur ve küme bu iki örneğe göre bölünür. Bu şekilde her alt küme için aynı işlem tekrarlanır.

Yoğunluk tabanlı kümeleme yöntemleri

Yoğunluk temelli kümeleme, kümelerin sayısı veya dağılımları hakkında herhangi bir varsayımda bulunmadığından parametrik olmayan bir yöntem olarak

düşünülebilir. Bu yöntem örneklerin ilk dağılımını bir yoğunluk fonksiyonu kullanılarak bulunmaya çalışılır ve belirli bir eşik yoğunluk değerini aşan bölgeleri küme olarak kabul eder. Ayrıca, gürültü alanlarındaki yoğunluğun herhangi bir kümedeki yoğunluktan daha düşük olduğu varsayılmaktadır.

Bu kategoride yer alan kümeleme yaklaşımlarından biri DBSCAN' dir. DBSCAN sabit yarıçaplı bir komşuluktaki nokta sayısını sayarak yoğunluğu tahmin eder ve örneklerin bağlantılı olma durumları (komşulukları) dikkate alarak birbirlerine komşu/ yakın olan iki noktayı bağlı olarak kabul eder.

Küme sayısı belirleme

Küme sayısının belirlenmesinde standart, objektif bir kural bulunmamaktadır. En iyi yöntem değişik küme sayılarına göre analizi yapmak ve içlerinden en mantıklı olan küme sayısını tercih etmektir. Ancak genel olarak dirsek kriterinin (elbow graph) kullanımı literatür çalışmalarında göze çarpmaktadır.

Dirsek kriteri: Bu yaklaşıma göre küme sayısı belirlenirken, mesafe belli bir değeri aştığında ya da ardışık iki mesafe arası belirsiz bir iyileşme sağlanıyorsa durulur. Küme sayısı hakkında teorik bir sınır belirli olabilir.

Kümelerin tanımlanması

Kümeleme analizi sonucunda elde edilen kümelerin tanımlanmasında her bir kümeyi etiketlemek amaçtır. Kümeler tanımlanırken analiz esnasında kıyaslanan değişkenler itibariyle kümeler gözlemlenir. En belirgin değişkenler ve konu spesifiğinde özellikle dikkat edilmesi gerekli olduğu tespit edilen değişkenler saptanır. Kümeler bu saptanan değişkenlere göre karşılaştırılır ve özelliklerine göre isimlendirilir.

Kümelerin yorumlanması

Belirli boyutlarda kümelerin nasıl farklılaştıklarını görebilmek için her bir kümenin nitelikleri tanımlanır. Burada kümelerin ortalama değerlerine bakılarak kümeler birbirleri ile karşılaştırılır. Bu aşamada her bir küme içindeki ortak faktörler aranır. Her bir kümedeki veri noktaları kullanılarak sınıflandırma yapılır.

Kümeleme analizinin farkı varyasyonlarından kazanılan çıkarımlar

Aggrawal ve Reddy (2014), kümeleme algoritmalarının çıktılarının değerlendirilmesinin yanı sıra ya daha titiz analizler kullanarak yahut ek teknikler ve veri girişleri aracılığı ile, genellikle daha derin öngörüler elde edilebileceğini belirtmişlerdir.

Bu yöntemler, kümeleme yaklaşımının sübjektif yapısı ve aynı veri setini kümeleyen birden çok yöntemin olması nedeniyle önemlidir. Kümeleme sonucu elde edilen kümelerin ne kadar iyi olduğu veya ele alınan problemi ne kadar iyi çözdüğü konularını araştırmaya yönelik sayısız yol vardır. Bu araştırma şu yollarla yapılabilir;

- etkileşimli görselleştirme (interactive visualization) ve insan etkileşimi,
- dış bilgi temelli denetim (external knowledge-based supervision),
- farklı olasılıkları değerlendirmek adına birden fazla çözüm yolunun açıkça araştırılması,
- daha sağlam topluluklar oluşturmak adına birden fazla çözüm yolunun birleştirilmesi,
- farklı çözüm yollarının kalitesinin farklı doğrulama (validasyon) kriterleri ile kontrol edilmesi.

Görsel Çıkarım (Visual Insights): Çok boyutlu verilerin görsel analizi, verinin altında yatan yapıyı keşfetmeye yönelik, insanlardan alınan geri bildirimlerinde muhtemelen sürece dahil edildiği sezgisel bir yoldur. Dolayısı ile bu yaklaşım enformal (resmi olmayan) kontrol olarak da kabul edilebilir. İnsanlar ile etkileşime girmenin en büyük avantajı, karmaşık, otomatikleştirilmiş bilgisayar programlarının sağlamasının mümkün olmadığı sezgisel çıkarımları sağlayabiliyor olmalarıdır.

Denetimli Çıkarım (Supervised Insights): Özellikle veri setinin boyutları büyük olduğunda ve alt uzay (subspace) yöntemleri kullanıldığında, veri seti birden fazla yol ile kümelenebilir. Farklı özellikler; yürütülen farklı uygulamalar ve elde edilmek istenen farklı çıkarımlar ile daha alakalı olabilir. Kümeleme çoğu veri madenciliği algoritmasında ara bir adım olarak kullanıldığından, yürütülen uygulamaya uygun özel bir kümelenme türü seçmek zordur. Kümenin öznelliği çok iyi bilinmektedir ve arka planda işleyen algoritmadaki ya da verideki küçük değişimler, alt kümelerde önemli değişikliklere yol açabilmektedir. Çoğu durumda, bu öznellik belirli bir kümelemenin diğerinden önemli ölçüde daha iyi olduğunu ifade etmenin de zor olduğuna işaret eder. Tam bu noktada denetim (supervise) etkin bir rol oynar çünkü analistin özel hedefini dikkate alır.

Çoklu Görüntü ve Topluluk Tabanlı Çıkarım (Multiview and Ensemble-Based Insights): Kümeleme sürecindeki ana konulardan biri farklı tip kümelerin var olma olasılığıdır. Eğer bir denetim yoksa (no supervision), kümeleme sürecinde

karşılaşılan sersemletici sayıdaki olasılık bazen analizi yapan kişi için problem teşkil edebilir. Bunlar alternatif kümeleri ifade eder ve teknik olarak farkı perspektiflerden elde edilen sonuçları temsil eder. Çoğu durumda birbirinden oldukça ayrışmış farkı kümeleme sonuçları elde etme kabiliyeti, analizi yapan kişinin verinin altında yatan anahtar kümeleme özellikleri hakkında çıkarım yapmasını sağlar. Bu geniş alan çok görüntülü kümeleme (multiview clustering) olarak ta adlandırılmaktadır. Çok görüntülü kümeleme için en sade metot, basitçe kümeleme algoritmasının birçok defa işletilmesi ve hangilerinin farklı olduğunu tespit etmek üzere farklı kümeleme sonuçlarının incelenmesidir. Alternatif kümelenme yöntemleri, topluluk temelli tekniklerin kullanımıyla daha sağlam çözümler oluşturmak için birleştirilebilir. Burada amaç, farklı kümeleme sonuçlarının çıktılarının kombinasyonunun, noktaların birbirleri ile nasıl ilişkili olduğuna dair daha sağlam bir resim ortaya koymasıdır.

Doğrulamaya Dayalı Çıkarım (Validation-Based Insights): Kümenin kalitesini anlama noktasında kullanılabilecek yöntemlerden biri girdi ve çıktı kümeleri arasındaki eşleşmeyi belirlemek için sentetik verileri kullanmaktır, sadece sentetik verilere dayanmak tam olarak yeterli değildir. Bunun nedeni sentetik veri üzerinden elde edilen sonuçların belirli bir algoritma odağında olabileceği ve rastgele veri setlerine kolayca genelleştirilemeyeceğidir. Bu nedenle, doğrulama kriterlerinin gerçek veri setleri üzerinde kullanılması tercih edilmektedir.

Kümeleme problemi konteksti içerisinde vurgulanan problem; kalite için belirlenen kriterlerin, etiket şeklinde harici doğrulama kriterlerinin mevcut olduğu sınıflandırma örneğinde olduğu gibi diğer birçok veri madenciliği tekniğindeki kadar net olmamasıdır.

Kümeleme, denetimsiz (unsupervised) doğası nedeniyle kesin ölçümün çoğu zaman mümkün olmadığı bir problemdir. Bununla birlikte, birçok teknik kümelerin altında yatanlar hakkında kısmi bir açıklama sağlamaktadır. Kümelerin doğrulanmasında yaygın olarak kullanılan bazı teknikler şunlardır:

- Literatürde yaygın olarak kullanılan yöntem, kümelerin kalitesini subjektif olarak göstermede vaka çalışmalarının kullanmasıdır. Vaka çalışmaları iyi sezgisel çıkarımlar sağlamaktadır ancak kalitenin daha titiz bir şekilde ölçülmesinde herhangi bir katkısı yoktur. Böyle bir yaklaşımın

kullanılmasıyla iki kümeleme yöntemini nicel bir bakış açısıyla karşılaştırmak genellikle zordur.

- Kalite ölçüsü sağlamak için; küme yarıçapı veya yoğunluk gibi kümelerin belirli ölçümleri kullanılabilir. Buradaki sorun, bu önlemlerin farklı algoritmaları farklı bir biçimde destekleyebilmesidir. Örneğin, bir k-means yaklaşımı tipik olarak ortalama küme yarıçapı cinsinden yoğunluğa dayalı bir kümeleme yöntemine göre daha üstün olacaktır, ancak yoğunluk-esaslı bir yöntem kümelerin tahmini yoğunluğu bakımından bir k-means algoritmalarından daha üstün olabilir. Bunun nedeni, aynı ölçüt kümeleme amacıyla kullanıldığında algoritmayı değerlendirmek için belirli bir ölçüt kullanmanın bir döngüselliği olmasıdır. Bu, değerlendirme sırasında yanlılığa neden olur. Kısaca faklı iki algoritmayı belirli bir kriter üzerinden kıyaslamak mümkün olmayacaktır. Öte yandan, bazen çok benzer bir tipteki iki farklı algoritmayı belirli bir kriter temelinde makul bir şekilde karşılaştırmak mümkün olabilir (örneğin, iki k-means varyasyonu).
- Birçok veri setinde, etiketler veri noktalarıyla ilişkilendirilebilir. Bu gibi durumlarda, küme kalitesi kümelerin veri etiketleri ile korelasyonlarına göre ölçülebilir. Bu, etiketleme kümeleme işleminde kullanılmamışsa, harici bir doğrulama kriteri (external validation criterion) sağlar. Bununla birlikte, böyle bir yaklaşım mükemmel değildir, çünkü sınıf etiketleri her zaman verideki doğal kümelerle uyumlu olmayabilir. Bununla birlikte, yaklaşım yukarıda tartışılan diğer iki yöntemden daha "tarafsız" olarak kabul edilir ve genellikle küme değerlendirmesinde kullanılır (Aggarwal ve Reddy, 2014).

Oluşturulan kümelerin değerlendirilmesi

Kümeleme modellerinde, modelin güvenilirliğinin değerlendirmesine yönelik, kümelerin kalitesinin objektif bir yöntem ile nasıl değerlendirileceği net değildir.

Kümelemede önceden belirlenmiş hedef değerler yer almadığından, kümeleme performansını ölçen performans ölçütleri, modelin, verilerin sergilediği benzerlik desenlerini ne kadar iyi yakaladığını ölçmeye odaklanmıştır. Bu, ilgili değişkenin tahmin edilen değerleri ve gerçek hedef değerinin kıyaslanması yoluyla yapılan sınıflama ve regresyon modellerinin tahmin performanslarını ölçmekten çok farklıdır. Bu temel farktan kaynaklı olarak kümeleme performans ölçümlerine kalite

performans ölçümleri de denmektedir. Burada hedef, eğitim verisi üzerinden bir çıkarım elde etmektir (Cichosz, 2015).

Doğrulamaya dayalı yaklaşım altında kümeleme modelleri değerlendirilirken temelde içsel (internal) ve dışsal (external) kalite ölçümleri şeklinde sınıflanmaktadır. İçsel ve dışsal kalite ölçütleri literatürde içsel ve dışsal doğrulama ölçütleri olarak ta geçmektedir. İçsel kalite ölçütleri ayrıca her küme için yapılan kalite ölçümleri (Per-cluster quality measures) ve toplu kalite ölçümleri (Overall quality measures) şeklinde ikiye ayrıştırılabilmektedir. Belirtilen kalite ölçüm yaklaşımları sırası ile incelenmiştir.

Küme Başına Kalite Ölçüleri (Per-cluster quality measures): Belirli bir kümenin uyum düzeyini ve diğer kümelerden ayrılma seviyesini tanımlamak için tasarlanmış kalite ölçümleridir (Cichosz, 2015).

- Çap (Diameter): Kompakt kümeler küçük diameter değerlerine sahip olurlar. Eğer aynı kümeleme modelinden oluşmuş kümelerden bazılarının diameter değerleri ortalamanın üstünde kalıyorsa, o zaman küme sayısının çok az olduğu anlaşılır.
- Ayırma (Separation): Kümenin ayırımı, onun diğer kümelerden ne derece ayrıldığını ölçer. Uniform ve yüksek küme ayrım değerleri istenir.
- İzolasyon (Isolation): İzolasyon, bir kalite ölçütünden ziyade, özellikle kompakt ve ayrık kümelerde beklenen bir özelliktir. İzolasyon özelliği kümenin diameter değeri separation değerleri ile ilişkilidir. İyi kümenin diameter değeri separation değerinden düşük olur buda kümenin izolasyon özelliğine işaret eder. Herhangi bir küme üyesi için aynı kümenin diğer herhangi bir üyesine maksimum farklılığı, diğer kümelerdeki herhangi bir örneğe olan minimum farklılığından daha az ise daha güçlü bir izolasyon oluşmuş demektir.
- Silhouette genişliği (Silhouette width): Kümeye atanan gözlemin küme ile ne kadar uyumlu olduğu göstermektedir. [-1,1] arasında değerler almaktadır. (1: Küme ile iyi eşleşmiş, 0: İki kümenin sınırları üstünde, -1: Komşu küme ile daha iyi eşleşmiş)

- Davies–Bouldin indeksi (Davies–Bouldin index): Tüm gözlem çiftlerinin farklılıklarının ortalaması yerine küme merkezlerinin farklılıklarının ortalamasını göstermektedir. İyi ayrışmış kümeler küçük değerler alır. Gürültülü verilere karşı dirençli bir ölçüttür.

Toplu Kalite Ölçüleri (Overall quality measures): Küme başına kalite ölçütleri, değerlendirilen kümelenme modelinin kalitesi hakkında faydalı bilgiler sunsa da, genellikle istenen, bir dizi alternatif modeli doğrudan karşılaştırmak için kullanılabilecek tek bir kalite göstergesidir. Bu isteri karşılama noktasında, Toplu Kalite Ölçütleri, tüm kümeleri göz önünde bulundurarak, bütünlük ve ayrılma göstergelerini anlamlı bir şekilde birleştirmeyi amaçlamaktadır (Cichosz, 2015).

- Dunn indeksi (Dunn index): Kümeleme modelinin kalitesini ölçer. Yüksek değerler tercih edilir. Uç değerlere hassasiyeti vardır.
- Ortalama Davies–Bouldin indeksi (Average Davies–Bouldin index): Küme başına elde edilen değerlerin ortalaması alınarak hesaplanır. Küçük değerler tercih edilir. Ortalama indeks değeri farklı modellerin kıyaslamasında kullanılabilir. Örneğin; k-means kümelemede farklı k değerlerinin kıyaslamasında kullanılabilir.
- C indeksi (C index): C indeksi kümeleme modellerini değerlendirirken, küme üyeliği hesaba katılmadan, küme içi çiftler arası farklılıkları doğrudan veri kümesindeki gözlenenlerle karşılaştırarak hesaplar. Sonuçlar [0,1] aralığında değişmektedir. Küçük değerler tercih edilir. Küçük değerler, kümeleme algoritmasının içten bağımlı (internally cohesive) kümeler elde edildiğine işaret eder.
- Ortalama silhoutte genişliği (Average silhouette width): Eğer çoğu gözlem bulunduğu kümede doğru şekilde konumlandırılmış ise yüksek değerler elde edilir.
- Loglikelihood: Değerlendirilen model ile değerlendirme için kullanılan veri kümesi arasındaki eşleşme derecesini olasılıkla değerlendirir. Yüksek değerler modelin data ile uyumlu olduğunu gösterir.

Dışsal Kalite Ölçümleri (External quality measures): Bu yöntem küme etiketlerinin olduğu dış bir data ile elde edilen kümeleme sonuçlarını kıyaslamakta ve modelin bu şekilde performansını ölçmeyi hedeflemektedir.

Bu ölçütler sırası ile Yanlış Sınıflandırma Hatası (Misclassification error), Rand indeksi ve Genel İlişkilerin Tespiti Ölçütlerinden (General relationship detection measures) oluşmaktadır.

3.1.1.3 Pekiştirmeli öğrenme

Etkileşimden öğrenme neredeyse bütün öğrenme ve zeka kuramlarının altında yatan temel fikirdir. Pekiştirmeli öğrenme temelde, etkileşimden öğrenmeye yönelik sunulmuş sayısal/ hesaplamalı bir yaklaşımıdır. Makine öğrenmesinde kullanılan diğer yaklaşımlar ile kıyaslandığında pekiştirmeli öğrenme yaklaşımında etkileşimden hedefe yönelik öğrenmeye daha fazla odaklanılmış bulunulduğu tespit edilmektedir.

Pekiştirmeli öğrenmede, öğrenene, çoğu makine öğrenmesi algoritmasında olduğu gibi hangi eylemlerin gerçekleştirileceği söylenmez. Bu yaklaşımda sistem, doğru kararlar için ödüllendirilirken, yanlış kararlar için cezalandırılır (ya da daha az ödüllendirilir). Sistemin hedefi sayısal ödül sinyallerini maksimum düzeye çıkarmaktır. Pekiştirmeli öğrenmede hangi eylemlerin olması gerektiği, deneyerek ve fazla ödülü vereni keşfederek bulunur. Dikkat edilmesi gereken bir diğer nokta eylemler sadece o andaki durumu etkilemek ile kalmayıp daha sonraki adımları ve ödül değerlerini de etkilemektedir. Pekiştirmeli öğrenmede; deneme yanılma ile tüm durumların araştırılması ve rötarlı ödüller (diğer adımlardaki ödülleri etkileme durumu), yöntemi diğer yöntemlerden ayıran iki temel karakteristik olarak tanımlanmıştır. (Sutton ve Barto, 1998).

Pekiştirmeli öğrenme yaklaşımının 4 temel bileşeni bulunmaktadır (Kılınç, 2018);

1. ‘Davranış prensipleri (politikalar): Sistemin gerçekleştirebileceği aksiyonları belirler.
2. Ödül: Gerçekleştirilen bir aksiyona karşılık alınan puandır.
3. Değer Fonksiyonu: En yüksek ödülü almak için uzun vadeli stratejilerin temelini oluşturur. Ödülü düşük bir aksiyonun değeri yüksek olabilir. Amaç toplamda maksimum ödülü alabilmektir.
4. Model: Simülasyon yapabilmek için oluşturulan bir çevre (ortam) modelidir.’

Bu öğrenme yaklaşımının günümüzde en çok kullanıldığı alanlar; oyun programlama, fabrika otomasyon kontrolleri ve robotiktir

3.1.1.4 Makine öğrenmesi modellerini geliştirme yolları

Kullanılan makine öğrenmesi algoritması, öğrenme işlemini tamamladığında, modelin test veri setinde vereceği sonuçlar merak edilir ve modelin daha iyi nasıl geliştirebileceği arayışına girilir. Mueller, J.P ve Massaron, L. makine öğrenmesi modellerinin geliştirilmesine yönelik 10 yol tanımlamıştır. Bunlar sırası ile (Frank ve Kotthoff, 2019);

1. Öğrenme eğrilerinin incelenmesi,
2. Doğru şekilde cross-validation uygulanması,
3. Doğru hata (error) ve skor (score) ölçütlerinin seçilmesi,
4. En iyi hiper parametrelerin araştırılması,
5. Birden fazla modelin test edilmesi,
6. Modellerin ortalamalarının alınması,
7. Modellerin yığılması,
8. Değişken türetilmesi (feature engineering),
9. Değişkenlerin ve örneklerin seçimi,
10. Data fazla veri bulunmasıdır.

Hiper Parametre Optimizasyonu (Hyperparameter optimization)

Model parametreleri ve hiperparametreler birbirlerinden ayrı kavramlardır. Model parametreleri, istenen çıktıyı elde etmek için girdi verilerinin nasıl kullanılacağını tanımlar ve eğitim sırasında öğrenilir. Hiperparametreler ise modelin nasıl yapılandırıldığını belirler. Makine öğrenmesi modellerinin ayarlanması bir tür optimizasyon problemidir. Modelde birtakım hiperparametreler mevcuttur ve bir fonksiyonun minimum veya maksimumlarını bulmamıza yardımcı olabilecek hiperparametre değerlerinin doğru kombinasyonunun bulunması amaçlanmaktadır. Hiperparametre optimizasyonuna yönelik kullanılan yöntemler sırası ile incelenmiştir.

Izgara arama (Grid Search): Izgara arama en basit hiperparametre optimizasyonu yöntemidir. Aynı zamanda tam etkenli tasarım (full factorial design) olarak da tanımlanmaktadır. Tam etkenli tasarım, yapılan denemelerde faktör setlerinin olası tüm kombinasyonlarının denendiği bir deneyi tanımlamaktadır.

Kullanıcı her hiperparametre için sınırlı sayıda değer kümesi belirler ve ızgara arama bu kümelerin kartezyen çarpımlarını (cartesian product) değerlendirir. (Yani burada da full factorial ile aynı mantıkta incelenmesi istenen tüm hiperparametler için belirlenen değerlerin tüm kombinasyonları tek tek denenmektedir). Ancak bu

yaklaşımda curse of dimensionality problemi ile karşılaşılır çünkü gerekli olan fonksiyon değerlendirmesi sayısı, konfigürasyon uzayının boyutu ile beraber üssel olarak (katlanarak) büyür.

Izgara arama yaklaşımının bir diğer problemi, kesikli hale getirme kararının (resolution of discretization) artması, gerekli olan fonksiyon değerlendirmesi sayısını önemli ölçüde artırır.

Rassal arama (Random Search): Izgara aramanın basit alternatifi rassal aramadır. İsminden de anlaşılacağı gibi araştırma için ayrılan bütçe (zaman vb.) tükenene kadar konfigürasyonlardan rastgele örneklem seçilir. Bu yaklaşım bazı hiperparametreler diğerlerinden çok daha önemli olduğu durumlarda ızgara aramadan daha işlevseldir.

Izgara arama ile kıyaslandığında avantaj sağladığı diğer noktalar, kolay paralelleşebilmesi ve kaynakların esnek bir şekilde düzenlenebilmesidir.

Rassal arama yararlı bir temel oluşturur, çünkü makine öğrenme algoritmasının optimize edildiğine dair hiçbir varsayımda bulunmaz ve yeterli kaynak verildiğinde, beklentiye göre keyfi olarak optimum seviyeye yakın bir performans elde eder.

Güdümlü arama (guided search) metotlar ile kıyaslandığında en yüksek performans gösteren hiperparametre konfigürasyonunu bulmada daha uzun süre harcamaktadır.

Rastgele aramanın (Random search), optimumunun bulunması için beklenen $2N - 1$ işlev değerlendirmesini gerektirirken, yönlendirmeli arama (guided search) $N + 1$ işlev değerlendirmesi ile en iyi sonucu bulabilir.

Güdümlü arama metotları genel olarak rassal arama metotlarından daha iyi performans gösterir (Frank ve Kotthoff, 2019).

Bayes arama (Bayesian Search): Bayes optimizasyonu, uzun zaman alan (pahalı) blackbox yöntemlerinin global optimumunu bulmaya yönelik ortaya konan en gelişmiş optimizasyon çerçevesidir.

Izgara arama veya rassal arama yaklaşımlarında hiperparametrelerin geçtikleri yolun çok akıllıca olmadığı söylenebilir. Her iki yöntemde de eğitim süresi boyunca doğrulanacak hiperparametre kümeleri önceden tanımlanır ve eğitim sırasında alabilecekleri bilgilerden faydalanmak üzere tasarlanmamıştır. Model performansına dayalı olarak önceki hiperparametre doğrulama yinelemelerinden öğrenmenin bir yolunun bulunması, hangi hiperparametre kümesinin bir sonraki yinelemede daha iyi

bir performans verebileceği hakkında bir fikir ortaya koyacaktır. yaklaşımlar Sıralı Mod Tabanlı Global Optimizasyon (SMBO) bu muhakemeden doğmuştur ve Bayes temelli hiperparametre optimizasyonu bu yaklaşımlardan biridir. Model tabanlı hiperparametre ayarı söz konusu olduğunda kullanılacak birkaç yaklaşım vardır ve bu yaklaşımlar SMBO altında bir araya gelir (S Das, n.d.).

3.2 Anomali Tespiti (Anomaly Detectıon)

Süreçlerde normalden sapmalar gözlemlenebilir ve bunun sebebi karşılaşılan anormal durumlar olabilir. Bu sapmalar, böyle bir proses sapması yaşanmadığında gözlemlenen değerlerden farklı gözlem değerleri elde edilmesine yol açar. Anomali tespitinin görevi, gözlenen veri değerlerinde (normdan) bu tür sapmaları bulmak ve buna sebep olan süreçteki farklılıkları ortaya çıkarmaktır (Mehrotra ve diğ., 2017).

Temel bir sorun, iki veri noktasının ne kadar benzer olduğunu değerlendirmemize izin veren basit bir tanımın olmaması ve dolayısıyla bir veri noktasının veri setindeki diğer veri noktalarından nasıl farklılaştığının bilinmemesidir (Papadimitriou ve diğ., n.d.). Çok sayıda veri noktası göz önüne alındığında, noktalardan hangilerinin ortalamadan önemli ölçüde farklı olduğunu bulmak anomali tespitinde bizleri yönlendirmektedir. Dağılıma iyi uyan veri noktaları yüksek uyum olasılıklarına sahipken, anomali durumlar çok düşük uyum olasılıklarına sahip olmaktadırlar.

Mehrotra K, 2017 kitabında anomali tespit yaklaşımlarını sırası ile Uzaklık Temelli Anomali Tespit Yaklaşımları (Distance-Based Anomaly Detection Approaches), Kümelenmeye Dayalı Anomali Tespit Yaklaşımları (Clustering-Based Anomaly Detection Approaches) ve Model Tabanlı Anomali Tespit Yaklaşımları (Model-Based Anomaly Detection Approaches) şeklinde belirtmiştir (Mehrotra ve diğ., 2017).

Kümelenmeye dayalı anomali tespit yaklaşımı kapsamında kümelerin tespitinde kullanılan yöntemler sırası ile;

- En Yakın Komşu Kümeleme (Nearest Neighbor Clustering)
- K-Ortalamalar Kümesi(k-Means Clustering)
- Bulanık Kümeleme (Fuzzy Clustering)
- Yığınsal Kümeleme (Agglomerative Clustering)
- Yoğunluk Temelli Kümeleme (Density-Based Agglomerative Clustering)

- Bölücü Kümeleme (Divisive Clustering)

şeklindedir.

Kümelerden yararlanarak anomali tespitinde (Anomaly Detection Using Clusters) dikkat edilen unsurlar;

- Küme Üyeliği Ya Da Büyüklüğü (Cluster Membership or Size)
- Diğer Noktalara Yakınlığı (Proximity to Other Points)
- En Yakın Komşuya Yakınlık (Proximity to Nearest Neighbor)
- Sınıra Mesafesi (Boundary Distance)
- Küme Boyutlarının Farklı Olması (When Cluster Sizes Differ)
- Birden Çok Noktadan Mesafeler (Distances from Multiple Points)

şeklinde sıralanmıştır.

Anomali tespiti; hastalık tespiti, kredi kartı sahtekârlığı tespiti ve mevcut paternler arasında herhangi bir yeni patern tespiti gibi geniş bir veri madenciliği problemleri havuzundan oluşur. Bu örneklere ek olarak, üretimde hataların veya anormallikleri tespit edilmesinde de kullanılabilir.

Kullanılabilir etiketlerin olmadığı durumlarda, sahtekârlık tespiti (fraud detection) bir anormallik (anomali) problemi olarak formülize edilebilir. Etiketsiz verilerde sahtekârlık tespitinde anomali, çoğunlukla sahtekârlık olarak tanımlanmayan gözlemlerden oluşan bir veri setindeki sahtekârlık durumlarını temsil eden gözlemlere karşılık gelecektir. Diğer bir nokta ise sahtekârlık yapanların, belirli bir tespit metodunun farkına vardıktan sonra kendi stratejilerini bu tespit metoduna yakalanmayacak şekilde değiştirmeleridir (Sparrow, 1996). Dolayısı ile mevcutta belirli sahtekârlık tanımlamaları olsa dahi her daim yeni tanımların oluşmuş olması olasılığını dikkate almak üzere denetimsiz öğrenme tekniklerinden de yararlanılmaktadır.

Standart denetimsiz anomali saptama yöntemleri, kümeleme analizini ve uç değerleri saptama yöntemlerini içerir (Friedman ve diğ., 2001). Literatürde çok sayıda uç değerin tespitine yönelik metot önerilmiştir; k-En Yakın Komşu (k-Nearest Neighbor-k-NN) outlier tespiti, Tek Sınıflı Destek Vektör Makinesi (One-Class Support Vector Machine - OC-SVM) (çekirdek temelli dahil) ve yoğunluk temelli

yöntemler Local Outlier Factor (LOF) gibi. Bu yöntemlerin etkinliği, ağa izinsiz girişlerin tespiti, kredi kartı sahtekârlığının tespiti ve elektronik ticarette anormal aktivitelerin tespiti dahil birçok uygulama alanında araştırılmıştır.

Bunun yanı sıra uç değerlerin tespitine yönelik birçok metot, global uç değerler olarak ifade ettiğimiz sadece tek bir global paterne işaret eden uç değerlerin tespiti için uygundur.

Yoğunluk temelli yöntemler hem global uç değerlerin hem de yerel uç değerlerin tespitinde etkili olmakta, ancak buradaki varsayım, yoğunluğun anormallik için tek ayrıştırıcı (discriminant) unsur olduğudur. Küçük yoğunluklu kümeler de anormallik oluştuğunda, yoğunluğa dayalı yöntemler başarısız olur. Ek olarak, yoğunluk temelli yöntemler, örneğin, LOF, genellikle kullanıcıların yoğunluğunu karşılaştırmak için bir komşuluk aralığı (neighborhood range) tanımlamasını gerektirir.

Yukarıda yer verilen paylaşımlardan da anlaşılabileceği üzere, denetimsiz öğrenme denetimli öğrenmeden çok daha zordur bunun temel sebebi öğrenme hedefleri öğrenme sürecine rehberlik etmek için uygun değildir. Uygulama aşamasında, denetimsiz öğrenmedeki zorluk düzeyi, denetimsiz öğrenme yöntemleri için ilgili özelliklerin (feauture) tanımlanmasındaki ilave zorluklarla daha da artmaktadır (Nian ve diğ., 2016).

3.3 K-Ortalamalar (K-Means)

K-ortalamalar kümeleme yöntemi, veri setini k farklı birbiri ile çakışmayan kümeye ayıran basit bir yaklaşımdır. K- ortalamalar kümelemesini gerçekleştirmek için, öncelikle verinin ayrıştırılmak istendiği küme sayısı (k) belirtilmelidir; devamında k-ortalamalar algoritması her gözlemin k kümeden birine atamasını gerçekleştirir. K-ortalamalar kümeleme yöntemi basit ve sezgisel bir matematiksel problemin sonucudur.

Bazı notasyonları tanımlayarak problemi inceleyecek olursak; C_1, C_2,…, C_k gözlemlerin ait oldukları kümeleri gösterir.

1. $C_1 \cup C_2 \cup \ldots \cup C_k = \{1,\ldots,n\}$. Her gözlem en az bir kümeye aittir.

2. $C_k \cap C_{k'}$ tüm $k \neq k'$ değerleri için. Bir gözlem birden fazla kümenin elemanı olamaz.

Örneğin eğer i. gözlem k. kümede ise $i \in C_k$ şeklinde gösterilir.

K-ortalamalar arkasındaki ana fikir, iyi kümeleme yapıldıysa küme içi değişkenliğin olabildiğince küçük olması gerektiğidir. C_k kümesi için küme içi değişkenlik, $\mathrm{W}(C_k)$ ile ölçülür. Çözülmek istenen problemin formülizasyonu şu şekildedir;

$$\underset{C_1, C_2, \ldots, C_k}{\text{minimize}} \left\{ \sum_{k=1}^{K} \mathrm{W}(C_k) \right\} \tag{3.1}$$

Bu formülün bize söylediği; gözlemleri toplam küme içi değişkenlikleri tüm K kümeleri için olabildiğince düşük olacak şekilde, K kümeye ayırmaktır.

1.denklemin çözülmesi mantıklı bir fikirdir ancak bunun çözülebilmesi için küme içi değişkenliğin tanımlanması gerekir. Bunun açıklanması noktasında çeşitli yaklaşımlar mevcuttur, ancak en fazla tercih edileni Öklid uzaklığının kareleridir. Şu şekilde tanımlanabilir;

$$\mathrm{W}(\mathrm{C_k}) = \frac{1}{|\mathrm{C_k}|} \sum_{\mathrm{i,i'} \in \mathrm{C_k}} \sum_{\mathrm{j=1}}^{\mathrm{p}} (\mathrm{x_{ij}} - \mathrm{x_{i'j}})^2 \tag{3.2}$$

$|\mathrm{C_k}|$ burada k kümede kaç adet gözlem olduğunu göstermektedir. Bir başka ifadeyle k. kümenin küme içi değişkenliği, k. kümede yer alan tüm gözlemlerin ikili Öklid uzaklıklarının karelerinin toplamıdır. 1. ve 2. formüllerin kombinasyonu k-ortalamalar kümeleme yönteminin optimizasyon problemini vermektedir.

$$\underset{C_1, C_2, \ldots, C_k}{\text{minimize}} \left\{ \sum_{k=1}^{K} \frac{1}{|\mathrm{C_k}|} \sum_{\mathrm{i,i'} \in \mathrm{C_k}} \sum_{\mathrm{j=1}}^{\mathrm{p}} (\mathrm{x_{ij}} - \mathrm{x_{i'j}})^2 \right\} \tag{3.3}$$

3 numaralı hedefi en aza indirilecek şekilde gözlemleri k kümeye ayıran bir algoritma k-ortalamalar tarafından yürütülmektedir. K-ortalamalar optimizasyon probleminin local optimumumu sağlayacak basit algoritmanın adımları aşağıda belirtildiği gibidir.

K-Ortalamalar Kümeleme Adımları

1. Her bir gözlem için rastgele 1'den K'ye bir sayı atar.

Bunlar gözlemler için ilk küme ataması görevi görür.

2. Küme atamaları değişmeyi bırakana kadar/sabit kalana kadar tekrar eder.

(a) k kümelerinin her biri için küme merkezleri (centroid) hesaplanır.

k kümesinin merkezi, p özelliğin k. kümedeki gözlemler için ortalamasını tutan vektördür.

(b) Her bir gözlem merkezini en yakın olan kümeye atar (en yakın küme Öklid mesafesi kullanılarak tanımlanır).

Algoritma her adımda, hedefin değerini düşürmeyi garanti eder. Bunun sebebini anlamak için aşağıdaki tanımlama yapılabilir:

$$\frac{1}{|C_k|}\sum_{i,i'\in C_k}\sum_{j=1}^{p}(x_{ij}-x_{i'j})^2 \;=\; 2\sum_{i,i'\in C_k}\sum_{j=1}^{p}(x_{ij}-\bar{x}_{kj})^2 \qquad (3.4)$$

$\bar{x}_{kj} = \frac{1}{|C_k|}\sum_{i\in C_k} x_{ij}$ burada C_k kümesindeki j özelliğinin ortalamasını temsil etmektedir.

Adım 2 (a) 'da, her bir özelliğin küme ortalaması, sapmaların kare toplamlarını minimize eden sabitleridir. Adım 2 (b) 4. denklemi iyileştirmek için gözlemleri yeniden konumlandırır. Bu, algoritma çalıştırıldığında, elde edilen kümelenmenin, sonuç artık değişmeyene kadar sürekli olarak iyileşeceği anlamına gelir; (3) asla artmayacaktır. Sonuç artık değişmediğinde, yerel bir optimum değere ulaşıldığı anlaşılır.

K-ortalamalar algoritmasının bulduğu değer global optimum değil yerel optimum değer olduğu için sonuçlar, 1.adımda gözlemlere rastgele yapılan küme atamalarına bağlı olmaktadır. Bu sebeple algoritmayı birden fazla defa, farklı rassal başlangıç atamaları ile çalıştırmak önem arz etmektedir. Devamında en iyi çözüm seçilir, yani amaç (3) en küçük olanı seçmektir(James ve diğ., 2013).

K-ortalamalar algoritmasının sahip olduğu avantaj ve dezavantajlar aşağıdaki gibi sıralanabilir (Kılınç, 2018);

Avantajlar

- “Anlaması ve gerçekleştirmesi oldukça kolaydır.

- Hesaplama karmaşıklığı diğer kategorideki yöntemlere göre daha azdır.
- Büyük veri kümelerinde hızlı çalışır.

Dezavantajlar

- Küme sayısının (k değerinin) baştan belirlenmesi gerekmektedir.
- Veri kümesi içerisinde aykırılıkların ve gürültünün olduğu ya da veri gruplarının şeklinin küresel olmadığı durumlarda iyi sonuç vermeyebilir.
- Küme merkezleri her seferinde rassal seçildiği için algoritma her çalıştığında farklı sonuçlar üretir."

3.4 Ağaç Temelli Modeller (Tree-Based Models)

3.4.1 Karar ağaçları (Decision trees)

Karar ağaları hem regresyon (Regression trees) hemde sınıflandırma (Classification trees) problemlerinde kullanılabilmektedir. Regresyon ağaçları ilk olarak tahmin uzayını birbiri ile çakışmayacak şekilde çok boyutlu dikdörtgenlere bölmektedir. İkinci olarak ilgili bölgeye düşen her gözlem için aynı tahmin gerçekleştirilmektedir. Burada temel hedef basitlik ve kolay yorumlanabilirliği sağlamaktır. Açgözlü bir algoritma (greedy algorithm) olarak tanımlanan karar ağaçları her adımda en iyi bölümü yaparak ilerler, ileriki adımlarda en iyi sonucu verecek olanı beklemez (James ve diğ., 2013).

3.4.2 Bagging

Karar ağaçlarının en büyük problemlerinden biri çok yüksek varyansa sahip olmalarıdır. Verideki küçük bir değişim bölünmelerde (split) büyük değişimlere sebep olmaktadır. Bu durum sonucunda model istikrarsız bir şekilde yorumlanabilmektedir. Bu problemi çözebilmek adına Bagging kullanılmaktadır. Bagging, Boostrap Aggregating'in kısaltılmış halidir. Bagging boostrap örneklemesini kullanır ve oluşan bireysel modellerin ortalamasını alarak toplulaştırır. Ortalama alma ensemle model oluşturmanın en basit yöntemlerindendir. Bagging yaklaşımı aşırı uyumu (overfitting) önler (James ve diğ., 2013).

Boostrap Sampling (Boostrapling): en basit şekli ile tanımı, eğitim (train) veri setinden yenileme (replacement) ile satırların rastgele örneklenmesidir. Yenileme ile örneklemlerin seçimi, bir eğitim verisi örnekleminin birden fazla defa seçilmesi olasılığını taşıdığı anlamına gelir. Bazı satırların birden fazla kullanıldığı ve bazı satırların hiç kullanılmadığı bu sonuçlar, eğitim veri setinin modifiye edilmiş

versiyonlarıdır. Bu yaklaşım başlangıç verisine benzer veriler türetme noktasında yardımcı olur. Böylece çok sayıda birbirine benzer model oluşturulabilmektedir.

Bagging Adımları:

- 1.Adım: Orijinal eğitim veri setindne B adet yinelemeli örneklem seç. (B<= n (eğitim veri setinde bulunan toplam gözlem sayısı))
- 2.Adım: Oluşturulan boostrap örneklemleri üzerine karar ağaçlarını oluştur.
- 3.Adım: 1. ve 2. adımları istenilen adette tekrarla.
- 4.Adım: Oluşturulan her ağacın tahmin değerinin ortalamasını al ve final tahmini elde et.

3.4.3 Rastgele Orman (Random Forest)

Rastgele ormanın boostrap'ten farkı ağaçları oluştururken tüm değişkenleri dikkate almaktansa bu değişkenlerin alt kümelerini (subset) dikkate alıyor olmasıdır. Bu yaklaşımda sadece birkaç değişken bölünme (split) için aday olmaktadır. Böylece ağacın bölünmesinde baskın olan değişkenleri her ağacı domine etmesinin önüne geçilmektedir (James ve diğ., 2013).

Çizelge 3.1 : Rastgele Orman hiper parametreleri.

ntree:	ağaç sayısı
mtry:	her bölünme için seçilen değişken sayısı (default: sınıflandırma modeli için > toplam değişken sayısının karekökü)
sampsize:	eğitilecek örneklemlerin büyüklüğü
nodesize:	son düğümde (terminal node) yer alacak minimum gözlem sayısı
maxnodes:	son düğümde (terminal node) yer alacak maksimum gözlem sayısı

3.4.4 Boosting

3.4.4.1 AdaBoost

Adaboost Adımları:

1. Adım: Karar ağacının eğitilmesine başlarken her gözleme eşit ağırlık atar.
2. Adım: İlk ağaç oluşturulduktan sonra sınıflandırılması kolay gözlemlere düşük ağırlık, sınıflandırılması zor gözlemlere yüksek ağırlık ataması gerçekleştirir.
3. Adım: İkinci ağaç ağırlıklandırılmış veri üzerinden büyütülür. Amaç ilk ağacın tahminlerini geliştirmektir.
4. Adım: Yeni model oluşturulan iki ağacın toplamı sonucu ortaya çıkar. Bu toplulaştırılmış modelin sınıflandırma hatası (classification error) hesaplanır. Oluşturulan bu iki yeni toplulaştırılmış model üzerinden revize edilmiş hataları tespit etmeye yönelik üçüncü ağaç büyütülür.
5. Adım: İşlemler belirlenen sayıda iterasyon için tekrar edilir.

Sonradan gelen ağaçlarda temel mantık, önceki ağaçlar tarafından iyi sınıflandırılmamış gözlemleri sınıflandırmaktır. Final toplulaştırılmış model daha önceki ağaç modellerinin tahminlerinin toplam tahminini verir.

3.4.4.2 Gradient Boosting Machine (GBM)

Boosting machine = Gradient Boosting Machine

Gradient Boosting = Gradient Descent + Boosting

Gradient boosting zayıf öğrenicilerin zayıf noktalarını, adaboost'ta kullanılan yüksek ağırlıklar ile öğrenmez. Onun yerine son fonksiyonun gradient'leri üzerinden öğrenir. Adaboost ta oluşturulan yüksek ağırlıklı veri noktaları ve GBM de oluşturulan gradient ler modelin nasıl geliştireceğini göstermektedir. Her aşamada zayıf öğreneni belirler, amaç zayıf öğrenende eksikleri telafi etmektir (James ve diğ., 2013).

Avantajları:

- Eğer GBM uygun şekilde ayarlanırsa (tune), diğer tüm algoritmalardan daha iyi performans gösterecektir.
- GBM kullanımının pratik faydası, kullanıcı tarafından belirlenen cost function'ı optimize etmesidir.

- Makine öğrenmesi algoritmalarının çoğu ‘lost function’ optimizasyonu üzerine çalışır ancak lost function üzerine kontrolümüz bulunmamaktadır ve gerçek dünya uygulamaları ile uyumlu olan ‘cost function’ ile uyumlu değildir.

Dezavantajı:

- GBM’lerin en büyük problemi aşırı uyum (overfitting) eğilimli olmalarıdır.
    - Dolayısı ile GBM ile eğitilirken verinin bir kısmı ayrı tutulmalıdır.
    - Bu problemin üstesinden gelebilmek için ‘early stopping’ kullanılması tavsiye edilir.
- Extreme değerlere ve gürültüye (noise) çok hassastır.

Çizelge 3.2 : GBM hiper parametreleri.

n.trees:	ağaç sayısı
bag.fraction:	her ağaçta örneklemi alınacak gözlemlerin oranı (RF> samp.size)
n.minobsinnode:	terminal düğümünde bulunacak minimum gözlem sayısı (RF> node.size)
interaction.depth:	her ağaçta bulunacak maksimum node sayısı
shrinkage:	öğrenme oranı

XGBoost

Çeşitli kaynaklarda XGBoost algoritmasının mevcut algoritmaların çoğunun üstünde performans gösterdiği vurgulanmaktadır. Algoritmanın başarısının arkasındaki en önemli özellik tüm senaryolara ölçeklenebiliyor olması şeklinde belirtilmiştir. XGBoost'un ölçeklenebilirliği birkaç önemli sistemden ve algoritmik optimizasyondan kaynaklanmaktadır.

XGBoost yaklaşımında, öğrenilen ağaçları tahmin etmede ağırlıklandırılmış quantile sketch ve seyrek verilere yönelik seyrekliği tespit eden bir algoritma oluşturulmuştur. Paralel ve dağıtılmış hesaplama, öğrenmeyi daha hızlı hale getirmekte ve bu da modelin keşfetme ve sonuca gitmesini hızlandırmaktadır. Diğer algoritmalar ile kıyaslandığında 10 kat daha hızlı çalıştığı belirtilmektedir.

Ağırlıklı Parçalı Sketch (Weighted Quantile Sketch): Tahmin algoritmalarında en önemli noktalardan biri aday ayırma noktalarının önerilmesi konusudur. Genellikle değişkenlerin yüzdelikleri (percentiles of a feature), adayların veriye eşit bir şekilde dağılmasını sağlamada kullanılır. Her örnek eşit ağırlığa sahip olduğunda, quantile

sketch adı verilen mevcut bir algoritma sorunu çözer. Ancak ağırlıklı veri kümeleri için mevcut bir quantile sketch yoktur. Bu sorunu çözmek için, ağırlıklı verileri işleyebilen ağırlıklandırılmış quantile sketch adında yeni bir algoritma oluşturulmuştur (Chen ve Guestrin, 2016).

Seyreklik Uyumu (Sparsity-aware Split Finding): Girdi değişkenlerinin değerlerinin kesikli olma durumu gerçek hayat problemlerinde sıkça karşılaşılan bir durumdur. Bunun temel sebeplerinden bazıları eksik verilerin bulunuyor olması, değer olarak çok sık sıfır değerinin girilmiş olması ve one-hot encoding gibi öznitelik mühendisliği sonuçları şeklinde sıralanabilir. Verideki bu seyreklik örüntülerinden (pattern) algoritmanın haberdar edilmesi önem arz etmektedir. Bunu sağlamak adına, her ağaç düğümüne saptanmış (default) bir yön eklenmesi önerilmiştir. 'X seyrek matrisinde bir değer eksikse, incelenen örnek, saptanmış yönde sınıflandırılır.' şeklinde tanımlama yapılmıştır. Her dalda iki saptanmış yön seçeneği vardır. En uygun olduğu saptanan yönler veriden öğrenilir.

Algoritmada yapılan önemli gelişme, sadece eksik olmayan verilerin ziyaret ediyor olmasıdır. Sunulan algoritma, var olmama durumunu (non-presence) eksik bir değer olarak ele almakta ve eksik verileri işlemek (eksik verilerle başa çıkabilmek) için en iyi yönü öğrenmektedir. Bu algoritma var olmayan verileri eksik veri olarak tanımlayıp çözümü devam ettirmenin yanı sıra kullanıcının belirlediği bir sınır değere göre de dikkate alınmaması gereken yönleri belirleyip algoritmanın bu şekilde en iyi yönü öğrenmesini sağlayabilir. Yani var olmama durumunun kullanıcı tarafında atanan bir değer ile belirlenmesi durumunda numaralandırmayı sadece tutarlı çözümler ile sınırlayarakta aynı algoritmayı uygulanabilmektedir.

Eta aşırı uyumu önlemede kullanılabilecek bir parametredir. Düşük eta değeri aşırı uyuma daha dayanıklı bir model elde edilmesini desteklemektedir. Aşırı uyumu engellemek adına kullanılan bir diğer parametre subsample'dır. Sumsample değerine 0.5 atanması durumu incelenecek olursa, burada XGBoost'un ağaçları büyütmede veri setindeki gözlemlerin rastgele yarısını alacağı anlamına gelmekte olup buda aşırı uyumun önüne geçecektir. Modelin overfit etmesini etkileyebilecek bir diğer parametre max_depth'dir. Eğer max_depth'e çok büyük değerler atanırsa ağaç daha kompleks bir form kazanır ve overfit etmeye yatkınlığı artar (Chen ve Guestrin, 2016).

Çizelge 3.3 : XGBoost hiper parametreleri.

eta :	Öğrenim oranı. (default: 0.3, range: [0,1])
gamma	Ağacın yaprak düğümünde daha fazla bölüm oluşturmak için gereken minimum loss reduction. Büyük değer atanırsa daha kontrollü/ korunumlu bir algoritma elde edilir. (default: 0, range: [0,∞])
min_child_weight	Yaprak düğümünde ihtiyaç duyulan minimum gözlem ağırlığı toplamıdır. Eğer yaprak düğümünde minimum gözlem ağırlığı toplamı min_child_weight değerinin altında oluşursa ağaç büyümeyi durdurur. (default: 1, range: [0,∞])
nrounds:	Final modeldeki ağaç sayısı
objective:	Modelin eğitim hedefi. (regresyon için 'reg:logistic', sınıflandırma için 'binary:logistic')
subsample	Her bir ağacı oluşturmak için alınan satır oranı. (default: 1, range: (0,1])
max_depth	Ağacın maksimum derinliği. (default: 6, range: [0,∞])
eval_metric	Doğrulama verileri için değerlendirme metriklerinin belirlenmesi gerekmektedir. (regresyon için 'rmse', sınıflandırma için 'error')
colsample_bytree	Her ağacı oluşturmak için alınan sütunların alt örnek oranı. (default:1, range: (0,1])

(*XGBoost*.; *XGBoost Parameters*.)

LightGBM

Son yıllarda büyük verinin ortaya çıkması ile Gradient boosting algoritması özellikle accuracy ve efficiency arasında bir ödünleşim olmak üzere belirli zorluklar ile karşı karşıya kalmıştır. Bu zorluklar ile mücadelede veri örneklerinin sayısını ve değişken sayısını düşürmek akla gelen ilk fikir/çözüm olmaktadır. Ancak bu yaklaşım da düşünüldüğü kadar kolay/ basit bir yaklaşım değildir. Örneğin GBDT için geri örneklemenin nasıl gerçekleştirileceği belirsizdir. Boosting in training sürecini hızlandırmak için bazı çalışmalar ağırlıklarına göre veriyi örneklemektedir, ancak bu GBDT ye direk uygulanamamaktadır çünkü örneklemlerin ağırlıkları GBDT te yoktur.

GBDT'nin en maliyetli/ zaman kaybettiren noktaları; karar ağaçlarının öğrenilmesi ve karar ağaçlarını öğrenmek için ağacın en iyi bölünme (split) noktalarını bulmak şeklinde tanımlanmaktadır.

Bu en iyi bölünme noktaları bulunurken pre-sorted algoritma kullanılmaktadır. Bu algoritmada tek tek tüm pre-sorted değişken değerlerinin olası tüm değerleri gözden geçirilmektedir. Ancak burada etkin olmayan (inefficient) noktalar sırası ile; öğrenim hızı (training speed) ve hafıza tüketimidir (memory consumption).

Bir diğer popüler yaklaşım histogram-based algoritmadır. Histogram based algoritma sürekli değişken değerlerini kesitli ambarlarda (discrete bin) depolar (buckets). Ve bu ambarları da değişkenlerin histogramlarının çiziminde kullanır. Bu yaklaşım hem hafıza kullanımı hem de training hızı yönünden daha etkindir.

LightGBM bu gibi zorlukları aşmaya yönelik iki temel tekniği bünyesinde barındırmaktadır. Bu teknikler sırası ile; Gradient-based One- Side Sampling (GOSS) ve Exclusive Feature Bundling (EFB). LightGBM, Gradient boosting algoritmasının geliştirilmiş versiyonu olarak kabul edilebilir.

LightGBM'de histogram tabanlı çalışmaktadır. Histogram temelli algoritma, en optimum bölünme noktasını değişkenlerin histogramlarına bakarak karar vermektedir.

Lightgbm leaf-wise tree growth strategy'i kullanmaktadır.

GOSS (Gradient-based One- Side Sampling): GBDT'te veri örnekleri (data instance) için bir ağırlık bulunmazken, farklı gradyanlara (gradient) sahip veri örneklerinin, bilginin kazanımına yönelik hesaplamalarda farklı roller oynadığı fark edilmiştir.

Şu fark edilmiştir ki küçük gradyan değerine sahip gözlemlerin eğitim hataları (training error) daha küçük olmaktadır yani bunlar iyi eğitilmiş gözlemler olmaktadırlar. Ama bu küçük gradyan değerli gözlemleri veriden çıkarmak verinin dağılımını bozacağından yeni bir yaklaşım sunulmuştur.

Buna yönelik GOSS, veri örneklerini örneklem ile seçerken accuracy değerini etkilememek adına büyük gradyan değerlerine sahip örneklerin hepsini veriye dahil ederken, küçük gradyan değerine sahip olanlardan örnekleme ile seçim yaparak veriye katmaktadır. Böyle bir yaklaşım ile aynı hedef örnekleme oranı üzerinden

uniformly random sampling e göre daha accurate gain estimationların yapıldığı belirlenmiştir (Ke ve diğ., 2017).

EFB (Exclusive Feature Bundling): Çok boyutlu veri genellikle çok kesiklidir. Gerçek uygulamalarda değişken uzayı kesiklidir. Kesikli değişken uzayında değişkenler nadiren 0 dışında değer alırlar. Örnek olarak one-hot- featurelar verilebilir. Değişken uzayının kesikli olması bize değişken sayısını düşürmede neredeyse sıfır kayıplı yaklaşımların tanımlanmasına olanak vermektedir.

Exclusive değişkenler bir araya getirilerek birleştirilebilir (bundle), bunlara exclusive feature bundle denmektedir. Nadiren 0 harici değer alan değişkenler bir araya getirilebilir (bundling).

Burada iki nokta önem arz etmektedir; ilki hangi değişkenlerin bir araya getirileceği, ikincisi bir araya getirmenin nasıl inşa edileceği/ yapılacağı.

Ele alınması gereken ilk konu: Graph coloring problemi bir NP-hard problemidir.

Sunulan yaklaşım good approximation algoritmasını bulmak adına optimal bundling problemi, graph coloring problemine indirgenmiştir ve sonrasında bundleları üretecek graph coloring için orta derecede iyi sonuçlar üretebilen greedy algoritması kullanılmıştır.

Ele alınması gereken ikici konu ise: değişkenlerin aynı bundle içerisinde birleştirilmesi ve bu birleştirmeye karşılık gelen eğitim karmaşıklığının/ kompleksliğini düşürecek iyi bir yaklaşıma ihtiyaç olmasıdır. EFB algoritması birçok exclusive değişkeni daha az sayıda ve yoğun değişkene birleştirebilir. Ki bu oluşturulan değişkenler ile sıfır değerine sahip gereksiz değişkenlerin hesaplamalarından uzak durulur.

Lightgbm; XGBoost (pre-sorted algorith), XGBoost (histogram-based algorith), GOSS ve EFM olamayan Lightgbm, sadece EFM olan Lightgbm algoritmaları ile kıyaslanarak farklı tipte veri setlerinde yapılan uygulamaların hepsinde hem tüketilen toplam hem de accuracy değerlerinde en iyi sonucu veren algoritma olmuştur (Ke ve diğ., 2017).

LightGBM 100’den fazla hiperparametreye sahiptir. Modelde düzenlenirken (tune), yapılmak istenen değişikliğe göre tercih edilecek parametre değişkenlik göstermektedir.

- Leaf-wise ile elde edilecek sonuçların iyi olması isteniyorsa dikkat edilmesi gereken parametreler; ‘num_leaves’ (70-80 dolayında tercih edilebilir), ‘min_data_in_leaf’ ve ‘max_depth’ parametreleridir.
- Daha hızlı bir model isteniyor ise düzenlenecek (tune) parametreler; ‘bagging_fraction’, ‘bagging_freq’, ‘feature_fraction’, ‘max_bin’ (küçük değer kullanılmalı) and ‘save_binary’.
- Daha iyi accuracy değeri elde etmek için ise; ‘max_bin’ (büyük değer kullanılmalı), ‘learning_rate’ (küçük değer kullanılmalı), ‘num_iterations’ (büyük değer kullanılmalı), ‘num_leaves’ (büyük değer kullanılmalı) ve ‘dart’ parametrelerine odaklanılmalı ve data büyük eğitim seti (training data) seçilmesi gerekmektedir.
- Aşırı uyum (overfitting) problemi yaşamamak için; max_bin (küçük değer kullanılmalı), ‘num_leaves’ (küçük değer kullanılmalı), ‘min_data_in_leaf’, ‘min_sum_hessian_in_leaf’, ‘bagging_fraction’, ‘bagging_freq’, ‘feature_fraction’, ‘lambda_l1’, ‘lambda_l2’, ‘min_gain_to_split’, ‘max_depth’ (büyük ağaçlardan kaçınılmalı), ‘extra_trees’ parametreleri için doğru değerlerin belirlenmesi ve daha büyük eğitim setinin (training data) seçilmesi gerekmektedir.

Çizelge 3.4 : LightGBM hiper parametreleri.

num_leaves:	Oluşturulacak ağacın maksimum yaprak sayısı. (default:31) 2^(max_depth)’den küçük yada eşit olmalıdır, büyükse overfitting. Büyük değerler daha derin ağaçların oluşması ile sonuçlanır.
min_data_in_leaf:	Leaf node da bulunacak minimum gözlem sayısı (default:20) Büyük değer atanırsa çok derin ağaç üretilmesinin önüne geçilir ama under-fitting olabilir.
max_depth:	Maksimum ağaç derinliği (her ağacın node sayısı) (default:-1) Küçük değerler zayıf tahmin algoritmasını temsil eder. Küçük veride overfitting ile baş etmede kullanılır.
max_bin:	EFB’de değişkenin değerinin (feature value) gireceği maksimum kutu (bucket) sayısını belirtir. (default:255) Küçük değer tercih edilirse modelin accuracy değeri etkilenir ancak overfitting ile baş edebilme gücü artabilir.
bagging_fraction:	Her bir iterasyon için kullanılacak veri miktarını/ oranını belirtir. (default:1) 0 ve 1 arasında değer alabilir. Genellikle eğitimi hızlandırmak ve overfittingi önlemek için kullanılır.
bagging_freq:	Bagging sıklığı (default:0) Bagging yapabilmek için bagging_fraction değerinin 1’den küçük olması gerekir.
feature_fraction:	Ağaçları oluşturmak için her iterasyon da parametrelerin yüzde kaçını rastgele seçeceğini belirler. (default:1) 1’den küçük olursa Lightgbm her iterasyonda değişkenlerin rastgele bir kısmını seçer.

learning_rate:	Öğrenme oranı. (default: 0.1) Her ağacın final çıktı üzerindeki etkisini belirler.
num_iterations:	Boosting iterasyon sayısı (default:100)
min_gain_to_split:	Bir ayrım (split) yapmak için gereken minimum kazancı belirtir. (default:0)
early_stopping_round:	Analizi hızlandırır. Validasyon verisinde bir metrikte gelişme/ iyileşme olmadığında model öğrenmeyi durdurur. (default:0)

(*LightGBM Parameters*.)

CatBoost

Sıralı boosting (ordered boosting) ve kategorik değişkenlerin proses edilmesine yönelik geliştirilen inovatif algoritma, Catboost ile sunulan iki kritik algoritmik ilerleme olarak belirtilmektedir.

Her iki teknik mevcutta bulunan gradient boosting algoritmalarında yaşanan tahmin kaymasıyla (predicting shift) mücadele etmek adına geliştirilmiştir. Tahmin kaymasına hedefe ulaşmada yaşanan kayıpların, firelerin sebep olduğu belirtilmektedir.

Bu sorunun ayrıntılı analizi gerçekleştirilmiş ve üretilen Catboost algoritmasının sorunu etkili bir şekilde çözdüğü görülmüştür.

CatBoost ("Categorical Boosting"), gradient boosting karar ağaçlarının mevcut en yeni uygulamalarından – XGBoost ve LightGBM'den daha iyi performans göstermektedir.

Tahmin Kaymasıyla (Predicting Shift): Catboost, tahmin kaymasını (prediction shift) önlemek için ordered boosting yöntemini kullanmaktadır. Türetilen ordered boosting, standart gradient boosting algoritmasının, hedefe ulaşmada yaşanan kayıpların önlenmesine yönelik modifiye edilmiş bir versiyonudur.

Tahmin kayması, bulmak istediğimiz değişkende (ör. GPA veya konut fiyatları) tahmin değerinin veri setinde yer alan değere yaklaşıp, gerçek değerden uzaklaştığında gerçekleşmektedir. Tahmin değeri ve gözlemlenen değer arasındaki fark artık (residual) olarak tanımlanmaktadır. Catboost algoritması temelde, artıklarda (residuals) kaymayı (shift) önlemeye çalışır (Prokhorenkova ve diğ., 2018).

Kategorik Değişkenlerin Prosesi: Makine öğrenmesi algoritmaları kategorik değişkenleri tanımlayamamaktadır. Dolayısı ile kategorik değişkenlerin nümerik değerlere dönüştürülmesi gerekmektedir. Standart algoritmalarda kategorik değişkenleri ön işlemede problemler oluşmaktadır. Buna yönelik kullanılan yaygın bir çözüm, tüm nominal değerleri ikili kukla değişkenlere dönüştürmektir (one-hot encoding). Ancak bu özellik kümenizi hızla büyütebilir/patlatabilir.

Gradient boosting'de kategorik değişkenleri kullanmanın en etkin yollarından biri, kategorileri hedef istatistiklere (target statistic) çevirmektir. Hedef istatistik basit bir

istatistiksel modelin kendisidir ve ayrıca hedef sızıntısına ve bir tahmin değişikliğine neden olabilir (Prokhorenkova ve diğ., 2018).

Catboost, nominal bir değişkenin (örn. Milliyet), her kategorisi için (örneğin, Hollandaca, Almanca, Belçika) bir değer (target-based statistic) hesaplamaktadır. Bu hesaplama için takip edilen adımlar;

1. Bir kategorik özellikle başlanır (örneğin, Uyruk). Buna x denir.
2. Rastgele seçilen bir satırda (training veri setindeki k-satırı), bu kategorik özelliğin (x'in i. Düzeyi) rastgele bir seviyesi bir sayıyla (örneğin, Felemenkçe 5 ile) değiştirilir.
3. Bu sayı (bizim örneğimizde 5) genellikle kategori düzeyine bağlı olarak hedef değişkeni (target variable-tahmin etmek istediğimiz) temel alır. Başka bir deyişle, hedef sayı beklenen sonuç değişkenini temel alır.
4. Bölme niteliği, iki eğitim verisi seti oluşturmak için kullanılır: ilk sette tüm kategoriler yer alacak olup (örneğin, Almanca, Fransızca, Hint vb.) bu kategoriler 3.adımda hesaplanan hedef değişkenlerden daha fazla olacaktır. Diğer set ise daha az/küçük hedef değişkenden oluşacaktır.

3. adımdaki sayı yalnızca rastgele bir değişken değildir, aynı zamanda hedef tabanlı bir istatistiğe (target-based statistic -TBS) dayanır.

Target-based statistic dört farklı şekilde hesaplanabilmektedir: Greedy, Holdout, leave-out-one ve ordered. Catboost ordered TBS'i kullanmaktadır. Ordered TBS yaklaşımı, geçmişin geleceği etkilediği varsayımına dayanmaktadır. Catboost'ta, odak satırından önceki satırların sırası randomize edilir ve hedef istatistikleri (TBS) hesaplamak için kullanılır. Bu birkaç kez tekrarlanır (*Get started Working with categorical data: Catboost,* n.d.).

Modelde aşırı uyum (overfitting) ile mücadelede kullanılan değişkenler l2 regularization ve random strenght'dir (*How to Win a Data Science Competition: Learn from Top Kagglers*), thread_count parametresi modelin performansını, hızını belirlemede yardımcı olmaktadır.

Çizelge 3.5 : CatBoost hiper parametreleri.

loss_function	Binary sınıflandırma için 'LogLoss' veya 'CrossEntropy' kullanılır. Çoklu sınıflandırma için 'MultiClass' kullanılır.
iterations	Oluşturulabilecek maksimum ağaç sayısı. Daha az kullanılabilir. (default:1000)
learning_rate	Gradyan adımlarını azaltmak için kullanılır. Genelde toplam eğitim süresini (time of training) etkiler: değer ne kadar küçük olursa, eğitim için daha fazla tekrar gerekir. (default: 0.03)
l2_leaf_reg	Cost function daki l2 regülasyon teriminin katsayısı. Düzenleyicinin mümkün olan en iyi değeri bulması için farklı değerler denenir. Herhangi bir pozitif değerler kullanılabilir. (default:3)
depth	Ağacın derinliği. 16'ya kadar değer alabilir. İyi değerler 1-10 arasındaki değerlerdir. (default: 6)
random_strength	Ağaç yapısı seçildiğinde bölümlerin (split) puanlanması için kullanılan rastgelelik miktarıdır. Overfitting i önlemek için kullanılan bir parametredir.
rsm	Random subspace (rastgele altuzay) yöntemi. Her rastgele seçimde değişkenler tekrar rastgele seçilirken, seçilecek/kullanılacak değişkenlerin yüzdesi. Rsm değeri (0,1] aralığında olmalıdır.
thread_count	Eğitim esnasında kullanılacak konu sayısını belirler. İşlemlerin yürütülme hızını optimize eder. Bu parametre sonucu etkilemez.
border_count	Sayısal değişkenler için bölünme sayısıdır. 1 ile 255 arasında değer alabilir.
logging_level	Stdout-standart çıktının günlük düzeyini belirler.
train_dir	Eğitim sırasında oluşturulan dosyaların saklandığı alanı belirtir.

(*R package training parameters*)

3.5 Model Sonuçlarının Yorumlanması ve Seçimi

İkili sınıflandırma modellerinde iki tip hata vardır. Hata tiplerinin açıklamasını anomali ve anomali olmayan gözlemler üzerinden tanımlayacak olursak; Sınıflandırma modeli ya anomali bir gözlemi anomali değil şeklinde yanlış sınıflandırabilir ya da anomali olmayan bir gözlemi de anomali şeklinde yanlış sınıflandırabilir. Bu iki hata tipinden hangisinin yapıldığının belirlenmesi temelde merak edilen noktayı ifade etmektedir. Confusion matrix (hata matrisi) bu bilgiyi göstermenin uygun bir yoludur.

Çizelge 3.6 : Hata matrisi.

n=Toplam Gözlem		TAHMİN		Toplam[1]
		Anomali Değil (-)	Anomali (+)	
GERÇEK	Anomali Değil (-)	TN	FP	N
	Anomali (+)	FN	TP	P
	Toplam[2]	N*	P*	

- Gerçek Pozitifler (TP): Anomali olarak tespit edilip, gerçekten anomali olanlar
- Gerçek Negatifler (TN): Anomali değil şeklinde tespit edilip, gerçekten anomali olmayanlar
- Yanlış Pozitifler (FP): Anomali olarak tespit edilip, gerçekte anomali olmayanlar
- Yanlış Negatif (FN): Anomali değil şeklinde tespit edilip, gerçekte anomali olanlar

İkili sınıflandırma (binary classification) için hata matrisi üzerinden hesaplanabilecek oranlar;

- Accuracy:
  - Genel olarak sınıflandırıcı ne sıklıkla doğru sınıflama yapmaktadır?
  - (TP + TN) / n
- Misclassification Rate:
  - Genel olarak sınıflandırıcı ne sıklıkla yanlış sınıflama yapmaktadır?
  - (FP + FN) / n
  - (1- Accuracy) değerine eşittir.
  - ‘Error Rate’ olarak ta adlandırılmaktadır.
- True Positive Rate:
  - Sınıflandırıcı gerçekte anomali olan gözlemleri, ne sıklıkla anomali olarak tahmin etmektedir?
  - TP/ P
  - ‘Sensitivity’, ‘Recall’ ya da ‘1-Type II error’ olarak ta adlandırılmaktadır.
- False Positive Rate:

---

[1] N: Gerçekte anomali olmayan gözlem sayısı toplamı, P: Gerçekte anomali olan gözlem sayısı toplamı

[2] N*: Anomali değil olarak tahmin edilen gözlem sayısı toplamı, P*: Anomali olarak tahmin edilen gözlem sayısı toplamı

    - Sınıflandırıcı gerçekte anomali olmayan gözlemleri, ne sıklıkla anomali olarak tahmin etmektedir?
    - FP / N
    - (1- Specificity) değerine eşittir.
    - ‘Type I error’ olarak ta adlandırılmaktadır.
- True Negative Rate:
    - Sınıflandırıcı gerçekte anomali olmayan gözlemleri, ne sıklıkla anomali değil olarak tahmin etmektedir?
    - TN / N
    - (1- False Positive Rate) değerine eşittir.
    - ‘Specificity’ olarak ta adlandırılmaktadır.
- Precision:
    - Sınıflandırıcı anomali olarak tahmin ettiğinde, ne sıklıkla doğrudur?
    - TP / P*
- Prevalence:
    - Anomali olma koşulu/ durumu aslında örneğimizde ne sıklıkla görülmektedir?
    - P/ n
- Null Error Rate:
    - Sınıflandırıcı çoğunlukta olan sınıfı her zaman tahmin ederse, ne sıklıkta yanılacağını gösterir.
    - Bu, sınıflandırıcıları karşılaştırmak için yararlı bir temel metrik olabilir.
- Cohen's Kappa:
    - Bu aslında sınıflandırıcının, şans eseri ne kadar iyi performans göstereceğine kıyasla esasen ne kadar iyi performans gösterdiğinin bir ölçüsüdür.
    - Diğer bir deyişle, accuracy ve null error rate arasında büyük bir fark varsa, bir model yüksek bir Kappa puanına sahip olacaktır.
- F Score:
    - Bu, true positive rate (recall) ve precision değerlerinin ağırlıklı ortalamasıdır.
- ROC Curve:
    - Bir sınıflandırıcının olası tüm eşikler (thresholds) üzerindeki performansını özetleyen grafiktir.
    - Belirli bir sınıfa gözlem atama eşiğini değiştirdiğinizde, Gerçek Pozitif Oranı (y ekseni) ve Yanlış Pozitif Oranı (x ekseni) ile çizilerek oluşturulur.

Sensitivity ve specificity terimlerinin sınıflandırmanın performansını karakterize ettiği ‘sınıfa özgü performans’ (anomali ya da anomali olmama durumunu tespit etme performansı) değerleri; bankacılık, tıp, biyoloji ve sigortacılıkta önem teşkil etmektedir.

Anomaliye düşenler için varsayılan durumu yanlış tahmin etmekten endişe edersek, eşiğin (threshold) düşürmesi düşünülebilir (Cichosz, 2015).

3.6 CRISP-DM Metodolojisi

90'lı yıllarda veri madenciliği sektörünün bir dar boğaz (chasm) içerisinde olduğu bilinmektedir. Veri madenciliğinin bu süreçten çıkabilmesi adına ortak bir sürecin olgunlaştırılması gerekliliği gündeme gelmiştir.

Belirlenecek modelin, veri madenciliğini tartışmak için ortak bir referans noktası olacağı ve özellikle müşteriler tarafı olmak üzere, tüm katılımcılar tarafından çok önemli veri madenciliği konularının anlaşılmasını artıracağı belirtilmiştir. Bu kapsamda zaman içerisinde çeşitli modeller geliştirilmiştir. SEMMA [Sample (Örnekleme), Explore (Keşfetme) Modify (Düzenleme), Model (Modelleme), Evaluate (Değerlendirme)] ve KDD (Knowledge and Data Discovery) bu amaca hizmet etmek üzere oluşturulan ilk modeller olarak ele alınabilir.

1997 yılında ise Avrupa Birliği projesi kapsamında ESPRIT fonu altında, 5 firmanın (DaimlerChrysler, AG, SPSS, NCR, and OHRA) ortak çalışması ile CRISP-DM (Cross Industry Standard Process for Data Mining) Çapraz Endüstri Standart Süreç Modeli geliştirilmiştir. 2006-2008 yıllarında modelde revizyona gidilmiştir.

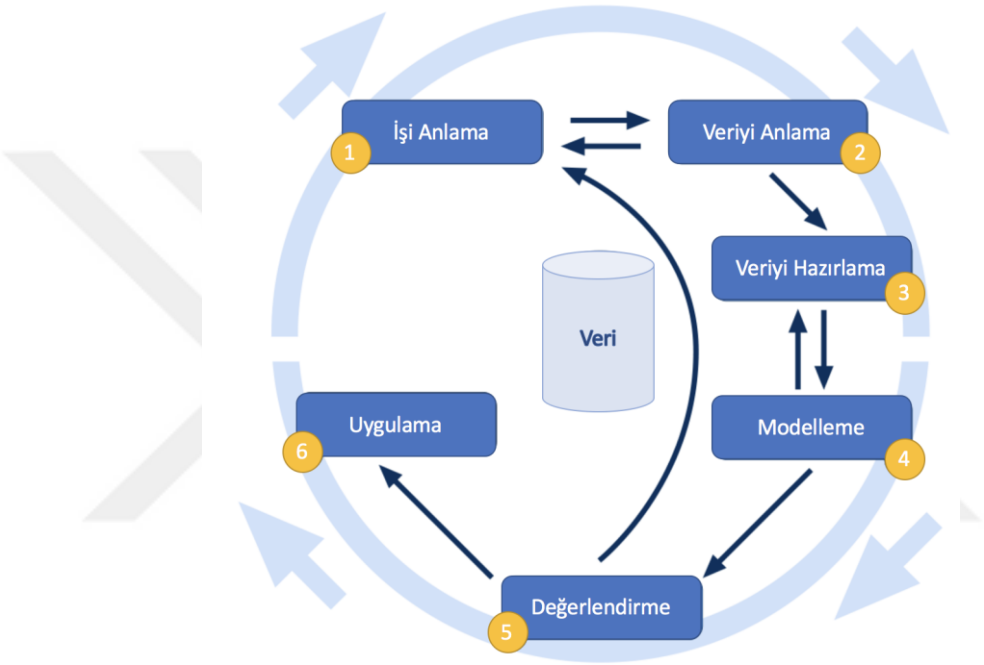
Veri madenciliğinde takip edilebilecek bir modelin varlığı; veri madenciliği kullanılmalı mı sorusundan ziyade veri madenciliği iş problemlerini çözmede nasıl kullanılmalı sorusuna geçiş sağlayarak,bir paradigma değişimi ortaya koymuştur.

Böyle bir model veri madenciliği projeleri gerçekleştiren analistler için birçok açıdan faydalıdır.

Özellikle yeni başlayanlar için süreç modeli, çalışmanın nasıl ilerletilmesi gerektiği noktasında rehberlik eder, projeyi yapılandırma noktasında yardımcı olur ve süreçte yer alan her görev için önerilerde bulunur. Deneyimli analistler ise bu modelden, analiz esnasında bir şeyi unutmadıklarından emin olmak için her görev için belirlenen kontrol listeleri aracılığı ile yararlanabilirler. Bunlara ek olarak ortak bir süreç modelinin en önemli rolü, sonuçların iletişimi ve belgelendirilmesidir/ raporlanmasıdır. Etkili ve efektif bir proje için farklı araçları kullanan, farklı

becerilere ve geçmişlere sahip insanları bir araya getirmeye yardımcı olur (Wirth ve Hipp, 2000).

Veri bilimi proje yönetimi yaklaşımından CRISP-DM, bir projesinin yaşam döngüsünü altı aşamaya ayırmaktadır. Bu adımlar sırası ile; işi anlama, veriyi anlama, veriyi hazırlama, modelleme, değerlendirme ve uygulamadır (Chapman P., Clinton J., 2000). Şekil 3.1'de Çapraz Endüstri Standart Süreç Model için belirlenen adımlar gösterilmektedir.



Şekil 3.1 : Çapraz Endüstri Standart Süreç Modeli (CRISP-DM).

3.6.1 İşi Anlama

İlk aşama olarak belirlenen işi anlama adımında, proje için belirlenen hedeflerin ilgili iş perspektifinden anlaşılması, elde edilen bilginin veri madenciliği problem tanımına dönüştürülmesi ve belirlenen hedeflere ulaşmak için bir ön plan geliştirilmesine odaklanır. İşi anlama adımında yer alan adımların bazıları şu şekildedir; iş hedeflerini belirleme, durumu değerlendirme, veri madenciliği hedeflerini belirleme veproje planını oluşturmaktır.

3.6.2 Veriyi Anlama

İlk olarak verilerin toplaması ile başlayan veriyi anlama aşamasında, analist verilere aşinalık kazanmaya ve veri kalitesine yönelik problemleri tanımlamaya çalışır.

Verilere ilişkin elde edilecek ilk öngörüleri keşfetmek ya da gizli bilgilere yönelik hipotezler oluşturmak için ilginç altkümelerin tespiti bu adım içerisinde yapılabilir. Kısaca veriyi anlama adımı dört adımdan oluşmaktadır; verilerin toplanması, verilerin tanımlanması, verilerin araştırılması ve veri kalitesinin doğrulanmasıdır.

3.6.3 Veriyi Hazırlama

Bu aşama, toplanmış olan ilk ham verilerden, nihai veri setini veya modelleme araçlarına beslenecek veri setlerini oluşturmak için yapılması gereken tüm faaliyetleri kapsar. Veri hazırlamadaki beş adım; verilerin seçilmesi, verilerin temizlenmesi, verilerin oluşturulması, verilerin entegre edilmesi ve verilerin biçimlendirilmesidir.

3.6.4 Modelleme

Modelleme aşamasında, incelenen problem ile ilişkili olan çeşitli modelleme teknikleri seçilip uygulanır ve parametrelerinin optimum değerleri belirlenir. Genel olarak, aynı veri madenciliği probleminde birden fazla teknik kullanılabilmektedir. Bazı tekniklerin özel gereksinimleri olabilmekte, bu gereksinimlere yönelik veri formları yeniden düzenlenmektedir. Bu gibi durumlarda veri hazırlama aşamasına geri dönülür. Modelleme adımları sırası ile; modelleme tekniğini seçme, test tasarımını oluşturma, modelleri oluşturma ve modelleri değerlendirmeyi içerir.

3.6.5 Değerlendirme

Analist tarafından oluşturulan modelin, uygulama aşamasına geçmeden önce, ayrıntılı bir şekilde değerlendirilmesi ve iş hedeflerine ulaşıp ulaşmadığının incelenmesi gerekir. Bu adımın sonuna gelindiğinde, analistin, yürütülen veri madenciliği çalışması sonucu elde edilen sonuçların nasıl kullanılacağına karar vermesi istenir. Buradaki anahtar adımlar; sonuçları değerlendirme, süreçleri inceleme ve ileriki adımların belirlenmedir.

3.6.6 Uygulama

Oluşturulan modeller genelde yürütülen projelerin sonu değildir. Müşterinin kullanabileceği şekilde, kazanılan bilginin düzenlenmesi ve sunulması istenir. Uygulama aşaması modelin, müşterinin kullanabileceği bir forma getirilmesini

kapsar. Buradaki anahtar adımlar; dağıtım planının oluşturulması, planın izlenmesi ve bakım, nihai rapor hazırlanması ve son olarak projenin gözden geçirilmesidir.

Belirlenen bu 6 aşama içerisinde takip edilmesi gereken adımlar aşağıdaki tabloda özetlenmiştir (Shearer, 2000) (Çizelge 3.7).

Çizelge 3.7 : CRISP-DM aşamaları içerisinde takip edilen adımlar.

İşi Anlama	Veriyi Anlama	Veriyi Hazırlama	Modelleme	Değerlendirme	Uygulama
• iş hedeflerini belirlemek • durumu değerlendirmek • veri madenciliği hedeflerini belirlemek • proje planını oluşturmak	• verilerin toplanması • verilerin tanımlanması • verilerin araştırılması • veri kalitesinin doğrulanması	• verilerin seçimi • verilerin temizlenmesi • verilerin oluşturulması • verilerin entegrasyonu • verilerin biçimlendirilmesi	• modelleme • tekniğinin seçimi • test tasarımının oluşturulması • modellerin oluşturulması • modellerin değerlendirilmesi	• sonuçların değerlendirilmesi • süreç incelemesi • sonraki adımların belirlenmesi	• dağıtım planı • plan izleme ve bakım • nihai raporun hazırlanması • projenin gözden geçirilmesi

4. UYGULAMA

Bu çalışma yürütülürken CRISP-DM yaklaşımı benimsenmiştir. Yapılan minimal düzeydeki modifikasyonlar ile modelin adımları sırasıyla takip edilerek çıktılar elde edilmiştir. Modelin son aşaması olan Uygulama aşaması tez çalışması için uygun olmadığından çalışma kapsamından çıkarılmıştır.

4.1 İşi Anlama

Sigortacılık datası üzerine yürütülecek bu çalışmanın iş boyutunu net bir şekilde anlamak adına sigortacılık sektöründe çalışan uzmanlar ile görüşmeler gerçekleştirilmiştir, sektör raporları, faaliyet raporları ve şirketlerin stratejik planları incelenmiştir. Ayrıca literatürde alana yönelik yapılmış çalışmalar taranmıştır.

Yapılan araştırmalar neticesinde yürütülecek çalışmada hangi parametreler üzerine odaklanılması gerektiğinin tespitine yönelik kapsamlı bilgi elde edilmiştir.

Özellikle sektör hedeflerine ulaşmada takip edilmesi tercih edilen göstergeler belirlenmiş olup, tespit edilen bu göstergeler çalışmanın ileriki aşamalarında modelin üzerinde çalıştırılacağı veri setinin oluşturmasında kullanılmıştır.

4.1.1 Sigorta sektöründe kullanılan göstergeler

Demirbaş P. (2010) çalışmasında, Müşteri Memnuniyeti ve Bağlılığı, Sürdürülebilir Rekabetçi Yapı, Karlılık ve Verimlilik, Büyüme, Çalışan Memnuniyeti ve Bağlılığı şeklinde belirlediği 5 stratejik yönelim altında göstergelerini sınıflandırmıştır (Çizelge 4.1).

Çizelge 4.1 : Stratejik yönelikler bazında göstergeler.

Karlılık ve Verimlilik	Brüt teknik kar/ prim oranı	**Büyüme**	Prim Üretimi
	Dosya başı maliyet		Pazar payı
	Hasar/ prim oranı		Büyüme oranı
	Teknik karlılık		Yenileme oranı
	Blokeli tahsilatların toplam tahsilat içerisindeki payı		Poliçeleştirme oranı
	Kredi kartı ile tahsilatların toplam tahsilat içerisindeki payı		Üretimin artış oranı
	Banka teminatı meblağının toplam teminat içindeki payı		Kaynak sayısındaki artış oranı
	Genel gider / prim oranı		Acente ve aday acente ziyaret sayısı
	Rücü / tahsilat oranı		Kampanya sayısı
	Şüpheli alacakların toplam alacak içindeki payı		Proje sayısı
	Öneri/ Uygulamaya alınma sayısı		Yenileme kaybı
	Banka sayısı		Uygulamaya alınan ürünler/ geliştirilen ürünler
	İnceleme sayısı		Reasürör/ broker sayısı
	Bulguların düzelme oranı		Belirlenen sürede planlanan çalışmanın yapılmış/ tamamlanmış olması
	Maliyet tasarrufu	**Müşteri Memnuniyeti ve Bağlılığı**	Haklı müşteri şikayet sayısı
	Hasar dosya kontrol süresi		Kampanya sayısı
	Peşin hasar ödeme süresi		Tanınırlık oranı
	Sigortalı sayısı oranı		Belirlenen sürede planlanan çalışmanın yapılmış/ tamamlanmış olması
	İskonto oranı	**Çalışan Memnuniyeti ve Bağlılığı**	Eğitim alan personel sayısı
	Turnover oranı		Kurum içi müşteri memnuniyet oranı
	Risk analizine gidilen iş sayısındaki artış		Belirlenen sürede planlanan çalışmanın yapılmış/ tamamlanmış olması
	Üretimdeki artış oranı	**Sürdürülebilir Rekabetçi Yapı**	Organizasyonel yapılanmalarının değerlendirme sayısı
	Belirlenen sürede planlanan çalışmanın yapılmış/ tamamlanmış olması		Belirlenen sürede planlanan çalışmanın yapılmış/ tamamlanmış olması

Ak R., Oztayşi B. çalışmalarında önerdikleri modelde Kurumsal Karne yaklaşımı üzerinde göstergeleri sınıflamışlardır (Ak ve Öztayşi, n.d.) (Çizelge 4.2).

Çizelge 4.2 : Kurumsal karne yaklaşımı ile belirlenen göstergeler.

Finansal Perspektif	Toplam gelir
	Maliyet azaltma yüzdesi
	Kar oranı (yüzde)
	Yatırımın geri dönüşü (yüzde)
Müşteri Perspektifi	Müşteri memnuniyeti oranı (yüzde)
	Çapraz satış yüzdesi
	Yeni müşteri kazanım oranı (yüzde)
	Müşteri tutma oranı (yüzde)
İç İş Süreçleri Perspektifi	Yeni ürün sayısı
	Servis hata oranı (yüzde)
	Bir servis çağrısına ortalama cevap (saat)
	Müşteri şikayet oranı (yüzde)
	Stratejik planlar için uygulama oranı (yüzde)
Öğrenme ve Büyüme Perspektifi	Çalışan başına yeniden beceri kazanma süresi (saat)
	Bilgi teknolojisi ve sistemlerinde yatırım oranı
	Müşteri memnuniyeti oranı (yüzde)

Roy ve Geore (2017) çalışmalarında sigorta taleplerinde sahtekârlık işlemleri tahmin etmede yararlanılacak makine öğrenmesi yaklaşımlarında kullanmak üzere veri dönüşümleri gerçekleştirmiştir. Bunlar sırası ile;

- Talep oluşma tarihi ile talep raporlama tarihi arasındaki fark
- Talep raporu tarihi ile talebin açılma tarihi arasındaki fark
- Poliçenin efektif olduğu tarih ile talep oluşma tarihi arasındaki fark
- Talep belgesi gönderildi mi?
- Parça maliyet farkı
- Kredi notu
- Poliçe primi
- Talep oluşumu başlangıç zamanı
- Müşteri iletişiminin sayısı
- Talebin gerçekleştiği yerin adı

Literatürde anomali tespitinde kullanıldığı görülen diğer göstergeler ise hasar prim oranı (Nagrecha ve diğ., 2018, Nagrecha, n.d.) , üretilen poliçeler, poliçelerin

maliyetleri (tazminat talebi tutarı), poliçelerin getirileri (ödenen prim) (Nagrecha, n.d.), sermaye oranı, primlerin karlılığı, kayıp oranı, beklenmedik sıklık ve gişe rekorları kıran sigortalardır (Benali ve Feki, 2017).

Poliçelerin oluşturulması aşamasında anomali potansiyeli taşıyan poliçelerin tespit edilmesi hedeflendiğinden, özellikle sınıflandırma modellerinde poliçe verisinde yer alan değişkenlerin kullanımına da yer verilmiştir.

4.2 Veriyi Anlama

4.2.1 Verilerin toplanması

Talep edilen çalışmanın şekillendirilmesi adına sigorta firması, poliçe ve hasar olmak üzere iki adet veri seti temin etmiştir. Veri setlerinde yer alan tüm değişkenler incelenmiş, çözümlenemeyen değişkenler firma ile gerçekleştirilen görüşmeler ile netleştirilmiştir. Görüşmeler neticesinde yanlış/eksik değerlerin yer aldığı tespit edilen değişkenler analize dahil edilmemiştir.

Sırası ile bu veri setlerindeki gözlem ve değişken sayısı Çizelge 4.3'te belirtildiği gibidir.

Çizelge 4.3 : Veri setinde yer alan gözlem ve değişken sayısı.

Veri Setinin Adı	Gözlem Sayısı	Değişken Sayısı
Poliçe	11.755.135 gözlem	47 değişken
Hasar	29279 gözlem	67 değişken

4.2.2 Verilerin tanımlanması

Veri setlerinde yer alan değişkenler Çizelge 4.4'te belirtildiği gibidir.

Çizelge 4.4 : Veri setlerinde yer alan değişkenler.

Veri Seti	Değişkenler
Hasar Verisi	Dosya Numarası, Teminat Kodu, Temıinat Adı, Hasar Nedeni, Hasar Nedeninin Açıklaması, Kaynak Kodu, Kaynak Adı, Cari Poliçe Numarası, Zeyl Sıra Numarası, Hasar Tarihi, Dosya Tarihi, Bölge Adı, Portfoy Adı, Acenta Numarası, Acenta Adı, Brans Kodu, Brans Adı, Tarife Kodu, Tarife Adı, Dosya Durumu, Dosya Durumunun Açıklaması, Sigortalı Numarası, Sigorta Ettiren Numarası, Ana Poliçe Başlama Tarihi, Tanzim Tarihi, Başlama Tarihi, Bitiş Tarihi, Segment, Plaka, Eksper Kodu, Tamirhane Kodu, Tamirhane Adı, Eksper Adı, Assıst Numarası, Kapanma Tarihi, Rezervasyon Tarihi, Yedek Parça (Kdv), İşçilik (Kdv), Rezervasyon Op Id, Dosya Sorumlusu, Dosyayı Açan Op Id, Dosyayı Açan Profil, Dosya türü, Kusur Yüzdesi, Rücu Tetkik, Hasar Yeri, İkinci Dosya Onaylayan, Onay Kontrol, Onay Tarihi, İkinci Onay Tarihi, Doğal Afet, Doğal Afet Nedeni, İhtiyaten, Açıklama, Açıklama 2, Tarih, Maddi Bedeli, Departman, Risk İli, İl Açıklaması, Kasa Tipi, Ödenen Dosya Adedi, Ödenen Tazminat, Sovtaj Alınan Rücu, Sovtaj, Alınan Rücu, Ödenen Masraf, Ödenen Eksper, Ödenen Avukat Masrafı, Muallak Dosya adedi, Muallak Hasar, Kasko Teminat Bedeli
Poliçe Verisi	Kaynak Kodu, Uzun Adı, Cari Poliçe Numarası, Zeyl Sıra Numarası, Zeyl Türü, Zeyl Adı, Zeyl Neden Kodu, Zeyl Neden Adı, Teklif Numarası, Tarife, Tarife Adı, Bölge Adı, Portfoy Adı, Acente İl Adı, Acente İlçe Adı, Acente Numarası, Acente Adı, Yeni Yenileme, Branş, Branş Adı, Segment, Tanzim Tarihi, Uwy (Underwritting Year), Başlama Tarihi, Bitiş Tarihi, Ana Poliçe Başlama Tarihi, Teminat Kodu, Teminat Adı, Sigortalı Numarası, Sigortalı İli, Sigortalı İlçesi, Cinsiyet, Gerçek Tüzel, Yaş, Müşteri İl Adı, Müşteri İl Kaynak Kodu, Uzun Adı, Cari Poliçe Numarası, Zeyl Sıra Numarası, Zeyl Türü, Zeyl Adı, Zeyl Neden Kodu, Zeyl Neden Adı, Teklif Numarası, Tarife, Tarife Adı, Bölge Adı, Portfoy Adı, Acente İl Adı, Acente İlçe Adı, Acente Numarası, Acente Adı, Yeni Yenileme, Branş, Branş Adı, Segment, Tanzim Tarihi, Uwy (Underwritting Year), Başlama Tarihi, Bitiş Tarihi, Ana Poliçe Başlama Tarihi, Teminat Kodu, Teminat Adı, Sigortalı Numarası, Sigortalı İli, Sigortalı İlçesi, Cinsiyet, Gerçek Tüzel, Yaş, Müşteri İl Adı, Müşteri İlçe Adı, Risk İl Adı, Risk İlçe Adı, Risk Latitude, Risk Longtitude, Prim

Verilerin tanımlanması aşamasında 2 veri setinde yer alan değişkenler incelenmiş ve analiz edilmiştir. Datalarda ortak başlık olarak yer alan Başlama tarihi, Tanzim Tarihi ve Ana Poliçe Tarihi aynı tarihleri göstermektedir. UWY, poliçe başlama tarihinin yılını göstermektedir. Ana police başlama tarihlerine bakılarak buradaki yıl değeri UWY ile kıyaslanmış ve uyumlu olduğu görülmüştür. Dolayısı ile poliçe yılı UWY olarak düşünülmüştür.

Kaynak Adı, 7 faktörlü bir değişkendir. Faktör seviyeleri sırası ile; Alınan Ihtıyarı; Broker; Direk İşler; Fıbabanka; Ing Bank A.Ş.;Teb Subelerı; Yetkılı Acente

Bölge Adı, 15 faktörlü bir değişkendir.

Segment, 5 segment tanımlamıştır. Faktör seviyeleri sırası ile; Corporate; Middle Market; Personal Lines; Receiving; Small Business

Dosya Türü, 4 faktörlü bir değişkendir. Faktör seviyeleri sırası ile; Anlaşmalı Tutanak; Beyan; Zabıt; Boş

Dosyayı Açan Profil, 4 faktörlü bir değişkendir. Faktör seviyeleri sırası ile; Call center, Inbound, Outbounds

4.2.3 Verilerin araştırılması

Bu aşamada veri seti anlaşılmaya çalışılmış, tez kapsamında yürütülecek faaliyetlerin odağının elde edilen çıktılar doğrultusunda şekillendirilmesi hedeflenmiştir. Veri araştırılırken çalışmanın ilerleyen aşamalarında kullanılması planlanan değişkenlerden yararlanılmış, yeni tanımlanan bu değişkenler özelinde, incelenen döneme dair mevcut durumu anlamak ve göze çarpan unsurları belirlemeye yönelik il, ilçe ve acenteler bazında analizler gerçekleştirilmiştir. Tanımlanan değişkenler;

- Dikkate alınan parametreler odağında gerçekleşen toplam hasar adetlerinin belirlendiği hasar frekansı değişkeni
    - il bazlı gerçekleşen toplam hasar adedinin belirlendiği hasar frekansı,
    - acente bazlı gerçekleşen toplam hasar adedinin belirlendiği hasar frekansı,
- Dikkate alınan parametreler odağında üretilen toplam poliçe adedinin belirlendiği üretim frekansı değişkeni
    - il bazlı üretilen toplam poliçe adedinin belirlendiği üretim frekansı,
    - acente bazlı üretilen toplam poliçe adedinin belirlendiği üretim frekansı,
- Üretilen poliçelerin yüzde kaçının hasara dönüştüğünün tespit edilmesine yönelik hasar adeti/ üretim adeti değişkeni
- Gerçekleşen hasarların kazanılan prime oranını veren hasar prim oranı değişkeni
- Poliçenin başlatıldığı tarih ve hasarın gerçekleştiği tarih arasında geçen süreyi bizlere gösteren geçen süre değişkeni

şeklinde tanımlanmıştır.

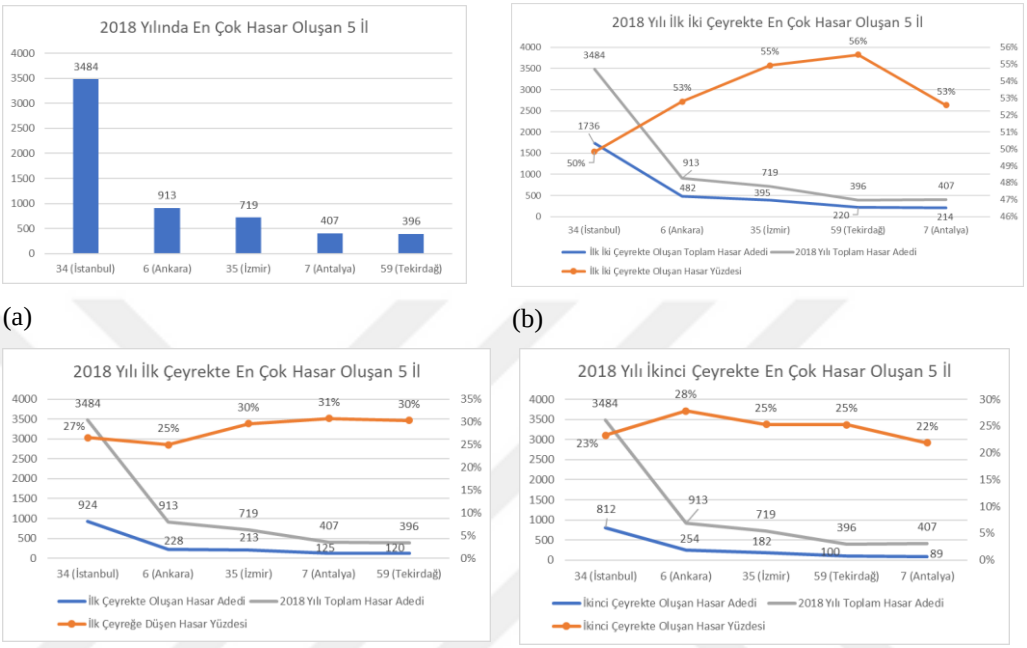
4.2.3.1 Hasar frekanslarının incelenmesi

Hasar verisi üzerinden hasar frekansları hesaplanmıştır. İlk etapta hasar frekansları, il ve acente bazlı hesaplanmıştır. Daha sonra ilk iki çeyrek kırılımlarına göre il ve

acente bazında hasar frekansları incelenmiştir. Burada hedef yılın belirli dönemlerinde herhangi bir değişimin gerçekleşip gerçekleşmediğinin tespit edilmesini sağlamaktır.

İl Bazında Hasar Frekansları Analizi

2018 yılı için il hesaplanan hasar frekansları Şekil 4.1'de verilmiştir.

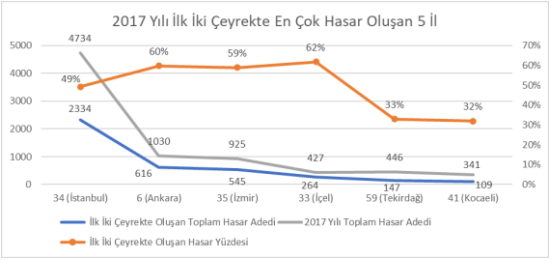


Şekil 4.1 : 2018 yılında en çok hasar oluşan 5 il (a) illerde gerçekleşen toplam hasar, (b) ilk iki çeyrek dağılımı, (c) ilk çeyrek dağılımı, (d) ikinci çeyrek dağılımı.

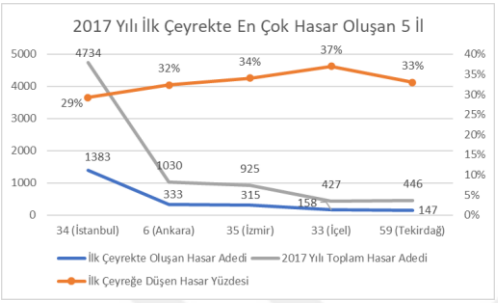
2018 yılında, yıllık toplam oluşan hasar miktarı ilk iki çeyreğe dengeli bir şekilde dağılmıştır.

2017 yılı için il hesaplanan hasar frekansları Şekil 4.2'de verilmiştir.

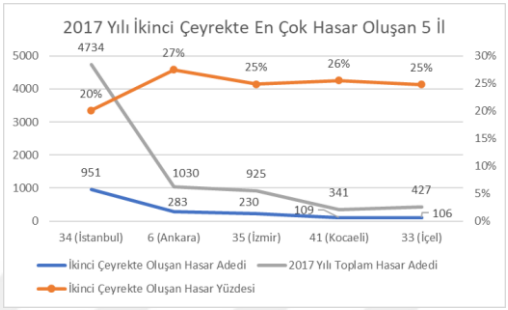
(a)



(b)



(c)



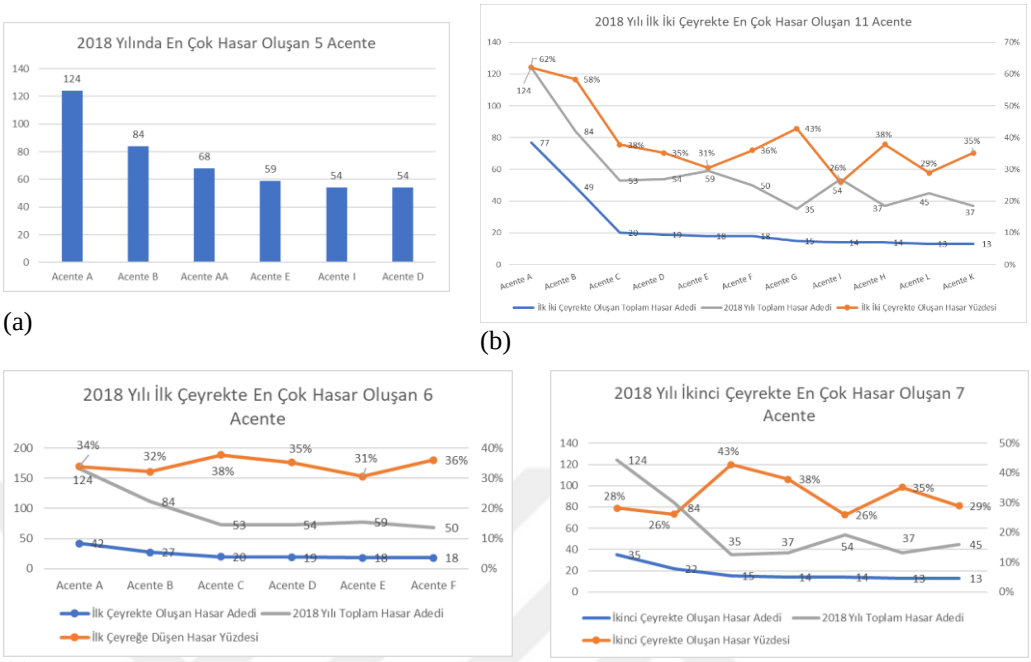
(d)

Şekil 4.2 : 2017 yılında en çok hasar oluşan 5 il (a) illerde gerçekleşen toplam hasar, (b) ilk iki çeyrek dağılımı, (c) ilk çeyrek dağılımı, (d) ikinci çeyrek dağılımı.

2017 yılında, yıllık toplam oluşan hasar miktarı ilk iki çeyreğe kısmen dengeli bir şekilde dağılmıştır. Ankara, İzmir ve İçel'de ilk iki çeyrekte oluşan hasar %50 baremini aşmış olup %60 civarında seyretmektedir. En çok hasarın gözlemlendiği diğer illerde ise (Tekirdağ ve Kocaeli) ilk iki çeyrekte gerçekleşen hasar miktarı %30 dolaylarındadır.

Acente Bazında Hasar Frekansları Analizi

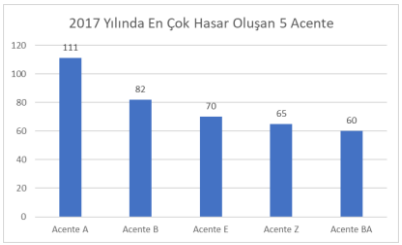
2018 yılı için acente bazlı hesaplanan hasar frekansları Şekil 4.3'de verilmiştir.



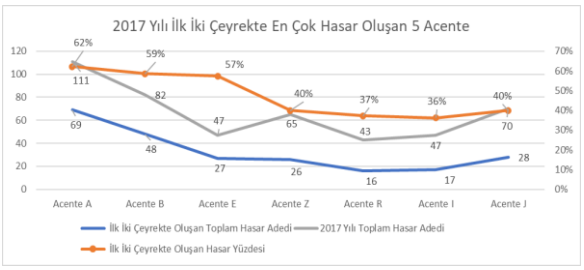
Şekil 4.3 : 2018 yılında en çok hasar oluşan 5 acente (a) acentelerde gerçekleşen toplam hasar, (b) ilk iki çeyrek dağılımı, (c) ilk çeyrek dağılımı, (d) ikinci çeyrek dağılımı.

2018 yılı verilerine bakıldığında, acente bazlı en yünsek hasarın Acente A'da olduğu görülmektedir. Acente bazında gerçekleşen hasarların yılın ilk iki çeyreğindeki dağılımı incelendiğinde ise Acente A ve Acente B de yıllık toplam hasarının %60'ının ilk iki çeyrekte olduğu gözlemlenirken, diğer acentelerde %30 civarında hasar gerçekleşme oranı tespit edilmiştir.

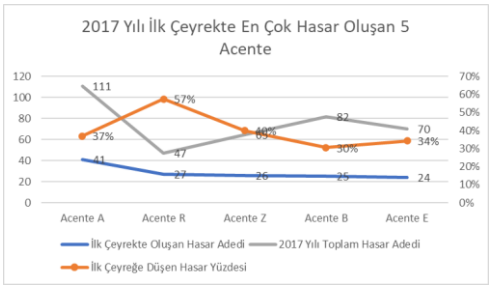
2018 yılı için acente bazlı hesaplanan hasar frekansları Şekil 4.4'de verilmiştir.



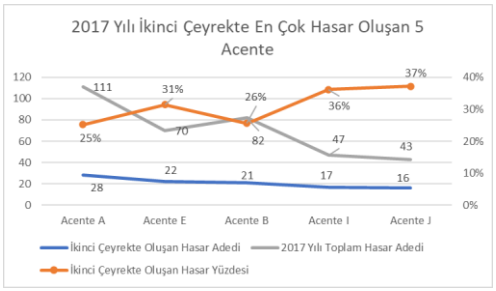
(a)



(b)



(c)



(d)

Şekil 4.4 : 2017 yılında en çok hasar oluşan 5 acente (a) acentelerde gerçekleşen toplam hasar, (b) ilk iki çeyrek dağılımı, (c) ilk çeyrek dağılımı, (d) ikinci çeyrek dağılımı.

2018 yılı ve 2017 yılında en çok hasar oluşan iki acentenin (Acente A ve Acente B) aynı acenteler olduğu tespit edilmiştir. 2017 yılı verileri incelendiğinde, acente bazında gerçekleşen hasarların yılın ilk iki çeyreğindeki dağılımında 2018 yılına benzer olarak Acente A ve Acente B de yıllık toplam hasarının %60'ının ilk iki çeyrekte olduğu gözlemlenirken, Acente E'de %57, diğer acentelerde ise daha düşük hasar gerçekleşme oranları tespit edilmiştir.

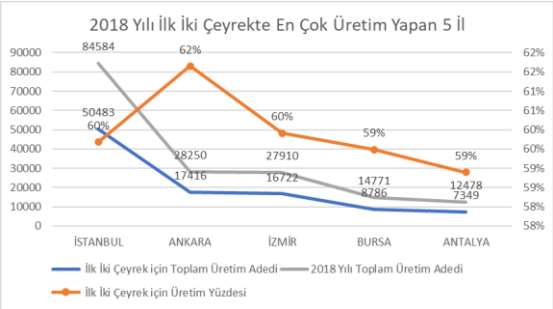
4.2.3.2 Üretim frekanslarının incelenmesi

Poliçe verisi üzerinden ilerlenerek üretim miktarları hesaplanmıştır. İlk etapta üretim frekansları (miktarları) il ve acente bazlı hesaplanmış olup, ilerleyen adımlarda ilk iki çeyrek kırınımlarına göre il ve acente bazında üretim frekansları detaylandırılmıştır.

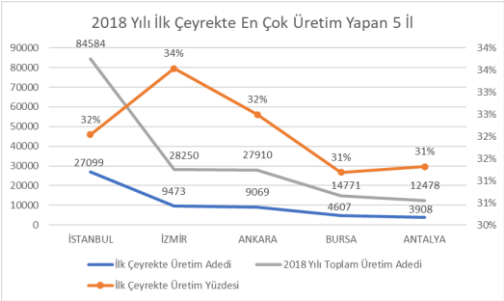
İl Bazında Üretim Frekansları Analizi

2018 yılı için il bazlı hesaplanan üretim frekansları Şekil 4.5'de verilmiştir.

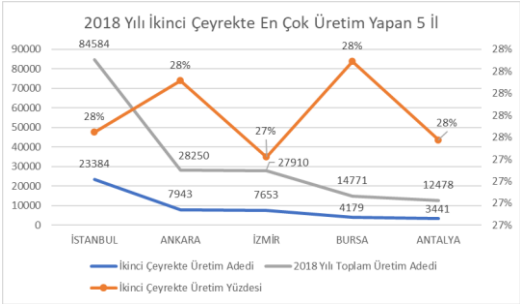
(a)



(b)



(c)



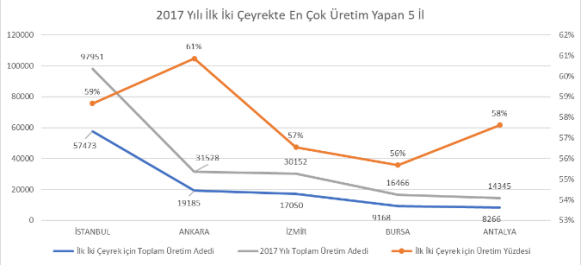
(d)

Şekil 4.5 : 2018 yılında en çok üretim yapılan 5 il (a) illerde gerçekleşen toplam üretim, (b) ilk iki çeyrek dağılımı, (c) ilk çeyrek dağılımı, (d) ikinci çeyrek dağılımı.

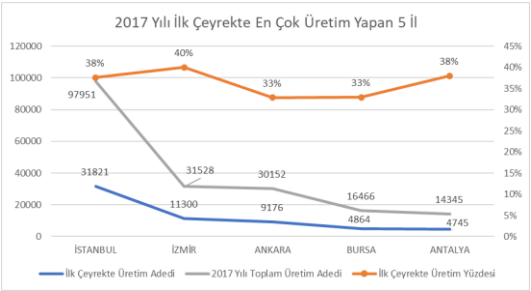
2018 yılında en fazla üretim 84584 adet ile İstanbul'da yapılmış. İstanbul'un yıllık toplam üretim miktarının ortalama %60'ı ilk iki çeyrekte gerçekleşmiştir.

2017 yılı için il bazlı hesaplanan üretim frekansları Şekil 4.6'da verilmiştir.

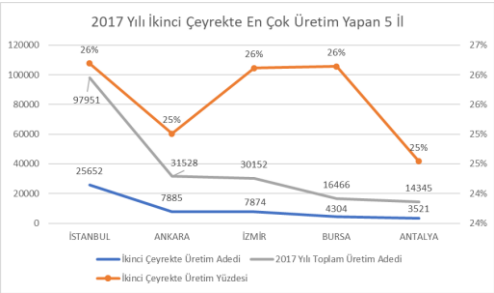
(a)



(b)



(c)



(d)

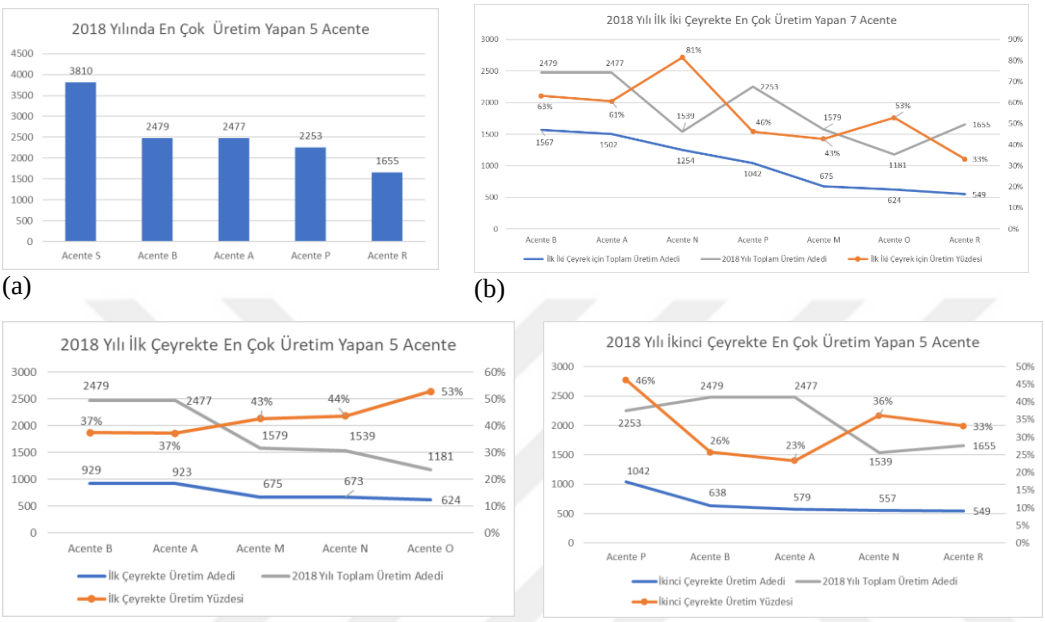
Şekil 4.6 : 2017 yılında en çok üretim yapılan 5 il (a) illerde gerçekleşen toplam üretim, (b) ilk iki çeyrek dağılımı, (c) ilk çeyrek dağılımı, (d) ikinci çeyrek dağılımı.

En yüksek üretimin gerçekleştiği iller sırası ile İstanbul, Ankara, İzmir, Bursa ve Antalya'dır. 2017 yılında, İstanbul'un yıllık toplam üretim miktarının ilk iki çeyreğe

dağılımı %60 oranlarında seyretmektedir. Ankara'da ilk iki çeyrekte oluşan hasar %60 oranını aşmıştır.

Acente Bazında Üretim Frekansları Analizi

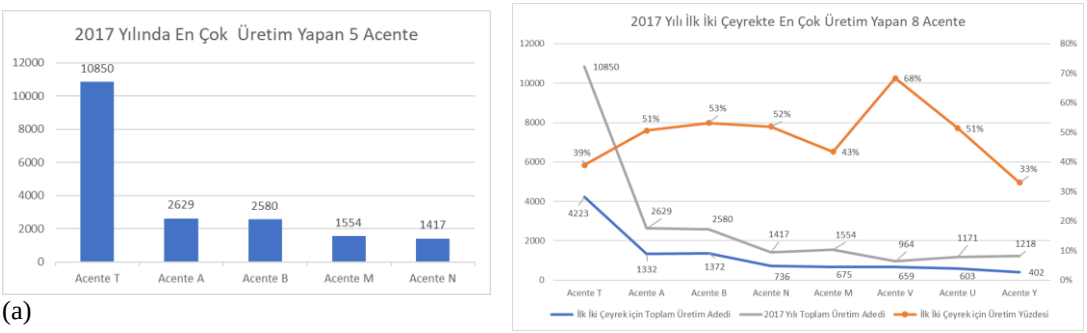
2018 yılı için il bazlı hesaplanan üretim frekansları Şekil 4.7'de verilmiştir. 2018 yılında en yüksek üretim yapan acente Acente S olarak tespit edilmiştir.



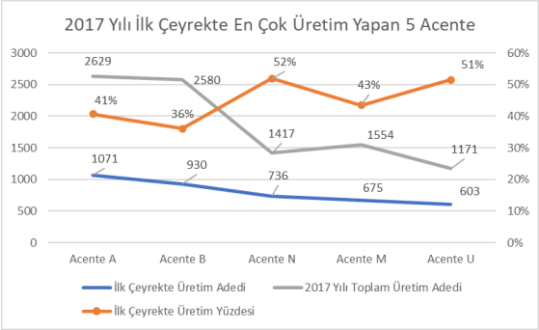
Şekil 4.7 : 2018 yılında en çok üretim yapılan 5 acente (a) acentelerde gerçekleşen toplam üretim, (b) ilk iki çeyrek dağılımı, (c) ilk çeyrek dağılımı, (d) ikinci çeyrek dağılımı.

2018 yılı verilerine bakıldığında, acente bazında yapılan üretimlerin yılın ilk iki çeyreğindeki dağılımı incelendiğinde Acente B ve Acente A'da yıllık toplam üretimin %60'ı civarında üretimin ilk iki çeyrekte tamamlandığı görülmektedir. İlk iki çeyrekte en yoğun üretimin yapıldığı üçüncü acente olan Acente N'nin, ilk iki çeyrekte yıllık gerçekleştirilen toplam üretim miktarının %81'inin tamamlandığı görülmektedir.

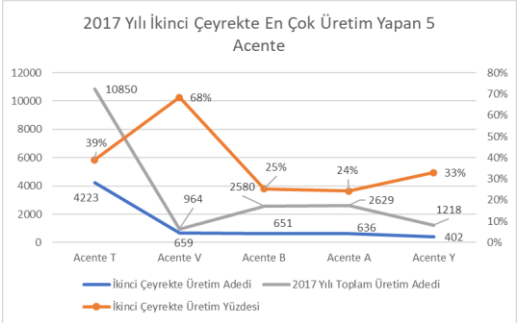
2017 yılı için il bazlı hesaplanan üretim frekansları Şekil 4.8'de verilmiştir.



(b)



(c)



(d)

Şekil 4.8 : 2017 yılında en çok üretim yapılan 5 acente (a) acentelerde gerçekleşen toplam üretim, (b) ilk iki çeyrek dağılımı, (c) ilk çeyrek dağılımı, (d) ikinci çeyrek dağılımı.

2017 yılında en fazla üretimin gerçekleştiği acente, Acente T'dir. Bu acenteyi 2018 yılında da en fazla üretimi gerçekleştirmiş olan Acente A ve Acente B takip etmektedir. Acente T kendisine en yakın sıralamadaki acenteden 4,12 kat daha fazla üretim gerçekleştirmiş olması dikkat çekmektedir.

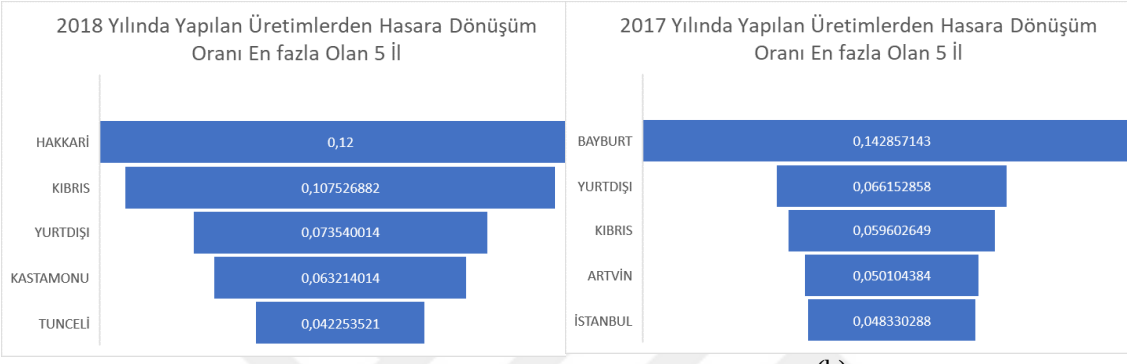
2017 yılı verileri incelendiğinde, 2017 yılı boyunca en fazla üretimi gerçekleştirmiş Acente T'nin yıllık toplam üretiminin sadece %39'unun ilk iki çeyrekte üretildiği, %61 oranındaki üretiminin ise yılın ikinci yarısında gerçekleştiği göze çarpmaktadır. Göze çarpan bir diğer nokta Acente V'nin yıllık toplam üretiminin %68'ini ilk iki yarıda tamamlamış olmasıdır. Öne çıkan acenteler incelendiğinde genel olarak ilk iki çeyrekte üretim tamamlanma oranının %50 dolaylarında olduğu görülmektedir.

Acente üretim adetlerindeki dalgalanmalar anomali tespitinde acente bazında üretim miktarlarınında değerlendirilmesi gerektiğine dair bilgi sunmaktadır.

4.2.3.3 Dönüşüm oranının incelenmesi

Belirlenen değişken ile araştırılan nokta il bazında yıllara sari üretilen poliçelerden yüzde kaçının hasara dönüştüğünün tespit edilmesidir. Dönüşüm oranı değerleri, hasar adetlerinin üretim adetlerine bölünmesi ile elde edilmiştir.

Sırası ile 2018 ve 2017 yıllarında dönüşüm oranı en yüksek olan ilk 5 il Şekil 4.9'da gösterilmiştir.



(a) (b)

Şekil 4.9 : Üretilen poliçelerin hasara dönüşüm oranı en fazla olan 5 il (a) 2018, (b) 2017.

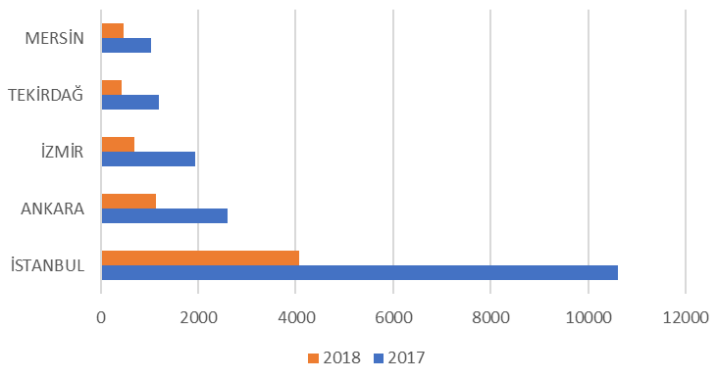
Üretim yüzde kaç oranında hasara dönüştüğü incelendiğinde; bazı illerde herhangi bir poliçe oluşturulmamış olmasına rağmen hasar girişi yapıldığı tespit edilmiştir. 2018 yılı içerisinde dönüşüm oranı en yüksek ilin Hakkari olduğu tespit edilmiştir. Ayrıca Kıbrıs'ta ve Yurtdışında üretilen poliçelerde de dönüşüm oranı yüksektir. 2017 yılı içerisinde dönüşüm oranı en yüksek ilin ise Bayburt olduğu tespit edilmiştir. 2017 yılında da Kıbrıs'ta ve Yurtdışında üretilen poliçelerde de dönüşüm oranı en yüksek dönüşüm oranları sıralamasında ilk üçte yer almıştır.

4.2.3.4 Hasar prim oranının incelenmesi

Bu oran kısaca; gerçekleşen hasarların kazanılmış prime olan oranı şeklinde tanımlanabilir.

İl Bazında Hasar Prim Oranı Analizi

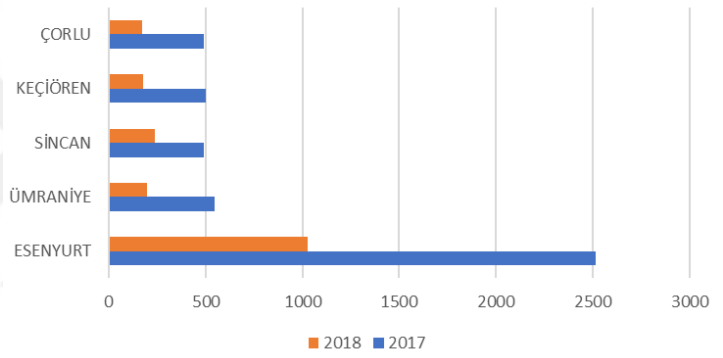
İl bazlı yapılan incelemede 2017 ve 2018 yıllarında en yükdek hasar prim oranı toplamına sahip ilk 5 ilin aynı olduğu sadece 4. ve 5. sırada yer alan illerin yıllara sarı sıralamalarının değiştiği görülmektedir. Şekil 4.10'da 2017 ve 2018 yıllarına ait ilk 5 ilçenin hpo oranları verilmektedir.



Şekil 4.10 : Hasar prim oranı toplamları en fazla olan 5 il.

İlçe Bazında Hasar Prim Oranı Analizi

İlçe bazlı yapılan incelemede 2017 ve 2018 yıllarında en yüksek hasap prim oranı toplamına sahip ilk 5 ilçenin aynı olduğu görülmektedir (Şekil 4.11).



Şekil 4.11 : Hasar prim oranı toplamları en fazla olan 5 ilçe.

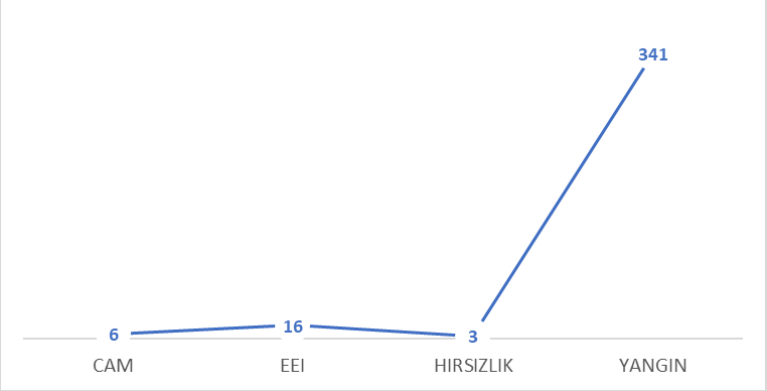
Hasar prim oranı toplamlarına göre sıralamada ilk 5'te yer alan illerde, hasar prim oranı en yüksek olan ilçeler incelenmiştir. Bu iller ve sırası ile ilde bulunan ilçeler aşağıdaki Çizelge 4.5'de verildiği üzeredir.

Çizelge 4.5 : Hasar prim oranı en yüksek olan iller ve ilçeler.

İlgili İlde En Yüksek HPO Oranı Gözlemlenen İlçeler				
İstanbul	**Ankara**	**İzmir**	**Tekirdağ**	**Mersin**
Esenyurt	Sincan	Karabağlar	Çorlu	Toroslar
Ümraniye	Keçiören	Çiğli	Çerkezköy	Mezitli
Küçükçekmece	Mamak	Buca	Kapaklı	Yenişehir
Eyüp	Etimesgut	Konak	Süleymanpaşa	Akdeniz
Maltepe	Çankaya	Torbalı	Merkez	Tarsus

4.2.3.5 Geçen sürenin incelenmesi

Geçen süre değişkeni ile poliçe başlama tarihi ve hasar tarihlerinin farkları alınarak, bu iki tarih arasında ne kadar süre geçtiği hesaplanmak istenmiştir.



Şekil 4.12 : Aynı gün içerisinde gerçekleşme frekansı.

Özellikle poliçenin oluşturulduğu gün gerçekleşen hasarlar dikkat çekmektedir. Toplamda 366 adet gözlemde poliçenin oluşturulduğu gün hasarın gerçekleştiği tespit edilmiştir. Poliçenin oluşturulduğu gün ile aynı gün içerisinde hasarın gerçekleşmesi durumunun en çok gözlemlendiği branş ise yangın branşıdır (Şekil 4.12).

Aynı gün içerisinde hasarın en yoğun gerçekleştiği iller sırası ile; İstanbul, İzmir, Ankara, Tekirdağ, Bursa, Antalya ve Kocaeli olarak tespit edilmiştir (Çizelge 4.6).

Çizelge 4.6 : Aynı gün içerisinde hasar gerçekleşmesinin en yoğun olduğu iller.

Ankara (6)	Antalya (7)	Bursa (16)	İstanbul (34)	İzmir (35)	Kocaeli (41)	Tekirdağ (59)
28	14	16	157	37	14	24

Acenteler bazında yapılan incelemede, aynı gün içerisinde üretilen poliçenin hasara dönüşümünün en yoğun gözlemlendiği acenteler, gözlemlenen durumun ilgili acentede hangi branşa kaçar adet dağıldığı ve hangi ilde yer aldığı cevap aranan sorular olmuştur. Aynı gün içerisinde hasarın gerçekleşmesi durumunun en fazla gözlemlenildiği acente İzmir'de yer alan Acente V'dir. İlgili acentede gözlemlenen durumun tespit edildiği toplamda 12 poliçenin tamamı yangın branşındandır. İlk 6 sıraya yerleşen acenteler Çizelge 4.7'de yer almaktadır.

Çizelge 4.7 : Aynı gün hasarın gerçekleşmesi durumunun acente, il ve branş bazında dağılımı.

Acente	Sigorta Branşı				İl
	CAM	EEI	HIRSIZLIK	YANGIN	
Acente B				5	Tekirdağ
Acente E				5	İçel
Acente CA				6	Mardin
Acente A				7	Tekirdağ
Acente V				12	İzmir
Acente DA				6	İstanbul

Verinin araştırılması aşamasında elde edilen bilgilere göre analizin hangi doğrultuda ilerletilmesine karar verilmiştir. Geçen süre değişkeni analiz boyunca odaklanılmak istenen branşın belirlenmesini sağlamıştır. Aynı gün içerisinde oluşturulan poliçelerin hasara dönüşümünün yoğunluklu olarak gözlemlendiği yangın branşı mülkiyet sigortası verisinde anomali tespiti çalışmasında odaklanılmaya karar verilen branşı belirlemiştir. Analizin ilerleyen aşamalarında kullanılmak üzere yangın branşına ait veriler hem poliçe hemde hasar verisinde filtrelenmiş ve bu veriler üzerinde kümeleme ve sınıflandırma algoritmaları işletilmiştir. Acente üretim adetlerindeki dalgalanmalar anomali tespitinde acente bazında üretim miktarlarınında değerlendirilmesi gerektiğine dair bilgi sunmaktadır. İl üretim ve hasar adetlerindeki sapmalar, bu değişkenlerin anormal durumların tespitinde kullanılabileceğini göstermiştir.

4.2.4 Veri kalitesinin doğrulanması

Veri kalitesi noktasında dikkat edilen hususlar belirlenmiş ve veri seti bu kurallar doğrultusunda düzenlenmiştir. Belirlenen kalite ölçütleri aşağıda belirtildiği gibidir;

- Başlangıç tarihi ile bitiş tarihi farkı negatif olanlar
- Hasar prim oranı negatif çıkanlar
- Prim rakamı 1'den büyük olanlar

4.3 Veriyi Hazırlama

4.3.1 Verilerin seçimi

Analizlerde kullanılacak verinin belirlenmesine yönelik, verilerin araştırılması aşamasında elde edilen çıkarımlar ve gerçekleştirilen birebir görüşmeler neticesinde elde edilen bilgiler doğrultusunda incelenen veriler düzenlenmiştir.
Aynı gün hasarın gerçekleşmesi durumunun en yoğun karşılaşıldığı branş Yangın branşı olduğundan, analize yangın branşı özelinde devam edilmesine karar verilmiştir. Veri setlerinde Yangın branşı filtrelendiğinde, toplam gözlem sayısı poliçe verisinde 3.349.962 gözleme, hasar verisi ise 23.744 gözleme düşmüştür.

Yıllara sari veri incelemesi yapıldığında en doğru veri girişlerin 2017 ve 2018 yıllarında yapıldığı, diğer yıllara ise, belirli ek işlemlerin rakamlarının yansıdığı görülmektedir. Çalışmaya dahil edilecek yıllar belirlenirken, tam veriye sahip yıllar üzerinden ilerlemek ve veri setlerinin dönem olarak eşitlenmesini sağlamak adına 2017 ve 2018 yıllarına ait gözlemler filtrelenmiştir. Belirlenen dönemlerin filtrasyonu sonucunda, toplam gözlem sayısı poliçe verisinde 3.271.518 gözleme, hasar verisi ise 13.428 gözleme düşmüştür.

Veriyi sağlayan sigorta şirketi ile yapılan görüşmelerde uzmanlar, daha doğru bir analiz gerçekleştirilmesine yönelik belirli değişkenlerin faktör seviyelerinden bazılarının çıkarılmasını gerekliliğini belirtmiştir. Bu doğrultuda 7 seviyeden oluşan kaynak adlarının yer aldığı değişkenin bir seviyesi, toplamda 5 seviyeden oluşan müşteri profiline yer verilen segment değişkeninin 2 seviyesi veri setinden çıkarılmıştır. Nihai durumda elde edilen verilerin detayları Çizelge 4.8'de belirtildiği gibidir.

Çizelge 4.8 : Düzenlenen veri setinde yer alan gözlem ve değişken sayısı.

Veri Setinin Adı	Gözlem Sayısı	Değişken Sayısı
Poliçe	3.267.344 gözlem	47 değişken
Hasar	13.154 gözlem	67 değişken

4.3.2 Verilerin temizlenmesi

Poliçe ve hasar datalarına belirli özel durumlar ile karşılaşılabilmektedir.

Poliçe verisi, cari poliçe numaraları bazında sıralanmaktadır, her cari poliçe numarası bir poliçeye karşılık gelmektedir. Veri setinde aynı cari poliçe numaralarının tekrar ettiği gözlemlenilmiştir. Bu tip tekrarların temel sebebi poliçeye sonradan yapılan eklemelerdir. Bunlara zeyil adı verilmektedir. Poliçe bazlı birden fazla teminatın oluşturulması ya da farklı işlemler sebebi ile veri satırlarında aynı cari poliçe numarası çok defa tekrar etmektedir. Teminatlatlar için ödenen primlerin ya da iptallerin tutarları, ilişkili oldukları cari poliçe numaraları özelinde veri gruplanmış ve poliçe bazlı, poliçeye yapılan tüm değişiklikleri kapsayan toplam prim tutarı hesaplanmıştır. Devamında tekrar eden poliçeler veriden çıkarılmıştır.

Elde edilen prim toplamları incelenmiş ve bazı gözlemlerde çok cüzi rakamların (örn: 0,00001) olduğu tespit edilmiştir. Bu tarz rakamlar ile neden karşılaşıldığı araştırıldığında, bu rakamların, belirli durumlarda (örn: poliçe iptali) banka ödemeleri sonrası geri ödemeler sonucu ilgili poliçe özelinde hesapta kalan artık miktarlardan kaynaklandığı tespit edilmiştir. Veri kalitesi adına prim miktarları minimum 1'den büyük olan poliçeler filtrelenmiştir. Bu işlem sonucu poliçe verisinin toplam gözlem sayısı 582.590'ye düşmüştür.

Hasar datasında kontrol sağlanmış ve burada da tekrar eden poliçe numaralarına rastlanmıştır. Bu veride poliçe numaralarının tekrar etmesinin temel sebebi bir poliçe özelinde birden fazla hasarın gerçekleşebiliyor olmasında kaynaklanmaktadır. Dolayısı ile her satırın bir hasar tazminat talebini resmettiği bu veri setinde cari poliçe numaraları üzerinden veriyi düzenlemek doğru bir yaklaşım olmayacaktır. Burada cari poliçe numaraları ve hasar tarihleri birleştirilerek yeni bir değişken tanımlanmıştır. Tanımlanan bu değişken özelinde toplam ödenen hasar tutarları toplanmıştır. Devamında hasar verisinden, tanımlanan yeni değişkende tekrar eden gözlemler çıkarılmıştır.

Hasar bazlı toplam ödenen hasar tutarı değişkeni belirlenirken bazı noktalara dikkat edilmiştir. Verilerin seçimi aşamasında, veri setlerinin dönem olarak eşitlenmesi adına 2017 ve 2018 yılları seçilmiştir. Ancak ilgili herhangi bir yıl için oluşturulan poliçede hasar gerçekleşmesi durumunda, ilgili poliçe için belirlenen hasar tutarının

ödemesi her zaman aynı yıl içerisinde gerçekleşmemektedir. Eksper muayeneleri, dava açılan dosyalar hasar ödemelerinde gecikmeye sebep olan bazı örneklerdir. Bu noktada poliçe bazlı toplam ödenen hasar tutarları belirlenirken sadece; ödenen tazminat, sovtaj alınan rücu, sovtaj, alınan rücu, ödenen masraf, ödenen eksper, ödenen avukat masrafı değişkenleri toplanarak elde edilen 'ödenen hasar' değişkeninin değerleri dikkate alınmamış, ayrıca hasar ihbarı esnasında tazminat talebinde bulunan sigortalı ve sigortacı arasında gerçekleşen görüşme neticesinde belirlenen muallak hasar bedelide analize dahil edilmiştir. Toplam ödenen hasar değişkeni oluşturulurken; hasar ödemesi gerçekleştirilen gözlemler için ödenen hasar tutarı, henüz ödemesi gerçekleşmemiş gözlemler için muallak hasar tutarı dikkate alınmıştır. Nihai durumda veri setlerinde bulunan gözlem sayıları Çizelge 4.9'da gösterildiği üzeredir.

Çizelge 4.9 : Düzenlenen veri setinde yer alan gözlem sayısı.

Veri Setinin Adı	Gözlem Sayısı
Poliçe	582.590 gözlem
Hasar	12.407 gözlem

4.3.3 Verilerin entegrasyonu

Yapılan işlemler neticesinde elde edilen poliçe ve hasar dataları cari poliçe numaraları üzerinden birleştirilmiştir. Elde edilen birleştirilmiş veri setinde toplam gözlem sayısı 584.226 adettir. Birleştirme öncesi hasar ve poliçe veri setlerine dahil edilen değişkenler seçilirken türetilmesi planlanan değişkenler dikkate alınmıştır. Nihai durumda elde edilen veri ile değişken türetme aşamasına geçilmiştir.

4.3.4 Verilerin oluşturulması (öznitelik çıkarımı)

Yapılan araştırmalar neticesinde, elde bulunan verilerden türetilmesi ve incelenilmesi gereken değişkenler belirlenmiştir. Kullanılmasının uygun olduğu tespit edilen göstergelerden, mevcut veri dolayında türetilebilecek (feature extraction) olanlar seçilmiştir. Seçilen göstergeler odağında türetilen yeni değişkenler sırası ile; hasar prim oranı, üretim frekansları (prim üretim adeti), hasar frekansları (hasar üretim adeti), hasar adetinin prim adetine oranı, geçen süredir (poliçe başlama tarihi ve

hasar tarihlerinin farkları). Belirlenen bu yeni değişkenler verinin incelenmesinde ve diğer analizlerde kullanılmıştır.

Hasar datasında yer alan; ödenen tazminat, sovtaj alınan rücu, sovtaj, alınan rücu, ödenen masraf, ödenen eksper, ödenen avukat masrafı değişkenleri toplanarak 'ödenen hasar' değişkeni tanımlanmıştır. Ayrıca hasar prim oranı hesabında kullanılmak üzere hasar ödemesi gerçekleştirilen gözlemler için ödenen hasar tutarının, henüz ödemesi gerçekleşmemiş gözlemler için muallak hasar tutarının dikkate alındığı 'toplam ödenen hasar tutarı' değişkeni tanımlanmıştır. Hasar ödemesi gerçekleştirilen gözlemlerde, ilk tazminat talebi esnasında belirlenen muallak hasar bedeli ile dosya nihayetlendirildiğinde ödenen hasar bedelinin farkları alınarak 'muallak tutar ve hasar tutarı farkı' değişkeni tanımlanmıştır. Burada hedef en başta çok yüksek bir bedel belirlenmesine rağmen, gerçek durum analiz edildiğinde ödenen bedel arasındaki farkı yakalayarak anomali durumlar ile ilişkilendirilebilirliğini incelemektir.

Kümeleme ve sınıflandırma modellerinde kullanılmak üzere ayrı değişkenler türetilmiştir.

Kümeleme analizinin odağı anomali olabilecek gözlemlerin tespitine yönelik olduğunda bu çalışma kapsamında türetilen değişkenler yoğunluklu olarak belirlenen anomali senaryosunu destekleyecek nitelikte olmuştur. Kümeleme analizine yönelik türetilen değişkenler; hasar prim oranı (hpo), poliçe üretim ve hasarın gerçekleşmesi arasında geçen süre (gecen_sure), belirtilen muallak hasar bedeli ve gerçekleşen hasar bedeli arasındaki fark (muallak_fark_hasar), ilçenin aylık toplam üretim adedi (aylık_uretim_adedi), ilçenin bir önceki aya ait üretim adedi (aylık_1_uretim_adedi), ilçenin haftalık üretim adedi (haftalık_uretim_adedi), ilçenin bir önceki haftaya ait üretim adedi (haftalık_1_uretim_adedi), ilçenin aylık hasar adedi (aylık_hasar_adedi), ilçede bir önceki ay gerçekleşen hasar adedi (aylık_1_hasar_adedi), ilçenin haftalık hasar adedi (haftalık_hasar_adedi), ilçenin bir önceki haftaya air hasar adedi (haftalık_1_hasar_adedi), acente aylık üretim adedi (acente_aylık_üretim), acentenin bir önceki aya ait üretim adedi (acente_aylık_1_üretim), acentenin haftalık üretim adedi (acente_haftalık_uretim), acentenin bir önceki haftaya ait üretim adedi (acente_haftalık_1_uretim), acentenin aylık hasar adedi (acente_aylık_hasar), acentenin bir önceki aya ait hasar adedi

(acente_aylık_1_hasar), acentenin haftalık hasar adedi (acente_haftalık_hasar) ve acentenin bir önceki haftata ait hasar adedi (acente_haftalık_1_hasar).

Sınıflandırma modellerinde temel hedef kümeleme ile tespit edilen anomali durumların, poliçe oluşturulması aşamasında tespit edilebilirliğini sağlamaktır. Burada kullanılan modellerin performanslarını geliştirmeye yönelik farklı denemelerde kullanılmak üzere değişkenler türetilmiştir.

Poliçe oluşturulurken sigorta ettirenden alınan sabit bilgilerden; sigorta için ödenen toplam prim tutarı (sum_prim), sigortanın oluşturulduğu kaynak (kaynak_adi), sigortayı oluştura acente (acente_no), sigortanın oluşturulduğu segment (segment), sigorta talebinin geldiği ilin tanımlandığı bölge adı (bölge_adi), sigortayı oluşturan potföy (portfoy_adi) ve sigortanın oluşturulduğu tarife (tarife_adi) değişkenleri analize dahil edilmiştir.

Bu değişkenlere ek olarak bir poliçe numarası kapsamında toplamda kaç adet zeyl gerçekleştirildiğini belirten 'zeyl sayısı' değişkeni tanımlanmıştır. İl ve ilçelerde faktör sayısı çok yüksek olduğundan, değişkenler direk kullanılmak istendiğinde belirlenen bazı modellerde sorun yaşanmaktadır. Bu noktada ilgili poliçe oluşturulurken sigortayı yapan sigorta şirketi/ acentenin bulunduğu il (sigorta ettiren il) ve sigorta talebinde bulunulan ilin (risk ili) aynı olup olmama durumu 'aynı şehir' değişkeni ile tanımlanmıştır. Aynı yaklaşım ile sigorta ettiren ilçe ile risk ilçesinin aynı olup olmama durumu 'aynı ilçe' değişkeni ile tanımlanmıştır.

Model performanslarını arttırmaya yönelik geliştirilen diğer değişkenler;

- Aşırı artışı takip eden değişkenler;
    - İl_ilçe_üretim_aşırı_artış: İl_ilçe bazında hem aylık üretim artışı için belirlenen sınırı hem de haftalık üretim artışı için belirlenen sınırı aşan gözlemleri belirten değişken
    - İl_ilçe_üretim_aşırı_artış_vol2: İl_ilçe bazında aylık üretim artışı için belirlenen sınır ya da haftalık üretim artışı için belirlenen sınırdan birini aşan gözlemleri belirten değişken (Burada hedef poliçe üretimlerinde bir önceki ay aşırı bir artış olmuş ama bir önceki hafta öyle hızlı bir artış olmadıysa onlarıda analize dahil edebilmek.)

- İl_ilçe_hasar_aşırı_artış: İl_ilçe bazında hem aylık hasar artışı için belirlenen sınırı hem de haftalık hasar artışı için belirlenen sınırı aşan gözlemleri belirten değişken
- İl_ilçe_hasar_aşırı_artış _vol2: İl_ilçe bazında aylık hasar artışı için belirlenen sınır ya da haftalık hasar artışı için belirlenen sınırdan birini aşan gözlemleri belirten değişken (Burada hedef hasar oluşum miktarlarında bir önceki ay aşırı bir artış olmuş ama bir önceki hafta öyle hızlı bir artış olmadıysa onlarıda analize dahil edebilmek.)
- acente_ üretim_aşırı_artış: Acente bazında hem aylık üretim artışı için belirlenen sınırı hem de haftalık üretim artışı için belirlenen sınırı aşan gözlemleri belirten değişken
- acente_ üretim_aşırı_artış_vol2: Acente bazında aylık hasar artışı için belirlenen sınır ya da haftalık hasar artışı için belirlenen sınırdan birini aşan gözlemleri belirten değişken (Burada hedef üretim oluşum miktarlarında bir önceki ay aşırı bir artış olmuş ama bir önceki hafta öyle hızlı bir artış olmadıysa onlarıda analize dahil edebilmek.)
- acente_hasar_aşırı_artış: Acente bazında hem aylık hasar artışı için belirlenen sınırı hem de haftalık hasar artışı için belirlenen sınırı aşan gözlemleri belirten değişken

- Günlük hasar prim oranını takip eden değişkenler;
  - tarife_gun_hpo: İlgili tarifenin ana poliçe başlama tarihlerinde hpo değerleri hesaplanmıştır. (Burada amaç poliçenin oluşturulmak istendiği tarife için, poliçenin oluşturulduğu tarihte ödenen primlerin hasara dönüşme oranını bulmaktır.)
  - il_gun_hpo: İlgili ilin ana poliçe başlama tarihlerinde hpo değerleri hesaplanmıştır. (Burada amaç poliçenin oluşturulmak istendiği il için, poliçenin oluşturulduğu tarihte ödenen primlerin hasara dönüşme oranını bulmaktır.)
  - ilce_gun_hpo: İlgili ilçenin ana poliçe başlama tarihlerinde hpo değerleri hesaplanmıştır. (Burada amaç poliçenin oluşturulmak istendiği ilçe için, poliçenin oluşturulduğu tarihte ödenen primlerin hasara dönüşme oranını bulmaktır.)
  - kaynak_gun_hpo: İlgili kaynağın ana poliçe başlama tarihlerinde hpo değerleri hesaplanmıştır. (Burada amaç poliçenin oluşturulmak istendiği

kaynak için, poliçenin oluşturulduğu tarihte ödenen primlerin hasara dönüşme oranını bulmaktır.)

- acente_gun_hpo: İlgili acentenin ana poliçe başlama tarihlerinde hpo değerleri hesaplanmıştır. (Burada amaç poliçenin oluşturulmak istendiği acente için, poliçenin oluşturulduğu tarihte ödenen primlerin hasara dönüşme oranını bulmaktır.)
- portfoy_gun_hpo: İlgili potföyün ana poliçe başlama tarihlerinde hpo değerleri hesaplanmıştır. (Burada amaç poliçenin oluşturulmak istendiği portföy için, poliçenin oluşturulduğu tarihte ödenen primlerin hasara dönüşme oranını bulmaktır.)

4.3.5 Verilerin biçimlendirilmesi

Kümeleme analizine geçmeden önce, kümeleme modelinin veri üzerinde doğru bir şekilde işletilebilmesi için, farklı değişkenlerin sahip olduğu farklı birimlerin kümeleme üzerindeki etkisini ortadan kaldırmak adına nümerik değerli değişkenler ölçeklendirilmiştir (scale).

5. MODELLEME

5.1 Kümeleme

5.1.1 K-Ortalamalar

Anomali tespitinde kümeleme yaklaşımı tercih edilmiştir. Kullanım yaygınlığı ve incelenen örneklerde gösterdiği başarılı performans dikkate alınarak k-means kümele yöntemi ile ilerlenmiştir.

5.1.1.1 Küme performansının ölçümü

Dışsal kalite ölçümleri (External quality measures) dışsal bilgiye (etiketli veri) ihtiyaç duyar, bu çalışma kapsamında herhangi bir etiketli veri bulunmamaktadır.

Dolayısı ile oluşturulan kümelerin performanslarının ölçümünde ki bu kalite ölçümü olarak ta geçmektedir, herhangi bir dışsal bilginin mevcut olmayışından kaynaklı, içsel kalite değerlendirme ölçütleri kullanılmıştır. Farklı k değerleri için kümeleme performanslarının kıyaslanmasında içsel kalite değerlendirme ölçütlerinden Ortalama silhoutte genişliği tercih edilmiştir.

5.1.1.2 Kullanılan değişkenler

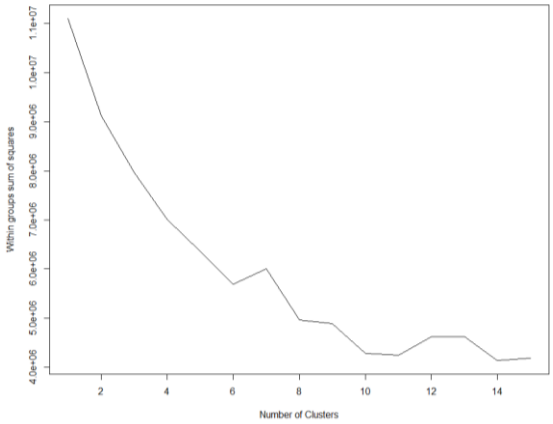
Belirlenen anomali senaryosunu yansıttığı düşünülerek kümelemede kullanılan değişkenler Çizelge 5.1'de belirtildiği gibidir.

Çizelge 5.1 : Kümeleme değişkenleri.

Kümelemede Kullanılan Değişkenler
Hasar prim oranı - hpo,
Poliçe üretim ve hasarın gerçekleşmesi arasında geçen süre - gecen_sure,
Belirtilen muallak hasar bedeli ve gerçekleşen hasar bedeli arasındaki fark - muallak_fark_hasar,
Aylık üretim adedi - aylık_uretim_adedi,
Bir önceki aylık üretim adedi - aylık_1_uretim_adedi,
Haftalık üretim adedi - haftalık_uretim_adedi,
Bir önceki haftalık üretim adedi - haftalık_1_uretim_adedi,
Aylık hasar adedi - aylık_hasar_adedi,
Bir önceki aylık hasar adedi - aylık_1_hasar_adedi,
Haftalık hasar adedi - haftalık_hasar_adedi,
Bir önceki haftalık hasar adedi - haftalık_1_hasar_adedi,
Acente aylık üretim adedi - acente_aylık_üretim,
Acente bir önceki aylık üretim adedi - acente_aylık_1_üretim,
Acente haftalık üretim adedi -acente_haftalık_uretim,
Acente bir önceki haftalık üretim adedi - acente_haftalık_1_uretim,
Acente aylık hasar adedi - acente_aylık_hasar,
Acente bir önceki aylık hasar adedi - acente_aylık_1_hasar,
Acente haftalık hasar adedi -acente_haftalık_hasar,
Acente bir önceki haftalık hasar adedi - acente_haftalık_1_hasar

5.1.1.3 Dirsek grafiği

En uygun küme sayısına karar verilirken dirsek grafiği çizdirilmiş ve average silloutte değerleri hesaplanmıştır (Şekil 5.1).



Şekil 5.1 : Dirsek grafiği.

k = 2, 3, 4 ve 6 için elde edilen ortalama silloutte değerleri Çizelge 5.2'de gösterildiği gibidir.

Çizelge 5.2 : Ortalama Silloutte değerleri.

k	Average Silloutte
2	0.4900723
3	0.2802728
4	0.3867878
6	0.2652201

Ortalama silloutte değerleri dikkate alındığında, a en yüksek değerin elde edildiği 'k=2' nin en uygun küme sayısı olduğuna karar verilmiştir. Mevcutta talep edilen anomali olma ve olmama durumlarını resmetmek adına uyumlu bir sonuca ulaşılmıştır.

K-means kümeleme ile tespit edilen iki kümeden ilkinde bulunan gözlem sayısı 1858 adet, ikincisinde ise 582.368 adettir.

Küme ortalamaları Çizelge 5.3'de belirtildiği gibidir.

Çizelge 5.3 : Oluşan kümelerin değişken bazında ortalama değerleri.

Değişken	1. Küme Ortalamaları	2. Küme Ortalamaları
Hasar prim oranı – hpo,	82.86	0.49
Poliçe üretim ve hasarın gerçekleşmesi arasında geçen süre - gecen_sure	138.12	361.85
Belirtilen muallak hasar bedeli ve gerçekleşen hasar bedeli arasındaki fark - muallak_fark_hasar	-7299.68	-43.69
Aylık üretim adedi - aylık_uretim_adedi	875.71	226.91
Bir önceki aylık üretim adedi - aylık_1_uretim_adedi	846.01	197.33
Haftalık üretim adedi - haftalık_uretim_adedi	184.86	63.97
Bir önceki haftalık üretim adedi - haftalık_1_uretim_adedi	185.26	43.77
Aylık hasar adedi - aylık_hasar_adedi	53.07	0.11
Bir önceki aylık hasar adedi - aylık_1_hasar_adedi	42.60	0.084
Haftalık hasar adedi - haftalık_hasar_adedi	17.01	0.04
Bir önceki haftalık hasar adedi - haftalık_1_hasar_adedi	10.19	0.02
Acente aylık üretim adedi - acente_aylık_üretim	85.97	170.35
Acente bir önceki aylık üretim adedi - acente_aylık_1_üretim	60.44	55.29
Acente haftalık üretim adedi -acente_haftalık_uretim	34.88	126.30
Acente bir önceki haftalık üretim adedi - acente_haftalık_1_uretim	13.82	11.92
Acente aylık hasar adedi - acente_aylık_hasar	5.89	0.04
Acente bir önceki aylık hasar adedi - acente_aylık_1_hasar	3.01	0.02
Acente haftalık hasar adedi -acente_haftalık_hasar	3.31	0.02
Acente bir önceki haftalık hasar adedi - acente_haftalık_1_hasar	0.78	0.004

5.1.1.4 Küme tanımları

Kümelerin ortalama değerleri dikkate alınarak küme tanımlamaları gerçekleştirilmiştir;

- 1.küme anomali gözlemler
- 2.küme normal gözlemler

Anomali olarak belirlenen 1.küme için tespitler;

- Bu kümede poliçenin hasara dönüşme oranı çok yüksektir
- Bu küme içerisindeki gözlemlerde hasarın oluşma süresi diğer kümeye nazaran daha kısa süreleri kapsamaktadır

- Bu kümede muallak olarak belirtilen hasar miktarı ile gerçekleşen hasar tutarı/ ödenen gerçek hasar tutarı arasında büyük farklar mevcuttur.
- Üretim adetlerinde ve hasar adetlerinde değişim olan gözlemler bu veri setinde yer almıştır

şeklindedir.

5.2 Sınıflandırma

5.2.1 Kullanılan araçlar

VPN bağlantısı ile 16 GB RAM kapasiteli bir bilgisayar üzerinden çalışmalar yürütülmüştür. R programlama dili kullanılmıştır. Çalışma kapsamında R'da kullanılan paketler ve fonksiyonlar Çizelge 5.4'de belirtilmiştir.

Çizelge 5.4 : Kullanılan araçlar.

Algoritma	Paket	Fonksiyon
K-Means	factoextra	kmeans()
Decision Tree	rpart	rpart()
Bagging Tree	ipred	bagging()
Random Forest	randomForest	randomForest()
Xgboost	xgboost	xgb.train()
Lightgbm	lightgbm	lightgbm()
Catboost	catboost	catboost.train()

5.2.2 Sınıflandırma modelleri

K-ortalamalar kümeleme yaklaşımı ile tespit edilen anormal durumların poliçelerin oluşturulması aşamasında tespitine yönelik geliştirilmesi hedeflenen modelde ağaç tabanlı algoritmaların performansları değelerndirilmiştir. Tez kapsamında üzerine çalışılan modeller sırası ile;

- Karar Ağaçları
- Bagging Ağacı
- Rastgele Orman
- Xgboost
- Lightgbm

- CatBoost

şeklindedir.

Belirlenen modeller ile gerçekleştirilen çalışma 5 aşamalı olarak yürütülmüştür. İlk olarak poliçelendirme aşamasında cevaplandırılan değişkenler üzerinden modeller çalıştırılmıştır.

İkinci aşamada, üretim ve hasarlarda aşırı artışların tespit edilmesine yönelik türetilen değişkenler modele dahil edilmiştir. Ancak türetilen yeni değişkenler ile model performansları istenen seviyelere ulaşmamıştır. Üçüncü aşamada model performanslarını geliştirmeye yönelik günlük hasar prim oranlarının hesaplandığı değişkenler türetilmiş ve analize dahil edilmiştir. Dördüncü adımda model performanslarını geliştirmeye yönelik hiper parametre optimizasyonu yapılmıştır. İlk dört adımda gerçekleştirilen faaliyetler neticesinde istenen seviyede performanlara erişilememesinden dolayı son olarak yeniden örnekleme yaklaşımları üzerine çalışılmıştır.

5.2.2.1 Birinci aşama: baz model değişkenleri üzerinden model performansı

İlk olarak poliçelendirme aşamasında cevaplandırılan değişkenler üzerinden modeller çalıştırılmıştır. İlk olarak, yüklenim aşamasında cevaplanan değişkenler üzerinde modeller çalıştırıldı. Burada modelin daha hızlı yanıt vermesini sağlamak için il, ilçe gibi çok faktörlü kategorik değişkenlerde özellik mühendisliği yapılmıştır. Tanımlanan 'aynı şehir' değişkeni ile sigortalı şehrin ve riskli şehrin aynı olup olmadığı belirlenmiş ve modele dahil edilmiştir. Aynı çalışma ilçe için de yapılmıştır. Bu aşamada ek olarak incelenen poliçede kaç adet değişiklik yapıldığının gösterildiği zeyl sayısı değişkenide modele dahil edilmiştir. Modelde kullanılan değişkenler sırası ile; anomali (y), sum_prim, tarife_adi, segment, bolge_adi, ayni_sehir, ayni_ilce, zeyl_sayısı şeklindedir. Belirtilen değişkenler özelinde elde edilen model performansları Çizelge 5.5’te belirtildiği gibidir.

Çizelge 5.5 : Birinci aşama model performansları.

Algoritma	Accuracy	Sensitivity	Specificity	F1	AUC
Decision Tree	0.9968	0.000000	1.0000	0.0000	0.5
Bagging Tree	0.9965	0.1328	0.9992	0.1922	0.6078
Random Forest	0.9969	0.0448	0.9999	0.0850	0.5995
Xgboost	0.9967	0.0679	0.9997	0.1174	0.5338
Lightgbm	0.9966	0.0609	0.9996	0.1046	0.5303
Catboost	0.9968	0.0520	0.9997	0.09250	0.5259

5.2.2.2 İkinci aşama: türetilen değişkenler üzerinden model performansı

İkinci aşamada, üretim ve hasarlarda aşırı artışların tespit edilmesine yönelik türetilen değişkenler modele dahil edilmiştir. Bu artışı takip eden modeller oluşturulurken, hem ilgili tarihten bir ay öncesine dair üretim ve hasar bilgileri dikkate alınmış hemde bir hafta öncesine ait bilgiler dikkate alınmıştır. Bu aşamada modele eklenen değişkenler sırası ile; il_ilçe_üretim_aşırı_artış, il_ilçe_üretim_aşırı_artış_vol2, il_ilçe_hasar_aşırı_artış, il_ilçe_hasar_aşırı_artış_vol2, acente_üretim_aşırı_artış, acente_üretim_aşırı_artış_vol2, acente_hasar_aşırı_artış şeklindedir. Bu değişkenler ile ani üretim artışları yada hasar artışlarının takip edilmesi hedeflenmiştir. Ancak türetilen yeni değişkenler ile model performansları istenen seviyelere ulaşmamıştır. Belirlenen değişkenler özelinde elde edilen model performansları Çizelge 5.6'te belirtildiği gibidir.

Çizelge 5.6 : İkinci aşama model performansları.

Algoritma	Accuracy	Sensitivity	Specificity	F1	AUC
Decision Tree	0.9968	0.0000	1.0000	0.00	0.5
Bagging Tree	0.9964	0.1831	0.9990	0.2457	0.6879
Random Forest	0.997	0.0736	0.9999	0.1346	0.6932
Xgboost	0.996	0.2038	0.9985	0.2497	0.6012
Lightgbm	0.9966	0.2038	0.9991	0.2792	0.6014
Catboost	0.9969	0.1274	0.9996	0.2046	0.5635

5.2.2.3 Üçüncü aşama: türetilen değişkenler üzerinden model performansı

Üçüncü aşamada model performanslarını geliştirmeye yönelik, bazı değişkenler özelinde günlük hasar prim oranlarını hesaplayan değişkenler türetilmiştir. Türetilen bu yeni değişkenler ile hedef, günlük hasar prim oranı hesaplanan değişken için incelenen gün özelinde üretilen poliçelerin hasara dönüşüm oranları takip edilmiştir. Bu aşamada modele eklenen değişkenler sırası ile; tarife_avg_hpo, ilce_avg_hpo, kaynak_avg_hpo, il_avg_hpo, acente_avg_hpo şeklindedir. Belirlenen değişkenler özelinde elde edilen model performansları Çizelge 5.7'te belirtildiği gibidir.

Çizelge 5.7 : Üçüncü aşama model performansları.

Algoritma	Accuracy	Sensitivity	Specificity	F1	AUC
Decision Tree	0.9968	0.0000	1.0000	0.0000	0.5
Bagging Tree	0.9971	0.2567	0.9994	0.3579	0.7906
Random Forest	0.9508	0.9174	0.9508	0.1058	0.9794
Xgboost	0.9962	0.2195	0.9987	0.2736	0.6091
Lightgbm	0.9968	0.2299	0.9992	0.3176	0.6146
Catboost	0.9969	0.1633	0.9995	0.2506	0.5814

5.2.2.4 Dördüncü aşama: hiper parametre optimizasyonu

2 ayrı aşamada türetilen yeni değişkenler sonucu elde edilen model performansları istenen seviyelere ulaşmadığı için hiper parametre optimizasyonuna (hyperparameter tuning) başvurulmuştur. R da bulunan caret paketi, makine öğrenme algoritmalarının parametrelerini ayarlamak için pratik bir olanak sağlamaktadır. Ancak caret paketi içerisinde tüm makine öğrenmesi algoritmaları yer almamaktadır. Dolayısı ile kullanılan bazı algoritmaları hiper parametre optimizasyonu için alternatif paketlerden yararlanılmıştır.

Yürütülen hiper parametre optimizasyonu çalışmasındaki asıl hedef; kullanılan algoritmaların birer kara kutu (black-box) olduğu kabulü üzerinden devam etmek yerine, algoritmaları parametre bazında daha iyi anlayıp, detaylandırıp incelemek ve algoritmayı mevcut probleme uygun forma getirmeye çalışmaktır.

Modellerin hiper parametre optimizasyonu işlemi ızgara arama (grid search) yaklaşımı benimsenerek gerçekleştirilmiştir. Her algoritma için belirlenen grid uzayı (hyperparameterların alması olası değerler) aşağıdaki Çizelge 5.8'te paylaşıldığı gibidir.

Çizelge 5.8 : Hiper parametre optimizasyonu için belirlenen değerler.

Algoritma	R Paketi	Hiper Parametre*	Izgara aramada incelenen değer kümeleri
LightGBM	Lightgbm Base	learning_rate min_data_in_leaf bagging_fraction max_bin	learning_rate= 0.001,0.02, 0.1, 0.2, 0.5 min_data_in_leaf = 5, 10, 20, 30 bagging_fraction = 1, 0.8, 0.6, 0.5, 0.4, 0.2 max_bin = 5,25,55,105,205,255
CatBoost	Catboost Caret	depth , learning_rate, l2_leaf_reg, rsm, border_count, iterations	depth = 1, 4, 6, 8 learning_rate = 0.001,0.02,0.1,0.2,0.5 l2_leaf_reg = 0.1,2,20 rsm = 0.95 border_count = 1,30,64,255 iterations = 10,1000
XGBoost	Xgboost Caret	nrounds, max_depth, colsample_bytree eta, gamma, subsample	nrounds = 50, 100, 250, 500, 1000 max_depth = 2, 4, 6, 8, 10 colsample_bytree = 0.5, 1 eta = 0.01, 0.001, 0.0001 gamma=1 subsample = 1
Random Forest	RandomForest Caret	mtry	mtry= 1:20

*Hiper parametrelerin açıklamalarına modellerin anlatımı kısmında yer verilmiştir.

Elde edilecek sonuçlarda aşırı uyumu (overfitting) engellemek adına 3-katmanlı çapraz doğrulama (3-fold cross validation) kullanılmıştır. Bu yaklaşım eğitim setindeki aşırı uyumu sınırlamak ve azaltmak için tasarlanmıştır ancak aşırı uyumu tamamen ortadan kaldırdığı söylenemez. Ayrıca bu yaklaşım arama işlemini yavaşlatmaktadır.

3.aşamada kullanılan tüm değişkenler dikkate alınarak yürütülen çapraz doğrulamalı hiper parametre optiimzasyonu (cross-validated hyper parameter tuning) sonrası elde edilen model performansları Çizelge 5.9'da paylaşılmıştır.

Çizelge 5.9 : Dördüncü aşama model performansları.

Algoritma	Accuracy	Sensitivity	Specificity	F1	AUC
Random Forest	0.997	0.2298	0.9994	0.3307	0.8837
Xgboost	0.9966	0.2770	0.9990	0.3509	0.6380
Lightgbm	0.9962	0.3118	0.9984	0.3496	0.6551
Catboost	0.9964	0.3052	0.9986	0.3530	0.6519

Sensitivity değerleri dikkate alındığında gerçekte anomali olan durumları anomali olarak tespit etmede en yüksek performansı Lightgbm'in gösterdiği görülmektedir. Ancak tespit yüzdeleri istenen seviyede değildir.

F1 score larına göre yürürütülen çalışma için en uygun modelin Catboost olduğu görülmüştür. Catboost sonucu belirlenen en önemli değişkenlerden ilk 5'i sırası ile;

- poliçenin kesildiği bölge (bolge_adi)
- poliçe için ödenen toplam prim(sum_prim)
- poliçenin kesildiği ilçenin günlük hasar prim oranı (ilce_gun_hpo)
- poliçeyi kesen acentedeki hasar artışlarında dikkat çeken artış (acente_hasar_aşırı_artış)
- poliçenin kesildiği ilin günlük hasar prim oranı (il_gun_hpo) ve poliçenin kesildiği tarife (tarife_adı) [il_gun_hpo ve tarife_adı biribirine çok yakın önem değerleri almıştır]

şeklindedir.

XGBoost'un belirlediği en önemli 3 değişken sırası ile toplam ödenen prim tutarı (sum_prim), poliçenin kesildiği ilçenin günlük hasar prim oranı (ilce_gun_hpo) ve poliçenin kesildiği ilin günlük hasar prim oranıdır (il_gun_hpo).

5.2.2.5 Beşinci aşama: yeniden örnekleme

Yeniden örnekleme aşamasında yerine koymadan rasgele örnekleme (randomly sample without replacement) ile undersampling yaklaşımı benimsenmiştir. Bu yaklaşımda yüksek sayıda bulunan sınıfta yer alan gözlemler belirlenen oranlara ya da adetlere göre indirgenmektedir. Çalışmada yapay olarak veri seti dengelendirilirken farklı sınıfların dağılımları denenmiş, model performanslarının belirlenen dağılımlarda gösterdikleri değişkenlik gözlemlenmiştir. En iyi performans gösteren modelin seçilmesi amaçlanmıştır.

Undersampling ilk etapta hiper parametre optimizasyonu sonucu elde edilen parametre değerlerine sahip modellere uygulanmıştır. Devamında aynı işlem hiper parametre optimizasyonu öncesinde belirlenen model parametre değerleri üzerine uygulanmıştır. Sonuçlar kıyaslandığında ikinci uygulamada daha iyi performanslar elde edildiği gözlemlenilmiştir. Çizelge 5.10'da, hiper parametre optimizasyonu

öncesinde kullanılan model parametreleri üzerinden, belirlenen sınıf dağılımları odağında yapılan yeniden örnekleme çalışmasının çıktıları yer almaktadır.

Çizelge 5.10 : Yeniden örnekleme sonrası model performansları.

Algoritma	Yeniden Örnekleme		Accuracy	Sensitivity	Specificity	F1	AUC**
Decision Tree	-		0.9968	0.0000	1.0000	0.00	0.5
	Undersampling Without resampling	**50:50**	0.9022	0.9497	0.8546	0.9065	0.9021
		60:40	0.8937	0.9533	0.8539	0.8776	0.9036
		70:30	0.9089	0.8743	0.9237	0.8521	0.8990
		80:20	0.9127	0.8420	0.9304	0.7942	0.8861
Bagging	-		0.997	0.2333	0.9994	0.3307	0.7844
	Undersampling Without resampling	**50:50**	0.9013	0.9210	0.8815	0.9031	0.9565
		60:40	0.908	0.9084	0.9078	0.8877	0.9622
		70:30	0.903	0.8384	0.9307	0.8384	0.9605
		80:20	0.9181	0.8043	0.9465	0.7971	0.9595
Random Forest	-		0.9508	0.9174	0.9508	0.1058	0.9794
	Undersampling Without resampling	**50:50**	0.921	0.9443	0.8977	0.9228	0.9648
		60:40	0.9195	0.9246	0.9162	0.9019	0.9738
		70:30	0.9245	0.8941	0.9376	0.8767	0.9719
		80:20	0.9299	0.8223	0.9569	0.8244	0.9667
XGBoost	-		0.9962	0.2195	0.9987	0.2736	0.6091
	Undersampling Without resampling	**50:50**	0.9048	0.9134	0.8965	0.9034	0.9049
		60:40	0.9073	0.8944	0.9158	0.9228	0.9050
		70:30	0.9084	0.8657	0.9278	0.8554	0.8967
		80:20	0.9213	0.7962	0.9538	0.8067	0.8749

Çizelge 5.10 (devam): Yeniden örnekleme sonrası model performansları.

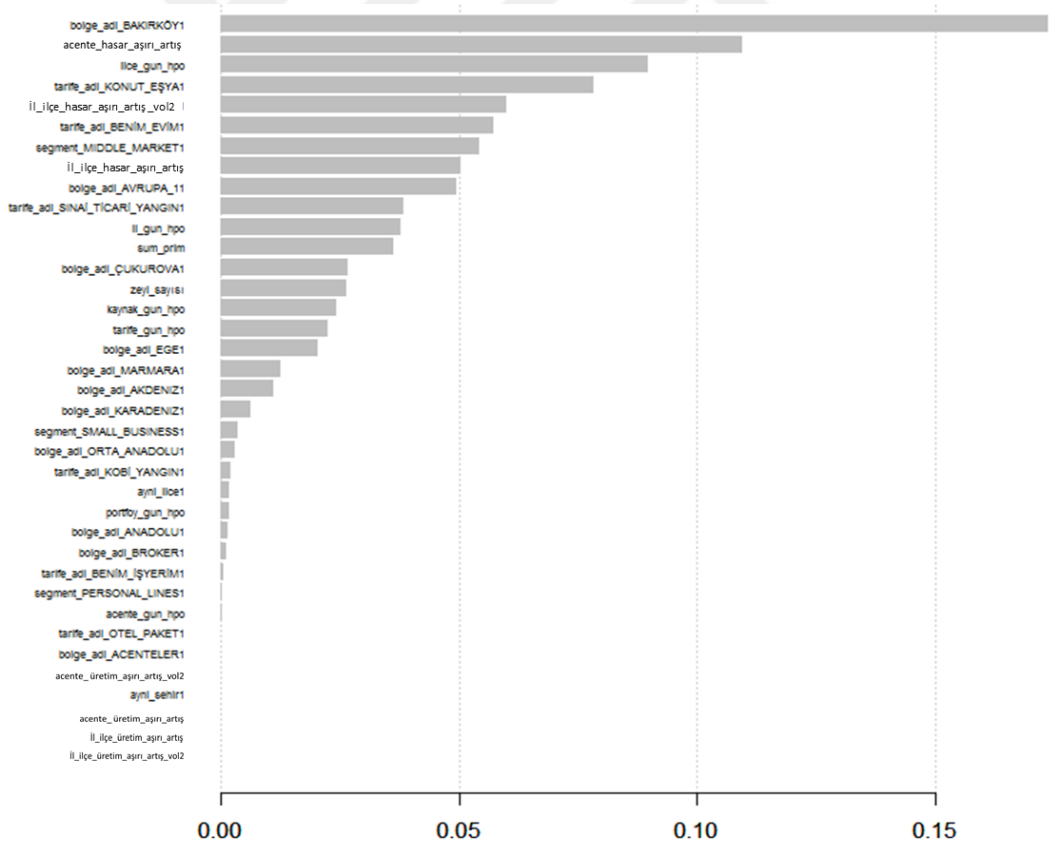
Algoritma	Yeniden Örnekleme		Accuracy	Sensitivity	Specificity	F1	AUC**
LightGBM	-		0.9968	0.2299	0.9992	0.3176	0.6146
	Undersampling Without resampling	**50:50**	0.8985	0.9116	0.8860	0.8975	0.8987
		60:40	0.8958	0.8871	0.9015	0.8704	0.8943
		70:30	0.8997	0.8503	0.9223	0.8415	0.8862
		80:20	0.9087	0.6603	0.9733	0.7490	0.8167
CatBoost	-		0.9969	0.1633	0.9995	0.2506	0.5814
	Undersampling Without resampling	**50:50**	0.912	0.9354	0.8887	0.9140	0.9120
		60:40	0.9131	0.8923	0.9269	0.8914	0.9096
		70:30	0.9127	0.8474	0.9407	0.8535	0.8940
		80:20	0.9109	0.7074	0.9618	0.7606	0.8345

Tabloda performansları kıyaslanan tüm algoritmaların farklı örnekleme oranlarındaki performansları ve örnekleme olmaksızın performansları gözlemlenilebilmektedir. Genel olarak incelendiğinde örneklemde az sayıda bulunan sınıf ile yoğunluklu bulunan sınıfın oranları birbirine yaklaştıkça basit sınıflandırma modellerinin performanslarında daha yüksek artışlar yakalandığı tespit edilmiştir. Genel olarak modellerin birçoğunun en yüksek performans gösterdiği dağılım oranı 50:50 olmuştur. Model performansları kıyaslanırken F1 skorlarına öncelik verilmiş olup sensitivity değerleri de incelenmiştir.

Verisetinin 50:50 oranında dengelendiği durumda F1 score değerleri dikkate alındığında en yüksek performans gösteren algoritmaların Rastgele Orman, Catboost ve Decision Tree olduğu, veri setinin 60:40 oranında dengelenmesi durumunda Xgboost, Rastgele Orman ve Catboost algoritmalarının, 70:30 oranında yine aynı kıyas parametresi ile Rastgele Orman, Xgboost ve Catboost algoritmalarının ve son olarak 80:20 oranında ile Rastgele Orman, Xgboost ve Bagging algoritmalarının ön plana çıktığı tespit edilmiştir.

Belirlenen algoritmaların performansları, veri setini dengelemeye yönelik seçilen tüm oranlar dikkate alınarak elde edilen F1 skorlarına göre kıyaslandığında, en

yüksek performans 60:40 oranında 0.922893 F1 skor ile XGBoost algoritmasında görülmüştür. XGBoost algoritmasında kullanılan değişkenlerin önemleri belirlenirken kazanç (gain), ağırlık (weight) ve cover olmak üzere üç değer hesaplanmaktadır. Kazanç temel olarak tüm ağaçlardan elde edilen bilgi kazancıdır. Bizlere kullanılan bölümlerin (split) ortalama kazancını göstermektedir. İlk olarak ana düğümün düğüm safsızlığını (node impurity) hesaplanır, devamında bölme için belirli bir özellik kullanılacaksa alt düğümlerin düğüm safsızlıkları hesaplanır. Son olarak, bilgi kazancı, alt safsızlıkların ana düğüm safsızlığından çıkarılmasıyla hesaplanır. Ağırlık bir değişkenin ağaçta kaç sefer görüldüğünü göstermektedir. Cover bölümlerin ortalamasıdır. Bölünmeden etkilenen örnek sayısı olarak tanımlanan kapsam (coverage) doğrultusunda kullanılan değişkenlerin bulunduğu bölümlerin kapsam ortalamasıdır (*XGBoost Feature Importance*). Bu Çalışmada değişken önemleri belirlenirken kazanç değeri üzerinden ilerlenmiştir. Bu model tarafından belirlenen en önemli değişkenler Şekil 5.2'de gösterildiği gibidir.



Şekil 5.2 : XGBoost tarafından belirlenen değişken önemleri.

6. SONUÇLAR

Bu çalışma, sigorta şirketinden temin edilen etiketsiz mülkiyet poliçesi verileri üzerinde gerçekleştirilmiştir. Çalışmanın ilk bölümü anomali gözlemlerin tespitine yönelik yürütülmüş olup devamında belirlenen anomali durumların tahminine yönelik sınıflandırma algoritmalarının performansları sorgulanmıştır.

Anomali gözlemlerin tespitine yönelik denetimsiz öğrenme tekniklerinden k-ortalamalar kullanılmıştır. Kümeleme işlemine dahil edilen değişkenler belirlenen anomali senaryosuna yönelik seçilmiştir. Elde edilen küme sonuçlarında anomali olarak belirlenen kümede özellikle yüksek hasar prim oranına sahip olma durumu göze çarpan unsurlardandır.

Çalışmanın ikinci kısmında, poliçelerin oluşturulması aşamasında elde edilen değişkenler ve türetilen yeni değişkenler dolayında denetimli öğrenme tekniklerinden yararlanılmıştır. Temelde cevap aranan problem ikili sınıflandırma problemi olduğundan ve poliçe verisine girdi teşkil eden değişkenlerin birçoğu kategorik değişken olduğundan bu alana hizmet eden algoritmalar tercih edilmiştir. Kullanılan algoritmalar sırası ile Karar Ağaçları, Bagging, Rastgele Orman, Xgboost, LightGBM ve CatBoost'tur. Kullanılan sınıflandırma algoritmalarının, kümeleme ile belirlenen anomalileri tahmin etme performansları kıyaslanarak, en yüksek performansa sahip algoritmanın tespit edilmesi hedeflenmiştir.

Model performansları ilk olarak baz değişkenler ile karşılaştırılmış ancak elde edilen sonuçlar çok düşük olduğundan yeni değişkenler türetilmiştir. İki ayrı aşamada türetilen yeni değişkenler sonucu model performansları yine istenen seviyelere ulaşamamış, bunun üzerine çapraz doğrulamalı hiper parametre optimizasyonuna gidilmiştir. Hiper parametre optimizasyonu sonucunda, sensitivity değerleri dikkate alındığında gerçekte anomali olan durumları anomali olarak tespit etmede en yüksek performansı Lightgbm algoritmasının gösterdiği, F1 score'larına göre

kıyaslandığında ise yürürütülen çalışma için en uygun modelin Catboost olduğu görülmüştür. Ancak tespit yüzdeleri istenen seviyede değildir.

Bu noktada literatürde pek çok çalışmada tercih edilen dengesiz verilerin yeniden örnekleme ile dengeli hale getirilerek model performanslarının yükseltildiği dikkate alınarak, yeniden örnekleme yaklaşımlarından undersampling metodu benimsenerek farklı denge seviyelerinde veri seti düzenlenerek model performansları tekrardan sorgulanmıştır. Elde edilen sonuçlarda ciddi bir iyileşme gözlemlenmiştir. Model çıktıları incelendiğinde, verisetinin 50:50 oranında dengelendiği durumda F1 score değerleri dikkate alındığında en yüksek performans gösteren algoritmaların Rastgele Orman, Catboost ve Decision Tree olduğu, veri setinin 60:40 oranında dengelenmesi durumunda Xgboost, Rastgele Orman ve Catboost algoritmalarının, 70:30 oranında yine aynı kıyas parametresi ile Rastgele Orman, Xgboost ve Catboost algoritmalarının ve son olarak 80:20 oranında ile Rastgele Orman Xgboost ve Bagging algoritmalarının ön plana çıktığı tespit edilmiştir.

Belirlenen algoritmaların performansları, veri setini dengelemeye yönelik seçilen tüm oranlar dikkate alınarak elde edilen F1 skorlarına göre kıyaslandığında, en yüksek performans 60:40 oranında XGBoost algoritmasında görülmüştür. XGBoost algoritmasının belirlediği en önemli değişkenler dikkate alındığında, sigorta şirketlerinin poliçelendirme aşamasında gelen taleplerin anomali olma durumunu incelerken ve bu durumlara yönelik stratejiler geliştirirken; poliçe talebinin hangi bölgeden geldiği, poliçeyi oluşturan acentede hasar oluşum adetlerinde dikkat çeken bir artışın olup olmadığı, poliçe talebinin geldiği ilçede günlük hasarın prime dönüşüm oranı ve poliçenin hangi tarife kapsamında oluşturulduğu noktalarına odaklanmaları gerektiği tespit edilmiştir.

7. GELECEK ÇALIŞMALAR

Çalışma kapsamında anomali tespiti aşamasında kümeleme algoritmalarından k-means kullanılmıştır. Sahtekârlık/ anomali tespitine yönelik yürütülen çalışmalarda temel problem verilerin dengesiz dağılımıdır. Anomali gözlemler genel olarak veri setlerinde az sayıda bulunmaktadır. Bu noktada farklı yoğunluğa sahip veri gruplarını tespit etmede tercih edilen yoğunluk temelli kümeleme yaklaşımları, ileriki çalışmalarda anomali tespitine yönelik unsupervised bir teknik olarak tercih edilebilir.

Çalışma kapsamında ağaç temelli algoritmaların performansları kıyaslanmıştır. Yapay sinir ağları ve derin öğrenme yöntemleri ileride yürütülecek çalışmaların kapsamına dahil edilebilir.

Özellikle ensemble metotlarda, belirlenen eşik değerleri (threshold) tahmin dağılımlarını etkilemektedir. Çalışma kapsamında tek bir eşik değeri üzerinden modellere tahminleme yaptırılmıştır. İleriki çalışmalarda farklı eşik değerleri için model tahminleme performansları değerlerdirilerek en uygun eşik değerin seçilmesi yaklaşımı benimsenebilir.

KAYNAKLAR

Aggarwal, C. C., & Reddy, C. K. (2014). *Data Clustering Algorithms and Applications*. Chapman & Hall/ CRC Press.

Ak, R., & Öztayşi, B. (n.d.). *Performance Measurement of Insurance Companies By*.

Aktan, C. C. (2008). Stratejik yönetim ve stratejik planlama. *Çimento İşveren Dergisi, 22(4), 4-21*.

Andrews, K. R. (1981). *The Concept of Corporate Strategy*. Dow Jones-Irwin Inc., Illinois.

Bauder, R. A., & Khoshgoftaar, T. M. (2017). Medicare fraud detection using machine learning methods. *IEEE International Conference on Machine Learning and Applications, In 16th (I.*

Bauder, R. A., Rosa, R., & Khoshgoftaar, T. (2018). Identifying Medicare Provider Fraud with Unsupervised Machine Learning. *IEEE International Conference on Information Reuse and Integration for Data ScienceAt: Salt Lake City*.

Benali, N., & Feki, R. (2017). The impact of natural disasters on insurers' profitability: Evidence from Property/Casualty Insurance company in United States. *Research in International Business and Finance*, *42*(December 2016), 1394–1400. https://doi.org/10.1016/j.ribaf.2017.07.078

Bentley, P. (2000). Investigating Committee-based Evolution of Fuzzy Rules for the Detection of Suspicious Insurance Claims. *Proc. of GECCO2000*. http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:"+Evolutionary+,+my+dear+Watson+"+Investigating+Committee-based+Evolution+of+Fuzzy+Rules+for+the+Detection+of+Suspicious+Insurance+Claims#1

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). "SMOTE: Synthetic minority over-sampling technique. *J. Artif. Intell. Res.*, *16*, *n*.

Cichosz, P. (2015). Data Mining Algorithms Explained Using R. In *John Wiley & Sons* (Vol. 66, Issue Book Review 2). John Wiley & Sons, Ltd. https://doi.org/10.18637/jss.v066.b02

Demirbaş, B. P. (2010). *Stratejik yönetim ve sigorta sektöründeki bir işle tmede uygulamasi*.

Domingues, R. (2015). *Machine Learning for Unsupervised Fraud Detection*. 66.

Frank, H., & Kotthoff, L. (2019). *Automatic machine learning: methods, systems,*

challenges. Springer. https://doi.org/https://doi.org/10.1007/978-3-030-05318-5

Friedman, J., Hastie, T., & Tibshirani, R. (2001). *The elements of statistical learning* ((Vol. 1, N). Springer series in statistics.

G**alar, M., Fernandez, A., Barrenechea, E., Bustince, H., & Herrera, F.** (2011). A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches. ,. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Ap*.

Get started Working with categorical data: Catboost. (n.d.). https://medium.com/whats-your-data/working-with-categorical-data-catboost-8b5e11267a37

Hassan, A. K. I., & Abraham, A. (2016). Modeling Insurance Fraud Detection Using Ensemble Combining Classification. *International Journal of Computer Information Systems and Industrial Management Applications*, *8*, 257–265.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning with Applications in R*. Springer. https://doi.org/10.1007/978-1-4614-7138-7_8

Joudaki, H., Rashidian, A., Minaei-Bidgoli, B., Mahmoodi, M., Geraili, B., Nasiri, M., & Arab, M. (2015). Using data mining to detect health care fraud and abuse: a review of literature. *Global Journal of Health Science*, *7*(1), 194–202. https://doi.org/10.5539/gjhs.v7n1p194

Kamu İdareleri İçin Stratejik Planlama Kılavuzu. (2018). Kalkınma Bakanlığı.

Khoshgoftaar, T. M., Van Hulse, J., & Napolitano, A. (2010). Comparing boosting and bagging techniques with noisy and imbalanced data. *IEEE Transactions on Systems, Man, and Cybernetics, Part A: Sy*.

Kirlidog, M., & Asuk, C. (2012). A fraud detection approach with data mining in health insurance. *Procedia-Social and Behavioral Sciences, 62, 989-994.*

Kılınç, D. (2018). *Uygulamalarla Veri Bilimi: Makine Öğrenmesi, Derin Öğrenme*. Abaküs.

Michalski, R. S., & Tecuci, G. (1994). *Machine Learning: A Multistrategy Approach* (vol. 4). Morgan Kaufmann.

Mkrtychev, S., & Enik, O. (2019). *Automated Underwriting Control in a Regional Insurance Company*. https://doi.org/10.2991/iscfec-18.2019.66

Moradi, H., & Hosseinkhani, J. (2017). *A Data Mining Approach for Analysis of Customer Behavior in order to Improve Policies in Insurance Industry based on Combination of Particle Swarm Optimization and k-Means Algorithm*. *6*(3), 31–39.

Murphy, K. P. (2012). Machine Learning: A Probabilistic Perspective. In *The MIT Press Cambridge, Massachusetts London, England*. https://doi.org/10.1007/978-94-011-3532-0_2

Nagrecha, S. (n.d.). *Fraudbuster|Predıctıng Expected Loss Ratıos To Reduce Auto Insurance Fraud*.

Nagrecha, S., Reid A. Johnson, & Chawla, N. V. (2018). *FraudBuster- Reducing Fraud in an Auto Insurance Market* (p. Volume 6 Number 1). Big Data.

Nian, K., Zhang, H., Tayal, A., Coleman, T., & Li, Y. (2016). Auto insurance fraud detection using unsupervised spectral ranking for anomaly. *The Journal of Finance and Data Science*, *2*(1), 58–75. https://doi.org/10.1016/j.jfds.2016.03.001

Onan, A. (2015). A fuzzy-rough nearest neighbor classifier combined with consistency-based subset evaluation and instance selection for automated diagnosis of breast cancer. *Expert Systems with Applications*, *42*(20), 6844–6852. https://doi.org/10.1016/j.eswa.2015.05.006

Palacio, S. M. (2019). Abnormal pattern prediction: Detecting fraudulent insurance property claims with semi-supervised machine-learning. *Data Science Journal*, *18*(1), 1–15. https://doi.org/10.5334/dsj-2019-035

Phua, C., Lee, V., Smith, K., & Gayler., R. (2010). *A comprehensive survey of data mining-based fraud detection research*. *arXiv prep*.

Phua, Clifton, Alahakoon, D., & Lee, V. (2004). Minority report in fraud detection. *ACM SIGKDD Explorations Newsletter*, *6*(1), 50–59. https://doi.org/10.1145/1007730.1007738

Rigby, D., & Bilodeau, B. (2017). Management Tools & Trends. In *Bain & Company -New York*. Bain & Company. file:///Y:/Gruppen/ben/00_Public/Lena/Alberto/BAIN_BRIEF-Management_Tools_and_Trends.pdf

Shan, Y., Jeacocke, D., Murray, D. W., & Sutinen, A. (2008). Mining medical specialist billing patterns for health service management. *In Proceedings of the 7th Australasian Data Mining Conference-Volume 87 (Pp. 105-110).*

Shearer, C. (2000). The CRISP-DM model: The New Blueprint for Data Mining. *Journal of Data Warehousing*, *5*(4), 13–22. www.spss.com%5Cnwww.dw-institute.com

Sparrow, M. K. (1996). Health care fraud control: understanding the challenge. *Journal of Insurance Medicine (New York, N.Y.)*, *28*(2), 86–96.

TSB Faaliyet Raporu. (2016).

TSB Faaliyet Raporu. (2017).

Viaene, S., Derrig, R. A., B. Baesens, & Dedene., G. (2002). A comparison of state- of-the-art classification techniques for expert automobile insurance claim fraud detection. ,. *Journal of Risk and Insurance, 69(3):373–*.

Wang, Y., & Xu, W. (n.d.). *Leveraging deep learning with LDA-based text analytics to detect automobile insurance fraud*.

Williams, G., & Huang, Z. (1997). Mining the Knowledge Mine: The Hot Spots Methodology for Mining Large Real World Databases. *Proc. of the 10th Australian Joint Conference on Artificial Intelligence.'*.

Wirth, R., & Hipp, J. (2000). CRISP-DM : Towards a Standard Process Model for Data Mining. *Proceedings of the Fourth International Conference on the Practical Application of Knowledge Discovery and Data Mining*, *24959*, 29–39. https://doi.org/10.1.1.198.5133

Automobile insurance.. https://en.wikipedia.org/wiki/Vehicle_insurance, erişim tarihi: 2020

How to Win a Data Science Competition*: Learn from Top Kagglers*. https://www.coursera.org/lecture/competitive-data-science/catboost-2-U61Ak, erişim: 2020

LightGBM Parameters. https://lightgbm.readthedocs.io/en/latest/Parameters.html, erişim: 2020

Property insurance. , erişim tarihi: 2020

R package training parameters. https://catboost.ai/docs/concepts/r-training-parameters.html, erişim tarihi: 2020

Sigorta Tanımları. Türkiye Sigorta Birliği. https://www.tsb.org.tr/sigorta-tanimlari.aspx?pageID=648, erişim tarihi: 2020

Sigorta Suistimalleri Bilgi Paylaşım Sistemi (SİSBİS) https://siseb.sbm.org.tr/tr/sisbis, erişim tarihi: 2020

XGBoost.. https://rpubs.com/omicsdata/145253, erişim tarihi: 2020

XGBoost Parameters. https://xgboost.readthedocs.io/en/latest/parameter.html, erişim tarihi: 2020

XGBoost Feature Importance. https://xgboost.readthedocs.io/en/latest/python/python_api.html, erişim tarihi: 2020

ÖZGEÇMİŞ

Ad-Soyad : **Ayşe Nurbanu Şahan**

Doğum Tarihi ve Yeri : **06.04.1993**

E-posta : nurbanusahan@gmail.com

ÖĞRENİM DURUMU:

- **Lisans** : 2015, Uludağ Üniversitesi, Mühendislik Fakültesi, Endüstri Mühendisliği
- **Yüksek Lisans** : 2020, İstanbul Teknik Üniversitesi, İşletme Mühendisliği Anabilim Dalı, İşletme Mühendisliği Program

MESLEKİ DENEYİM:

- 3 seneye yakın yönetim danışmanlığı sektöründe yer almış olup günümüzde çalışmalarını savunma sanayiinde faaliyet gösteren bir firmada devam ettirmektedir.

DİĞER YAYINLAR:

- Cavdur, F., Yagmahan, B., Oguzcan, E., Arslan, N., & Sahan, N. (2018). Lean service system design: a simulation-based VSM case study. *Business Process Management Journal*.