FİNANSAL TABLO HİLESİ RİSKİ TAŞIYAN ŞİRKETLERİN VERİ MADENCİLİĞİ İLE BELİRLENMESİ*

Dr. Öğr. Üyesi Kadir KIRDA[a]
Dr. Öğr. Üyesi Münevver KATKAT ÖZÇELİK[b]

Ampirik Araştırma
(Empirical Research)

Muhasebe ve Vergi Uygulamaları Dergisi
Temmuz 2021; 14 (2): 609-639

ÖZ

Finansal tablo hilesi, şirketlerin finansal tablolarındaki verileri kendi çıkarları doğrultusunda değiştirerek yayınlamalarıdır. Kurumlara, paydaşlara ve ekonomik yapıya ciddi zararlar veren finansal tablo hilelerinin tespit edilmesi önemli bir problemdir. Bunun için çeşitli denetim mekanizmaları bulunmaktadır. Ancak zaman içerisinde geliştirilebilecek hile yöntemlerine karşı yenilikçi denetim yöntemlerine ihtiyaç duyulmaktadır.

Veri madenciliği, finansal tablo hilelerinin tespitinde umut vadeden bir alandır. Veri madenciliğinin sınıflandırma analizinde sınıflandırma metotlarıyla mevcut verilerden örüntüler elde edilir ve bunlar görülmemiş birimlerin sınıflandırılmasında kullanılır. Bu çalışmada veri madenciliğinin sınıflandırma metotları ile finansal tablo hilesi riski taşıyan şirketlerin tespiti üzerine bir araştırma yapılmıştır.

Veriler Borsa İstanbul'da 2014-2018 arasında işlem gören şirketlerin yayınladıkları finansal tablolardan elde edilmiştir. İlk olarak yedi sınıflandırılma metodu kullanılmış, en başarılı üçü seçilmiştir. Sonraki aşamada başarım değerlerinin geliştirilmesi amacıyla hiper parametre optimizasyonu yapılmıştır.

Sınıflandırma metotlarından K-Nearest Neighbor ile yüzde 91,73, Random Forest ile yüzde 90,51 ve XGBoost ile yüzde 90,37 doğruluk oranlarına ulaşılmış, en iyi tahmin oranı K-Nearest Neighbor ile elde edilmiştir. Son kısımda rasgele alt örnekleme yöntemiyle yapılan karşılaştırmalarda da en iyi performans değerleri K-Nearest Neighbor ile elde edilmiştir.

Anahtar Sözcükler: Finansal Tablo Hilesi, Veri Madenciliği, Sınıflandırma.

JEL Kodları: M42, H83, C38.

APA Stili Kaynak Gösterimi:

Kırda, K., Özçelik, M. K. (2021). Finansal Tablo Hilesi Riski Taşıyan Şirketlerin Veri Madenciliği ile Belirlenmesi. *Muhasebe ve Vergi Uygulamaları Dergisi*. 14 (2), 609-639.

* *Makalenin gönderim tarihi: 30.09.2020; Kabul tarihi: 12.02.2021, iThenticate benzerlik oranı %7*

[a] *Artvin Çoruh Üniversitesi, İktisadi ve İdari Bilimler Fakültesi, kadirkirda@artvin.edu.tr*
ORCID: 0000-0003-0779-0175.
[b] *Artvin Çoruh Üniversitesi, İktisadi ve İdari Bilimler Fakültesi, mkatkat@artvin.edu.tr*
ORCID: 0000-0001-7299-7952.

IDENTIFYING THE COMPANIES WITH THE RISK OF FINANCIAL STATEMENT FRAUD BY DATA MINING

ABSTRACT

Financial statement fraud is when companies change and publish the data in their financial statements in line with their interests. Detecting financial statement fraud that causes serious damage to organizations, stakeholders, and the economic structure is an important problem. There are various control mechanisms for this. However, there is a need for innovative control methods against new fraud methods that may be developed over time.

Data mining is a promising field for detecting financial statement fraud. In the classification analysis of data mining, patterns are obtained from existing data with classification methods, and these are used in the classification of unseen units. In this study, research was carried out on the identification of companies with the risk of financial statement fraud through the classification methods of data mining.

The data were obtained from the financial statements published by companies traded in Borsa Istanbul between 2014 and 2018. Seven classification methods were used first, and the three most successful ones were selected. In the next stage, hyperparameter optimization was carried out to improve the performance values.

Accuracy rates of 91.73 percent with K-Nearest Neighbor, 90.51 percent with Random Forest, and 90.37 percent with XGBoost were obtained from the classification methods, and the best prediction rate was obtained with K-Nearest Neighbor. In the last part, the best performance values were also obtained with K-Nearest Neighbor in the comparisons using random sub-sampling method.

Keywords: Financial Statement Fraud, Data Mining, Classification.

JEL Codes: M42, H83, C38.

EXTENDED ABSTRACT

Introduction

There are many definitions in the literature about financial statements fraud, and different definitions have also been made by various institutions. One of these definitions has been made by the Capital Markets Board of Turkey (CMB). In the Communiqué on Independent Auditing Standards in the Capital Markets published by CMB, fraud is defined as “deliberate acts of deception by those in the management of the enterprise, persons responsible for management, business employees or third parties in order to gain a fair or illegal interest”. Financial statements fraud occurs for variety purposes such as hiding corruption, evading taxes and showing the real situation of the business differently. Financial statements fraud, regardless of what its purpose, is a crime factor that causes serious losses when it cannot be revealed.

The subject of this study is to investigate the determination of the companies at risk of financial statements fraud by using of the classification methods in data mining. The study aims to design a comprehensive data mining application that basically includes feature selection, comparison of

classification methods and hyper parameter optimization processes in order to identify companies prone to commit financial statements fraud with the classification analysis of data mining. As a data source, financial statements published by 19 fraudulent and 20 regular companies operating in Borsa Istanbul between 2014-2018 were used. Financial ratios were selected as the input variable, and whether company has committed financial statements fraud or not was selected as the class variable. The results of the analysis made with the classification methods in data mining were presented comparatively.

Literature on Research

There are various academic studies at home and abroad on the use of data mining for financial statements fraud. One of the studies in Turkey was carried out by Ata and Seyrek (2009) in order to detect fraudulent financial statements using data mining with 100 companies traded in IMKB and operating in manufacturing sector. In the research, decision trees 67.92 percent and artificial neural networks have reached 77.36 percent accurate classification success.

In a study conducted by Terzi and Sen (2012), using data from IMKB in manufacturing sector, 25 companies with a risk of financial statements fraud and 25 companies with no risk were used, artificial neural networks and decision tree methods were selected for classification. In another study conducted by Terzi and Sen (2015), the use of artificial neural networks was investigated in the determination of financial statements fraud. 12 financial ratios were used in the data set consisting of 26 company information in Borsa Istanbul.

There are various studies abroad using different methods such as feature selection (Ravisankar et al., 2011; Rizki et al., 2017; Yao et al., 2018), neural networks (Jan, 2018; Kirkos et al., 2007; Lin et al., 2015; Perols, 2011; Ravisankar et al., 2011; Rizki et al., 2017), genetic algorithms (Hoogs et al., 2007; Ravisankar et al., 2011), decision trees (Bai et al., 2008; Jan, 2018; Lin et al., 2015; Perols, 2011; Yao et al., 2018), fuzzy (Lenard et al., 2007), logistic regression (Lin et al., 2015; Perols, 2011) and support vector machines (Perols, 2011; Rizki et al., 2017).

Operations such as feature selection, comparison of multiple methods and models, and hyper parameter optimization in data mining are observed in practical applications. In the literature, there is no comprehensive study including all of these transactions in data mining studies on financial statements fraud. Furthermore, innovative methods such as XGBoost classification method, RobustScaler, RandomizedSearch, which have been successfully applied recently and attracted attention, have also been included in the study.

Method of The Research

The analysis process consists of six stages. First of all, data obtained from companies' financial reports were compiled. The data source used in the study is the quarterly financial statements shared publicly by companies traded in Borsa Istanbul between 2014 and 2018 on the access page of the Public Disclosure Platform (kap.org.tr). The research unit is a one-time financial statement of a company. The properties (variables) of the unit consist of financial ratios and a class variable which is indicating fraud status.

In the second stage, the variables to be evaluated in the model development process were determined. These variables were selected from financial ratios that are frequently encountered in the literature. The values obtained as a result of the calculations using the values in the balance sheet and income tables were added to the data set as the properties of the research units. Recursive Feature Elimination (RFE) method was used for feature selection, and 11 features obtained as a result of the process were used as inputs.

The third stage covers the implementation of the transformation processes required in order to make the data suitable for the analysis. The basic data transformation process performed at this stage was the scaling of the data. The most common methods used in data scaling are normalization and standardization. Since standardization is extremely sensitive to outliers in the data set, RobustScaler scaling method, which is not affected by outliers, was used.

In the fourth stage, the first comparison was made with the default settings of seven classification methods, and then three methods with the best accuracy scores were selected. Random Forest (RF), K-Neighbors (KNN), XGBoost (XGB), Support Vector Machine (SVM), Multi-Layer Perceptron (MLP), Decision Tree (CART), and Gaussian Naive Bayes (NB) classification methods were used for the first comparison. RF, KNN and XGB, the first three methods with the highest accuracy, were chosen for the next stage.

In the fifth stage, hyper parameter optimization was performed in order to increase the predictive power of the selected methods. Each change in hyper parameter value generates a new model that needs to be evaluated. In the case of few options, this process does not take long, but there are many combinations of hyper parameters in the study. In these conditions, the number of models that need to be evaluated increases exponentially, and consequently technology and time sources limit the procedure. In order to overcome this handicap, hyper parameter optimization was done by using the "RandomizedSearch" method developed by Scikit-Learn and optimum hyper parameter values were obtained.

In the last stage, optimized models were compared with different training-test ratios and the results were evaluated. In the second comparison, models were compared with different training-test ratios by using random subsampling method. The training-test ratios chosen are 80-20, 70-30, 60-40 and 50-50. The random data set selection was repeated 10 times for each model, and the calculated evaluation criteria were compiled for interpretation.

Findings of The Research

In the first classification analysis, models were compared with 10-fold cross-validation method using the default hyper parameter values of the classification methods. Accordingly, the highest classification score is 89.83 percent obtained by RF method. Then, 88.87 percent and 86.44 percent correctly classification ratios were achieved with KNN and XGB methods respectively.

After that, hyper parameters choices were created, and then hyper parameter optimization was performed with RandomizedSearch method to evaluate the success of the models consisting of diverse hiper parameter combinations.

As a result of the hyper parameter optimization; KNN, RF and XGB reached 91.73, 90.51 and 90.37 accuracy scores respectively. It was observed that the performances of all of the models were improved as a consequence of the hyper parameter optimization process.

Then, the models were compared with different training-test ratios using the random subsampling method. It was observed that the success of RF and XGB methods has decreased somewhat with the decrease in training data ratios, however, the correct classification success of KNN method remains high despite the decrease in training data ratios. In addition, KNN method was found to have higher accuracy scores and lower standard deviations compared to other methods.

It was discovered that the most successful results for all methods were obtained in calculations with training-test ratios of 80-20. It has been observed that higher number of units in the training data set made the classification model more successful. When the learning ratios was gradually reduced in the study, the success ratios also decreased progressively.

As a result, it has been seen that the classification analysis of data mining for financial statements fraud can be successfully applied. In this study, variety applications and methods were used from the first stage to the final stage, but the appropriate applications and methods may vary according to the analysis conditions.

Conclusion

Detection of financial statements fraud is an important issue for companies, stakeholders and economies of countries. Conventional audit methods may be insufficient in detecting financial statements fraud, so innovative and more effective methods may be required. Data mining, which has been successfully applied in many areas, is also promising in this area. In this study, data mining classification analysis was conducted to determine the companies that have the risk of financial statements fraud. Financial statements published between 2014-2018 by companies traded on Borsa Istanbul were used as data source, and as input, 16 financial ratios for each financial statements of each company in each period were calculated and added to the data set. In order to determine which financial variables will be included in the analysis, feature selection was made with the Recursive Feature Elimination algorithm, and 11 financial attributes were selected as input variables.

The first comparison was made using seven different classification methods with default settings. In order to use in the next stage, K-Nearest Neighbor (KNN), Random Forest (RF) and XGBoost (XGB) methods with the highest performance values were selected. As a result of the hyper parameter optimization performed with RandomizedSearch method, correct classification ratios were achieved at 91.73 percent with KNN, 90.51 percent with RF and 90.37 percent with XGB. In the second comparison, the models were compared using the best hyper parameter values in various training-test ratios. Accordingly, KNN models were found to have the best scores in all training-test ratios. In addition, KNN was found to be the least affected classification method by the reduction in the proportion of training data.

The study has two main constraints. The first is about data collection and accuracy. Each outlier was not examined in detail, since the data was shared by the companies on their own initiative. The other constraint is the classification of companies. The companies that are reported to have not committed financial statements fraud were classified in this manner as no fraudulent status could be detected.

As a result, in order to be protected from the damages caused by financial statements fraud, it must be revealed first. Therefore, it is necessary to benefit from innovative and effective solutions. With this study conducted with data mining, a high classification scores were obtained, and new methods rarely found in the literature were included. In terms of its contribution to academic research, this is an original study, as it is a comprehensive data mining classification study aimed at identifying companies at risk of financial statements fraud and is intended to be enlightening for future studies.

MUVU/ 2021-2

1. GİRİŞ

Hile ile ilgili olarak literatürde pek çok tanım bulunmakta olup, ayrıca çeşitli kurumlar tarafından da değişik tanımlar yapılmıştır. Bu tanımlardan biri Sermaye Piyasası Kurulu (SPK) tarafından yapılan tanımdır. SPK tarafından yayımlanan Sermaye Piyasasında Bağımsız Denetim Standartları Hakkında Tebliğ'de hile, "işletme yönetimindekiler ile yönetimden sorumlu kişilerin, işletme çalışanlarının veya üçüncü şahısların kasıtlı olarak adil veya yasal olmayan bir menfaat sağlamak amacıyla aldatma içeren davranışlarda bulunmaları" şeklinde tanımlanmıştır. Association of Certified Fraud Examiners (ACFE) tarafından yayınlanan rapora göre de muhasebe hilesi; varlık suistimali (asset appropriation), yolsuzluk (corruption) ve finansal tablo hilesi (financial statement fraud) olarak üç temel kategoriye ayrılmıştır (ACFE, 2020, s. 10). Finansal tablo hileleri, şirketlerin aldatıcı finansal tablo bilgileri yayınlayarak haksız şekilde kazanç elde etmesiyle gerçekleşen bir suç unsurudur (Cotton, 2002). Hileler, yolsuzlukları gizleme, vergi kaçırma ve işletmenin gerçek durumunu farklı gösterme gibi amaçlarla yapılmaktadır.

Finansal tabloların güvenilir olmaması ve bu tablolarda manipüle edici işlemlerin varlığı mikro düzeyde yatırımcı, borç veren, satıcılar ve çalışanların zarar görmesine neden olurken, makro düzeyde ise kaynakların yanlış yere aktarılmasına ve tüm toplumun zararına neden olmaktadır (Hatunoğlu, Koca & Kıllı, 2012, s. 177). ACFE tarafından yayınlanan raporda, finansal tabloların manipüle edilmesi yoluyla gerçekleşen hilelerin, tüm muhasebe hilelerine oranının sadece yüzde 10 olmasına rağmen, verdiği finansal zararın oldukça yüksek olduğu belirtilmiştir (ACFE, 2020, s. 10). Dünya çapında büyük şirketlerde yapılan finansal tablo hileleri büyük zararlar vermiştir. Örneğin; Enron, WorldCom, Qwest, Global Crossing ve Tyco şirketlerindeki finansal tablo hileleri sebebiyle hissedarlar 460 milyar dolar zarara uğramışlardır (Cotton, 2002). Dünyada büyük çaplı finansal tablo hilelerini gerçekleştiren şirketler ve bu şirketlerin izlediği yöntemler Tablo 1'de gösterilmiştir. Tabloda görüldüğü gibi hileler, belli bir menfaat temini amacıyla belgeler üzerinde bilinçli olarak yapılmaktadır. Bu nedenle, muhasebe sistemi içerisinde kendiliğinden ortaya çıkarılması mümkün değildir (Hatunoğlu, Koca & Kıllı, 2012, s. 177). Finansal tabloların ve bu tabloların dayanağını oluşturan kayıt ve belgelerin güvenilir olup olmadığı sistematik olarak incelenerek kamunun aydınlatılması amacıyla muhasebe denetimi yapılmaktadır (Erol, 2008, s. 232). Geçmişte yaşanan büyük çaplı finansal tablo hileleri, denetime tabi tutularak yayınlanan finansal tablolara olan güvenin sarsılmasına ve bağımsız denetimden geçmiş finansal tablolara şüphe ile bakılmasına neden olmuş, finansal tablo hilelerinin önceden tespitini sağlayacak erken uyarı sistemlerinin gerekliliğini ortaya çıkarmıştır (Uğurlu & Sevim, 2015, s. 66).

Tablo 1: Dünyada Büyük Çaplı Finansal Tablo Hilelerinden Örnekler

Şirketler	Finansal Tablo Hilelerinde İzlenen Yöntemler
Aurora Foods, Inc.	Kazançların olduğundan yüksek, ticari pazarlama harcamalarının ise olduğundan düşük gösterilmesi
Cendant Corporation	1995-1997 arasındaki kazancın abartılarak 500 milyon doların üzerine çıkarılması yoluyla kazanç yönetimi yapılması
Enron Corporation	Borcu gizlemek, ortak Özkaynak yaratmak ve kazançları şişirmek için özel amaçlı kuruluşların (ortaklıkların) oluşturulması
Global Crossing	Hatalı ve yanıltıcı finansal tabloların açıklanması, piyasa değerini yükseltmek amacıyla içeriden bilgi sızdırılması
HBO & Company	1997'den Mart 1999'a kadar kazanç yönetimi yapılması
KnowledgeWare	Sahte yazılım satışı yapılarak rapor edilen kazançların şişirilmesi
MicroStrategy, Inc.	Gelirlerin olduğundan fazla gösterilmesi
Sunbeam Corporation	Gerçekleşmemiş satışların gelir kaydedilmesi yoluyla kazanç yönetimi yapılması
WorldCom	Finansal kayıtların yasa dışı tahrif edilerek gelirlerin 7 milyar dolardan fazla şişirilmesi

Kaynak: Rezaee (2005, s. 281)

Muhasebe hilelerinin tespitinde reaktif ve proaktif olmak üzere iki yaklaşım bulunmaktadır. Reaktif yaklaşımda hile belirtileri ortaya çıktıktan ya da bir şikâyet alındıktan sonra ayrıntılı hile incelemesi yapılırken, proaktif yaklaşımda bir durum ve olasılık önceden tahmin edilerek ona göre belirli çıkarımlar ve duruşlar hazırlanır (Abdioğlu, 2007, s. 122). Bu yaklaşımda hile belirtileri ortaya çıkmadan teknolojik imkânlardan yararlanılarak işletmede yapılabilecek hileler analiz edilir ve hilelerin yapılmaması için önlemler alınır (Ertikin, 2017, s. 72). Proaktif yaklaşımda, istatistiki yöntemler, mali analiz yöntemleri, bilgisayar destekli denetim teknikleri, yapay zekâ, bulanık mantık, uzman sistemler, dijital analiz yöntemleri (Benford Kanunu) ve veri madenciliği yöntemi gibi yöntemler kullanılır (İşgüden Kılıç & Anadolu, 2018, s. 62).

Veri Madenciliği teknikleri diğer yöntemlere göre daha az hata payı içermesi nedeniyle muhasebe denetiminde yaygın kullanım alanı bulmuştur (Ulucan, Özkul & Pertekin, 2009, s. 72). Bu çalışmanın konusu da veri madenciliğinin sınıflandırma metotları ile finansal tablo hilesi riski taşıyan şirketlerin tespitini araştırmaktır. Çalışma, finansal tablo hilesi yapma ihtimali yüksek olan şirketleri veri madenciliğinin sınıflandırma analizi ile

tespit edebilmek için temel olarak özellik seçimi, sınıflandırma metotların kıyaslanması ve hiper parametre optimizasyonu işlemlerini içeren kapsamlı bir veri madenciliği uygulaması tasarlamayı amaçlamaktadır. Veri seti olarak Borsa İstanbul'da 2014-2018 arasında faaliyet gösteren 19 hileli ve 20 düzenli şirketin yayınladıkları finansal tablolar kullanılmıştır. Finansal oranlar girdi değişkeni, şirketin finansal tablo hilesi yapıp yapmadığı ise sınıf değişkeni olarak seçilmiştir. Veri madenciliğinin sınıflandırma metotları ile yapılan analizlere ilişkin sonuçlar karşılaştırmalı olarak sunulmuştur.

Veri madenciliğinde yer alan özellik seçimi, birden çok metodun karşılaştırılması, hiper parametre optimizasyonu gibi işlemler pratik uygulamalarda gözlenmektedir. Literatürde finansal tablo hileleri ile ilgili veri madenciliği çalışmalarında bu işlemlerin tümünü içeren kapsamlı bir çalışmaya rastlanmamıştır. Bunun dışında son zamanlarda başarıyla uygulanan ve dikkatleri üzerine çeken XGBoost sınıflandırma metodu, RobustScaler, RandomizedSearch gibi yenilikçi metotlara yer verilmiştir.

2. LİTERATÜR TARAMASI

Finansal tablo hileleri için veri madenciliğinin kullanılması konusunda yurt içinde ve yurt dışında çeşitli akademik çalışmalar mevcut olmakla birlikte yurt içinde bu alanda az sayıda çalışma bulunmaktadır. Yurt içindeki çalışmalardan biri Ata ve Seyrek (2009) tarafından yapılmıştır. Yazarların İMKB'de işlem gören ve imalat sektöründe faaliyet gösteren 100 şirketin verilerini kullanarak hileli finansal tabloları veri madenciliği ile tespit etmek amacıyla yaptıkları çalışmada karar ağaçları yüzde 67,92 ve yapay sinir ağları yüzde 77,36 doğru sınıflandırma başarısına ulaşmıştır. Yazarlar ayrıca kaldıraç oranı ve aktif karlılık oranının finansal tablo hilelerinin tespitindeki önemini vurgulamışlardır.

Terzi ve Şen (2012) tarafından yapılan çalışmada İMKB'de imalat sektöründe işlem gören, hile riski bulunan 25 ve hile riski bulunmayan 25 şirket verileri kullanılmış, sınıflandırma için yapay sinir ağları ve karar ağacı metotları seçilmiştir. Stokların vadeli borçlara oranı, aktif karlılık oranı ve duran varlıkların aktiflere oranının girdi değişkenleri olarak seçildiği çalışmada; yapay sinir ağlarının riskli şirketleri belirlemede karar ağacına göre daha başarılı olduğu bulunmuştur. Terzi ve Şen (2015) tarafından yapılmış bir diğer çalışmada adli muhasebe hilelerinin tespitinde yapay sinir ağlarının kullanımı araştırılmıştır. Borsa İstanbul'daki 26 şirket bilgilerinden oluşan veri setinde 12 finansal oran kullanılmış ve elde edilen bulgular paylaşılmıştır.

Yurt dışındaki çalışmalarda ise; Kotsiantis vd. (2006) hileli finansal tablo yayınlayan şirketlerin tespitinde makine öğrenme yöntemlerini test

etmişlerdir. Kaynak olarak 164 Yunanlı şirketin 2001-2002 yıllarında paylaştıkları veriler kullanılmıştır. Bunun için bir dizi çözüm yöntemini kapsayan melez (hybrid) karar destek sistemi oluşturulmuştur. Çalışmada finansal hile tespitinde finansal oranların önemi vurgulanmıştır.

Hoogs vd. (2007) potansiyel finansal tablo hilelerini belirlemek için bir genetik algoritma modeli geliştirmişlerdir. Genetik algoritma, sınıf özelliklerinin seçiminde sınıf sınırlarını belirlemek için kullanılmıştır. Çalışmanın sonucunda yüzde 63 gerçek pozitif (TP) ve yüzde 5 hatalı pozitif (FP) oranları elde edilmiştir. Bu tekniğin finans ve üretim dahil olmak üzere çok boyutlu davranış analizi için bir temel teşkil ettiği belirtilmiştir.

Kirkos vd. (2007) çalışmalarında hileli finansal tablolar yayınlayan şirketlerin tespiti için veri madenciliği sınıflandırma tekniklerinin verimliliğini araştırmışlardır. Sınıflandırma yöntemlerinden karar ağaçları, sinir ağları ve bayes ağlarının kullanışlılığı incelenmiş, üç modelin performansları karşılaştırılmıştır.

Lenard vd. (2007) çalışmalarında halka açık bilgilerle teknolojik hizmet şirketlerindeki hileli finansal raporları tespit etmeye yönelik bir araştırma yapmışlardır. Finansal ve finansal olmayan değişkenler kullanılmış, bulanık mantık metoduyla hileli şirketler yüzde 76,7 doğrulukta tahmin edilmiştir.

Bai vd. (2008) hatalı finansal tabloları tanımlamak ve tahmin etmek için CART yöntemini tercih etmişlerdir. Veri seti olarak 24 hatalı ve 124 hatasız finansal rapor kullanmışlardır. Yapılan karşılaştırmada logit regresyon yöntemiyle daha başarılı sonuçlar elde edildiği görülmüştür.

Deng ve Mei (2009) kümeleme geçerlilik ölçüsüne dayalı özdüzenleyici harita (self organizing map) ve k-ortalama kümelemeyi birleştiren bir model tasarlamışlardır. Veri seti olarak 1996-2006 yılları arasında işlem gören halka açık 100 Çinli şirketin finansal tabloları kullanılmıştır. Değişken olarak 47 finansal oran seçilmiş, yapılan modelden elde edilen deneysel sonuçlar sunulmuştur.

Gaganis (2009) çalışmasında, tahrif edilmiş finansal tabloların belirlenmesi için logit analiz kullanarak 10 alternatif sınıflandırma modeli geliştirmiştir. Yunanlı şirketlere ait 2001–2004 yılları arası yarısı hileli 398 finansal tablo kullanılmış, finansal ve finansal olmayan değişkenlerle yapılan analizlerin karşılaştırma sonuçları ayrıntılı olarak sunulmuştur.

Perols (2011) çalışmasında altı popüler istatistik ve makine öğrenme modelini farklı sınıflandırma maliyeti varsayımları altında karşılaştırmıştır. Sonuçlara göre lojistik regresyon ve destek vektör makinesi performansının yapay sinir ağları, torbalama, C4.5 ve istiflemeye (stacking) göre daha iyi olduğu saptanmıştır.

Ravisankar vd. (2011) çalışmalarında finansal tablo sahtekarlığına başvuran

şirketleri belirlemek için çeşitli veri madenciliği tekniklerini kullanmışlardır. 202 Çinli şirkete ait finansal bilgilerin veri seti olarak kullanıldığı araştırmada özellik seçimi bulunup bulunmamasına göre karşılaştırmalar yapılmıştır. Özellik seçiminin yapıldığı durumda genetik programlama ve olasılıklı sinir ağlarının her ikisinin, özellik seçimi yapılmadığı durumda ise olasılıklı sinir ağlarının daha başarılı olduğu belirlenmiştir.

Lin vd. (2015) çalışmalarında veri madenciliği yöntemleri ve uzman görüşlerini kullanarak sahtekârlık analizi yapmışlardır. Veri madenciliği için lojistik regresyon, karar ağaçları ve yapay sinir ağları kullanılmıştır. Çalışmada yapay sinir ağları ve karar ağaçları ile diğer yöntemlerden daha başarılı sonuçlar elde edilmiş, buna ek olarak uzman değerlendirmelerinin katkısı incelenmiştir.

Rizki vd. (2017) yaptıkları analizde finansal tablo hilelerinin veri madenciliği ile tespitinde destek vektör makinesi ve yapay sinir ağlarını kullanmışlardır. Yapılan analiz sonucunda özellik seçimi yapılarak destek vektör makinesi ile yüzde 88,37, özellik seçimi yapılmadan ise yapay sinir ağları ile yüzde 90,97 doğru sınıflandırma başarısına ulaşılmıştır.

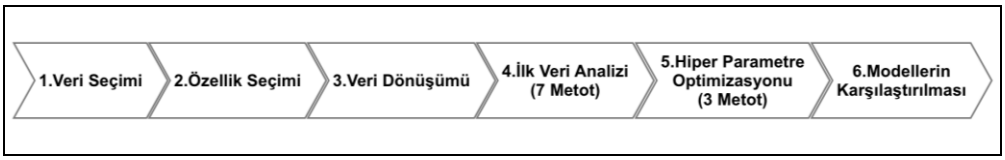
Jan (2018) şirketlerin ve finansal piyasaların sürdürülebilir gelişimi için finansal tablo hilelerini tespit etmeye yönelik bir model geliştirmiştir. Tayvan'daki 160 şirketin 2004-2014 arası verileri kullanılmış, analiz için çoklu veri madenciliği teknikleri benimsenmiştir. Çalışmada yapay sinir ağları ve CART metodunun birlikte kullanıldığı model finansal tablo hilelerinin tespitinde yüzde 90,83 doğruluk ile en iyi sınıflandırma sonucunu vermiştir.

Yao vd. (2018) özellik seçimi ile makine öğrenimi sınıflandırmasını birleştiren optimize edilmiş bir finansal hile tespit modeli önermişlerdir. 120 hileli ve 120 hileli olmayan şirket ile yapılan analizde, 17 finansal değişken ve 5 finansal olmayan değişken kullanılmıştır. Sınıflandırmada Random Forest, özellik seçiminde ise XGBoost kullanılarak en iyi performans değerleri elde edilmiştir.

3. METODOLOJİ

Analiz süreci altı aşamadan oluşmaktadır (Şekil 1). Başlangıç olarak şirketlerin finansal raporlarından elde edilen veriler derlenmiştir. İkinci aşamada model geliştirme sürecinde değerlendirmeye alınacak değişkenler belirlenmiştir. Üçüncü aşama, verilerin sınıflandırma metotlarına uygun hale getirilmesi için gerekli olan dönüşüm işlemlerinin uygulanmasını kapsamaktadır. Dördüncü aşamada metotların varsayılan yapısal ayarları ile ilk kıyaslama işlemi yapılmış, en iyi doğruluk oranına sahip üç sınıflandırma metodu seçilmiştir. Beşinci aşamada, seçilen metotlar için

hiper parametre optimizasyonu yapılmıştır. Altıncı ve son aşamada ise optimize edilmiş modeller farklı eğitim-test oranlarıyla karşılaştırılmış ve elde edilen sonuçlar değerlendirilmiştir.



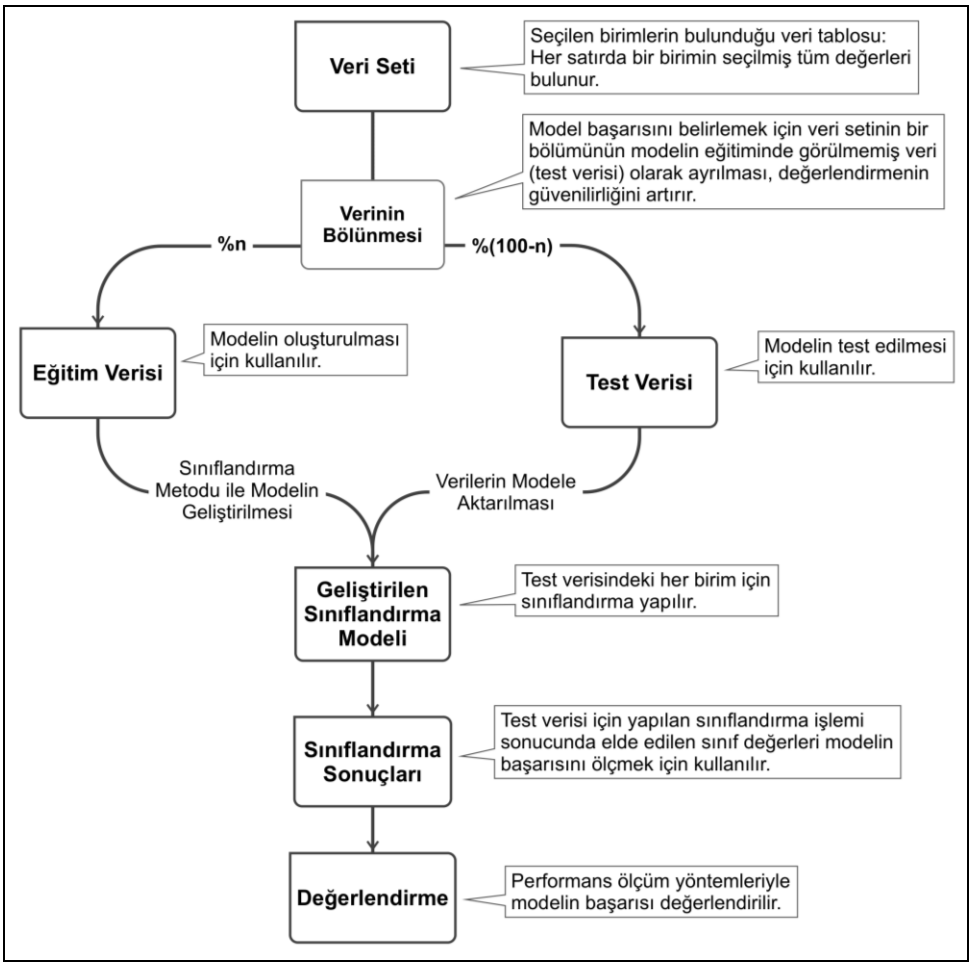
Şekil 1: Araştırma Sürecinin Modeli

3.1. Sınıflandırma Analizi

Veri madenciliği, kısaca verideki kalıpları keşfetme süreci olarak tanımlanabilir (Witten vd., 2011, s. 5). Veri madenciliğinin uygulama alanlarından biri olan sınıflandırma ise önemli veri sınıflarını tanımlayan modelleri çıkaran bir veri analizi biçimidir. Bu tür modeller, kategorik sınıf etiketlerini tahmin eder (Han vd., 2011, s. 327).

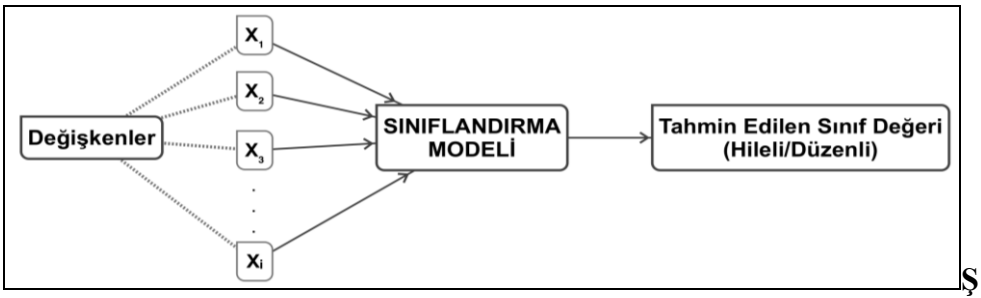
Bu çalışmada veri madenciliğinin sınıflandırma analizi uygulanmıştır. Sınıflandırma analizinde sınıflandırma metotları kullanılarak sınıflandırma modeli geliştirilir. Sınıflandırma modelinin geliştirilmesi ve değerlendirilmesi süreci açıklamalarıyla birlikte görselleştirilmiştir (Şekil 2). Geliştirilen model, bağımsız değişkenleri kullanarak seçilen birimi sınıflandırır (Şekil 3). Şekilde X'ler ilgili birimin bağımsız değişkenlerini temsil etmektedir.

MUVU/ 2021-2



Şekil 2: Sınıflandırma Modelinin Oluşturulması ve Değerlendirilmesi Süreci

Kaynak: Yazarlar



Şekil 3: Sınıflandırma Modelinin İşleyişi

Kaynak: Yazarlar

3.2. Veri Kaynağı

Çalışmada kullanılan veri kaynağı, Borsa İstanbul'da 2014-2018 arasında süreklilik gösteren şirketlerin Kamuyu Aydınlatma Platformu'nun erişim

sayfasında (kap.org.tr) halka açık paylaştıkları çeyrek dönemlik finansal tablolarıdır. Araştırma birimi bir şirketin bir dönemidir. Birimin özellikleri (değişkenleri) finansal oranlar ve sınıf değişkeninden oluşmaktadır.

Sınıf değişkeninin değerleri "hileli" ve "düzenli" isimleri ile kodlanmıştır. Hileli şirketler, SPK haftalık bültenleri, denetim raporları ve denetçi görüşleri taranarak belirlenmiştir. Düzenli şirketler, aynı sektörlerden tesadüfi olarak seçilen şirketlerdir. Hileli şirket sınıfında, finansal kuruluşlar ve yatırım şirketleri dışında kalan 19 şirket bulunmaktadır (Tablo 2). Seçilen düzenli şirket sayısı ise 20'dir. Veri setinde 19 hileli şirkete ait toplam 337 dönem, 20 düzenli şirkete ait toplam 400 dönem bulunmaktadır. Temel sınıfta 337 birimin bulunma sebebi, bu sınıftaki şirketlerin bir kısmının finansal tablolarını yıl içinde tüm dönemlerde değil sadece 2 dönemde yayınlamış olmalarıdır.

Tablo 2: Hileli ve Düzenli Şirketlerin Sektörlere Göre Dağılımı

Sektörler	Hileli Şirketler	Düzenli Şirketler
Ana Metal Sanayi	1	1
Elektrik, Gaz ve Su	1	1
Gıda, İçecek ve Tütün	3	3
Holdingler	3	3
İdari ve Destek Hizmet Faaliyetleri	2	1
İnşaat ve Bayındırlık	1	2
Kimya İlaç Petrol Lastik ve Plastik Ürünler	1	1
Metal Eşya Makine Elektrikli Cihazlar ve Ulaşım Araçları	1	1
Tarım, Ormancılık ve Balıkçılık	2	1
Taş ve Toprağa Dayalı	1	2
Tekstil, Giyim Eşyası ve Deri	1	1
Toptan ve Perakende Ticaret, Lokantalar ve Oteller	1	2
Ulaştırma, Depolama ve Haberleşme	1	1
Toplam	19	20

3.3. Uygulama Ortamı

Verilerin derlenmesi, düzenlenmesi ve analizler için Python programlama dili ve Jupyter-Python Notebook editörü kullanılmıştır. Analizde kullanılan sınıflandırma metotları ve verilerin işlenmesi için Python diliyle geliştirilmiş kütüphanelerden yararlanılmıştır.

MUVU/ 2021-2

3.4. Değişkenler

Analize dahil edilecek finansal değişkenler belirlenmeden önce başlangıç olarak literatürde sıkça karşılaşılan finansal oranlar ve bunların nasıl hesaplandığını gösteren formüllerden oluşan bir tablo hazırlanmıştır (Tablo 3). Daha sonra bilanço ve gelir tablolarındaki değerler kullanılarak, verilen formüllere göre yapılan hesaplamalar sonucunda elde edilen değerler, araştırma birimlerinin özellikleri olarak veri setine eklenmiştir.

Tablo 3: Finansal Değişkenler

	Değişken Adı	Formül
1	Cari Oran	$\frac{Dönen\ Varlıklar}{Kısa\ Vadeli\ Yabancı\ Kaynaklar}$
2	Stoklar/Dönen Varlıklar	$\frac{Stoklar}{Dönen\ Varlıklar}$
3	Öz Kaynaklar/Yabancı Kaynaklar	$\frac{Öz\ Kaynaklar}{Yabancı\ Kaynaklar}$
4	Toplam Borçlar/Aktifler	$\frac{Toplam\ Borçlar}{Aktif\ Toplamı}$
5	Aktif Devir Hızı	$\frac{Net\ Satışlar}{Aktif\ Toplamı}$
6	Alacak Devir Hızı	$\frac{Net\ Satışlar}{Yabancı\ Kaynaklar}$
7	Çalışma Sermayesi/Aktifler	$\frac{Çalışma\ Sermayesi}{Aktif\ Toplamı}$
8	Alacaklar/Net Satışlar	$\frac{Ticari\ Alacaklar}{Net\ Satışlar}$
9	Net Kâr/Net Satışlar	$\frac{Net\ Kar}{Net\ Satışlar}$
10	Net Kâr/Aktifler	$\frac{Net\ Kar}{Aktif\ Toplamı}$
11	Brüt Kâr/Aktifler	$\frac{Brüt\ Kar}{Aktif\ Toplamı}$
12	Stoklar/Net Satışlar	$\frac{Stoklar}{Net\ Satışlar}$
13	Finansman Giderleri/Faaliyet Giderleri	$\frac{Finansman\ Giderleri}{Faaliyet\ Giderleri}$
14	Ödenecek Vergi/Net Satışlar	$\frac{Ödenecek\ Vergi}{Net\ Satışlar}$
15	Satılan Malın Maliyeti/Net Satışlar	$\frac{Satılan\ Malın\ Maliyeti}{Net\ Satışlar}$
16	Kısa Vadeli Borç Artışı/Satışların Artışı	$\frac{K.V.B.(Cari\ Yıl) - K.V.B.(Önceki\ Yıl)}{Satışlar(Cari\ Yıl) - Satışlar(Önceki\ Yıl)}$

Veri madenciliği analizinde hangi değişkenlerin analizde bulunması gerektiği önemli bir konudur. Değişkenler arasındaki ilişkiler, geliştirilecek modelin başarısını düşürebilir. Bunun dışında, bazı değişkenlerin veri setinde bulunması sadece gereksiz yere fazladan işlem süresine ve kaynak tüketimine sebep olabilir. Özellik seçimi genel olarak daha az sayıda değişkenle işlem yapılmasını sağladığı için hesaplama süresi ve işlem miktarı tasarrufu sayesinde verimliliği artırmaktadır. Bu gibi sebeplerden ötürü analizde en etkili değişkenler seçilip diğerleri elenir, yani özellik seçimi yapılır.

Çalışmada özellik seçimi için Özyinelemeli Özellik Eliminasyonu (Recursive Feature Elimination - RFE) metodu kullanılmıştır. RFE, tekrarlamalı olarak daha az sayıda özellik seçimini amaçlar. Bunun için öncelikle tahminleyici başlangıçtaki özellikler seti ile eğitilir ve her özelliğin önemi belirlenir. Daha sonra en az öneme sahip özellikler veri setinden çıkarılır. Bu prosedür istenen değişken sayısına ulaşıldığında sona erer (Scikit-Learn, 2020a). RFE algoritmasının sözde kodu bu metodun işleyişini özetlemektedir (Şekil 4). Özellik seçiminin özyinelemeli yapılması gereklidir, çünkü bazı ölçütler için her bir özelliğin göreceli önemi, aşamalı eleme işlemi sırasında (özellikle yüksek korelasyona sahip özellikler açısından) farklı özellik alt kümeleri için önemli ölçüde değişebilir (Granitto, Furlanello, Biasioli & Gasperi, 2006, s. 84). Çalışmada RFE işleminde tahminleyici olarak Lojistik Regresyon fonksiyonu kullanılmıştır. Özelliklerin sayısı için tüm seçenekleri içeren bir döngü kullanılarak en uygun değişken sayısı 11 olarak bulunmuştur. Bu işlem sonucunda seçilen değişkenler tablo halinde verilmiştir (Tablo 4).

Girdiler:

Eğitim seti T

F'nin p özellik kümesi = $\{f_1,\ldots,f_p\}$

Sıralama yöntemi $M(T, F)$

Çıktılar:

Nihai sıralama R

Kod:

{1:p} içinde i'yi tekrarla

$M(T, F)$ ile F kümesini sırala

$f^* \longleftarrow F$ içinde son sıradaki özellik

$R\ (p - i + 1) \longleftarrow f^*$

$F \longleftarrow F - f^*$

Şekil 4: Özyinelemeli Özellik Seçimi Algoritmasının Sözde Kodu

Kaynak: Granitto (2006, s. 84)

MUVU/ 2021-2

Tablo 4: Özellik Seçimi İşlemi Sonucunda Seçilen Değişkenler

Değişken	Değişken Adı
X_1	Aktif Devir Hızı
X_2	Net Kâr/Net Satışlar
X_3	Net Kâr/Aktifler
X_4	Çalışma Sermayesi/Aktifler
X_5	Brüt Kâr/Aktifler
X_6	Stoklar/Net Satışlar
X_7	Toplam Borçlar/Aktifler
X_8	Ödenecek Vergi/Net Satışlar
X_9	Stoklar/Dönen Varlıklar
X_{10}	Alacaklar/Net Satışlar
X_{11}	Satılan Malın Maliyeti/Net Satışlar

Yapılan incelemelerde aykırı değerlere sahip birimlerin çıkarılmasının, birim sayısını analizin sağlıklı yapılmasına engel olacak miktarda düşürdüğü görülmüş, bu sebeple aykırı değerler analizde tutulmuştur. Tanımlayıcı istatistikler, aykırı değerlerden etkilenmemesi amacıyla her değişken için %5-%95 aralığındaki veriler kullanılarak hesaplanmıştır (Tablo 5).

Tablo 5: Sınıflara Göre Tanımlayıcı İstatistikler

	Minimum		Maksimum		Ortalama ± Standart Sapma	
	Hileli	**Düzenli**	**Hileli**	**Düzenli**	**Hileli**	**Düzenli**
X_1	0,0275	0,0146	1,5753	0,8911	0,4214±0,3418	0,3066±0,2297
X_2	0,0673	0,0105	2,1413	1,6843	0,5011±0,4018	0,4565±0,3767
X_3	-0,0548	-0,022	0,2988	0,2128	0,0432±0,0673	0,0609±0,0499
X_4	-0,1684	-0,059	0,185	0,1547	0,0041±0,0535	0,0314±0,0442
X_5	-0,3856	-0,996	0,481	1,2324	0,0118±0,1519	0,0642±0,2892
X_6	-1,0062	-1,038	-0,6008	-0,3352	-0,8298±0,0961	-0,7758±0,136
X_7	0	0	0,6064	0,5388	0,2549±0,1608	0,1818±0,1542
X_8	0	0	1,5491	1,9393	0,3449±0,3293	0,3402±0,4082
X_9	0,1539	0,0302	2,9908	0,7561	0,4725±0,344	0,4339±0,1785
X_{10}	-1,5956	-0,086	0,6293	0,4991	0,1145±0,2817	0,1762±0,1451
X_{11}	0	0	0,043	0,0583	0,0048±0,0088	0,0097±0,0132

3.5. Veri Ölçeklendirme

Ölçeklendirme, verinin analize hazırlanması için yapılması gereken işlemlerden biridir. Ölçeklendirme algoritmadaki hesaplamaların hızlandırılmasına yardımcı olur (Thara, PremaSudha & Xiong, 2019). Ayrıca değişkenler arasındaki ortalama ve varyans farklarının çok olması bazı değişkenlerin diğerlerine göre gereğinden fazla baskın olmasına neden olabilir. Ölçeklendirme, değişkenlerin belirli bir ölçeğe göre yeniden düzenlenmesini sağlayan işlemdir. Ölçeklendirmede en çok kullanılan yöntemler normalizasyon ve standardizasyondur. Normalizasyon işleminde özellikteki minimum ve maksimum değer 0 ve 1 olacak şekilde değişkenin tüm değerleri orantılı olarak ölçeklendirilir. Standardizasyon işleminde ilgili değişkenin ortalaması 0 ve standart sapması 1 olacak şekilde tüm değerler doğrusal orantılı olarak değiştirilir. Tipik olarak bu işlem, ortalamanın çıkarılması ve birim varyansına göre ölçeklendirmesiyle yapılır. Bununla birlikte, aykırı değerlerin bulunması ortalama ve varyansı olumsuz yönde etkileyebilir (Scikit-Learn, 2020b). Standardizasyon, veri setindeki aykırı değerlere aşırı duyarlıdır. Bu sebeple çalışmada aykırı değerlerden etkilenmeyen RobustScaler ölçeklendirme metodu kullanılmıştır.

RobustScaler metodu ortancayı kaldırır ve ölçeklendirmeyi kartil aralığına göre yapar. Varsayılan özelliğe göre kartil aralığı IQR (Interquartile Range) olarak geçer ve birinci kartil ile üçüncü kartil arasındaki aralıktır. Merkezleme ve ölçeklendirme, eğitim veri setinden elde edilen örneklerin istatistiklerinin her özellik için ayrı ayrı hesaplanması ile yapılır (Scikit-Learn, 2020b).

3.6. Sınıflandırma Metotları

Veri madenciliği, makine öğrenmesi, istatistik, örüntü tanıma, veri tabanları gibi farklı disiplinlerle etkileşim içinde olan kapsamlı bir alandır. Dolayısıyla sınıflandırma amacına yönelik çok sayıda metot bulunmaktadır. Birinci karşılaştırma için yaygın olarak kullanılan sınıflandırma metotlarından yedisi seçilmiştir (Tablo 6). Tabloda sınıflandırma metotları, kısaltmaları ve kütüphane bilgileri bulunmaktadır. Benzer işlemler için aynı kodları tekrar tekrar yazmak çok zaman alabilen, hatalara açık ve zahmetli bir süreçtir. Bunun için belirli prosedürleri ve işlemleri gerçekleştirebilen, özelleştirilebilen fonksiyonlar bulunan ve genellikle topluluklar tarafından iş birliğiyle geliştirilen kod kütüphaneleri bulunmaktadır. Kullanıcılar benzer işlemleri yeni baştan kod yazmayla uğraşmadan, kod kütüphanelerinden yararlanarak hızlıca yapabilmektedirler. Çalışmada kullanılan sınıflandırma metotları için bu tür kütüphanelerden yararlanılmış, ilgili kütüphanelerin bağlantıları tabloda verilmiştir. Bu kütüphaneler herkesin kullanımına açıktır ve ayrıca paylaşım sayfalarında hem geliştiriciler hem de kullanıcılar için bilgilendirici kaynaklar bulunmaktadır.

Tablo 6: Sınıflandırma Metotları ve Kullanılan Kod Kütüphaneleri

Metot	Python Kütüphanesi
Random Forest (RF)	sklearn.ensemble.RandomForestClassifier *https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html*
K-Neighbors (KNN)	sklearn.neighbors.KNeighborsClassifier *https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html*
XGBoost (XGB)	xgboost.XGBClassifier *https://xgboost.readthedocs.io/en/latest/python/python_api.html*
Support Vector Machine (SVM)	sklearn.svm.SVC *https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html*
Multi Layer Perceptron (MLP)	sklearn.neural_network.MLPClassifier *https://scikit-learn.org/stable/modules/generated/sklearn.neural_network.MLPClassifier.html*
Decision Tree (CART)	sklearn.tree.DecisionTreeClassifier *https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html*
Gaussian Naive Bayes (NB)	sklearn.naive_bayes.GaussianNB *https://scikit-learn.org/stable/modules/generated/sklearn.naive_bayes.GaussianNB.html*

3.7. Modellerin Değerlendirilmesi

Farklı metotlar kullanılarak oluşturulan sınıflandırma modellerinin birbirleriyle kıyaslanabilmesi için sınıflandırma başarıları ölçülmelidir. Bu sebeple iki konuda karar vermek gerekir. Bunlardan birincisi kullanılacak performans ölçütlerinin seçilmesi, diğeri ise bu ölçütlerin hesaplanabilmesi için kullanılacak veri setinin (eğitim-test) kullanım yönteminin belirlenmesidir.

3.7.1. Değerlendirme Ölçütleri

Uygulama ikili sınıflandırma problemidir ve performans ölçütleri buna göre hesaplanmalıdır. Hesaplamalar için gerçek sınıfların, tahmin edilen sınıfların ve doğru-yanlış sınıflandırılan örnek sayılarının gösterildiği Karışıklık Matrisi (Confusion Matrix) kullanılmıştır (Tablo 7). Karışıklık matrisinde pozitif ve negatif olarak isimlendirilen iki sınıf bulunur. Çalışmada pozitif, finansal tablo hilesi tespit edilen şirketlerin sınıfıdır, yani

hileli sınıfındaki birimlerdir. Negatif ise finansal tablo hilesi tespit edilmeyen şirketlerin sınıfıdır, yani *düzenli* sınıfındaki birimlerdir.

Tablo 7: Karışıklık Matrisi (Confusion Matrix)

	Tahmin Edilen →	**+ / hileli**	**- / düzenli**
Gerçek Sınıf	**+ / hileli**	TP - True Positive (Gerçek Pozitif)	FN - False Negative (Hatalı Negatif)
	- / düzenli	FP - False Positive (Hatalı Pozitif)	TN - True Negative (Gerçek Negatif)

Kaynak: Witten vd. (2011, s. 164)

Modellerin genel performansı için temel karşılaştırma kriteri olarak doğruluk (accuracy) değerlerine bakılır. Doğruluk, karışıklık matrisindeki veriler kullanılarak şöyle hesaplanır (Han, Kamber & Pei, 2011, s. 366):

$$\text{Doğruluk} = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

Diğer performans ölçütleri TP-Oranı (Gerçek Pozitif Oranı), TN-Oranı (Gerçek Negatif Oranı) ve F-Ölçütüdür. TP-Oranı tahmin sonucunda tespit edilen gerçek pozitif örneklerin oranını vermektedir. Çalışmada bu oran hileli şirketlerin doğru tahmin oranı olarak düşünülebilir. Hesaplanışı aşağıdaki gibidir (Han vd., 2011, s. 368):

$$\text{TP-Oranı} = \frac{TP}{TP+FN} \quad (2)$$

TN-Oranı gerçek negatiflerin örneklerin oranını vermektedir. Çalışmadaki karşılığı, düzenli şirketlerin doğru tahmin oranıdır ve şöyle hesaplanır (Hajek & Henriques, 2017, s. 145):

$$\text{TN-Oranı} = \frac{TN}{TN+FP} \quad (3)$$

F-Ölçütü (F-Measure) değeri kesinlik ve hassasiyet değerlerinin armonik ortalamasıdır. Dengelenmemiş sınıfların olduğu problemlerde F-Ölçütü değeri daha belirgin farklılıklar gösterebilir. Karışıklık matrisi verileri kullanılarak şöyle hesaplanır (Lipton, Elkan & Naryanaswamy, 2014, s. 227):

$$\text{F-Ölçütü} = \frac{2TP}{2TP+FP+FN} \quad (4)$$

Modellerin karşılaştırılmasında performans ölçütlerine ek olarak hız, ölçeklenebilirlik ve sağlamlık (robustness) kriterleri de göz önünde bulundurulabilir. Hangi kriterlerin ne derecede dikkate alınacağı problemin niteliğine ve araştırmanın amacına bağlı olarak değişkenlik gösterir.

3.7.2. Modellerin Doğrulanması

Sınıflandırma problemlerinin temel konularından biri modelin oluşturulması için kullanılan eğitim verisi ve geliştirilen modelin doğrulanması için kullanılan test verisinin belirlenmesi işlemidir. Eğitim ve test için aynı veri seti kullanılırsa sonuçlar güvenilir olmayabilir, test verisinin eğitim amacıyla kullanılmamış olması yani modelin görülmemiş veriyle test edilmesi gerekir (Witten vd., 2011, s. 148). Eğitim ve test verisinin belirlenmesi için farklı yöntemler geliştirilmiştir. Bu çalışmada bu amaçla 10-kat çapraz doğrulama ve rasgele alt örnekleme yöntemleri kullanılmıştır.

10-kat çapraz doğrulama yönteminde veri 10 parçaya ayrılır ve 10 farklı denemede her bir parça modelin test edilmesi için, geri kalanı ise modelin eğitilmesi için kullanılır. 10 deneme sonucunda tüm yinelemelerin başarı ortalaması alınarak modelin başarısı belirlenir.

Holdout yönteminde, veri seti seçilen bir oranla eğitim ve test verisi olarak iki bölüme ayırır. Rasgele alt örnekleme yöntemi, holdout yönteminin birden fazla kez uygulanması ile elde edilir. Doğruluk tahmini, yinelemelerin tümünden elde edilen doğruluk değerlerinin ortalaması alınarak bulunur (Han vd., 2011, s. 370). Rasgele alt örnekleme yönteminde yapılan yinelemenin, sonuçların güvenilirliğini artırması sebebiyle, rasgele alt örnekleme yöntemi her model için 10 kez yinelemeyle uygulanmıştır.

4. UYGULAMA SONUÇLARI

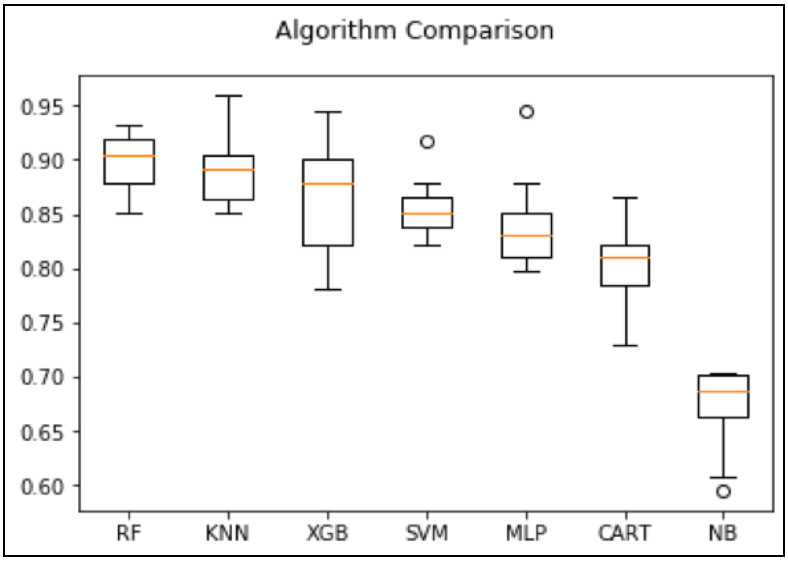
4.1. Birinci Karşılaştırma

İlk olarak, seçilen sınıflandırma metotlarının varsayılan hiper parametre değerleri kullanılarak, 10-kat çapraz doğrulama yöntemiyle modeller karşılaştırılmıştır (Tablo 8, Şekil 5). Buna göre en yüksek doğru sınıflandırma oranı RF metodu ile elde edilen yüzde 89,83 performans değeridir. Ardından sırasıyla KNN ve XGB metotlarıyla yüzde 88,87 ve yüzde 86,44 doğru sınıflandırma oranlarına ulaşılmıştır.

Tablo 8: Sınıflandırma Metotlarının Karşılaştırması

Metot	Doğruluk (%)	Std. Sapma (%)
RF	89,83	2,41
KNN	88,87	3,19
XGB	86,44	5,40
SVM	85,49	2,71
MLP	84,00	4,29
CART	80,33	4,06
NB	67,17	3,78

Veri madenciliğinin zorluklarından bir tanesi kullanılan metot için en uygun hiper parametre değerlerini bulmaktır. Bunun sebebi, en uygun değerleri bulmak için deneme yanılmadan başka bir yöntemin bulunmamasıdır. Uygulayıcılar için geliştirilen kütüphanelerde metotların varsayılan hiper parametre değerleri bulunur. Her ne kadar varsayılan değerler genel kullanım için en uygun hale getirilmeye çalışılsa da doğru sınıflandırma başarısı veri setine, seçilen özelliklere ve problemin türüne göre değişkenlik gösterebilir. Buna göre yapılan karşılaştırma sonucunda en yüksek doğruluk oranına sahip üç metot seçilmiş, bu üç metot için en uygun hiper parametre değerlerinin bulunması amacıyla bir sonraki adımda optimizasyon işlemi yapılmıştır.



Şekil 5: Sınıflandırma Metotlarının Doğruluk Değerlerinin Kutu Grafiği

4.2. Hiper Parametre Optimizasyonu

Model hiper parametresi, modelin dışında olan ve değeri verilerden tahmin edilemeyen bir yapılandırma değeridir. Belirli bir problem için en iyi hiper parametre değeri bilinemez. Benzer problemlerden daha önceden elde edilmiş yakın değerler kullanılabilir ancak, bu seçenek en iyi sonucu garanti etmez. En iyiye yakın hiper parametre değerlerine, sadece deneme yanılma yöntemiyle ulaşılabilir. Hiper parametrelerin temel özellikleri şöyle sıralanabilir (Brownlee, 2017):

- Model parametrelerinin tahmin edilmesi için kullanılır.
- Genellikle uygulayıcı tarafından belirlenir.
- Genellikle sezgisel tarama kullanılarak ayarlanabilir.
- Genellikle modelleme problemi için uyarlanırlar.

MUVU/ 2021-2

Birinci karşılaştırma bütün metotların varsayılan hiper parametre değerleri kullanılarak yapılmıştır. Bu değerlerin değiştirilmesi doğru sınıflandırma oranlarını da etkilemektedir. Ancak seçenekleri deneyerek görmek dışında en uygun hiper parametre değerlerine ulaşma olanağı yoktur. Bu amaçla birinci karşılaştırmada seçilen metotların hiper parametre değerleri için seçenek değerler listeleri oluşturulmuştur. Metotlarda uygulanan her bir hiper parametre değeri değişikliği, değerlendirilmesi gereken yeni bir modeli temsil eder. Seçeneklerin az olduğu durumda bu işlem uzun sürmez ancak, çalışmada çok sayıda hiper parametre kombinasyonu bulunmaktadır. Bu durumda kurulması gereken model sayısı katlanarak artmakta, altyapı ve zaman kaynağı bu olanağı sınırlandırmaktadır. Bu zorluğu aşabilmek için Scikit-Learn tarafından geliştirilen "RandomizedSearch" metodu kullanılarak uygun hiper parametre değerlerine ulaşılmaya çalışılmıştır.

Scikit-Learn kütüphanesinin bir metodu olan RandomizedSearch, seçilen hiper parametreler için oluşturulan değerlerin tüm kombinasyonlarının değerlendirilmesinin mümkün olmadığı durumlarda, bu kümeden belirli sayıda model seçerek hesaplama zamanı ve donanım kaynaklarında tasarruf sağlar (Sjardin, Massaron & Boschetti, 2016). Örnek olarak, seçilen üç metottan biri olan KNN için "leaf_size", "metric", "n_neighbors" ve "weights" hiper parametrelerinin değer kümeleri oluşturulmuştur (Tablo 9). Bu değerlerle oluşturulabilecek tüm kombinasyonların analiz edilebilmesi için 12 x 3 x 30 x 2 = 2.160 farklı modelin kurulması gerekir. Bunun yerine RandomizedSearch metodu ile 80 modelin değerlendirilmesi yeterli görülmüştür. Bu denemelerden elde edilen sonuçlar ve bu sonuçların hangi hiper parametre değerleriyle alındığı bir sonraki tabloda verilmiştir (Tablo 10). Buna göre hiper parametre değerleri; param_weights: "distance", param_n_neighbors: "2", param_metric: "manhattan" ve param_leaf_size: "20" olduğunda, yüzde 2,35 standart sapma ile ortalama yüzde 91,73 tahmin başarısı elde edilmektedir. Tablodaki "fit_time" değeri modelin oluşturulması için harcanan süreyi, "score_time" ise modelin test edilmesi için harcanan süreyi göstermektedir.

Tablo 9: KNN için Hiper Parametre Seçenekleri

Hiper Parametre	Değerler	Seçenek
leaf_size	5; 10; 15; 20; 25; 30; 35; 40; 45; 50; 55; 60	12
metric	euclidean; manhattan; minkowski	3
n_neighbors	1; 2; 3; 4; 5; 6; 7; 8; 9; 10; 11; 12; 13; 14; 15; 16; 17; 18; 19; 20; 21; 22; 23; 24; 25; 26; 27; 28; 29; 30	30
weights	uniform; distance	2

Tablo 10: KNN için Hiper Parametre Optimizasyonu Sonuçları (En İyi 5)

rank_test_score	**1**	**2**	**2**	**4**	**4**
mean_fit_time	0,00300	0,00300	0,00250	0,00300	0,00320
std_fit_time	0,00001	0,00001	0,00050	0,00000	0,00040
mean_score_time	0,00190	0,00200	0,00210	0,00220	0,00240
std_score_time	0,00030	0,00000	0,00030	0,00040	0,00049
param_weights	distance	distance	distance	distance	distance
param_n_neighbors	2	4	4	4	4
param_metric	manhattan	minkowski	euclidean	manhattan	manhattan
param_leaf_size	20	25	35	15	10
mean_test_score	0,917327	0,917290	0,917290	0,915939	0,915939
std_test_score	0,023543	0,027863	0,027863	0,032312	0,032312

Aynı işlem RF ve XGB metotları için de yapılmış, araştırılan hiper parametre setleri ve en uygun hiper parametre değerleri ilgili tablolarda verilmiştir (Tablo 11-Tablo 14). Hiper parametre optimizasyonu sonucunda KNN yüzde 91,73, RF yüzde 90,51 ve XGB yüzde 90,37 doğru sınıflandırma oranlarına ulaşmıştır.

Tablo 11: RF için Hiper Parametre Seçenekleri

Hiper Parametre	**Değerler**	**Seçenek**
bootstrap	True; False	2
max_depth	10; 20; 30; 40; 50; 60; 70; 80; 90; 100; None	11
max_features	auto; sqrt	2
min_samples_leaf	1; 2; 4	3
min_samples_split	2; 5; 10	3
n_estimators	50; 100; 200; 500; 1000	5

Tablo 12: RF için Hiper Parametre Optimizasyonu Sonuçları (En İyi 5)

rank_test_score	**1**	**2**	**3**	**4**	**5**
mean_fit_time	0,266038	0,525508	1,317636	0,262877	1,315487
std_fit_time	0,003290	0,003174	0,011304	0,001857	0,008386
mean_score_time	0,011521	0,021599	0,053018	0,011940	0,053345
std_score_time	0,000334	0,000240	0,001457	0,000578	0,001655
param_n_estimators	100	200	500	100	500
param_min_samples_split	5	2	2	5	2
param_min_samples_leaf	1	1	1	1	1
param_max_features	sqrt	sqrt	sqrt	auto	auto
param_max_depth	None	50	20	60	20
param_bootstrap	False	False	False	False	False
mean_test_score	0,905128	0,903776	0,903758	0,903721	0,902407
std_test_score	0,037095	0,041478	0,039253	0,041987	0,037990

Tablo 13: XGB için Hiper Parametre Seçenekleri

Parametre	**Değerler**	**Seçenek**
learning_rate	0,05; 0,10; 0,15; 0,20; 0,25; 0,30	6
max_depth	3; 4; 6; 8; 10	5
min_child_weight	1; 3; 5	3
gamma	0,0; 0,1; 0,2; 0,3; 0,4	5
colsample_bytree	0,3; 0,4; 0,5; 1	4

Tablo 14: XGB için Hiper Parametre Optimizasyonu Sonuçları (En İyi 5)

rank_test_score	**1**	**2**	**3**	**4**	**5**
mean_fit_time	0,108376	0,071300	0,107259	0,128274	0,148616
std_fit_time	0,004404	0,002040	0,007742	0,001565	0,007425
mean_score_time	0,001771	0,001589	0,001980	0,001684	0,001808
std_score_time	0,000303	0,000051	0,000338	0,000032	0,000115
param_min_child_weight	3	3	1	5	3
param_max_depth	4	4	6	10	6
param_learning_rate	0,3	0,2	0,2	0,2	0,15
param_gamma	0	0	0,1	0,2	0,3
param_colsample_bytree	1	0,5	0,5	1	1
mean_test_score	0,903739	0,903721	0,903702	0,902369	0,902351
std_test_score	0,026472	0,035904	0,016292	0,025429	0,022280

Veri madenciliği analizinde performans ölçütleri tek başına başarı kriteri olarak yeterli görülmeyebilir. Örneğin, kimi zaman birbirine çok yakın doğruluk değerlerine sahip metotların işlem süreleri de önem taşıyabilir. Bu durum uygulamadaki ihtiyaca göre değişmektedir. Hiper parametre optimizasyonu sonuç tabloları incelendiğinde KNN metodunun diğer iki metoda göre daha iyi işlem sürelerine sahip olduğu görülmektedir.

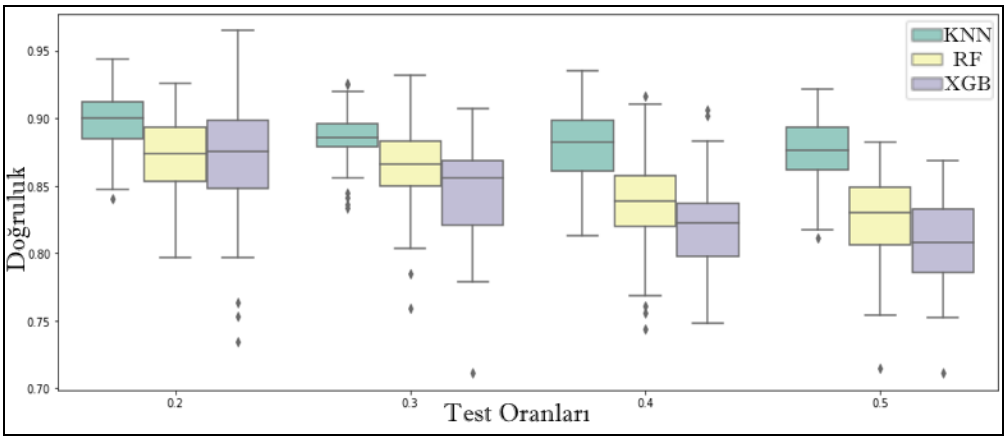
Bu aşamada, seçilen üç sınıflandırma metodu ile elde edilebilecek en iyi sınıflandırma oranları ve bu oranlara ulaşabilmek için kullanılabilecek en uygun hiper parametre değerleri belirlenmiştir. Sonraki aşamada kurulan modellerde buradaki hiper parametre değerleri kullanılmaktadır.

4.3. İkinci Karşılaştırma

Sınıflandırma analizinin ikinci karşılaştırma aşamasında, rasgele alt örnekleme yöntemi kullanılarak, modeller farklı eğitim-test oranlarıyla kıyaslanmıştır (Tablo 15). Seçilen eğitim-test yüzde oranları; 80-20, 70-30, 60-40 ve 50-50'dir. Rassal veri seti seçimi her model için 10 kez tekrarlanmış, hesaplanan değerlendirme ölçütleri "ortalama±standart sapma" formatında gösterilmiştir. Ek olarak, her bir model için 10 kez yinelemeli doğruluk oranlarıyla kutu grafiği oluşturulmuştur (Şekil 6).

Tablo 15: Farklı Eğitim-Test Oranlarıyla Modellerin Karşılaştırılması

Eğitim-Test	Metot	Doğruluk	TP-Oranı	TN-Oranı	F-Ölçütü
%80-%20	KNN	90,41±1,59	89,21±2,90	91,50±2,41	89,54±1,68
	RF	88,18±2,30	85,68±4,16	90,23±2,63	86,88±3,06
	XGB	88,38±3,20	83,09±6,97	92,86±3,05	86,67±4,15
%70-%30	KNN	89,23±0,98	86,74±2,79	91,45±2,30	88,29±1,05
	RF	87,39±1,99	83,56±4,09	90,77±3,33	86,09±2,26
	XGB	85,50±2,30	81,13±4,56	89,31±1,99	83,91±2,86
%60-%40	KNN	88,64±1,48	85,20±2,39	91,64±2,50	87,46±1,81
	RF	85,02±2,29	79,93±4,39	89,43±3,33	83,18±2,99
	XGB	83,66±1,32	78,02±1,72	88,59±3,10	81,65±1,35
%50-%50	KNN	88,27±1,83	85,73±2,90	90,38±2,62	87,05±2,08
	RF	83,85±1,64	78,76±4,05	88,21±2,34	81,74±2,27
	XGB	82,30±1,92	77,39±2,99	86,49±2,36	80,09±2,40



Şekil 6: Farklı Eğitim-Test Oranlarının Başarı Grafiği

Tablo ve kutu grafiği incelendiğinde, RF ve XGB metotlarının başarısının, eğitim verisi oranlarının düşüşüyle bir miktar azaldığı, bununla birlikte KNN metodunun eğitim verisi oranındaki düşüşe rağmen doğru sınıflandırma başarısının yüksek seyrettiği görülmektedir. Kutu grafiği, her model için yapılan 10 yinelemedeki aykırı değerleri nokta şeklinde göstermektedir. Buna göre KNN, TN-Oranı, TP-Oranı ve F-Ölçütlerinde de görüldüğü üzere diğer metotlara göre daha yüksek doğruluk ve daha düşük standart sapmaya sahip olduğu görülmüştür.

Tüm metotlar için en başarılı sonuçların %80-%20 eğitim-test oranıyla yapılan hesaplamalarda gerçekleştiği belirlenmiştir. Eğitim veri kümesindeki birim sayısının yüksek olması sınıflandırmanın daha başarılı olmasını sağlamıştır. Test veri kümesinin daha çok birimden oluşması istendiğinde yani test oranının daha yüksek olduğu durumda eğitim kümesinden taviz verilmesi, daha az birimle modelin geliştirilmesi gerekir. Çalışmada eğitim oranı kademeli olarak düşürüldüğünde, başarı oranları da kademeli olarak düşmüştür. Bu durumda en iyi sınıflandırma başarısı için yapılan sınıflandırma analizinde %80-%20 oranının tercih edilmesinin diğerlerine göre daha uygun olacağı söylenebilir.

Sonuç olarak, finansal tablo hilelerine yönelik veri madenciliğinin sınıflandırma analizinin başarıyla uygulanabileceği görülmüştür. Sınıflandırma analizinde sadece bir metoda bağlı kalınmamalı, birden çok yöntemin karşılaştırması yapılmalıdır. Veri madenciliği geniş bir uygulama alanıdır. Bu alanda yapılan çalışmalarda konuya göre her adımda farklı işlemlerin ve yöntemlerin uygulanması gerekebilir. Bu çalışmada, önişleme aşamasından model doğrulamasına kadar farklı yöntemler bulunmaktadır, ancak ihtiyaca göre gerekli adımlar ve metotlar değişkenlik gösterebilir.

5. SONUÇ

Finansal tablo hilelerinin tespiti şirketler, paydaşlar ve ülke ekonomileri için büyük önem arz etmektedir. Klasik denetim yöntemleri finansal tablo hilelerinin tespitinde yetersiz kalabilmekte, yeni ve daha etkin yöntemlere ihtiyaç duyulabilmektedir. Birçok alanda başarıyla uygulanan veri madenciliği bu alanda da umut vadetmektedir. Bu çalışmada finansal tablo hile riski taşıyan şirketlerin belirlenmesine yönelik veri madenciliği sınıflandırma analizi yapılmıştır. Veri kaynağı olarak Borsa İstanbul'da işlem gören şirketler tarafından 2014-2018 yılları arasında yayınlanmış finansal tablolar kullanılmış, girdi olarak her bir şirketin bir dönemine ait 16 finansal oran hesaplanıp veri setine eklenmiştir. Hangi finansal değişkenlerin analize dahil edileceğini belirlemek için Özeyinelemeli Özellik Eliminasyonu algoritması ile özellik seçimi yapılmış, 11 finansal değişken girdi değişkeni olarak belirlenmiştir.

Birinci karşılaştırma, seçilen yedi farklı sınıflandırma metodunun varsayılan değerleri kullanılarak yapılmış, sonraki aşama için en yüksek sınıflandırma başarısına sahip K-Nearest Neighbour (KNN), Random Forest (RF) ve XGBoost (XGB) metotları seçilmiştir. RandomizedSearch yöntemiyle yapılan hiper parametre optimizasyonu sonucunda KNN ile yüzde 91,73, RF ile yüzde 90,51 ve XGB ile yüzde 90,37 doğru sınıflandırma başarısına ulaşılmıştır. İkinci karşılaştırmada modeller, belirlenen hiper parametre değerleri ile farklı eğitim-test oranlarında kıyaslanmıştır. Buna göre KNN metodunun tüm eğitim-test oranlarında en iyi skorlara sahip olduğu görülmüştür. Ayrıca KNN, eğitim verisi oranının düşürülmesinden en az etkilenen sınıflandırma metodu olmuştur.

Çalışmanın iki temel kısıtı bulunmaktadır. Birinci kısıt verilerin elde edilmesi ve doğruluğu ile ilgilidir. Veriler şirketlerin kendi paylaştıkları veriler olduğu için aykırı değer konusunda her değerin ayrıntılı araştırılması yolu izlenmemiştir. Diğer kısıt, şirketlerin sınıflandırılması kısıtıdır. Finansal tablo hilesi yapmadığı belirtilen şirketler, hakkında herhangi bir hile durumu tespit edilemediği için bu şekilde sınıflandırılmıştır.

Sonuç olarak, finansal tablo hilelerinin verdiği zararlardan korunabilmek için öncelikle tespit edilmeleri gerekir. Bunun için de güncel ve etkin çözümlerden yararlanmak gerekir. Veri madenciliği ile yapılan bu çalışma ile yüksek sınıflandırma başarısı elde edilmiş, literatürde az bulunan yeni yöntemlere yer verilmiştir. Akademik araştırmalara katkısı açısından, bu çalışma finansal tablo hilesi riski taşıyan şirketlerin tespit edilmesine yönelik geniş kapsamlı bir veri madenciliği sınıflandırma araştırması olması dolayısıyla özgün bir çalışmadır ve sonraki çalışmalar için aydınlatıcı olması amaçlanmıştır.

KAYNAKÇA

Abdioğlu, H. (2007). Hilelerin Önlemesi ve Ortaya Çıkarılmasına Yönelik Proaktif Yaklaşımlar. *Muhasebe ve Denetime Bakış*, 22, 119-137.

ACFE. (2020). *Report to The Nations: 2020 Global Study on Occupational Fraud and Abuse*. Association of Certified Fraud Examiners: https://www.acfe.com/report-to-the-nations/2020/

Ata, H. A. & Seyrek, I. H. (2009). The Use of Data Mining Techniques in Detecting Fraudulent Financial Statements: An Application on Manufacturing Firms. *Suleyman Demirel University Journal of Faculty of Economics & Administrative Sciences, 14*(2), 157-170.

Bai, B., Yen, J. & Yang, X. (2008). False Financial Statements: Characteristics of China's Listed Companies and CART Detecting Approach. *International Journal of Information Technology & Decision Making, 7*(2), 339-359. doi:Doi 10.1142/S0219622008002958

Brownlee, J. (2017). *What is the Difference Between a Parameter and a Hyperparameter?* https://machinelearningmastery.com/difference-between-a-parameter-and-a-hyperparameter/

Cotton, D. L. (2002). *Fixing CPA Ethics Can Be an Inside Job.* https://www.washingtonpost.com/archive/opinions/2002/10/20/fixing-cpa-ethics-can-be-an-inside-job/b7441564-e0a6-431b-9280-8c27c6267ebc/

Deng, Q. & Mei, G. (2009). *Combining Self-Organizing Map and K-means Clustering for Detecting Fraudulent Financial Statements.* 2009 IEEE International Conference on Granular Computing.

Erol, M. (2008). İşletmelerde Yaşanan Yolsuzluklara (Hata Ve Hileler) Karşı Denetimden Beklentiler. *Süleyman Demirel Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi*, 13(1), 229-237.

Ertikin, K. (2017). Hile Denetimi: Kırmızı Bayrakların Tespiti İçin Kullanılan Proaktif Yaklaşımlar. *Muhasebe ve Finansman Dergisi*, 75, 71-93.

Gaganis, C. (2009). Classification Techniques for the Identification of Falsified Financial Statements: A Comparative Analysis. *Intelligent Systems in Accounting, Finance & Management: International Journal, 16*(3), 207-229.

Granitto, P. M., Furlanello, C., Biasioli, F. & Gasperi, F. (2006). Recursive Feature Elimination with Random Forest For ptr-ms Analysis of Agroindustrial Products. *Chemometrics and Intelligent Laboratory Systems, 83*(2), 83-90. doi:https://doi.org/10.1016/j.chemolab.2006.01.007

Hajek, P. & Henriques, R. (2017). Mining Corporate Annual Reports for Intelligent Detection of Financial Statement Fraud – A Comparative Study

of Machine Learning Methods. *Knowledge-Based Systems, 128*, 139-152. doi:10.1016/j.knosys.2017.05.001

Han, J., Kamber, M. & Pei, J. (2011). *Data Mining: Concepts and Techniques*. Elsevier.

Hatunoğlu, Z., Koca, N. & Kıllı M. (2012). İç Kontrolün Muhasebe Sistemindeki Hata Ve Hilelerin Önlenmesindeki Rolü Üzerine Bir Alan Çalışması. *Mustafa Kemal Üniversitesi Sosyal Bilimler Enstitüsü Dergisi*, 9(20), 169-189.

Hoogs, B., Kiehl, T., Lacomb, C. & Senturk, D. (2007). A Genetic Algorithm Approach to Detecting Temporal Patterns Indicative of Financial Statement Fraud. *Intelligent Systems in Accounting, Finance & Management: International Journal, 15*(1-2), 41-56.

İşgüden Kılıç, B., Anadolu, Z. (2018). Dijital Çağın Yarattığı Muhasebe Uygulamalarının Muhasebe Hilelerinin Önlenmesine Etkisi. *Muhasebe ve Vergi Uygulamaları Dergisi*, Özel Sayı: 55-97.

Jan, C.-l. (2018). An Effective Financial Statements Fraud Detection Model for the Sustainable Development of Financial Markets: Evidence from Taiwan. *Sustainability, 10*(2), 513.

Kirkos, E., Spathis, C. & Manolopoulos, Y. (2007). Data Mining Techniques for the Detection of Fraudulent Financial Statements. *Expert Systems with Applications, 32*(4), 995-1003. doi:https://doi.org/10.1016/j.eswa.2006.02.016

Kotsiantis, S., Koumanakos, E., Tzelepis, D. & Tampakas, V. (2006). Forecasting Fraudulent Financial Statements using Data Mining. *International Journal of Computational Intelligence, 3*(2), 104-110.

Lenard, M. J., Watkins, A. L. & Alam, P. (2007). Effective Use of Integrated Decision Making: An Advanced Technology Model for Evaluating Fraud in Service-Based Computer and Technology Firms. *Journal of Emerging Technologies in Accounting, 4*(1), 123-137.

Lin, C.-C., Chiu, A.-A., Huang, S. Y. & Yen, D. C. (2015). Detecting The Financial Statement Fraud: The Analysis of the Differences Between Data Mining Techniques And Experts' Judgments. *Knowledge-Based Systems, 89*, 459-470.

Lipton, Z. C., Elkan, C. & Naryanaswamy, B. (2014). *Optimal Thresholding of Classifiers to Maximize F1 Measure*, Berlin, Heidelberg.

Perols, J. (2011). Financial Statement Fraud Detection: An Analysis of Statistical and Machine Learning Algorithms. *Auditing: A Journal of Practice & Theory, 30*(2), 19-50. doi:10.2308/ajpt-50009

Ravisankar, P., Ravi, V., Rao, G. R. & Bose, I. (2011). Detection of

MUVU/ 2021-2

Financial Statement Fraud and Feature Selection Using Data Mining Techniques. *Decision Support Systems, 50*(2), 491-500.

Rezaee, Z. (2005). Causes, Consequences, and Deterence of Financial Statement Fraud. *Critical Perspectives on Accounting*, *16*(3), 277-298.

Rizki, A. A., Surjandari, I. & Wayasti, R. A. (2017). *Data Mining Application to Detect Financial Fraud in Indonesia's Public Companies.* 3rd International Conference on Science in Information Technology (ICSITech).

Scikit-Learn. (2020a). *sklearn.feature_selection.RFE.* https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.RFE.html

Scikit-Learn. (2020b). *sklearn.preprocessing.RobustScaler*. https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.RobustScaler.html

Sjardin, B., Massaron, L. & Boschetti, A. (2016). *Large Scale Machine Learning with Python*. Packt Publishing, ISBN: 9781785887215.

Terzi, S. & Şen, İ. K. (2012). Finansal Tablo Hilelerinin Veri Madenciliği Yardımıyla Tespit Edilmesi: Üretim Sektöründe Bir Araştırma. *Determination of Fraudulent Financial Statements Using Data Mining: A Research in Manufacturing Sector, 5*(2), 25-40.

Terzi, S. & Şen, İ. K. (2015). Adli Muhasebede Hilelerin Tespitinde Yapay Sinir Ağı Modelinin Kullanımı. *International Journal of Economic and Administrative Studies, 7*(14), 477-490.

Thara, D. K., PremaSudha, B. G. & Xiong, F. (2019). Auto-Detection of Epileptic Seizure Events Using Deep Neural Network with Different Feature Scaling Techniques. *Pattern Recognition Letters, 128*, 544-550.

Uğurlu, M. & Sevim, Ş. (2015).Finansal Tablolardaki Hile Riskinin Tahmin Edilmesinde Karma Modellerin Nispi Başarısı Üzerine Karşılaştırmalı Bir Analiz. *Gaziantep University Journal of Social Sciences*, 14(1), 65-88.

Ulucan Özkul.F & Pertekin, P. (2009). Muhasebe Yolsuzluklarının Tespitinde Adli Muhasebecinin Rolü ve Veri Madenciliği Tekniklerinin Kullanılması. *MÖDAV*, 2009/4, 57-88.

Witten, I. H., Frank, E. & Hall, M. A. (2011). *Data Mining: Practical Machine Learning Tools and Techniques* (3rd ed.). Burlington, MA: Morgan Kaufmann.

Yao, J., Zhang, J. & Wang, L. (2018). *A Financial Statement Fraud Detection Model Based on Hybrid Data Mining Methods.* International Conference on Artificial Intelligence and Big Data (ICAIBD).

MUVU/ 2021-2