Sigorta Sektöründe Sahte Hasarların Tahmini İçin Geliştirilen Makine Öğrenmesi Modellerinin Kıyaslanması

Araştırma Makalesi/Research Article

Özgür Erkut ŞAHİN[1], Serkan AYVAZ[2], Engin ÇALIMFİDAN[3]

[1]Eğitim Bilimleri Fakültesi, Bahçeşehir Üniversitesi, İstanbul, Türkiye
[2]Mühendislik Fakültesi, Bahçeşehir Üniversitesi, İstanbul, Türkiye
[3]Sosyal Bilimler Enstitüsü, Bahçeşehir Üniversitesi, İstanbul, Türkiye
erkut.sahin@es.bau.edu.tr, serkan.ayvaz@eng.bau.edu.tr, engincalimfidan@gmail.com
(Geliş/Received:22.03.2019; Kabul/Accepted:25.10.2020)
DOI: 10.17671/gazibtd.543265

Özet— Araştırmanın amacı, sigorta sektöründe kasko sigortası için sahte hasarların tespitinde hasar dosyası incelemelerine yardımcı olabilecek makine öğrenmesi modelleri geliştirmektir. Bu çalışmada özel bir sigorta şirketinin kasko sigortasına ait hasar verileri kullanılmıştır. Model oluşturulmasında k-en yakın komşuluk, karar ağaçları, lojistik regresyon, yapay sinir ağ algoritmaları denenmiştir. Elde edilen sonuçlar doğrultusunda makine öğrenimi yöntemlerinin kullanımının suistimali hasarların tespiti için hasar ekiplerine ve sigorta şirketlerine yardımcı olabileceği düşünülmektedir.

Anahtar Kelimeler— Sigorta sektörü, kasko sigortası, suistimal tespiti, sahte hasar tahmini, veri analizi, makine öğrenmesi, yapay sinir ağları.

Comparison of Machine Learning Models for Predict Fraudulent Claims in Insurance Sector

Abstract— The aim of this research is to develop machine learning models that can assist in the investigation of automobile insurance claims by detecting counterfeit damages filed in the insurance industry. In this study, automobile insurance claims data belonging to a private insurance company is used for analysis. The k-nearest neighborhood, decision trees, logistic regression, and artificial neural network algorithms have been explored in data modeling. Based on the research results, it is observed that the use of machine learning methods can help claims investigation teams and insurance companies to detect fraudulent activities.

Keywords— Insurance Sector, automobile insurance, fraud detection, fraud prediction, data analysis, machine learning, artifical neural network.

1. GİRİŞ (INTRODUCTION)

Sigorta sektöründe şirketler suistimalli durumlar ile birçok alanda karşı karşıya kalmaktadırlar. Suistimal, kişilerin haksız kazanç elde etmek için şirketlere önemli ölçüde maddi kayıp yaşattıkları ve itibarlarını zedeledikleri kasıtlı olarak gerçekleştirilen davranışlardır [1]. Suistimal tespitine proaktif yaklaşım, suistimalin maddi zararlarını azaltma konusunda şirketlere büyük ölçüde fayda sağlamaktadır [2]. Şirketler, ileri seviye analitik yöntemler kullanarak, suistimali önceden tespit edebilme kabiliyeti kazanmakta ve suistimale karşı önlemler alabilmektedirler.

Geçmişte, suistimal tespit sistemleri genel olarak iş kurallarına dayanmaktaydı. Bu yöntem ile şirketler daha önceden öğrenilmiş suistimal tiplerini tespit edebiliyordu. Makine öğrenmesi teknikleri sayesinde ise şirketler karmaşık örüntüleri ve daha önce bilinmeyen suistimal tiplerini de tespit etmeye başladılar [3].

Makine öğrenmesi teknikleri ile birlikte, şirketler daha çok suistimali ortaya çıkarma yeteneğine sahip olabilmektedirler. Burada en önemli etken diğer veri madenciliği çalışmalarında olduğu gibi, verinin kalitesi ve erişilebilir olmasıdır [3].

Bu araştırmada, "Hasar ekiplerinin manuel olarak yaptığı sahte hasarların tespit edilmesi sürecine destek olacak nitelikte bir makine öğrenmesi modeli geliştirilebilir mi?" ve "Suistimali yapan kişilerin davranışları arasında benzerlikler var mı?" soruları üzerinden çalışma yapılmıştır.

Bu çalışmada belirtilen sorulara makine öğrenmesi yöntemleri kullanılarak cevap aranmıştır.

2. LİTERATÜR TARAMASI (LITERATURE REVIEW)

2.1. Benzer Çalışmalar (Related Work)

Bu çalışmaya benzer olarak, sigorta sektöründe suistimal ve makine öğrenmesi konularında daha önceden yayımlanmış bazı tez ve makaleler aşağıda sıralanmıştır.

Evren Kasap'ın çalışmasında, sigortacılık sektöründe müşteri ilişkileri yönetimi incelenmiştir. Bu çalışma, veri madenciliği teknikleri ile müşterileri sınıflandırma, kümeleme ve davranış olasılıklarını tahmin eden analizlerden oluşmaktadır. Yapılan çalışmada özellikle bankacılık sektöründe yaygın olarak kullanılan müşteri ilişkileri yönetimi ve veri madenciliği teknikleri sigortacılık sektöründe de uygulanmaya çalışılmıştır. Ürün-müşteri, şirket-müşteri arasındaki ilişkileri ortaya koyarak, müşterilerin tercihlerine göre poliçe satışında artış sağlanmaya yardımcı olacak yollar araştırılmıştır [4].

Duygu Muslu'nun çalışmasında, sigortacılık sektörü için kurulmuş olan hasar ihbar veri tabanında bulunan verilerden suistimal riski tahmin etme çalışması yapılmıştır. Çalışmada, hedef değişken ve onu etkileyecek nitelikler belirlenmiştir. Makine öğrenmesinde sınıflandırma algoritmaları arasında yer alan karar ağacı kullanılarak bir modelleme çalışması yapılmıştır. Daha sonra sonuçlar değerlendirilerek risk maddesi oluşturup oluşturmadığına dair yorumlar geliştirilmiştir [5].

Yasin Kaya'nın çalışmasında, motokaravan sigortası yaptırabilecek müşterilerin tahmin edilmesi üzerine bir araştırma yapılmıştır. Karar ağaçları, lojistik regresyon ve yapay sinir ağları algoritmaları kullanılarak modeller geliştirilmiştir. Böylece hangi özelliklere sahip müşterilerin motokaravan sigortası yaptırabileceğine ilişkin bir tahmin çalışması gerçekleştirilmiştir [6].

İbrahim Şişaneci, araştırmasında sağlık sektöründe uygulanan performansa dayalı ek ödeme sisteminde suistimal tespiti probleminin, otomatik olarak çözülüp çözülemeyeceği incelemiş ve çözülmesi için veri madenciliği modelleri geliştirmiştir [7].

Ahmet Yılmaz'ın araştırmasında, sigortacılıkta sahte hasarları tespit eden bir lojistik regresyon modeli geliştirmek amaçlanmıştır. Sahte hasar olma ihtimaline göre artan sırada hasarlar 1 ile 9 arasında risk gruplandırması yapılmıştır. Modelin başarısını test etmek için risk grubu 6-9 arasında olan hasar dosyalarındaki sahte hasarlarının dağılımı incelendiğinde, modelin sahte hasarların %84,94'ünü 76318 dosya yerine sadece 7404 dosya incelenerek bulunmasını sağladığı görülmüştür. Kurulan model sahte hasar yakalamak için binlerce hasar dosyası incelemek zorunda kalan sigorta şirketlerinin işlerini azaltmak, doğru karar verme oranını yükseltmek ve karar verme sürecini istatistik tahmin modeli kullanarak sistematik hale getirdiği için faydalı gözükmektedir [8].

Rekha Bhowmik araştırmada, rapor ve verileri kullanarak otomobil sigortasında sahtekarlıkları tespit etmek için modeller oluşturmayı amaçlamıştır. Modelleri eğitmek için, naive bayes sınıflandırıcısı ve karar ağacı algoritması kullanılmıştır. Doğruluk, hatırlama, hassasiyet ve karışıklık matrisi gibi performans ölçütleri ile modellerin başarısı ölçülmüştür. Çalışmada, sahtekarlık tespitine yardımcı olabilecek modeller geliştirilmiştir [9].

Richard Bauder ve Taghi Khoshgoftaar araştırmalarında, ABD'deki sağlık hizmetlerinde (Medicare) dolandırıcılığı engellemek için makine öğrenmesi modeli geliştirmeyi amaçlamışlardır. Bu araştırmada, C4.5 karar ağacı ve lojistik regresyon ile modeller eğitilmiştir. Çalışmanın ABD'deki sağlık hizmetlerinde, sahtekarlıkların tespitine yardımcı olabileceği düşünülmektedir [10].

Bu çalışmada yukarıda hakkında bilgiler verilen çalışmaların bazılarıyla sigortacılık sektöründe çalışılmasından dolayı, bazılarıyla da model kurma aşamasında k en yakın komşuluk algoritması, yapay sinir ağları, karar ağaçları ve lojistik regresyon kullanılmasından dolayı benzerlikler vardır.

Bu araştırmada, kasko sigortası için sigortalıların hasar ve poliçe verileri kullanılarak, sahte hasarları tahmin etmede dosya incelemecilerine yardımcı olacak bir makine öğrenmesi modeli geliştirmek amaçlanmıştır. Model geliştirmeleri yapılırken açık kaynak (open source) yazılım teknolojileri kullanılmıştır. K en yakın komşuluk, yapay sinir ağı, karar ağaçları ve lojistik regresyon algoritmaları kullanılarak modeller geliştirilmiştir ve modellerin sahte hasarları tahmin etme başarıları kıyaslanmıştır. Araştırmanın kasko branşı üzerinden yapılmış olması, farklı makine öğrenmesi algoritmalarının kullanılması ve modeller geliştirilirken açık kaynak yazılım teknolojileri kullanılması bakımından diğer araştırmalardan ayrılmaktadır.

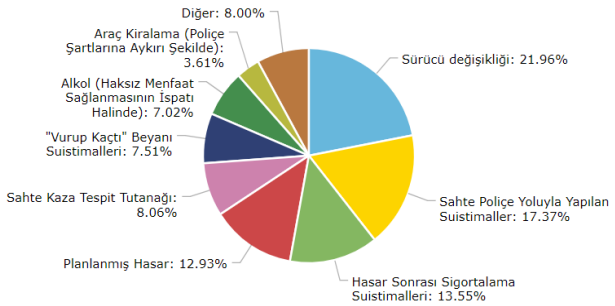
2.2. Sigortacılıkta Suistimal Kavramı (The Concept of Fraud in Insurance)

Sigortacılıkta suistimal, suistimali yapan kişiye veya üçüncü şahıslara haksız, hileli çıkar sağlamaya yönelik eylemler olarak tanımlanabilir [11].

Sigorta sektöründe suistimal; aynı araçları sürekli birbirleri ile çarptırarak, sahte hasar ile sigorta şirketinden talep edilen paralar ile veya kasıtlı olarak yanlış beyanda bulunma, hasar ile ilgili bilgilerin gizlenmesi şeklinde kendini göstermektedir [12].Sigorta suistimallerinin en çok yapıldığı alanların başında trafik, kasko, makine kırılması, konut, işyeri hırsızlık sigortaları ve özel sağlık sigortaları yer almaktadır [13].

Sigorta suistimalleri birçok ülkede vergi kaçakçılığından sonra en yaygın ekonomik suç olarak kabul ediliyor. Türkiye'de, sigorta sektörünün ödediği toplam hasarın yüzde 10-25'lik kısmında suistimal olduğu tahmin ediliyor. Uygulanan yöntemler açısından da dikkat çeken suistimallerin çok büyük kısmı ise oto sigortalarında görülür [14].

Şekil 1'de görüldüğü gibi Sigorta Sahtecilikleri Engelleme Bürosu (SİSEB) tarafında açıklanan Ekim 2017 - Ekim 2018 suistimal yöntemleri verilerine göre en çok kullanılan yöntem yüzde 21,96 ile sürücü değişimi/sürücü firari, ikinci olarak yüzde 17,37 ile sahte poliçe yoluyla yapılan suistimaller, üçüncü olarak yüzde 13,55 ile hasar sonrası sigortalama suistimalleri, dördüncü olarak ise yüzde 12,93 ile planlamış hasar (Organize hasar) gelmektedir.



Şekil 1. Suistimal yöntemleri ekim 2017- ekim 2018 dağılımı [15]
(Insurance fraud october 2017-october 2018 distribution [15])

2.3. Kasko Sigortası için Sahtekarlık (Fraud for Automobile Insurance)

Kişilere maddi olarak güvence sunan sigorta şirketleri, zaman zaman sigortalıların bu güvenceyi kötüye kullanmak istemeleri nedeni ile güç durumda kalmaktadırlar. "Sigorta sahtekarlığı" olarak adlandırılan bu tür vakalar her sigorta branşında karşılaşılsa da en çok otomobil, yangın ve sağlık sigortalarında yaşanmaktadır [16].

Otomobil sigortalarının amacı sigortalının zararını karşılamak, maddi açıdan hasar gerçekleşmemiş gibi hayatına devam etmesini sağlamaktır [17]. Sigorta suistimal konusunun dünya genelinde bir sorun olduğu, İspanya'da oto sigorta suistimali analizi üzerine yapılan bir çalışmada da belirtilmiştir [18].

Otomobil sigortalarında genel olarak suistimal, sigortalının bilerek hasar yapması veya hasarı olduğundan büyük göstermeye çalışması şeklinde meydana gelmektedir. Örnek olarak, araba çalınmalarının yüzde 10'unun gerçek çalınma vakası olmadığı görülmüştür [19].

Accenture'nin 2003 yılında yayınlamış olduğu bir ankette, birçok insan sigorta sahtekarlığına tamamen karşı değil ve belli bir ölçüde tolerans gösteriyor. Anket sonucuna göre hasar değerlerinin yüzde 24 abartılması kabul edilebilir bulunmaktadır. Yüzde 11'i olmadıkları tedavilerin masraflarını talep etmenin yine kabul edilebilir olduğunu bildiriyorlar. Yüzde 30'u ise ekonomilerinin bozulduğu durumlarda sahte hasar yapma eğiliminde olduklarını, yüzde 49'u ise sigorta suistimallerinden uzak durduklarını söylüyorlar [8].

2.4. Sigorta Hasar Süreci (The Insurance Claim Process)

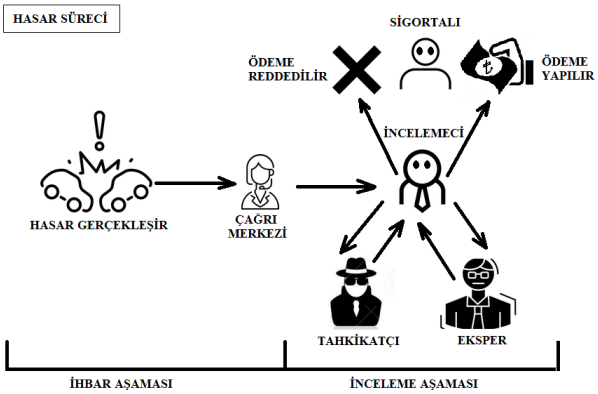
Sigorta şirketlerinde hasar süreçlerini Şekil 2'de görüldüğü gibi genel olarak ihbar aşaması ve inceleme aşaması olarak ikiye ayırabiliriz.

İhbar aşaması, hasar gerçekleştikten sonra sigorta şirketine bildirilmesine kadar süren aşamadır. Sigorta şirketine bildirim yapıldıktan sonra hasar ile ilgili bilgiler alınır ve bir dosya oluşturulur.

Dosya oluşması aşamasından sonra inceleme aşaması başlamaktadır. Sigorta şirketlerinde hasar dosyalarını inceleyen ekipler ve bu ekiplerde görevli incelemeciler vardır. İncelemeciler kendilerine atanmış dosyaları incelerler. Bu aşamada hasarın nasıl gerçekleştiği anlaşılmaya çalışılır, hasar tutarının tespiti yapılır. Dosyaya eksper atanır ve ödeme yapılıp yapılmayacağına karar verilir.

Hasarın nasıl gerçekleştiği ile ilgili araştırma yapılırken dosyada bir anormallik olduğu düşünülür ise incelemeci tarafından dosyaya tahkikatçı atanır. Tahkikatçılar, olay yerini inceler, görgü tanıkları ile konuşur, kazanın beyan edildiği gibi gerçekleşip gerçekleşmediğini araştırırlar. Araştırma sonucunda bir rapor hazırlayarak incelemeciyi bilgilendirirler. Tahkikatçılar, sigorta şirketleri için çalışan dedektifler gibidir. Tahkikatçının hazırladığı rapor sonucunda incelemeci dosya ile ilgili sigortalıya tazminat ödemesi yapıp yapmamaya karar verir.

Sigorta şirketlerine her gün onlarca hasar ihbarı yapılmaktadır. Bu dosyalar incelenirken her dosya için bir tahkikatçı atanması sigorta şirketi açısından maliyete sebep olmaktadır. Bu sebepten dolayı, sadece incelemeci tarafından şüpheli bulunan dosyalara tahkikatçı atanmaktadır. Bu süreçte genel olarak şüpheli bulunan dosyaların tespiti, incelemeci tarafından manuel olarak veya geçmiş iş tecrübelerinden elde edilen bilgiler ışığında oluşturulmuş iş kuralları ile yapılmaktadır.



Şekil 2. Sigorta hasar süreci gösterimi
(Illustration of İnsurance claim process)

3. VERİ VE YÖNTEM (DATA AND METHODS)

3.1. Araştırmanın Amacı (The Purpose of Research)

Bu araştırmanın amacı, sigorta sektörü için maliyetlerin artmasına ve itibar kaybına sebep olan sahte hasarlar ile ilgili çalışan hasar ekiplerine, manuel olarak inceledikleri dosyanın sahte hasar olup olmadığı ile ilgili ayırt etmede yardımcı olmak ve sahte hasarların tespit oranında başarı düzeyini artırmaya katkıda bulunmak için bir makine öğrenmesi modeli geliştirilmesidir.

3.2. Araştırma Problemi ve Hipotezler (Research Problem and Hypothesis)

Türkiye'de sigorta sektörünün ödediği toplam hasarın yüzde 10 ile 25'lik kısmında suistimal olduğu tahmin edilmektedir. Sigortalının poliçe ve hasar verilerinde, sahte hasar yapanlarla yapmayanlar arasında farklar olabilmektedir. Bunlara ek olarak, sahte hasar yapan ile yapmayan sigortalıların davranışları arasında da farklar olması gerektiği düşünülmektedir. Bu çalışmada, sahte hasar yapanların davranış benzerliklerini açıklayabilen değişkenlerle, sahte hasar yapanların, yapmayanlardan ayırt edilebileceği savunulmaktadır.

- H0: Sahte hasar yapanların davranışları arasında anlamlı bir fark yoktur.
- H1: Sahte hasar yapanların davranışları arasında anlamlı bir fark vardır.

Bu hipotezin geçerliliğini araştırmak amacıyla bu çalışma yapılmıştır.

3.3. Araştırmanın Evreni ve Örneklemi (Universe of Research)

Araştırma, Türkiye Sigorta Sektörü içinde öncü olan şirketlerden birinden alınan hasar, dosya ve poliçe bilgileri ile elde edilen veriler ile yapılmıştır. Örnekler Sigorta şirketinin veri tabanındaki tablolardan alınmıştır. İhbar tarihi 01.01.2013 – 01.04.2016 arasında olan suistimalli/suistimalsiz şeklinde etiketlenmiş 524260 tane dosya bulunmaktadır. Bu dosyaların 4519 tanesi suistimal yakalanmış dosyalardır. Eğitim için örneklem alınırken toplamda 4519 tane suistimalli dosyanın tamamı kullanılmıştır. Suistimal olmayan dosyalardan da suistimalli dosyaların 9 katı kadar rastgele bir şekilde veri toplanmıştır. Toplamda 45190 tane dosya alınarak modeller eğitilmiştir.

3.4. Araştırmada Kullanılan Değişkenler (Variables Used in Research)

Bu çalışmada veri ön işleme aşamasında yapılan incelemeler sonundan bazı değişkenlerin çok sayıda kayıp değer içermesi veya modelle ilişkili bulunmaması nedeniyle kullanılan değişkenler arasından çıkarılmıştır. Sonuç olarak, araştırmada 20 bağımsız değişken ve 1 bağımlı değişken (hedef) olmak üzere 21 değişken kullanılmıştır. 45190 dosya için Tablo 1'de verilen değişkenler doldurularak veri seti hazırlanmıştır.

Tablo 1. Araştırmada kullanılan tüm değişkenler
(All variables used in the research)

Değişken Adı	Değişken Tipi	Açıklama
ALKOLLU_MU	Nominal - Boolean	Sürücü Alkollü mü bilgisi. (EVET:1 / HAYIR:0)
RUCU_VAR_MI	Nominal - Boolean	Rücu var mı bilgisi. (EVET:1 / HAYIR:0 / KAYIP DEĞER: 2)
DOSYA_BOLGE	Multinominal - Kategorik	Dosya bölgesi poliçenin kesildiği yer olarak alınır.
HASAR_NEDENI	Multinominal - Kategorik	Hasar gerçekleşme nedeni
GECIKMIS_ODEME_MI	Nominal - Boolean	Gecikmiş ödememi. (EVET:1 / HAYIR:0)
SUPELI_LISTESINDE_MI	Nominal - Boolean	Şüpheli listesinde olması. (EVET:1 / HAYIR:0)
HASAR_GEC_BILGIRILMIS_MI	Multinominal - Kategorik	Hasar geç bildirilmiş mi.
HASARPOLICE_TARIHFARK_FLG	Multinominal - Kategorik	Hasar Tarihi ile Poliçe Tarihi arasındaki fark kategorize edilmiştir.

Değişken Adı	Değişken Tipi	Açıklama
GECE_YARISI_FLG	Multinominal - Kategorik	00:00 – 05:00 arası yapılmış kaza mı? (EVET:1 / HAYIR:0 / KAYIP DEĞER: 2)
CINSIYET	Multinominal - Kategorik	Cinsiyet bilgisi. (ERKEK:3 / KADIN:2 / KAYIP DEĞER: 1)
YAPTIRILMIS_HASAR_MI	Nominal - Boolean	Yaptırılmış hasar mı bilgisi. (EVET:1 / HAYIR:0)
HIZLI_ISLEM_DOSYASI_MI	Nominal - Boolean	Hızlı işlem dosyası mı bilgisi. (Hasar Tutarı 2500 TL altı olan dosyalar.)
POLIS_GELMIS_MI	Nominal - Boolean	Polis gelmiş mi bilgisi. (EVET:1 / HAYIR:0)
S_FRAUDLUDOSYASAYISI_FLG	Multinominal - Kategorik	Sigortalının suistimalli dosya sayısı.
SONBIRYILHASARSAYISI_FLG	Multinominal - Kategorik	Sigortalının son bir yıldaki hasar sayısı.
ARAC_MARKASI	Multinominal - Kategorik	Araç markası bilgisi.
KAZA_YERI	Multinominal - Kategorik	Kaza yeri.
BLOKLU_POLICE_MI	Nominal - Boolean	Bloklu poliçe mi bilgisi. (EVET:1 / HAYIR:0)
SATIS_KANALI	Multinominal - Kategorik	Satış kanalı bilgisi.
CALINMIS_ARABA_MI	Nominal - Boolean	Araba çalınmış mı bilgisi. (EVET:1 / HAYIR:0)
INCELENMELI_MI (Hedef Değişken)	Nominal - Boolean	İncelenmeli mi. (EVET:1 / HAYIR:0) Bağımlı değişken.

3.5. Regresyon Analizi (Regression Analysis)

Regresyon analizi, aralarında neden sonuç ilişkisi bulunan iki veya daha fazla değişken arasındaki ilişkiyi belirlemek ve bu ilişkiyi kullanarak o konu ile ilgili tahminler ya da kestirimler yapabilmek amacıyla yapılır [20].

Bir bağımlı değişken ve birden fazla bağımsız değişkenin yer aldığı regresyon modellerine çok değişkenli regresyon analizi denir [21]. Çok değişkenli regresyon analizinde bağımsız değişkenler eş zamanlı olarak bağımlı değişkendeki değişimi açıklamaya çalışmaktadır.

Bu araştırmada değişkenleri analiz etmek için, çok değişkenli regresyon analizi ve geriye doğru eleme yöntemi kullanılmıştır.

Geriye doğru eleme yöntemi; regresyon analizi sonucu anlamlılık seviyesi eşik değerinin üzerinde kalan değişkenlerin, veri setinden sıra ile çıkarılarak ikinci bir veri seti oluşturulması aşamasında kullanılan yöntemdir [22].

Bu yöntemler ile bağımlı değişkeni daha iyi açıklayabilecek bağımsız değişkenlerden oluşan bir veri seti oluşturulur.

3.6. Makine Öğrenmesi Yöntemleri (Machine Learning Methods)

Makine öğrenmesi yöntemleri; gözetimli öğrenme (supervised learning), gözetimsiz öğrenme (unsupervised learning), yarı gözetimli öğrenme (semi-supervised learning) ve pekiştirmeli öğrenme (reinforcement learning) olarak 4 ana başlıktan oluşmaktadır [23].

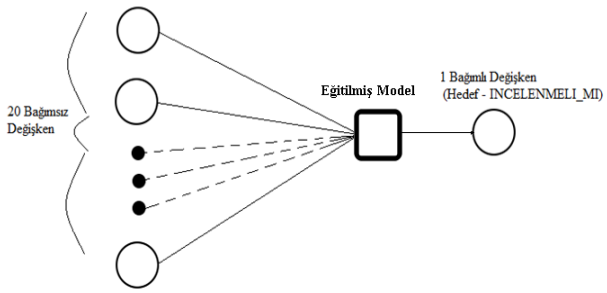
Bu araştırmada, makine öğrenmesi yöntemlerinden gözetimli öğrenme türüne (supervised learning) ait algoritmalar kullanılmıştır.

Gözetimli öğrenme, etiketli veri seti bulunduğu durumlarda sınıflandırma veya regresyon problemleri için model geliştirmeyi sağlayan öğrenme yöntemidir [22]. Gözetimli öğrenmedeki temel unsur daha önceki gözlemlerden ve onlara ait sonuçlardan oluşan bir veri setine sahip olma gerekliliğidir.

3.7. Araştırma Modeli (Research Model)

Araştırmadaki modeller, gözetimli öğrenme algoritmalarından olan k en yakın komşuluk (k nearest neighbor), lojistik regresyon (logistic regression), karar ağacı (decision tree) ve yapay sinir ağları (multilayer perceptron) algoritmaları kullanılarak eğitilmiştir.

Şekil 3'de kurulan modelimiz şekil olarak ifade edilmiştir. 20 bağımsız değişkenden oluşan giriş (input) değerlerimiz, eğitilmiş modelimiz ve "incelenmeli mi" bağımlı değişkenimiz (output) görülmektedir.



Şekil 3. Bağımsız değişkenden hedef değişkene ulaşma modeli
(Insurance fraud October 2017-October 2018 distribution [5])

3.8. Modelleri Eğitmek İçin Kullanılan Araçlar (The Tools Used for Training The Models)

Modellerin geliştirmesinde python programlama dili ve python kütüphanelerinden (numpy, pandas, matplotlib, sklearn, keras) faydalanılmıştır.

K en yakın komşuluk (k nearest neighbor), lojistik regresyon (logistic regression) ve karar ağacı (decision tree) algoritmaları ile eğitilen modellerde numpy, pandas, matplotlib ve sklearn kütüphaneleri ile oluşturulan python kodları kullanılmıştır.

Yapay sinir ağları (multilayer perceptron) algoritması için eğitilen modellerde numpy, pandas, matplotlib ve sklearn kütüphanelerine ek olarak keras kütüphanesi kullanılmıştır.Yapay sinir ağı modelinin en öenmli özelliği hızlı, güçlü olmaso ve makine öğrenmesi alanında karşılaşılan problemleri etkin bir şekilde çözmesidir [24]. Modellerin eğitimleri, bu kütüphaneler kullanılarak oluşturulan python kodları ile yapılmıştır.

3.9. Veri Toplama Araçları (Data Collection Tools)

Sigorta şirketinde oracle veritabanı kullanılmaktadır. Araştırmada tablolardaki verileri toplayıp özetlemek için Yapılandırılmış Sorgu Dili (SQL, Structured Query Language) ile çeşitli sorgular yazılmıştır. Veri tabanına bağlanmak ve SQL sorgularını çalıştırmak için Toad for Oracle ve Oracle Client gibi programlar kullanılmıştır.

4. BULGULAR (RESULTS)

4.1. Regresyon Analizi Bulguları (Regression Analysis Results)

Bu çalışmada, bağımlı değişkeni (“INCELENMELI_MI”), 20 bağımsız değişkenin ne kadar iyi açıkladığını anlayabilmek adına regresyon analizi yapılmıştır. Regresyon analizi ile değişkenlerin anlamlılık seviyeleri belirlenmiştir. Tablo 2'de analiz sonucu detaylıca gösterilmiştir. Geriye doğru eleme yöntemi ile anlamlılık seviyesi 0.05 ve üzeri olan değişkenler veri setinden çıkarılmıştır. Bu yöntem ile birlikte 15 bağımsız değişkenden oluşan ikinci bir veri seti oluşturulmuştur. Bu iki veri seti ile algoritmalar çalıştırılmış ve modeller eğitilip karşılaştırılmıştır.

Tablo 2. Regresyon analizi ile değişkenlerin anlamlılık seviyeleri
(Significance of variables with regression analysis)

Coefficient	Std Error	t statistic	Değişken Adı	Anlamlılık Seviyesi (P>|t|)
0,8114	0,020	41,355	ALKOLLU_MU	0.000
-0,0964	0,002	-39,658	RUCU_VAR_MI	0.000
0,0003	0,000	0,615	DOSYA_BOLGE	0.538
0,0004	3,97e-05	9,914	HASAR_NEDENI	0.000
0,0602	0,018	3,277	GECIKMIS_ODEME_MI	0.001
0,3693	0,013	28,423	SUPELI_LISTESINDE_MI	0.000
-0,0012	0,001	-0,796	HASAR_GEC_BILGIRILMIS_MI	0.426
-0,0677	0,003	-26,790	HASARPOLICE_TARIHFARK_FLG	0.000
0,0513	0,002	21,042	GECE_YARISI_FLG	0.000
0,0062	0,002	3,592	CINSIYET	0.000
-0,0724	0,010	-7,426	YAPTIRILMIS_HASAR_MI	0.000
-0,0742	0,004	-17,157	HIZLI_ISLEM_DOSYASI_MI	0.000

Coefficient	Std Error	t statistic	Değişken Adı	Anlamlılık Seviyesi (P>|t|)
0,0284	0,004	7,795	POLIS_GELMIS_MI	0.000
0,7361	0,011	67,481	S_FRAUDLUDOSYASAYISI_FLG	0.000
0,0014	0,002	0,614	SONBIRYILHASARSAYISI_FLG	0.539
2,413e-05	1,17e-05	2,063	ARAC_MARKASI	0.039
6,395e-05	7,35e-05	0,870	KAZA_YERI	0.384
-0,0514	0,008	-6,508	BLOKLU_POLICE_MI	0.000
-0,0146	0,006	-2,291	SATIS_KANALI	0.022
-0,1104	0,108	-1,021	CALINMIS_ARABA_MI	0.307

Coefficient kolonu regresyon algoritmasının değişkenler için atadığı katsayıları göstermektedir. Standart Error değerinin küçük olması gerçek katsayıdaki sapmanın az olduğu anlamına gelir. Tablo 5'de gösterildiği üzere R-squared değeri ise, bağımsız değişkenlerin bağımlı değişkendeki değişimin %22,2'sini açıklayabildiği anlamına gelmektedir.

Tablo 3. Sıradan en küçük kareler yöntemi
(Ordinary least squares)

R squared	Adjusted R-squared	F statistic
0,222	0,222	645,2

4.2. Eğitilen Modellerin Kıyaslanması (Comparison of Trained Models)

Araştırmadaki modeller, 20 bağımsız değişkenli veri seti ve 15 bağımsız değişkenli veri seti kullanılarak ayrı ayrı eğitilmiştir. Veri kümemizin yüzde 10'u test verisi olarak ayrılmıştır, geri kalan veriler eğitim amacı ile kullanılmıştır. Modeller doğruluk (accuracy), hassasiyet (precision), geri çağırma (recall), f1 skoru (f1 score) ve karışıklık matrisine (confusion matrix) göre değerlendirilmiştir. Sonuçlar aşağıdaki tablolarda gösterilmektedir.

Belirtilen ölçüm değerlerinden bizim için en önemlileri geri çağırım (recall) ve karmaşıklık matrisi'dir (confusion matrix) diyebiliriz. Bu araştırma için önemli olan suistimalli dosyaların başarılı tahmin edilmesidir. Recall değeri bize suistimalli olan dosyaların ne kadar iyi tahmin edildiğini göstermektedir [25]. Karmaşıklık matrisi (confusion matrix) ise modelin genel başarısını yorumlamamızı sağlar.

Tablo 4 ve tablo 5'de sonuçları görülen ilk model, 20 bağımsız değişkenli veri seti ve K en yakın komşuluk algoritması kullanılarak geliştirilmiştir. K değeri 3, 5, 7 şeklinde arttırılarak eğitilen modelin doğruluk (accuracy), hassasiyet (precision), geri çağırım (recall) ve karmaşıklık matrisi (confusion matrix) ölçümleri kıyaslanmıştır.

Tablo 4. 20 bağımsız değişkenli veri seti ile K en yakın komşuluk için sonuçlar
(K nearest neighborhood results with 20 independent variables data set)

K Değeri	Accuracy(%)	Precision(%)	Recall(%)	F1 skoru(%)
K = 3	88.3	28	11	16
K = 5	89.1	34	6.4	11
K = 7	89.5	42	4.4	8

Tablo 5. 20 bağımsız değişkenli veri seti ile K en yakın komşuluk için karmaşıklık matrisi sonuçları
(Confusion matrix results for k nearest neighborhood with 20 independent variables data set)

K Değeri	Confusion Matrix		
K = 3		Gerçek Değerler	
		Evet	Hayır
	E	49	121
	H	403	3946
K = 5		Gerçek Değerler	
		Evet	Hayır
	E	29	55
	H	423	4012
K = 7		Gerçek Değerler	
		Evet	Hayır
	E	20	27
	H	432	4040

Tablo 4'de görüldüğü gibi K değeri arttıkça recall değeri düşmüş ve precision değeri artmıştır. Karmaşıklık matrisini (confusion matrix) değerlendirmek gerekirse; k değeri 3 seçildiğinde (k=3), 452 tane suistimalli olan kayıttan 49'u incelenmeli olarak tahmin edilmiştir, geri kalan 403 kayıt ise incelenmelerine gerek yok şeklinde

tahmin edilmiştir. Recall değeri bu sebepten yüzde 11, yani çok düşük çıkmıştır.

K değeri 5 olarak değiştirildiğinde recall değeri daha da düşmüş, precision değeri yükselmiştir. Karmaşıklık matrisini incelediğimizde 452 tane suistimalli dosyanın sadece 29 tanesi yakalanmıştır. 4067 tane suistimalli olmayan dosyanın ise 4012 tanesi incelenmeye gerek yok olarak tahmin edilmiştir. Bir önceki modele göre incelenmesi gerekenlerin başarı oranı azalmış, incelenmemesi gerekenler daha başarılı tahmin edilmiştir. Bu sebeplerden dolayı recall değeri düşmüş ve precision değeri artmıştır. K değerini 7 yaptığımızda eğitilen modelde diğer iki modeldeki değişime benzer bir değişim görülmüştür.

Tablo 6. 15 bağımsız değişkenli veri seti ile K en yakın komşuluk için sonuçlar
(K nearest neighborhood results with 15 independent variables dataset)

K Değeri	Accuracy(%)	Precision(%)	Recall(%)	F1 skoru(%)
K = 3	89.6	42	20	27
K = 5	90	57	15	24
K = 7	89.9	61	13	21

Tablo 7. 15 bağımsız değişkenli veri seti ile K en yakın komşuluk için karmaşıklık matrisi sonuçları
(Confusion matrix results for k nearest neighborhood with 15 independent variables dataset)

K Değeri	Confusion Matrix		
K = 3		Gerçek Değerler	
		Evet	Hayır
	E	91	127
	H	361	3940
K = 5		Gerçek Değerler	
		Evet	Hayır
	E	70	53
	H	382	4014
K = 7		Gerçek Değerler	
		Evet	Hayır
	E	57	37
	H	395	4030

Tablo 6 ve tablo 7'de sonuçları görülen modeller, 15 bağımsız değişkenli veri seti ve K en yakın komşuluk algoritması kullanılarak geliştirilmiştir. K değeri 3, 5, 7 şeklinde arttırılarak eğitilen modelin doğruluk (accuracy), hassasiyet (precision), geri çağırım (recall) ve karmaşıklık matrisi (confusion matrix) ölçümleri kıyaslanmıştır.

K değeri 3 seçildiğinde (k=3), bir önceki modelde 452 tane suistimalli olan kayıttan 49'u incelenmeli olarak tahmin edilmişken, bu modelde 91 tanesi incelenmeli olarak tahmin edilmiştir. 4067 tane suistimalli olmayan dosyanın ise 3940 tanesi incelenmeye gerek yok olarak tahmin edilmiştir. Bu sonuçlara göre diğer veri setinde geliştirilen modellerden daha başarılı bir model geliştirilmiştir denebilir.

K değeri arttıkça diğer veri seti ile geliştirilen modellerdekine benzer bir şekilde recall değerleri düşmüş ve precision değeri artmıştır. K değeri arttıkça suistimalli dosyaların tahmin edilmesindeki başarı oranı azalırken, suistimalsiz dosyaların tahminindeki başarı oranı artmıştır.

Tablo 8. 20 bağımsız değişkenli veri seti ile lojistik regresyon için sonuçlar
(Logistic regression results with 20 independent variables dataset)

Accuracy (%)	Precision (%)	Recall (%)	F1 skoru (%)	Confusion Matrix		
90	70	14	24		Gerçek Değerler	
					Evet	Hayır
				E	67	28
				H	385	4039

Tablo 8'de 20 bağımsız değişkenli veri seti ve lojistik regresyon algoritması ile eğitilmiş modele ait sonuçlar görülmektedir. 452 tane suistimalli dosyanın 67 tanesi incelenmeli olarak doğru tahmin edilmiştir. 4067 tane suistimalli olmayan dosyanın ise 4039 tanesi incelenmeye gerek yok olarak tahmin edilmiştir. Recall değeri yüzde 14, f1 skoru'da yüzde 24 olarak hesaplanmıştır.

Tablo 9. 15 bağımsız değişkenli veri seti ile lojistik regresyon için sonuçlar
(Logistic regression results with 15 independent variables dataset)

Accuracy (%)	Precision (%)	Recall (%)	F1 skoru (%)	Confusion Matrix		
91	84	24	37		Gerçek Değerler	
					Evet	Hayır
				E	109	20
				H	343	4047

Tablo 9'da 15 bağımsız değişkenli veri seti ve lojistik regresyon algoritması ile eğitilmiş modele ait sonuçlar görülmektedir. 452 tane suistimalli dosyanın 109 tanesini incelenmeli olarak doğru tahmin edilmiştir. 4067 tane suistimalli olmayan dosyanın ise 4047 tanesi incelenmeye gerek yok olarak tahmin edilmiştir. Recall değeri yüzde 24 ve f1 skoru yüzde 37 olarak hesaplanmıştır ve bir önceki modele göre daha iyi sonuçlar alındığı gözlemlenmiştir.

Bu aşamaya kadar geliştirilen modeller arasında en iyi model, 15 bağımsız değişkenli veri seti ve lojistik regresyon algoritması kullanılarak eğitilen model olmuştur.

Tablo 10. 20 bağımsız değişkenli veri seti ile karar ağacı için sonuçlar
(Decision tree results with 20 independent variables dataset)

Accuracy (%)	Precision (%)	Recall (%)	F1 skoru (%)	Confusion Matrix
87	37	33	35	

	Gerçek Değerler	
	Evet	Hayır
E	150	255
H	302	3812

Tablo 10'da 20 bağımsız değişkenli veri seti ve karar ağacı algoritması ile eğitilmiş modele ait sonuçlar görülmektedir. Bu modelde 452 tane suistimalli dosyanın 150 tanesi incelenmeli olarak doğru tahmin edilmiştir. 4067 tane suistimalli olmayan dosyanın ise 3812 tanesi incelenmeye gerek yok olarak tahmin edilmiştir. Precision değeri yüzde 37, recall değeri yüzde 33, f1 skoru'da yüzde 35 olarak hesaplanmıştır.

Tablo 11. 15 bağımsız değişkenli veri seti ile karar ağacı için sonuçlar
(Decision tree results with 15 independent variables data set)

Accuracy (%)	Precision (%)	Recall (%)	F1 skoru (%)	Confusion Matrix
90	61	27	37	

	Gerçek Değerler	
	Evet	Hayır
E	121	76
H	331	3991

Tablo 11'de 15 bağımsız değişkenli veri seti ve Karar ağacı algoritması ile eğitilmiş modele ait sonuçlar görülmektedir. 452 tane suistimalli dosyanın 121 tanesi incelenmeli olarak doğru tahmin edilmiştir. 4067 tane suistimalli olmayan dosyanın ise 3991 tanesi incelenmeye gerek yok olarak tahmin edilmiştir. Precision değeri yüzde 61, recall değeri yüzde 27, f1 skoru'da yüzde 37 olarak hesaplanmıştır. Tablo 10 ve tablo 11'deki eğitilen modelleri kıyasladığımızda, ilk model suistimalli dosyaları daha iyi tahmin ederken, genel olarak modellerin performansı açısından değerlendirildiğinde yüzde 37'lik f1 skoru ile ikinci model daha iyi sonuç vermiştir.

Tablo 12'de 20 bağımsız değişkenli veri seti ve yapay sinir ağı (multilayer perceptron) algoritması ile eğitilmiş modele ait sonuçlar görülmektedir. Bu modelde 452 tane suistimalli dosyanın 102 tanesi incelenmeli olarak doğru tahmin edilmiştir. 4067 tane suistimalli olmayan dosyanın ise 4036 tanesi incelenmeye gerek yok olarak tahmin edilmiştir. Precision değeri yüzde 77, recall değeri yüzde 22, f1 skoru'da yüzde 35 olarak hesaplanmıştır.

Tablo 12. 20 bağımsız değişkenli veri seti ile yapay sinir ağı için sonuçlar
(Multilayer perceptron results with 20 independent variables dataset)

Accuracy (%)	Precision (%)	Recall (%)	F1 skoru (%)	Confusion Matrix
91.2	77	22	35	

	Gerçek Değerler	
	Evet	Hayır
E	102	31
H	350	4036

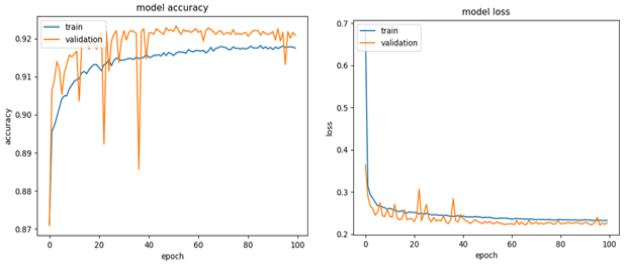
Tablo 13. 15 bağımsız değişkenli veri seti ile yapay sinir ağı için sonuçlar
(Multilayer perceptron results with 15 independent variables dataset)

Accuracy (%)	Precision (%)	Recall (%)	F1 skoru (%)	Confusion Matrix
91.5	79	22	35	

	Gerçek Değerler	
	Evet	Hayır
E	100	26
H	352	4041

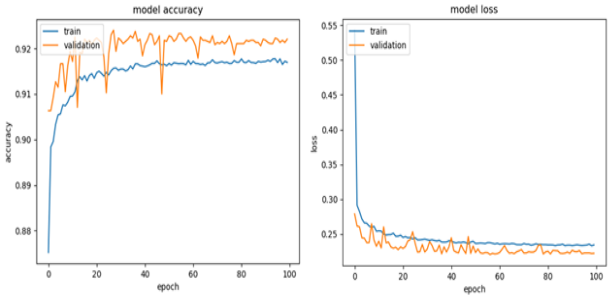
Tablo 13'de 15 bağımsız değişkenli veri seti ve yapay sinir ağı (multilayer perceptron) algoritması ile eğitilmiş modele ait sonuçlar görülmektedir. 452 tane suistimalli dosyanın 100 tanesi incelenmeli olarak doğru tahmin edilmiştir. 4067 tane suistimalli olmayan dosyanın ise 4041 tanesi incelenmeye gerek yok olarak tahmin edilmiştir.

Tablo 13'de 15 bağımsız değişkenli veri seti ve yapay sinir ağı (multilayer perceptron) algoritması ile eğitilmiş modele ait sonuçlar görülmektedir. 452 tane suistimalli dosyanın 100 tanesi incelenmeli olarak doğru tahmin edilmiştir. 4067 tane suistimalli olmayan dosyanın ise 4041 tanesi incelenmeye gerek yok olarak tahmin edilmiştir. Precision değeri yüzde 79, recall değeri yüzde 22, f1 skoru'da yüzde 35 olarak hesaplanmıştır.



Şekil 4. 20 bağımsız değişkenli veri seti ile yapay sinir ağı accuracy & loss
(Multilayer perceptron accuracy & loss with 20 independent variables dataset)

Yapay sinir ağı modelleri, 20 giriş nöronu, 20 gizli nöron ve 1 çıkış nöronu kullanılarak tasarlanmıştır. Hedef değişkenimiz, boolean (doğru (1) veya yanlış (0)) türünde bir değişken olduğu için çıkış nöronunda aktivasyon fonksiyonu olarak sigmoid kullanılmıştır. Ağ ağırlıklarını güncellemek için adam optimizer algoritması kullanılmıştır. Hiper parametreler, makalelerde tavsiye edildiği gibi varsayılan (default) değerleri ile bırakılmıştır. Öğrenme hızı (Learning rate) 0.001 olarak seçilmiştir. Aynı anda eğitilecek veri miktarı (batch size) 60 olarak, epoch değeri ise 100 olarak seçilmiştir. Şekil 4 ve Şekil 5'de görüldüğü gibi her epoch sonunda modelin doğruluk (accuracy) ve kayıp (loss) değerleri ile performansı ölçülmüştür.



Şekil 5. 15 bağımsız değişkenli veri seti ile yapay sinir ağı accuracy & loss
(Multilayer perceptron accuracy & loss with 15 independent variables dataset)

Şekil 4 ve Şekil 5'de doğruluk (accuracy) oranı ve kayıp (loss) değerlerin grafiklerini incelediğimizde, eğitim (train) verileri ile doğrulama (validation) verileri arasında, değişimin orantılı bir şekilde olduğunu, doğruluk (accuracy) açısından yükselen eğilimde, kayıp (loss) değerleri açısında düşüş eğilimde olduğunu ve aralarında çok fark olmadığını gözlemliyoruz. Bu da geliştirdiğimiz modellerimizin ezberlemeye (overfitting) başlamadığını, başarılı bir şekilde eğitildiğini göstermektedir.

5. TARTIŞMA VE SONUÇ (DISCUSSION AND CONCLUSIONS)

Bu çalışmada; sigorta sektörü için önemli suistimal türleri arasında yer alan sahte hasarların hasar ekiplerince manuel incelenerek tespiti dışında, bu sürece destek olacak nitelikte bir makine öğrenmesi modeli ile tespit edilme ihtimallerini artırmak amaçlanmıştır.

Veri seti özel bir sigorta şirketine ait veriler kullanılarak hazırlanmıştır. Sigorta şirketine ait veriler, ihbar tarihi 01.01.2013 - 01.04.2016 arasında olan suistimalli/suistimalli değil şeklinde etiketlemiş toplamda 45190 tane dosyadan oluşmaktadır. Toplamda elimizde bulunan verilerin 4519 tanesi suistimalli olduğu bilinen dosyalardır. İlk veri setimiz 20 tane bağımsız değişkenden ve 1 tane hedef değişkenimizden oluşmaktadır. Bu değişkenler için regresyon analizi yapılıp anlamlılık seviyeleri ölçülmüştür. 0.05 anlamlılık seviyesinin altında olan değişkenler geriye doğru eleme yöntemi kullanılarak 15 bağımsız değişkene düşürülmüştür ve 15 bağımsız değişkenli ikinci bir veri seti oluşturulmuştur. Bu iki veri seti ile modeller eğitilmiş ve sonuçları kıyaslanmıştır.

Modellerin başarısını etkileyen en önemli unsur veri kalitesidir. Etiketlenmiş veriler içinde 4519 tane suistimalli dosyanın olması ve bu örnekler dışında daha fazla verinin olmamasından dolayı modelin daha çok veri ile başarısının artırılması sağlanamamıştır. Veri sayısının azlığı, güncel veri olmaması ve kayıp veriler sebebi ile modellerin başarı oranları azalmıştır. Sigorta şirketleri hatalı ve kayıp veriye sebep olan unsurları incelemeli ve kayıp verilerin azaltılması için çalışma yapmalıdırlar. Hasar ekiplerinin dosya incelemesinde kullandıkları ve veri girişi yaptıkları yazılımları için hatalı veya eksik bilgi girmelerini önleyecek geliştirmeler yapılması ve hata önleyici kontrollerin koyulması veri kalitesini arttıracaktır. Her yılın sonunda yeni ve güncel verilerle modelin tekrar eğitilmesi modelin güncel kalmasını sağlayacaktır. Yeni gelen verilerde yakalanmış sahte hasara ait bilgiler ile etiketli verimizin sayısı artacaktır. Bu sayede yeni gelen verilerin modele katılmasıyla ve doğru tutulan kayıpsız veriler ile modelin başarısının artacağı düşünülmektedir.

Araştırmada elde edilen sonuçlara göre *15 bağımsız değişkenli veri seti ile* eğitilen lojistik regresyon, karar ağacı ve yapay sinir ağı modelleri iyi sonuçları vermiştir. *20 bağımsız değişkenli veri seti ile* sadece yapay sinir ağı modelinde diğerlerinden daha iyi sonuç elde edilmiştir. Yüzde 91 başarı oranı ve yüzde 84 hassasiyet (precision) değeri ile sahte hasar olmayan kayıtları en iyi bulan model lojistik regresyon olmuştur. Yüzde 87 başarı oranı ve yüzde 33 geri çağırma (recall) değeri ile sahte hasarları en iyi tahmin eden model karar ağaçları modeli olmuştur. Yapay sinir ağında ise yüzde 91.2 başarı oranı, yüzde 77 hassasiyet (precision) ve yüzde 35 f1 skoru ile diğer başarılı modellere yakın sonuçlar elde edilmiştir. Bu modellerin hasar ekiplerine sahte hasarların tahmininde yardımcı olabileceği görülmüştür.

KAYNAKLAR (REFERENCES)

[1] L. Šubelj, Š. Furlan, M. Bajec, "An expert system for detecting automobile insurance fraud using social network analysis", *Expert Systems with Applications*, 38(1), 1039-1052, 2011.

[2] S. Erol, **Hile Denetiminde Proaktif Yaklaşımlar**, Yüksek Lisans Tezi, İstanbul Ticaret Üniversitesi, Sosyal Bilimler Enstitüsü, 2016.

[3] M. Ö. Dolgun., B.Cenk, A. A. Koç, "Sigortacılık Sektöründe Araç Sigortalarında Suistimal Tespit Sistemi", 2. **Ulusal Sigorta ve Aktüerya Kongresi**, Karabük, 28-29 Eylül, 2015.

[4] E. Kasap, **Sigortacılık Sektöründe Müşteri İlişkileri Yönetimi Yaklaşımıyla Veri Madenciliği Teknikleri ve Bir Uygulama,** Yüksek Lisans Tezi, Marmara Üniversitesi, Bankacılık Ve Sigortacılık Enstitüsü, 2007.

[5] D. Muslu, **Sigortacılık Sektöründe Risk Analizi: Veri Madenciliği Uygulaması**, Yüksek Lisans Tezi, İstanbul Teknik Üniversitesi, Fen Bilimleri Enstitüsü, 2009.

[6] Y. Kaya, **Motokaravan Sigortacılığı Tahmin Modellemesi ve Uygulanan Yöntemlerin Karşılaştırılması**, Yüksek Lisans Tezi, Beykent Üniversitesi, Fen Bilimleri Enstitüsü, 2017.

[7] İ. Şişaneci, **Sağlık Sisteminde Veri Madenciliği ile Suistimal Tespiti**, Yüksek Lisans Tezi, Gebze Yüksek Teknoloji Üniversitesi, Fen Bilimleri Enstitüsü, 2009.

[8] A. Yılmaz, **Sahte Hasarların Lojistik Regresyon Analizi ile Tahmini**, Yüksek Lisans Tezi, İstanbul Üniversitesi, Sosyal Bilimler Enstitüsü, 2014.

[9] R. Bhowmik, "Detecting Auto Insurance Fraud by Data Mining Techniques", *Journal of Emerging Trends in Computing and Information Sciences, Computer Sciences*, 2(4), 156-162, 2011.

[10] A. R. Bauder, M. T. Khoshgoftaar, "The Detection of Medicare Fraud Using Machine Learning Methods with Excluded Provider Labels", **The Thirty-First International Florida Artificial Intelligence Research Society Conference**, College of Engineering & Computer Science Florida Atlantic University, FLAIRS-31, A.B.D. 2018.

[11] T. Martin, J. Biegelman, T. Bartow, **Executive Roadmap to Fraud Prevention and Internal Control**, John Wiley & Sons, New Jersey, A.B.D. 23, 2006.

[12] Y. Wang, W. Xu, "Leveraging deep learning with LDA-based text analytics to detect automobile insurance fraud", *Decision Support Systems*, (105), 87-95, 2018.

[13] Y. Li, C. Yan, W. Liu, M. Li, "A principle component analysis-based random forest with the potential nearest neighbor method for automobile insurance fraud identification", *Applied Soft Computing*, (70), 1000-1009, 2018.

[14] Internet: S. Aligil, Cumhuriyet Gazetesi Ekonomi Bölümü, https://www.cumhuriyetarsivi.com/monitor/index.xhtml, 11.11.2018.

[15] İnternet: Sigorta Sahteciliklerini Engelleme Bürosu, Suistimal Yöntemleri Grafiği, https://siseb.sbm.org.tr/tr/istatistikler, 25.10.2018.

[16] S. Hipgrave, "Smarter fraud investigations with big data analytics", *Network Security*, (12), 7-9, 2013.

[17] S. Subudhi, S. Panigrahi, "Use of optimized Fuzzy C-Means clustering and supervised classifiers for automobile insurance fraud detection*", Journal of King Saud University - Computer and Information Sciences*, 32(5), 568-575, 2020.

[18] M. Artis, M. Ayuso, M. Guillen, "Detection of Automobile Insurance Fraud with Discrete Choice Models and Misclassified Claims", *Journal of Risk & Insurance*, (69), 325-340, 2002.

[19] L. Bai, J. Cai, M. Zhou, "Optimal reinsurance policies for an insurer with a bivariate reserve risk process in a dynamic setting", *Insurance: Mathematics and Economics*, 53(3), 664-670,2013.

[20] S. Şahinler, "En Küçük Kareler Yöntemi ile Dogrusal Regresyon Modeli Oluşturmanın Temel Prensipleri", *Mustafa Kemal Üniversitesi Ziraat Fakültesi Dergisi,* (5), 57-73, 2000.

[21] C. Yan, Y. Li, W. Liu, M. Li, J. Chen, L. Wang, "An artificial bee colony-based kernel ridge regression for automobile insurance fraud identification", *Neurocomputing*, (393), 115-125, 2020.

[22] M. K. Ayyüce, B. Bolat, "Derin Öğrenme ile Kalabalık Analizi Üzerine Detaylı Bir Araştırma", *Bilişim Teknolojileri Dergisi*, 11(3), 263-286, 2018.

[23] M. Eminağaoğlu , A, Vahaplar, "Turnaround Time Prediction for a Medical Laboratory Using Artificial Neural Networks", *Bilişim Teknolojileri Dergisi*, 11(4), 357-368, 2018.

[24] E. Seninç, "The Effect of Hidden Neurons in Single-Hidden Layer Feedforward Neural Networks", *Bilişim Teknolojileri Dergisi*, 12(4), 277-286, 2019.

[25] L. Guelman, "Gradient boosting trees for auto insurance loss cost modeling and prediction", *Expert Systems with Applications* 39(3), 3659-3667, 2012.