Alanya Akademik Bakış Dergisi
Yıl: 2023, C:7, S:1, s.511-528

Alanya Academic Review Journal
Year: 2023, Vol:7, No:1, p.511-528

Karar Ağacı Destekli Hile Tespiti ve Bir Uygulama

(Araştırma Makalesi)

Decision Tree Supported Fraud Detection and an Application

Doi: 10.29023/alanyaakademik.1196078

Önder GÜR
Öğr. Gör. Dr., Kırklareli Üniversitesi, Babaeski Meslek Yüksekokulu, Finans, Bankacılık ve Sigortacılık Bölümü
ondergur@klu.edu.tr
Orcid No: 0000-0003-3249-4300

Bu makaleye atıfta bulunmak için: *Gür, Ö. (2023). Karar Ağacı Destekli Hile Tespiti ve Bir Uygulama. Alanya Akademik Bakış, 7(1), Sayfa No.511-528.*

Anahtar kelimeler:
Hile Tespiti, Hileli Ödemeler, Makine Öğrenmesi, Kara Ağacı

Makale Geliş Tarihi: 28.10.2022
Kabul Tarihi: 26.12.2022

ÖZET

Çalışmada, Sertifikalı Hile Denetçileri Birliği'nin (ACFE) hile ağacında yer alan ve işletmelerde sıklıkla karşılaşılan hileli ödemelerin verdiği zararı azaltmak için makine öğrenmesi yönteminin kullanıldığı bir uygulama ile hile tespit sürecine katkının sağlanması amaçlanmıştır. Bu amaçla, elde edilmek istenen çıktılar için Python'da bir uygulama sistemi tasarlanmıştır. Çalışmada, bir bankaya ait normal işlemler ile hileli işlemlerin yer aldığı yapay veri setinden yararlanılmıştır. Yöntem olarak kullanılmasına karar verilen Karar Ağacı tekniğiyle önce sınıf etiketleri bilinen bir veri setiyle ana model oluşturulmuş, sonra etiketsiz bir veri seti üzerinde modelin test edilmesi sağlanmıştır. Karar ağacı tekniğinin modeli, %97,1 doğruluk, %98,4 f1-skor, %98,9 kesinlik ve %98 duyarlılık değerlerini elde etmiştir. Çalışma, karar ağacı tekniğinin tahmin aşamasında ürettiği hatalı sınıf etiketlerinin azaltılması açısından iyileştirmeye açık olup, diğer tekniklerle karşılaştırılarak da geliştirilebilir.

Keywords:
Fraud Detection, Fraudulent Disbursements, Machine Learning, Decision Tree

ABSTRACT

In the study, it is aimed to contribute to the fraud detection process with an application in which machine learning method is used to reduce the damage caused by fraudulent disbursements, which is included in the fraud tree of the Association of Certified Fraud Examiners (ACFE). To achieve the desired outcomes, a Python application system is developed for this purpose. In the study, an artificial data set containing normal transactions and fraudulent transactions of a bank, was used. Using the Decision Tree technique, which was selected as the chosen method, the main model was developed using a data set with known class labels, and then the model was evaluated using unlabeled data. The model of the decision tree technique achieved 97,1% accuracy, 98,4% f1-score, 98,9% precision and 98% sensitivity. The study is open to improvement in terms of reducing the erroneous class labels produced by the decision tree technique during the estimation phase and can be improved by comparing it with other techniques.

1. GİRİŞ

Günümüzde hileli faaliyetler, yatırımcının güvenini azaltmakta, işletmeleri istikrarsızlaştırmakta ve insanların yaşam maliyetlerini etkilemektedir. Ekonomik olarak, finansal dolandırıcılık giderek daha ciddi bir sorun haline gelmektedir (Ngai vd., 2011:559). Bugün dünya ölçeğinde hile üzerine araştırma yapan Sertifikalı Hile Denetçileri Birliğinin (ACFE) 2022 yılında hazırlamış olduğu rapora göre; 133 ülkede 2.110 vakanın neden olduğu kayıp yaklaşık 3,6 milyar USD olmuştur (ACFE, 2022:4). Hile kaynaklı kayıpların kesin büyüklüğü tam olarak bilinmemekle beraber tahmini olarak hesaplandığında çoğu kuruluş gelirlerinin %0,5 ile %2'sini kaybetmektedir (Coderre, 2009:1).

Özellikle geleneksel denetim yaklaşımlarının faaliyetleri manuel olarak yürütüldüğünden karmaşık yapıda olan hileli faaliyetler karşısında verimsiz ve güvenilmez olduğu kabul edilmektedir (West ve Bhattacharya, 2015:47). Bu doğrultuda araştırma raporları incelendiğinde, pasif olarak yürütülen hile tespit çalışmalarının aktif yöntemlere kıyasla çoğunun daha uzun sürdüğü ve bu sebeple işletmelerin kayıplarının daha da büyüdüğü tespit edilmiştir (ACFE, 2022:22). Muhasebe departmanlarının birçoğu, genel yönergeleri ve normal denetim prosedürlerini kullanarak tahrif edilmiş finansal tabloları ve diğer hileli faaliyetleri tespit etmekte zorlanmaktadır (Gaganis, 2009:208). Bu nedenle denetçiler aynı anda birçok finansal veriyi tarayarak denetim görevlerini basitleştirecek yeni araçlara ve tekniklere ihtiyaç duymaktadır. Bu bağlamda teknoloji destekli araçlar kullanarak zamandan ve paradan tasarruf sağlayabilirler (Ata ve Seyrek, 2009:158).

Bu mücadelede gelişmiş sınıflandırma ve tahmin yeteneklerine sahip olan veri madenciliği tekniklerinin kullanılması hem iç hem de dış denetçilerin işlerini kolaylaştırabilir (Kirkos vd.,2009:1002). Veri madenciliği, son yıllarda finans dünyasında artan bir popülerlik kazanmıştır. Yapılan çalışmalara bakıldığında başarılı veri madenciliği uygulamaları dikkat çekmiş ve veri madenciliği kullanımının arttığını ortaya koymuştur. Profesyonel muhasebe kuruluşları da veri madenciliğini yeni yüzyıl için önemli bir teknoloji olarak belirlemiştir (Zhou ve Kapoor, 2011:574). Hilekârlar, hile yöntemlerini sürekli olarak geliştirmektedir. Bu nedenle hile tespit yöntemlerinin de o ölçüde gelişmesi gerekmektedir.

Geçmişte hileli faaliyet gerçekleştikten sonra yürütülen hile tespit çalışmalarının aksine günümüzde hileyi tespit ve önleme uygulamalarının neredeyse tamamı eşzamanlı olarak yürütülmektedir. İşletmeler, hile araştırma sistemleriyle hileye işaret eden kalıpları tespit etmek için yıllara dayanan işlem verilerini kullanarak hızlı bir şekilde model oluşturmaya ve eşzamanlı veriler üzerinde hileli faaliyetleri önlemeye yönelik şekilde tasarlamaktadır (Kudyba, 2014:206).

Hile tespitinde yaygın olarak kullanılan makine öğrenmesi algoritmaları, verilerdeki kalıpları hızla keşfetmekte ve insanların yürüttüğü faaliyetlere göre daha verimli sonuçlar elde etmektedir. Hileye ilişkin araştırma sürecinde elde edilen bilgiler ile hileli işlemleri tespit etmek ve önlemek daha kolay olmaktadır. Denetim süreçlerine dahil edilen yapay zekâ gibi teknolojik araçlar, veri toplamayı ve işlemeyi otomatikleştirerek zaman alıcı manuel faaliyetlerden kurtarmaktadır. Bu sayede hile riski ve hileden kaynaklanan önemli yanlışlıklar daha kolay belirlenebilmektedir (Hacıhasanoğlu vd., 2021:85-86).

Bu çalışmada, hile tespit sürecinde makine öğrenmesi algoritmalarından karar ağacını (DT, Decision Tree) kullanarak veri setlerindeki hileli olma ihtimali olan işlemleri tahmin etmeye yönelik bir uygulama yapılmıştır. Çalışmada elde edilmek istenen çıktılar için çalışmaya özel Python'da bir uygulama sistemi tasarlanmıştır. Uygulama sistemiyle yapılan hile tespiti,

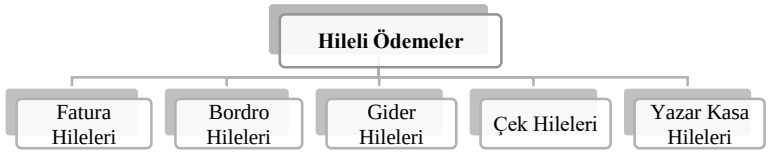
kamuya açık normal ödemeler ile hileli ödemelerin yer aldığı etiketli bir veri seti üzerinde gerçekleştirilmiştir. Uygulama sisteminde teknoloji temelli teknikleri kullanılarak özellikle iç denetimcilerin periyodik olarak hileyi ortaya çıkarmak ya da önlemek için yaptığı taramalarda yaşadıkları problemlerin çözülmesine ve hileli ödemelere yönelik fatura, çek, gider makbuzu, bordro gibi evrakları bazı manuel kurallara bağlı veya rastlantısal olarak belirlemek yerine inceleyeceği evrakların makine öğrenimi sürecinden geçmiş anlamlı çıktılar içinden seçimler yapmasına yardımcı olmayı hedeflemektedir. Bununla birlikte uygulama sisteminin işlemleri hızlı gerçekleştirmesi sebebiyle hilenin daha hızlı ve kolay bir şekilde ortaya çıkartılmasına olanak vererek, finansal açıdan faydalı sonuçların elde edileceği düşünülmektedir. Çalışmaya hileli ödemeler hakkında bilgi verilerek başlanmış, literatür incelemesiyle devam edilmiştir sonra araştırmanın yöntemi anlatılmış, uygulama yapılmış ve elde edilen bulgular paylaşılmıştır.

2. HİLELİ ÖDEMELER

Hile kavramı, Hileli Finansal Raporlama Ulusal Komisyonun (Treadway Commission) Ekim 1987'de raporunu yayınladığından beri denetim mesleği mensupları için giderek daha önemli bir konu haline gelmiştir. Komisyon, hilekarlığın giderek arttığını ve önlenmesi için tedbirler alınması yönünde kamuoyuna çağrıda bulunmuştur (Coderre, 2009:1-2). Özellikle 2000'li yılların başlarına gelindiğinde alınan tedbirlerin yetersiz olmasından dolayı büyük dolandırıcılık olayları (Enron gibi) meydana gelmiştir. Sonrasında muhasebe denetim işletmelerinin sorumluluğu arttırılmış ve hükümetler Sarbanes-Oxley Yasası gibi doğru finansal raporlamayı sağlamak için yeni kurallar ve düzenlemeler geliştirmiştir (Aksoy, 2021:33).

Hile, kişinin diğer kişilere karşı avantaj elde etmek amacıyla insani marifetlerini kullanarak tasarladığı aldatmacalı yöntem ve araçlara başvurmasına verilen genel bir terimdir (addır). Çünkü hile, yaratıcılık, kurnazlık ve haksızlık gibi tüm yolları kapsar, hiçbir şekilde onu tanımlayabilecek kesin, değişmez kural ve önerme ortaya konulamamaktadır. Onu anlatabilecek tek sınır insanın hile yaratıcılığıdır (Albrecht vd, 2011:6). Başka bir çalışma işletmedeki işlemlere ait açıklamalarda önemli ölçüde hataların bulunması, işletme kurallarının bilerek ihlal edilmesi veya işletme sahibine bir yanlışlığın gerçek olarak kabul ettirilmesi gibi nedenlerden kaynaklanarak işletmeleri finansal kayıplara uğratan bir yapı olarak tanımlamıştır (Chen, 2016:3).

Literatürde hile üzerine yapılan çalışmalara bakıldığında uzmanlar, ACFE'nin belirlediği hilekârlık kategorisi konusunda birleşmiştir. İşletmelerde karşılaşılan hile türlerini; 1) Varlıkların kötüye kullanımı, 2) Yolsuzluk ve 3) Hileli finansal raporlama şeklinde sınıflandırmıştır (Craja vd., 2020:139). Varlıkların kötüye kullanılması nakit hilesi yoluyla doğrudan nakde el koyma, hasılatı eksik gösterme ve hileli ödemeler gibi hile faaliyetleri içermektedir. Bu hile türünde çalışanlar, işletmeye mal satan satıcılarla ya da rekabet ettiği işletmelerle iş birliğine girerek işletmeyi zarara uğratabilmektedirler. Ayrıca işletme varlıklılarının hilekârın kendi çıkarları için kullanması, yanlış raporlamalar ve kendi adına yaptığı harcamalar da bu alanda değerlendirilen bir hileli faaliyettir (Golden vd, 2011:5).

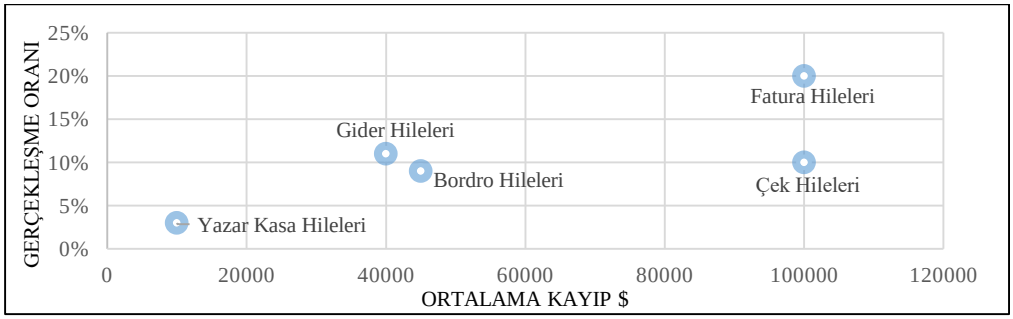


Şekil 1. Hileli Ödeme Türleri

Kaynak: *Nejat Bozkurt, İşletmelerin Kara Deliği Hile, 3.Baskı, İstanbul: Alfa Yayınları, 2009, s.212.*

Varlıkları kötüye kullanma hilelerinin en yaygın grubu hileli ödemelerdir. Hileli ödemeler, bazı işletme hesaplarında işlemlerin normal gibi görünmesine rağmen aslında hileli bir şekilde fon dağıtımının yapıldığı planlı işlemlerdir. Fonların elde edilişi bazen sahte bir çek, bazen sahte bir faturayla ya da sahte bir zaman çizelgesiyle yapılmaktadır. Hileli ödemelerin 5 ana kategorisi vardır: a) Fatura hilesi, b) Bordro hilesi, c) Gider hilesi d) Çek hileleri e) Yazar kasa hileleri (Singleton ve Singleton, 2010:86).

Hileli ödemelerde kullanılan yöntemlerin içinde en çok kullanılanı ve zarar vereni fatura hileleridir. Bu hile türünde, hileli olarak satın alma esası yattığından işlemin gerçekleştirilmesinde gerçeği yansıtmayan satın alma belgelerinin oluşturulması gerekmektedir. Aynı zamanda faturalar, satın alma istekleri, satın alma sipariş formları ve alış raporları örnek olarak verilebilir. Bu süreç, işletmeye bir şekilde ödeme yaptırıldıktan sonra bu ödemenin hileyi yapan kişinin cebine girmesiyle sonuçlanmaktadır (Bozkurt, 2009:215-216). Fatura hileleri, varlıkları kötüye kullanmanın en yaygın biçimi olması ve aynı zamanda en yüksek kayba neden olması nedeniyle işletmelere önemli bir risk oluşturmaktadır. Bu hile türü, işletmeleri ortalama 100.000 USD'lik bir zarara uğratmaktadır (ACFE, 2022:12).



Şekil 2. Hileli Ödeme Türlerinin Gerçekleşme Oranı ve Ortalama Kayıp Tutarı

Kaynak: *Association of Certified Fraud Examiners (ACFE), "2022 Global Study on Occupational Fraud and Abuse", Report to The Nation, USA: 2022, s.12.*

ACFE (2022:12-14) işletmelerde gerçekleşen 2.110 vakanın 416 tanesi fatura hilesi, 232 tanesi gider hilesi, 208 tanesi çek hilesi, 198 tanesi bordro hilesi ve 58 tanesi yazar kasa hilesi şeklinde meydana geldiğini tespit etmiştir. Böylelikle toplam vakaların yarısından fazlasının hileli ödeme türlerinden kaynaklandığını ortaya koymuştur. Ayrıca işletmelerde fatura hileleri, gider hileleri, çek hileleri ve bordro hileleri ortalama 18 ay, yazar kasa hilelerinin ise ortalama 12 ay boyunca devam ettiğini belirtmiştir. Bu tür hileli faaliyetler, genellikle gizli anlaşmalara müsait ortamın oluştuğunun veya kontrollerin yetersiz kaldığının göstergesidir. Bu nedenle sürekli izleme ve denetim gibi faaliyetler bu hile türleri için kontrolün anahtarıdır (Singleton ve Singleton, 2010:86). Bu yüzden, sadece hükümetler ve kuruluşlar tarafından yapılan yasal

düzenlemeler hilenin ekonomik olarak verdiği zararın azaltılmasında yeterli olmamaktadır. Hilenin tespit edilebilmesi için hile faktörlerinin sınırsız ve çok sayıda olduğu bilinerek hareket edilmelidir. Hileye karşı verilen mücadele yeni yöntemlerin geliştirilmesine olan ihtiyacı belirginleştirmiştir (Golden vd, 2011:7) Hile denetçileri, teknolojinin gelişmesiyle kullanıcı odaklı araçlar kullanarak hileye ait işaretleri aramak için büyük veri kümelerinin tamamına erişilebilir hale gelmiştir (Albrecht vd, 2011:168). Artık hilenin göstergelerini tespit etmede teknoloji kaynaklı sistemlerin kullanımı artarak devam etmektedir. Bu tür sistemler, yatırımcıların yatırım kararlarını kolaylaştırmasına, denetim işletmelerinin denetim doğruluğunu arttırmasına, işlemleri hızlandırmasına ve kamu otoritesinin soruşturmalarını daha etkin yapmasına yardım etmektedir (Craja vd., 2020:140).

3. LİTERATÜR ARAŞTIRMASI

Son araştırmalar gösteriyor ki denetim faaliyetleri teknoloji destekli araçlarla yürütülmeye başlamıştır. Birçok alanda olduğu gibi veri madenciliği araçlarının kullanımı giderek artmaktadır. Veri madenciliğinin makine öğrenmesi yöntemlerinin hile tespitinde kullanımı ön plana çıkmaktadır. Finansal dolandırıcılığa karşı Lojistik Regresyonun (LR, Logistic Regresion), Yapay Sinir Ağlarının (ANN, Artificial Neural Network), Bayes İnanç Ağının (BBN, Bayesian Belief Network) ve Karar Ağaçlarının (DT) sıkça kullanıldığı görülmektedir. Bu tekniklerin tümünün, hileli verilerin tespiti ve sınıflandırılmasıyla ilgili sorunlara kapsamlı çözümler sunduğu anlaşılmıştır (Ngai vd,2011:559).

West ve Bhattacharya (2015:55-60) çalışmalarında, 2004-2014 yılları arasındaki hile tespitine yönelik hem istatistiksel hem de matematiksel yaklaşımları kullanan araştırmaları incelemişler ve BBN, LR, ANN, DT, Destek Vektör Makinesi (SVM, Support Vector Machine), Genetik algoritmaları, metin madenciliği gibi her bir tekniğin çeşitli finansal dolandırıcılık biçimlerini tespit etmede yeterli düzeyde olduğunu söylemişlerdir. Özellikle yapay sinir ağları ve destek vektör makineleri, hilenin kompleks yapısını araştırmaya olan uyum sağlama yeteneği hilekarların geliştirdiği taktikler üzerinde oldukça etkili olduğunu belirtmiştir.

Gaganis (2009:207) makalesinde, denetçilerin hileli finansal tabloları tespit etmesine yardımcı olmak amacıyla finansal ve finansal olmayan verileri kullanarak SVM, ANN, Olasılıksal Sinir Ağları, En Yakın Komşular (k-NN, k-Nearest Neighbors), Logit Analizi, Diskriminant analiz teknikleriyle sınıflandırma modeli geliştirmeye çalışmıştır.

Zhou ve Kapoor (2011:570) çalışmalarında hilenin tespitinde veri madenciliği araçlarından regresyon, DT, ANN ve BBN tekniklerini kullanmışlardır. Geçmiş finansal verilere dayanan geleneksel finansal dolandırıcılık tespit tekniklerinden farklı olarak, olası dolandırıcıların önüne geçmek amacıyla finansal tablo dolandırıcılığına ait planlarını keşfetmeye yönelik bilgisayar destekli otomatik hile tespit mekanizmalarının oluşturulabileceğini belirtmişlerdir.

Liou (2008:650) çalışmasında, hileli ve hatalı raporlamanın tespiti için 52 finansal değişkeni LR, ANN ve DT tekniklerinde kullanarak tespit/tahmin modelleri oluşturmuştur. Kurduğu modeller, hem hileli finansal raporlamayı tespit etmede hem de iş başarısızlıklarını tahmin etmede başarılı olmuştur. Genel doğruluk açısından lojistik regresyon, hileli finansal raporlamayı tespit etmede diğer iki teknikten daha iyi performans göstermiştir.

Tatar ve Kıymık (2021:1700) çalışmalarında 2015-2019 yılları arasında Borsa İstanbul (BİST) tekstil, giyim eşyası ve deri sektöründe işlem gören işletmelerin finansal oranları aracılığıyla veri madenciliğine dayalı yöntemleri kullanarak hile riskini tespit etmeyi amaçlamışlardır. Modellerin, hile riski olan ile hile riski olmayan finansal tabloları doğru sınıflandırmada

%70'in üzerinde başarı sağladığı ve en başarılı yöntemlerin J48 ve Derin Öğrenme (DL, Deep Learning) yöntemleri ile kurulan modellerin olduğu sonucuna ulaşmıştır.

2014 yılındaki makalelerinde Ceyhan ve Kırlıoğlu (2014:13), denetçilerin karar almasına yardımcı olmak amacıyla BİST'e kayıtlı 40 işletmenin finansal tablo verilerini kullanarak finansal açıdan başarılı ya da başarısız şeklinde sınıflandırarak ön analitik inceleme yapmayı hedeflemişlerdir. k-NN algoritması ve 10 kat çapraz doğrulama tekniğiyle %95 gibi yüksek bir oranda doğru finansal sınıflama tahmini elde ettiğini belirtmişlerdir.

Literatürdeki diğer çalışmalarda kullanılan tekniklere, ölçütlere ve temel bulgulara ilişkin özet bilgi aşağıda Tablo 1'de gösterilmiştir.

Tablo 1. Literatür Özet Tablosu

Yazarlar	**Çalışma Konusu**	**Kullandığı Yöntemler**	**Kullanılan Ölçütler**	**Temel Bulgular**
Kirkos vd. (2007)	Hileli Finansal Raporlama	DT, ANN ve BBN	Tip I Hata ve Tip II Hata (FN, FP)	Bayesian İnanç Ağı modeli, diğer modellere göre daha iyi sınıflandırma performansı göstermiştir.
Lægreid (2007)	Sigortacılık	LR, DT, ANN ve BBN	Hosmer–Lemeshow Testi	Lojistik Regresyon modelini kullanmış iyi sınıflandırma performansı göstermiştir. Ayrıca diğer modellerin kullanılabileceğini de belirtmiştir.
Ata ve Seyrek (2009)	Hileli Finansal Raporlama	DT, ANN	T-testi	Yapay Sinir Ağının modeli, karar ağacının modeline göre daha iyi sınıflandırma performansı göstermiştir.
Chen (2016)	Hileli Finansal Raporlama	DT, BBN, SVM ve ANN	T-testi ve Wilcoxon sıra toplamı testi	Karar Ağacı algoritması olan CHAID–CART ikilisinin modeli, diğer ikili modellere göre daha iyi sınıflandırma performansı göstermiştir.
Dutta vd. (2017)	Hileli Finansal Rapolama	DT, ANN, Naïve Bayes (NB), SVM ve BBN	Doğruluk, Duyarlılık, Özgüllük, Kesinlik	Yapay Sinir Ağı modeli, diğer modellere göre daha iyi sınıflandırma performansı göstermiştir.

Jan (2018)	Hileli Finansal Raporlama	DT, ANN ve SVM	Doğruluk ve Ortalama Tip I – II Hata Oranı (FN, FP)	Yapay Sinir Ağı ile Karar Ağacı CART algoritmasının oluşturduğu model, diğer ikili modellere göre daha iyi sınıflandırma performansı göstermiştir.
Lakshmi ve Kavila (2018)	Kredi Kartı Dolandırıcılığı	LR, DT ve Rastgele Orman (RF)	Doğruluk, Duyarlılık, Özgüllük, Hata oranı	Rastgele Ormanın Modeli, Karar Ağacı Modelleri ve Lojistik Regresyon modeline göre daha iyi sınıflandırma performansı göstermiştir.
Kurien ve Chikkamannur, (2019)	Kredi Kartı Dolandırıcılığı	Benford Yasası, DL, Otomatik Kodlayıcı, LR ve RF	Kesinlik, Duyarlılık, F1-Skoru	Benford Yasası ve Derin Öğrenme Otomatik Kodlayıcılarının oluşturduğu model, diğer modellere göre daha iyi sınıflandırma performansı göstermiştir.
Craja vd. (2020)	Hileli Finansal Raporlama	ANN, LR, SVM, DL ve RF	Doğruluk, Duyarlılık, Özgüllük, F1-Skoru, F2-Skoru	Derin Öğrenme modeli, diğer modellere göre daha iyi sınıflandırma performansı göstermiştir.
Aksoy (2021)	Hileli Finansal Raporlama	ANN, DT, SVM ve LR	Doğruluk, Duyarlılık, Kesinlik, F-Skor, Kappa	Yapay Sinir Ağı ve Karar Ağacı Modelleri, diğer modellere göre daha iyi sınıflandırma performansı göstermiştir.

Kaynak: *Tablo1, yazar tarafından yapılan alan araştırması sonucu derlenerek hazırlanmıştır.*

4. YÖNTEM

Daha önce belirtildiği üzere, ekonomik açıdan zarar veren hilelerin tespitinde makine öğrenmesi yöntemlerinin kullanımı her geçen gün artmaktadır. Hilenin tespitinde karar ağacı, yapay sinir ağları, derin öğrenme, rastgele orman, lojistik regresyon, bayes inanç ağları ve destek vektör makinesi sınıflandırma tekniklerinin alanda sıklıkla kullanıldığı görülmüştür.

Sınıflandırma teknikleri, farklı sınıflardaki nesneleri ayırt etmek ve sınıfları bilinmeyen nesnelerin etiketlerini tahmin etmek için kullanılmaktadır (Ngai vd., 2011:562) Hilenin tespit edilmesinde de hile ile hilesiz sınıf tahminleri sınıflandırıcı tekniklerle yapılmaktadır. Tahminler, problemlerin çözümünde bilinen sınıflandırılmış verilerin değişken öznitelik

değerlerine dayandırılarak yapılmaktadır. Elde edilen tahmin sonuçları daha sonra nihai sonuçlara ulaşmak için sınıflandırılmamış verilere uygulanmaktadır (Jan, 2018:513).

Bu çalışmada, hile tespiti yapmaya çalışan kişilere katkı sağlayabilmek amacıyla makine öğrenmesi yöntemleri kullanılarak bir uygulama yapılmıştır. Ayrıca bu çalışmanın diğer çalışmalardan farkı:

1) Veri setlerinde hile araştırmasını manuel yapmak yerine, oluşturulacak teknoloji tabanlı bir uygulama sistemiyle hile tahminlemesinin yapılabileceğini göstermeye çalışmasıdır,
2) Denetçilerin, denetim esnasında seçeceği evrakları rastgele seçmek yerine makine öğrenimi sürecinden geçmiş çıktılar arasından anlamlı seçimler yapabilmesine yardımcı olmasıdır,
3) Hile tespitinde zaman kaybını önlemeye yardımcı olmasıdır,
4) Zaman kaybına bağlı olarak ortaya çıkan büyük kayıpları engellemeye çalışmasıdır.

4.1. Kullanılan Teknik

Karar ağaçları, makine öğreniminin gelişimine önemli katkı sağlayan tekniklerden biri olmuştur. Fonksiyonel ilişkiyi doğrudan ifade etmek zorunda kalmadan girdi değerlerini birkaç bölüme ayırarak değerlendirilebileceği kuralları bulacak şekilde tasarlanmışlardır (Nisbet vd., 2018:15). Karar ağaçları, en sezgisel olan ve en sık kullanılan tekniklerinden biridir (Kotu ve Deshpande, 2019: 66). Analizciler tarafından kolayca anlaşılabilen ve birçok durumda iyi performans gösteren bir sınıflandırma tekniğidir (Shmueli vd., 2017:183). Genel olarak karar ağacı modellerinin eksik verilerden az etkilenmesi, özniteliklerin seçiminde gereksiz olanları kullanmaması ve yüksek boyutlu verilerle iyi çalışması olumlu yanlarıdır (Kudyba, 2014:86).

Karar ağacı tekniğinde kullanılan çeşitli algoritmalar bulunmaktadır. Yaygın olarak kullanılan algoritmalar ID3 (Iterative Dichotoiser 3), C4.5, CART (Classification and Regression Trees) ve CHAID (Chi-squared Automatic Interaction Detection)'dır. Kategorik ve sürekli değerlerle çalışabilen karar ağacı tekniğinin CART algoritması, hem sınıflandırma hem de regresyon ağacına yönelik iki işlevi birden yerine getirme özelliği bulunmaktadır (Akpınar, 2017:248-267). CART, yalnızca ikili bölünmeleri olan ağaçları oluşturmaktadır. Bu kısıtlama (budama), bölme kriterini basitleştirmeye yöneliktir. Eğer eğitim setindeki verilerin etiketi ikili ise, CART'a daha uygun olduğundan kategorik olan özniteliklerin alt dallara en iyi şekilde bölünmesine imkân vermektedir (Kantardzic, 2020:219). Büyük verilerde rahat çalışması, yorumlanmasının kolay olması, farklı veri tipleri ile çalışabilmesi ve hile varlığına ya da yokluğuna ilişkin ikili bir sınıflandırma yapılması amaçlandığından karar ağacı tekniğinin CART algoritması tercih edilmiştir.

Hile tespitinde karar ağacı tekniğini kullanan çalışmalardan bazıları: Aksoy (2021), Jan (2018), Lakshmi ve Kavila (2018), Dutta vd. (2017), Chen (2016), Zhou ve Kapoor (2011), Ata ve Seyrek (2009) Liou (2008), Kirkos vd. (2007).

4.2.Kullanılan Ölçütler

Sınıflandırma algoritmaları eğitim kümesinden öğrenmeyi gerçekleştirdikten sonra sınıfın bilinmediği başka bir test kümesine uygulanır. Sonuca göre model daha sonra yeni vakalarda ilgilenilen durumu sınıflandırmak veya tahmin etmek için kullanılmaktadır (Shmueli vd., 2017:16). Çalışmada kullanılan karar ağacı algoritması kaynak veri setinin %69'u ile eğitilmiş, %31 ile test edilmiştir. Karar ağacı tekniğinin sınıflandırma performansı doğruluk (accuracy), kesinlik (precision), duyarlılık (recall), f-skoru (f-score) metrikleri ile ölçülmüştür.

4.2.1.Doğruluk

Sınıflandırıcının doğruluk (accuracy) ölçütü, toplamda doğru sınıflandırılmış pozitiflerin ve negatiflerin, toplam örnek sayısına bölünmesiyle hesaplanmaktadır (Memiş vd., 2019:3).

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

4.2.2. Kesinlik

Sınıflandırıcının kesinlik (precision) ölçütü, doğru sınıflandırılmış pozitif örneklerin, toplam pozitif tahmin edilmiş örneklere bölünmesiyle hesaplanmaktadır (Memiş vd., 2019:3).

$$Kesinlik = \frac{TP}{TP + FP} \quad (2)$$

4.2.3. Duyarlılık

Sınıflandırıcının duyarlılık (recall) ölçütü, doğru sınıflandırılmış pozitiflerin, toplam gerçek pozitif sınıfa bölünmesiyle hesaplanmaktadır (Memiş vd., 2019:3).

$$Duyarlılık = \frac{TP}{TP + FN} \quad (3)$$

4.2.4. F Skoru

Sınıflandırıcının f-skoru (f-score), kesinlik ve duyarlılık değerlerinin harmonik ortalaması alınarak hesaplanmaktadır (Memiş vd., 2019:3).

$$f\text{-}skoru = \frac{2\,(Precision \times Recall)}{Precision + Recall} \quad (4)$$

4.3. Kullanılan Veri

Çalışmada kullanılan kaynak veri, kamuya açık makine öğrenimi ve veri bilimi topluluğu olan '**kaggle.com**' adlı internet sitesinden sağlanmıştır. Kullanılan veri, ödemelerde aracılık yapan bir bankanın bir dönemine ait işlemleri simüle edilerek hazırlanmış yapay bir veri setidir. Veri seti 594.643 adet satır, 9 adet değişken/öznitelik ve 1 adet sınıflandırıcı sütundan oluşmaktadır. Veri setindeki 587.443 tanesi normal işlemken, 7.200 tanesi hileli işlemdir (Kaggle, 2017). Veri setindeki dolandırıcılık işlemleri tüm işlemlerin yaklaşık %1,21'ni oluşturmaktadır.

Uygulama sisteminin tahmin aşamasında kullanmak üzere kaynak veri setinden 1.000 tane kayıt ayrılmıştır. Bu yeni veri seti, 987 tanesi hileli olmadığını gösteren "0" etiketliden ve 13 tanesi hileli olduğunu gösteren '1' etiketliden olacak şekilde rastgele seçilerek bir araya getirilmiştir. Daha sonra karar ağacı tekniğiyle tahmin yapmak için bu yeni veri setinin sınıf etiketleri kaldırılmıştır. Uygulama sistemini hızlandırmak ve kaynak veri setindeki hileli kayıtların hileli olmayan verilere olan oranını (%1,21) yakalamak için hileli kayıtların 13 tanesi alınmıştır. Kaynak verinin öznitelikleri aşağıda tanımlamıştır.

- **İşlem Basamağı:** Bu değişken, işlemlerin 0 ile 179 adımdan oluştuğunu gösteren bir değişkendir. İşlemi yapanın 0-179 arası adımın hangi adımında gerçekleştirdiğini gösteren numerik bir değerdir.
- **Müşteri No:** Bu değişken, işlemi yapan müşterinin numarasıdır. Müşteri olduğunu belirtmek amacıyla numarasının başına 'C' harfi vardır.
- **Yaş:** Bu değişken, müşterilerin 0'dan 6'ya kadar olan yaş grubunu gösteren numerik bir değerdir. Ancak yaş grubu içinde tanımlanamayan 1.178 adet 'U' ifadesi vardır.
- **Cinsiyet**: Bu değişken, müşterilerin cinsiyetlerini 'M' ve 'F' harfleriyle gösteren metin bazlı bir değere sahiptir.
- **Müşteri Posta Kodu**: Bu değişken, müşterilerin posta kodlarını gösteren numerik bir değerdir.
- **Üye İşyeri No:** Bu değişken, işlemi yapan satıcının üye işyeri numarasıdır. Satıcı olduğunu belirtmek amacıyla numarasının başına 'M' harfi vardır.
- **Üye İşyeri Posta Kodu:** Bu değişken, satıcılara ait üye işyerlerinin posta kodlarını gösteren numerik bir değerdir.
- **Kategori:** Bu değişken müşterilerin sağlık, eğitim, taşımacılık, bar ve restoran, giyim, pazarlama, güzellik ve bakım, yiyecek, ev, otel, seyahat, spor ve eğlence, teknoloji, tasarım ve diğer hizmetler gibi sektörlerde yaptıkları harcamaları gösteren metin bazlı değere sahiptir.
- **Tutar:** Bu değişken, müşterilerin 15 farklı sektörlerde yaptıkları harcamalara ilişkin tutarlarını göstermektedir.
- **Hile Durumu:** Bu değişken, yapılan işlemin hileli olup olmadığını gösteren sınıflandırıcı bir etikettir. Bu etikete göre hilesiz işlemler '0' ile hileli işlemler '1' ile tanımlanmıştır.

Tablo 2. Veri Setinin Genel Görünümü

İşlem Basamağı	Müşteri No	Yaş	Cinsiyet	Posta Kodu	Üye İşyeri No	Posta Kodu	Kategori	Tutar	Hile Durumu
0	'C1093826151'	'4'	'M'	'28007'	'M348934600'	'28007'	'es_transportation'	4,55	0
0	'C352968107'	'2'	'M'	'28007'	'M348934600'	'28007'	'es_transportation'	39,68	0
0	'C2054744914'	'4'	'F'	'28007'	'M1823072687'	'28007'	'es_transportation'	26,89	0
0	'C1760612790'	'3'	'M'	'28007'	'M348934600'	'28007'	'es_transportation'	17,25	0
0	'C757503768'	'5'	'M'	'28007'	'M348934600'	'28007'	'es_transportation'	35,72	0
…	…	…	…	…	…	…	…	…	…
90	'C1350963410'	'5'	'F'	'28007'	'M2122776122'	'28007'	'es_home'	80,45	1
90	'C1307336396'	'2'	'F'	'28007'	'M2122776122'	'28007'	'es_home'	818,88	1
…	…	…	…	…	…	…	…	…	…
179	'C616528518'	'4'	'F'	'28007'	'M1823072687'	'28007'	'es_transportation'	26,93	0

Kaynak: *Kaggle, Synthetic Data From A Financial Payment System, 2017, https://www.kaggle.com/ealaxi/banksim1 (16 Şubat 2022).*

5. UYGULAMA

Günümüzde geliştirilen veri madenciliği araçları çeşitli yazılımlar ile kullanılmaktadır. Bu yazımlardan en bilinen ve kamuya açık olanları DataLab, DBMiner, Knime, RapidMiner, Weka, R ve Orange vd. sıklıkla kullanılmaktadır (Kantardzic, 2020:574-576). Ancak uygulama sistemi, kamuya açık veri madenciliği yazılım programlarının sınırlı ölçüde kullanıma imkân vermesi ve çalışma için önemli olan çıktıların üretilememesi nedeniyle Python üzerinde tasarlanmıştır. Aşağıda Python programlamasında kullanılan kütüphanelerin paylaşımı yapılmıştır.

pandas in /usr/local/lib/python3.7/dist-packages (1.3.5) ,
scikit-learn in /usr/local/lib/python3.7/dist-packages (1.0.2),
numpy>=1.17.3 in /usr/local/lib/python3.7/dist-packages (from pandas) (1.21.6),
joblib>=0.11 in /usr/local/lib/python3.7/dist-packages (from scikit-learn) (1.2.0),
scipy>=1.1.0 in /usr/local/lib/python3.7/dist-packages (from scikit-learn) (1.7.3).

Python programlama dilinin seçilmesinin amacı, programlama açısından çok kullanışlı olması, geniş bir kütüphaneye sahip olması ve veri madenciliğinde kullanımının da hızla artmasından kaynaklanmıştır (Layton, 2015:1). Ayrıca PYPL programlama dillerinin popülerlik indeksine göre Eylül 2022 itibariyle en çok tercih edilen programlama dili olmuştur (PYPL, 2022).

5.1. Kaynak Veriyi Hazırlama ve Dönüştürme

Veri hazırlama aşaması, ham veriden başlayarak nihai veri dizisine ulaşıncaya kadar gereken tüm faaliyetleri kapsamaktadır. Verilerin temizlenmesi ve dönüştürülmesi için görselleştirme, kayıt, öznitelik seçimi gibi faaliyetleri içerebilmektedir. Verinin kullanışlı hale getirilmesi ve modelin kurulması aşamalarını sağlayan işlemler toplam sürecin büyük bir kısmını oluşturmaktadır (Akpınar, 2017:79-82). Farklı nedenlerden kaynaklı olarak veri kümesi için uygun olmayan (eksik, hatalı veya tutarsız) değerler bulunabilmektedir. Tespit edilen bu değerlerin modelin tahmin gücünü azaltacağından düzeltilmesi, değiştirilmesi ya da silinmesi gerekebilir (Nisbet vd., 2018:62-63).

Çalışmanın kaynak veri setinde az da olsa hatalı olarak adlandırılan uygun olmayan değerler bulunduğu anlaşılmıştır. Veri setinin "age" sütununda diğer değerlerle benzeşmeyen "U" aykırı değerlerine rastlanmıştır. Bu kayıtların hileli işlemler olması ihtimali nedeniyle veri kümesinden çıkartmak yerine makul bir değer ataması yapılmak üzere içinde bırakılmıştır. Yukarıda Tablo 2'de görüldüğü üzere veri seti içinde metin bazlı olan değerler de bulunmaktadır. Diğer yandan, değişkenlerin tanımlanması ya da belirtilmesi amacıyla kullanılan karakterlerden temizlenmesi gerekmektedir.

Tablo 3. Temizlenmiş ve Dönüştürülmüş Kaynak Veri Seti

İşlem Basamağı	Müşteri No	Yaş	Cinsiyet	Posta Kodu	Üye İşyeri No	Posta Kodu	Kategori	Tutar	Hile Durumu
0	210	4	2	0	30	0	12	455	0
0	2753	2	2	0	30	0	12	3968	0
0	2258	4	1	0	18	0	12	2689	0
0	1650	3	2	0	30	0	12	1725	0

0	3585	5	2	0	30	0	12	3572	0
…	…	…	…	…	…	…	…	…	…
179	3564	2	2	0	18	0	12	5117	0
179	2639	3	1	0	18	0	12	2053	0
179	3369	4	1	0	18	0	12	5073	0
179	529	2	1	0	31	0	2	2244	0

Bir veri setindeki değişkenler, sayısal veya kategorik gibi farklı türlerde olabilir. Bir müşteri grubuna ait kredi puanı, kategorik değerle (zayıf, iyi, mükemmel) veya rakamsal değerle temsil edilebilmektedir. Veri madenciliğinden iyi bir sonuç alınmak isteniyorsa, örneğin doğrusal regresyon modeli için verilerin sayısal, karar ağacı algoritması için kategorik değere dönüştürülmesi daha faydalı olmaktadır (Kotu ve Deshpande, 2019:27). Veri seti temizlendikten sonra içinde farklı yapılarda bulunan veriler, karar ağacı algoritması için uygun hale getirilmesi gerekmektedir. Uygulama sistemiyle numerik olan ama kategorik olmayan veriler önce kategorik bir yapıya dönüştürülmüştür. Sonra kategorik olan ama numerik olmayan veriler, numerik veri tipine dönüştürülmüştür.

5.2.Ana Modelin Kurulması ve Test Edilmesi

Model kurma sürecinde oluşturulmak istenen modelin eğitim faaliyetleri öğrenme kümesi üzerinde gerçekleştirilir. Test kümesi olarak ayrılan bölümle oluşturulan modelin test edilmesi sağlanmalıdır. Modelin test edilmesinin amacı, modelin temsil edildiği veri kümesi üzerinden yapılan öğrenmenin başarının ölçülmesi içindir. Bazı projelerde test kümesine ek olarak doğrulama kümesi adı altında bir üçüncü küme oluşturulmaktadır. Doğrulama kümesi ile model bir kez daha test edilmiş olduğundan ikna edici ve güvenilir sonuçlara ulaşılmaktadır (Olson ve Delen, 2008:16). Modeller, eğer parametre ayarları değiştirilmezse varsayılan ayarlarla işlem yapmaktadır. Problemin çözümüne yönelik özel bir parametre gerekmesi durumunda ayarları değiştirilerek tekrar model oluşturulabilir (Nisbet vd., 2018:49). Verilerin anlaşılmasını kolaylaştıracak ve uygulanabilirliği arttıracak parametrelerin belirlenmesi de önemlidir. Seçilen modelle belirli bir veri setindeki ilişkinlerin soyut temsili ancak bu şekilde sağlanabilmektedir (Kotu ve Deshpande, 2019:29).

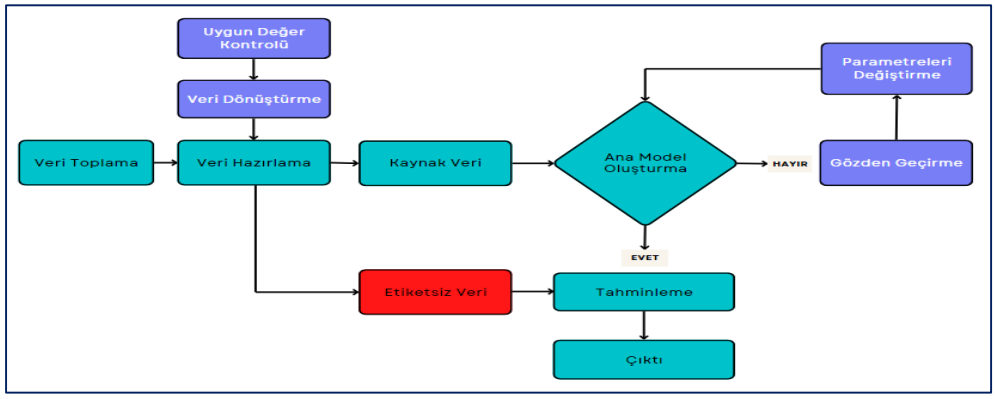
Bu aşamada, uygulama sisteminin modülleri vasıtası ile ana bir model oluşturma işlemi yapılmıştır. Eğitim ve test veri setinin belirlendiği komut kısmına manuel olarak giriş yapabilmekte mümkündür. Böylelikle model kurulma aşamasında kaynak veri setinin bölümlendirilmesi kullanıcıya bırakılmıştır. Yapılan denemeler sonucunda elimizdeki kaynak veri seti için en başarılı model %69 eğitim ve %31 test seti ayrımı ile oluşturulmuştur. Uygulama sistemiyle karar ağacı tekniği kullanılmış ve yapılan tahminlemeye ilişkin istatistik sonuçları elde edilmiştir. Sonuçlar incelendiğinde, modelin doğruluk oranı %99,42, f1-skoru %75, kesinlik oranı %74 ve duyarlılık oranı %76 hesaplanmış, eğitimi ve tahmini 1,32 saniyede tamamlanmıştır. Elde edilen metrik performansları oluşturulan ana modelin kullanılabilir olduğunu göstermiştir.

Tablo 4. Modelin Test İstatistikleri

Karar Ağacı					
Karışıklık Matrisi			**Doğruluk**	**:**	0,994
	Gerçek Pozitifler	**Gerçek Negatifler**	**F1-Skoru**	**:**	0,754
Tah. Ed. Pozitifler	181.353	552	**Kesinlik**	**:**	0,748
Tah. Ed. Negatifler	505	1.620	**Duyarlılık**	**:**	0,762

5.3 Ayrılan Etiketsiz Veri Setinin Sınıflandırılması

Uygulama sistemine girişi yapılan sınıf etiketi olmayan veri seti, önceki aşamalarda kaynak veri setine uygulanan işlemlerden geçirilerek hazırlanmalıdır. Etiketsiz veri setindeki belirtme amacıyla kullanılan belirtme ifadelerinden temizlenmiş ve uygun veri tipinin sağlanması adına dönüştürme işlemi yapılmıştır. Etiketsiz veri seti hazırlandıktan sonra karar ağacı algoritması kaynak veri setiyle eğitilmiş etiketsiz veri seti üzerinde tahmin yapması sağlanmıştır. Sürecini tamamlayan uygulama sistemi kısa bir süre sonra algoritmanın tahminlerini gösteren dosyayı çıktı olarak üretmiştir.



Şekil 3. Uygulama Sisteminin Akışı

Çalışmadaki uygulama sistemi, makine öğrenimi yöntemleriyle yapılan tahmin çıktılarının analizciler tarafından incelemesine ve doğrulamasına dayandırılmaktadır. Makine öğrenimi yöntemleri her ne kadar başarılı tahminler yapsa da bir kalıp ya da örüntünün sonuçlarını değerlendirmek ve model tahminlerinin makul olduğunu doğrulamak için analistler hala gerekmektedir. Çünkü bu yöntemler, ilgili ve alakasız korelasyon arasındaki farkı tanımak için gereken insan deneyiminden ve sezgisinden yoksundur. Bu nedenle analizcilere olan ihtiyaç devam etmektedir (Fernandez, 2003:Chapter 1,s.5).

5.4. Bulgular

Çalışmada önce ana model oluşturulmuş ve sonradan sisteme eklenen sınıf etiketsiz veri setinin, sınıf etiketlerinin oluşturulması sağlanmıştır. Daha sonra karar ağacı tekniğinin, etiketsiz veri seti üzerindeki sınıflandırma başarısı ölçülmüştür. Uygulama sistemi tarafından üretilen dosya Microsoft Excel programıyla açılmıştır. Karar ağacı algoritmasının ürettiği sınıf etiketleri ile kontrol amaçlı ayrılan veri setinin sınıf etiketleri bir araya getirilmiştir. Karar ağacı algoritmasının etiketleri veri setinde '**Tahmin Edilen**' sütununda, kontrol veri setinden alınan sınıf etiketleri '**Hile Durumu**' sütununda gösterilmiştir.

Tablo 5. Sınıf Etiketlerinin Bir Araya Getirilmesi

İşlem Basamağı	Müşteri No	Yaş	Cinsiyet	Posta Kodu	Üye İşyeri No	Posta Kodu	Kategori	Tutar	Tahmin Edilen	Hile Durumu
0	155	3	1	0	21	0	4	939	0	1
0	688	1	1	0	8	0	12	463	0	0
0	479	2	0	0	17	0	12	189	0	0
0	656	4	1	0	17	0	12	139	0	0
…	…	…	…	…	…	…	…	…	…	…
16	101	1	0	0	8	0	12	546	1	0
16	724	2	1	0	17	0	12	705	0	0
16	205	3	0	0	1	0	14	941	1	1
…	…	…	…	…	…	…	…	…	…	…
17	606	2	0	0	8	0	12	212	0	0

Sınıflandırmanın başarısı doğru pozitif (TP), doğru negatif (TN), yanlış pozitif (FP), yanlış negatif (FN) ile ifade edilen değerlere bağlıdır. Yapılan sınıflandırmada hilesiz işlemler uygulama sistemi tarafından da hilesiz olarak tahmin edilmişse doğru pozitif (TP) ve hileli olan işlemler uygulama sistemi tarafından hileli olarak tahmin edilmişse doğru negatif (TN) olarak adlandırılmaktadır. Diğer yandan hilesiz işlemler uygulama sistemi tarafından hileli olarak tahmin edilmişse yanlış pozitif (FP), hileli işlemler uygulama sistemi tarafından hilesiz olarak tahmin edilmişse yanlış negatif (FN) olarak adlandırılmaktadır (Kurien ve Chikkamannur, 2019:1034). Yukarıda gösterilen Tablo 5'teki "**Tahmin Edilen**" ve "**Hile Durumu**" sütunları filtrelenerek karışıklık matrisine ulaşılmıştır.

Tablo 6. Karışıklık Matrisi

	Doğru Pozitif	Doğru Negatif	Toplam
Tahmin Edilen Pozitif	968 (TP)	10 (FP)	978
Tahmin Edilen Negatif	19 (FN)	3 (TN)	22
Toplam	987	13	1.000

Karışıklık matrisi, sınıflandırma modelinin tahminlerinin doğruluğunu özetleyen bir (N × N) tablodur. N sınıf sayısını temsil etmektedir. Bir ikili sınıflandırma probleminde N = 2'dir. Diğer bir ifadeyle, gerçek sınıf etiketleri ile modelin tahmin ettiği sınıf etiketleri arasındaki korelasyonu ölçmeye yarayan bir tablodur. Karışıklık matrisinin bir ekseni modelin öngördüğü etiket, diğer eksen ise gerçek etiketten oluşmaktadır (Bhatia, 2019:120).

Karışıklık matrisi oluşturulduktan sonra model istatistikleri için gereken parametreler hazırlanmıştır. Modelin sonuçlarını analiz etmek gerekirse, 1.000 tane kayıtta yer alan 987 tane hilesiz kaydın 968 tanesini hilesiz olarak etiketleyerek doğru tahmin yapmıştır. Diğer taraftan 13 tane hileli kaydın 3 tanesini hileli olduğunu etiketleyerek doğru tahmin yapmıştır. Tahmin başarısının ölçülmesi için daha önce ana modelin başarısını ölçmede kullanılan metrikler hesaplanmıştır.

Tablo 7. Değerlendirme Ölçütlerinin Hesaplanması

Metrikler	Oran	Hesaplama
Doğruluk/ Accuracy	0,971	TP+TN/ TOPLAM
f1-skoru/f1-score	0,984	(2*Precision*Recall) / (Precision +Recall)
Kesinlik /Precision	0,989	TP/TP+FP
Duyarlılık /Recall	0,980	TP/TP+FN

Böylelikle etiketsiz veri seti üzerinde kullanılan karar ağacı tekniğiyle oluşturulan modelin istatistikleri; doğruluk skoru (0,971), f1-skoru (0,984), kesinlik (0,989) ve duyarlılık (0,980) sağlıklı bir şekilde oluşturulmuştur. Elde edilen sonuçlara göre uygulama sistemi başarılı bir sınıflandırma işlemi gerçekleştirmiştir. Böylelikle uygulama sistemiyle yapılacak olan sınıflandırmanın kullanım amacına uygun çıktıları ürettiği söylenebilir.

6. SONUÇ VE ÖNERİLER

Uygulamanın yapılabilmesi için önce verinin anlaşılması ve eksikliklerinin tespit edilmesi sağlanmış daha sonra Python programlama dilinde uygulama sistemi tasarlanmıştır. Bu süreç fazlaca vakit alsa da tasarımı sayesinde kullanımı kolay ve işlemleri oldukça hızlı yapan bir ürün ortaya konmuştur.

Makine öğrenimine dayalı yaptığımız hile tespitinde sağlıklı sonuçlar alabilmek adına karar ağacı tekniğiyle önce başarılı ana model oluşturulmuştur. Değerlendirmede, karar ağacı tekniği metriklerinin doğruluk %99,42'lik, f1-skor %75'lik, kesinlik %74'lük ve duyarlılık %76'lık skoruyla anlamlı istatistik elde edilmiştir. Karar ağacı tekniği, ayırdığımız 1.000 tane kaydın yer aldığı etiketsiz veri seti üzerinde kullanılarak tekniğin tahmin başarısına bakılmıştır. Tahmininde işlemlerini 1,17 saniyede bitirmiş ve model metriklerinin doğruluk %97,1'lik, f1-skor %98,4'lük, kesinlik %98,9'luk ve duyarlılık %98'lik skor yakalama başarısını göstermiştir.

Çalışmadaki uygulama sistemi, işlem kayıtlarının makine öğrenimi sürecinden geçirildikten sonra hileli işlem kayıtlarının tespit edilmesine ve hilenin önlenmesine yardımcı olmaya yöneliktir. Çalışmaya özel tasarlanan uygulama sisteminin tasarımı sayesinde denetçilere, karar ağacı tekniğinin tahminlerine dayalı hileli olma ihtimali olan kayıtları tespit etme imkânı sağlanmıştır. Ayrıca işletmelerin iş akışı içinde ya da hemen sonrasında oluşan kayıtların sınıf etiketleri olmayacağı için makinenin hileli olarak etiketlediği kayıtlar süreç içerisinde araştırılabilir. Diğer yandan iç denetçilere, hileyi önleme çalışmalarında kullanılan veri setinin sicil numarası, tutar, yaş, cinsiyet gibi özniteliklerine ve tahmin edilen hileli sınıf etiketine göre çapraz sorgulamalarla farklı bakış açılarının da kazandırılabileceği düşünülmektedir. Ayrıca sisteminin çalıştırılması, veri setinin yüklenmesi, modelin kurulması ve tahminin yapılmasına ait işlemlerin tamamı 10 dakika gibi bir sürenin altında bitirilmiştir. Buradan hareketle uygulama sisteminin gerekli hazırlıkları yapıldıktan sonra verilerin makine öğreniminden geçmesi ve verilere ilişkin tahmin sürecinin hızlı olması, denetçilere hem anlamlı seçimler yapmasını sağlayacak hem de büyük bir vakit kazandıracaktır. Süreçlerin kısa sürede yapılması işletmedeki hileli yapıların ortaya çıkarılmasını hızlandıracağı için zararın büyümesi engellenecektir.

Bu çalışma, genel itibariyle hilenin ortaya çıkarılmasına yönelik bir uygulama tasarımı gerçekleştirildiği için daha önce tespit edilmiş veriler kullanılmıştır. Süreçte, uygulama

sistemiyle hile ihtimali olan kayıtlar ortaya çıkarılmış olunacak ve inceleme sonucu hilenin doğrulanması durumunda sistemin kullanabileceği kaynak verilere ulaşılmış olunacaktır. Böylelikle kalıpların toplanması durumunda sistem kendi hile veritabanını beslemiş olacaktır. Bu düşünce sonraki yapılacak çalışmalar için bir kaynak oluşturabilir ve bir sistem önerisiyle açıklanabilir.

Çalışmada kullanılan karar ağacı tekniğinin tahminlerinde hataların olabileceği unutulmayarak yaptığı tahminlere kesin gözüyle bakılmadan incelenmesi oldukça önemlidir. Ayrıca diğer tekniklerle deneme yapılarak karşılaştırma yapılabilir. Uygulamada kaynak verinin tanınma ve eksiklerinin tespit edilmesi manuel olarak yapıldığı için oldukça vakit harcanmıştır. Bu süreci kısaltacak şekilde çalışmalar yapılarak uygulama sistemi yeniden tasarlanabilir. Böylelikle uygulama sistemi daha kullanışlı hale getirilebilir.

KAYNAKÇA

ACFE, Association of Certified Fraud Examiners, "2022 Global Study on Occupational Fraud and Abuse", Report to The Nation, https://acfepublic.s3.us-west-2.amazonaws.com/2022+Report+to+the+Nations.pdf , 23.10.2022

AKPINAR, H. (2017). Data, Veri Madenciliği Veri Analizi. 2.Baskı. Papatya Yayıncılık Eğitim, İstanbul.

AKSOY, B. (2021)."Finansal Tablo Hilelerinin Makine Öğrenmesi Yöntemleri ve Lojistik Regresyon Kullanılarak Tahmin Edilmesi: Borsa İstanbul Örneği", Maliye ve Finans Yazıları, 115:29-60.

ALBRECHT, W.S., ALBRECHT, C.O., ALBRECHT, C.C., & ZIMBELMAN M. F. (2011). Fraud Examination. Fourth Edition. South-Western, United States.

ATA, H.A., & SEYREK, İ.H. (2009). "The use of data mining techniques in detecting fraudulent financial statements: an applicatıon on manufacturing firms", Süleyman Demirel Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi, 14(2):157-170.

BHATIA, P. (2019). Data Mining and Data Warehousing Principles and Practical Techniques, First Published. Cambridge University Press, United Kingdom.

BOZKURT, N. (2009). İşletmelerin Kara Deliği Hile. 3.Baskı. Alfa Yayınları, İstanbul.

CHEN, S. (2016). "Detection of fraudulent financial statements using the hybrid data mining approach", SpringerPlus, 5(89):1-16.

CODERRE, D. (2009). Computer-Aided Fraud Prevention and Detection: A Step-by-Step Guide. Wiley & Sons, New Jersey.

CRAJA, P., KIM, A., & LESSMANN, S. (2020). "Deep learning for detecting financial statement fraud", Decision Support Systems, 139:1-13.

DUTTA, I., DUTTA, S., & RAAHEMI, B. (2017). "Detecting Financial Restatements Using Data Mining Techniques", Expert Systems With Applications, 90:374-393.

FERNANDEZ, G. (2003). Data Mining Using SAS Applications. Chapman & Hall/ CRC Press, United States.

GAGANIS, C. (2009). "Classification Techniques for the Identification of Falsified Financial Statements: A Comparative Analysis", Intelligent Systems In Accounting, Finance And Management, 16:207–229.

GOLDEN, T.W., SKALAK, S.L., CLAYTON, M. M., & PILL, J. S. (2011). A Guide to Forensic Accounting Investigation. Second Edition. John Wiley & Sons, New Jersey.

HACIHASANOGLU, T., ASLAN, T., & DALKILIC, E. (2021). Hile: İşletme Bilimi Perspektifinden Genel Bir Bakış. Paradigma Akademi, Çanakkale.

JAN, C. (2018). "An effective financial statements fraud detection model for the sustainable development of financial markets: evidence from Taiwan", Sustainability, 10(513):1-14.

KAGGLE, Synthetic Data From A Financial Payment System, 2017, https://www.kaggle.com/ealaxi/banksim1, 16.02.2022

KANTARDZIC, M. (2020). Data Mining: Concepts, Models, Methods, and Algorithms. Third Edition. John Wiley & Sons, New Jersey.

KIRLIOĞLU, H., & CEYHAN, I.F. (2014). "Mali Tablo Denetiminde Ön Analitik İnceleme Tekniği Olarak Veri Madenciliğinin Kullanımı: Borsa İstanbul Uygulaması", Akademik Yaklaşımlar Dergisi, 5(1):13-36.

KIRKOS, E., SPATHIS, C., & MANOLOPOULOS, Y. (2007). "Data mining techniques for the detection of fraudulent financial statements", Expert Systems with Applications, 32:995-1003.

KOTU, V., & DESHPANDE, B. (2019). Data Science Concepts and Practice. Second Edition. Morgan Kaufmann Publishers, United States.

KUDYBA, S. (2014). Big Data, Mining, and Analytics: Components of Strategic Decision Making. CRC Press, Taylor & Francis Group, Florida.

KURIEN, K.L., & CHIKKAMANNUR, A.A. (2019). "Benford's Law and Deep Learning Autoencoders: An approach for Fraud Detection of Credit card Transactions in Social Media", 2019 4th International Conference on Recent Trends on Electronics, Information, Communication & Technology, India: IEEE, 1030-1035.

LÆGREID, I. (2007)."Automatic Fraud Detection — Does it Work?", Annals of Actuarial Science, 2(2):271–288.

LAKSHMI, S.V.S.S., & KAVILA, S.D. (2018). "Machine Learning For Credit Card Fraud Detection System", International Journal of Applied Engineering Research, 13(4):16819-16824.

LAYTON, R. (2015). Learning Data Mining with Python. First Edition. Packt Publishing Ltd., Birmingham.

LIOU, F.M. (2008). "Fraudulent Financial Reporting Detection and Business Failure Prediction Models: A Comparison", Managerial Auditing Journal, 23(7):650-662.

MEMIS, S., ENGINOGLU, S., & ERKAN, U. (2019). "A Data Classification Method in Machine Learning Based on Normalised Hamming Pseudo-Similarity of Fuzzy Parameterized Fuzzy Soft Matrices", Bilge International Journal of Science and Technology Research, Özel Sayı (3):1-8.

NGAI, E.W.T., HU, Y., WONG, Y.H., CHEN, Y., & SUN, X. (2011). "The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literatüre", Decision Support Systems, 50:559–569.

NISBET, R., MINER, G., & YALE, K. (2018). Handbook Of Statistical Analysis And Data Mining Applications. Second Edition. Academic Press – Elsevier, London.

OLSON, D. L. (2008). DELEN, D. Advanced Data Mining Techniques. Springer-Verlag, Berlin.

PYPL, Programlama Dilleri Popülerlik İndeksi, https://pypl.github.io/PYPL.html, 15.09.2022

SHMUELI, G., BRUCE, P.C., STEPHENS, M.L., & PATEL, N.R. (2017). Data Mining For Business Analytics: Concepts, Techniques, And Applications With JMP Pro. First Edition. John Wiley & Sons, New Jersey.

SINGLETON, T.W., & SINGLETON, A.J. (2010). Fraud Auditing and Forensic Accounting. Fourth Edition. John Wiley & Sons, New Jersey.

TATAR, B., & KIYMIK, H. (2021). "Finansal Tablolarda Hile Riskinin Tespit Edilmesinde Veri Madenciliği Yöntemlerinin Kullanılmasına Yönelik Bir Araştırma", Journal of Yasar University, 16(64):1700-1719.

WEST, J., & BHATTACHARYA, M. (2016). "Intelligent Financial Fraud Detection: A Comprehensive Review", Computers & Security, 57:47-66.

ZHOU, W., & KAPOOR, G. (2011). "Detecting evolutionary financial statement fraud" Decision Support Systems, 50:570–575.

EKLER

Veri Setine Erişim Linki
https://www.kaggle.com/ealaxi/banksim1

Uygulamada Kullanılan Veri Setlerine Erişim Linki
https://drive.google.com/drive/folders/1E7s3WbrV9ghzszktaRcaQ8bceRb5xYRc?usp=sharing