Dicle University
Journal of Engineering
https://dergipark.org.tr/tr/pub/dumf
duje.dicle.edu.tr

Araştırma Makalesi

Banka Ödemelerinde Dolandırıcılığın Çizge Madenciliği ve Makine Öğrenimi Algoritmalarıyla Tespiti

Detection of Fraud in Bank Payments Using Graph Mining and Machine Learning Algorithms

Hande ÇAVŞİ ZAİM[1,*], Esra Nergis YOLAÇAN[2], Eyyüp GÜLBANDILAR[3]

[1] Eskişehir OsmanGazi Üniversitesi, Bilgisayar Mühendisliği Bölümü, Eskişehir, ORCID iD 0000-0002-9032-5145
[2] Eskişehir OsmanGazi Üniversitesi, Bilgisayar Mühendisliği Bölümü, Eskişehir, ORCID iD 0000-0002-1655-0993
[3] Eskişehir OsmanGazi Üniversitesi, Bilgisayar Mühendisliği Bölümü, Eskişehir, ORCID iD 0000-0001-5559-5281

MAKALE BİLGİLERİ

Makale Geçmişi:

Geliş 8 Agustos 2021
Revizyon 14 Eylül 2021
Kabul 25 Eylül 2021
Online 28 Eylül 2021

Anahtar Kelimeler:

Dolandırıcılık Tespiti, Çizge Madenciliği, Veri Analizi, Makine Öğrenimi Algoritmaları, Çizge Madenciliği Algoritmaları

ÖZ

Günümüzde, şirketler gelecekte yapmayı planladıkları işleri içeren çok sayıdaki önemli verilerini elektronik ortamlarda saklamaktadırlar. Saldırı durumunda ise hem şirkete hem de bireylere zarar verebilecek finansal bilgiler hedef alınmaktadır. Bu saldırı türlerinden biri de banka ödemelerinde meydana gelen dolandırıcılık saldırılarıdır. Grafik veri bilimi kullanılması, mevcut analitik ve makine öğrenimi ardışık düzenlerini güçlendirerek, var olan dolandırıcılık tespit yöntemlerinin doğruluğunu ve uygulanabilirliğini arttırmaktadır. Bu çalışmada İspanya'daki bir banka ödeme bilgi simülasyonundan oluşturulan BankSim veri kümesi kullanılmıştır. BankSim üzerinde bulunan normal ödemeler ve sahte veriler sınıflandırılarak dolandırıcılık tespiti gerçekleştirilmesi amaçlanmıştır. Sınıflandırma için Python dilinde RandomForest (RF), Support Vector Machine SVM, XGBoost (XGB), K-Nearest Neighbors (k-NN) sınıflandırma algoritmaları kullanılmıştır. Performans değerlendirmeleri için K-katlamalı çapraz doğrulama kullanılmıştır. Çizge madenciliği için Neo4j veritabanı kullanılmış ve Neo4j sorgu dili olarak CypherQL kullanılmıştır. Bu dolandırıcılık tespitinin uygulanması ile daha az hileli işlem ve daha güvenilir bir gelir akışı elde edilmiştir. Çizge madenciliği aşamasında PageRank, Community, degree gibi çizge algoritmaları ile birlikte standart makine öğrenimi yöntemi ile elde edilen sonuçlar optimize edilmiştir. Bu açıdan çizge madenciliği ve makine öğrenimi algoritmalarının birlikte kullanılmasının diğer yöntemlere kıyasla doğruluk oranlarının daha yüksek olduğu ve daha hızlı sürede hesap yapan bir yöntem olduğu ispatlanmıştır.

ARTICLE INFO

Article history:

Received 08.08.2021
Received in revised form 09.08.2021
Accepted 14.09.2021
Available online

Keywords:

Fraud Detection, *Graph Mining, Data Mining, Machine Learning Algorithms, Graph Mining Algorithms.*

Doi: 10.24012/dumf.1002110

* Sorumlu Yazar

ABSTRACT

Today companies keep a large number of important data, including the work they plan to do in the future, electronically. In many cases, financial information is stolen that can harm the entire company or individual. One of these types of fraud occurs in bank payments. The use of graph data science augments existing analytics and machine learning pipelines, increasing the accuracy and applicability of existing fraud detection methods. In this study, BankSim dataset created from a bank payment information simulation in Spain was used. It is aimed to detect fraud by classifying normal payments and injected fraud signatures on BankSim. RandomForest (RF), SVM, XGBoost (XGB), K Nearest Neighbors (k-NN) classification algorithms in python language were used for classification. K-fold cross validation was used for performance evaluations. Neo4j database was used for graph analysis and CypherQL was used as Neo4j query language. The implementation of this fraud detection has resulted in fewer fraudulent transactions and a more reliable revenue stream. The performances of SVM, RF, XGB, k-NN algorithms were evaluated for fraud detection in bank payments, and the performance of the algorithms was compared according to the K-Fold cross-validation results in terms of performance. In the graph mining phase, the results obtained with the standard machine learning method were optimized together with graph algorithms such as PageRank, Community, and degree. In this respect, it has been proven that the use of graph mining and machine learning algorithms together has higher accuracy rates compared to other methods and is a method that calculates in a faster time.

Giriş

Bankalar ve sigorta şirketleri her yıl dolandırıcılık nedeniyle milyonlarca dolar zarara uğramaktadırlar. Geleneksel dolandırıcılık tespit yöntemleri, bu kayıpların en aza indirilmesinde önemli bir rol oynasa da giderek daha karmaşık hale gelen dolandırıcılık saldırıları hem birlikte çalışarak hem de sahte kimlikler inşa etmenin çeşitli yollarını kullanarak tespit edilmesi zor hale gelmiştir [1]. Hiçbir dolandırıcılık önleme yöntemi mükemmel olarak çalışmasa da tek tek veri noktalarının ötesinde onları birbirine bağlayan bağlantılara bakılarak önemli iyileştirme fırsatları yakalanabilir [2]. Çizge veri madenciliği, alt grafikler üzerinde yoğunlaşan ve sık model madenciliğinden türemiştir ve çizge alt grafik madenciliği, çizge madenciliğinin popüler bir uzantısıdır [3, 4, 5]. Standart makine öğrenimi yöntemleri çeşitli veri kaynaklarından veri toplanması ve bunun sonucunda artan veri boyutunu işlemek ve görselleştirmek konusunda yetersizdir [6, 7]. Ayrıca gelişen dolandırıcılık saldırıları karşısında standart makine öğrenimi yöntemleri bu saldırıdaki sahte işlemlerin bazılarını yakalayamayabilir. Bu problemlere çözüm olarak bu çalışmada çizge veri tabanlarının ve çizge algoritmalarının kullanılmasıyla dolandırıcılık tespitinde mevcut olarak kullanılan makine öğrenimi yöntemlerinin doğruluğunun arttırılması ve çalışmanın uygulanabilirliğinin kolaylaştırılması hedeflenmiştir. Bu çalışmada BankSim veri seti üzerinde bulunan normal ödemeler ve sahte ödemeler RF, SVM, XGB, k-NN sınıflandırma algoritmalarıyla sınıflandırılmıştır. Dolandırıcılık tespiti gerçekleştirilmesi amacıyla klasik makine öğrenimi algoritmaları ve çizge veri tabanları kullanılarak geleneksel dolandırıcılık tespiti yöntemleriyle tespit edilmesi zor olan dolandırıcılık işlemlerinin tespit edilmesi ve veriler arasındaki ilişkinin diğer yöntemlere kıyasla daha iyi ifade edilmesi amaçlanmıştır.

Bu çalışmada, literatürdeki diğer çalışmalardan farklı olarak, makine öğrenimi tekniklerinin çizge madenciliği algoritmalarıyla birlikte kullanılması sonucu dolandırıcılık tespiti işlemlerinin performansı üzerinde nasıl bir etkisi olduğu ayrıntılı olarak incelenmiştir. Sunulan çalışma bu alanda çizge madenciliği ve makine öğrenimi algoritmalarının birlikte kullanılarak daha iyi sonuçlar elde edebileceğini gösterme konusunda literatüre katkı sağlamaktadır.

Çalışmanın literatür taraması bölümü kredi ve banka kartı işlemleri için dolandırıcılık tespiti yapan literatürdeki güncel çalışmaları sunmaktadır. Metodoloji bölümünde çalışmada kullanılan veri setine, sınıflandırma algoritmalarına, veri ön işleme yöntemlerine ve kullanılan araçlara genel bir bakış açısıyla yer verilmektedir. Neo4j ve veri seti bölümü veri setinin detaylarını, Neo4j aracının tanımını ve veri setinin görselleştirilmesinde kullanılması konularını içermektedir. Veri ön işleme bölümü veri ön işleme aşamalarının detaylarını içermektedir. Makine öğrenimi ve çizge algoritmaları bölümü kullanılan makine öğrenimi sınıflandırma algoritmaları ve çizge algoritmalarının detaylarını içermektedir. Son olarak Dolandırıcılık tahmininde yanlış pozitifleri azaltmak amacıyla Roy ve ark. [12] otomatik özellik mühendisliği tabanlı bir yaklaşım önermiştir. Bu çalışmada yazarlar işlemlerle ilişkili yapılan deney sonuçlarına deney sonuçları bölümünde yer verilmiştir ve bu çalışmada elde edilen sonuçlar literatürde aynı ön işleme ve veri seti kullanan makine öğrenimi yöntemlerinin sonuçlarıyla birlikte listelenmiştir.

Literatür taraması

Son yıllarda dijital ödemelerdeki benzeri görülmemiş büyüme, dolandırıcılık ve mali suçlarda önemli değişiklikleri tetiklemiştir. Bu yeni ortamda kural tabanlı gibi geleneksel tespit yaklaşımları ise, büyük ölçüde etkisiz hale gelmiştir ve yapay zekâ makine öğrenimi çözümleri büyük ilgi görmüştür. Kurshan ve ark. [8] dolandırıcılık ve mali suç tespitinde grafik tabanlı çözümlerin karşılaştığı yaygın uygulama hususlarına ve kapsamlı uygulama zorluklarına genel bir bakış açısı sunmuştur. Çalışmada ödeme dolandırıcılığı, kimlik hırsızlığı, mali ve eski dolandırıcılar, hesap devralma, sentetik kimlik ve hesap dolandırıcılığı, kara para aklama ve diğer dolandırıcılık türlerine değinilmiştir. Dolandırıcılık algılama tekniklerinden veri madenciliği, grafik anormalliği algılama, denetimli ve denetimsiz alt grafik analizi, akış ve yol analizi, grafik tabanlı makine öğrenimi yöntemleri incelenmiştir. Dolandırıcılık tespitinde manuel yöntemlerin kullanılması özellik çıkarma işlemi aşamasında alan bilgisine ihtiyaç duymaktadır. Bu durum ise, çevrim içi dolandırıcılık tespiti sistemindeki dolandırıcılık davranışlarına en yakın dolandırıcılık davranış modellerine odaklanmayı gerektirir. Bu nedenle [9] kredi kartı dolandırıcılık tespiti için mekansal-zamansal dikkat tabanlı bir grafik ağı (STAGN) modeli önerilmiştir. Daha sonra 3 boyutlu bir evrişim ağına beslenen öğrenilmiş temsillerinin üzerine uzamsal-zamansal dikkat kullanılmıştır. Dikkat ağırlıkları evrişim ve algılama ağları ile uçtan uca olacak şekilde birlikte öğrenilir. Gerçek kelime kart işlemi veri seti üzerinde kapsamlı deneyler yapılmıştır ve STAGN'nin hem AUC (Area Under the Curve) hem de hassas geri çağırma (recall) eğrilerinde diğer son teknolojilere kıyasla daha iyi performans gösterdiği görülmüştür. STAGN öngörücü model olarak dolandırıcılık tespit sistemine entegre edilmiş ve sistemdeki her modelin uygulama detayı sunulmuştur.

Konum tabanlı grafik gösterimini öğrenmeye ilişkin son çalışmalar yan bilgi ile veriyi gömme ve gelişmiş bilginin korunması olarak iki yönde incelenmiştir [10]. Düşük boyutlu temsillerini öğrendiklerinde düğümlerle ilişkili zengin bilgileri (ör; metin, konum kodları) hesaba katan TADW'yi önermişlerdir. Bu ağ, ağ hakkındaki yan bilgileri (düğüm içerikleri ve kenar içerikleri) içermektedir.

Xie ve Yin [11]'de konuma dayalı bir öneri sistemi için çizge tabanlı point of interest (POI) yerleştirmeyi öğrenen bir sistem önermişlerdir. Konum tabanlı grafiklerin temsillerini öğrenebilen gelişmiş ağ yerleştirme yöntemleri kullanılarak ilk çabalar gerçekleştirilmiş olsa da konum tabanlı grafik temsillerini koruyarak dolandırıcılık işlemlerini tespit etmeye odaklanan çok az çalışma bulunmaktadır.

kartın geçmiş verilerine dayalı olarak davranış özelliklerini otomatik olarak türetmek için derin özellik sentezi algoritmasını kullanmışlardır. Oluşturulan özellikler rastgele

orman tekniği kullanılarak bir sınıflandırıcı tarafından öğrenilmiştir. Bu çalışmanın sonucunda yanlış pozitif oranında %54 düşüş gözlemlenmiştir.

Vidanelage ve ark. [13] tarihle sınırlı olmayan sentezlenmiş veri seti kullanarak çoklu makine öğrenme teknikleri yürütmüş ve açık kaynak veri bilimi araçları kullanarak beklentilerin üzerinde sonuçlar elde etmişlerdir. Bu çalışmada veri seti olarak İspanya'daki bir bankanın verileri baz alınarak oluşturulmuş BankSim veri seti kullanılmıştır ve k-NN, Çok katmanlı algılayıcı (MLP), Gauss Naive Bayes (GNB), Multinomial Naive Bayes (MNB) gibi 4 farklı makine öğrenimi algoritması dolandırıcılık tespiti için kullanılmıştır. MLP en yüksek doğruluktaki sonucu vermiştir.

Dolandırıcılık tespitini iyileştirmek için Carcillo ve ark., [14]'de hem denetlenen hem de denetimsiz yöntemleri birleştiren karma bir teknik önermişlerdir. Araştırmacılar dolandırıcılık modellerinin geçmişin analizinden öğrenilebileceği varsayımından yararlandıkça denetimli bir öğrenmenin mümkün olacağı beklenmiştir. Ancak müşteri davranışındaki ve dolandırıcıların yeteneklerindeki değişiklikler yeni dolandırıcılık kalıpları ürettiğinde zorluklar ortaya çıkmaya başlamıştır. Bu durumda denetimsiz öğrenme bu zorluklara yardımcı olabilir. Yazarlar çözüm olarak bu iki tekniği birleştiren hibrit bir yöntem kullanmışlardır ve sonuçlar verimli olmuştur.

Lebichot ve ark., [15]'da dolandırıcılık tespitinde kullanılmak üzere iki orijinal transfer öğrenme modeli önermişlerdir. Çalışma, sahtekarlık algılama alanında daha gelişmiş teknikleri incelemiştir. E- ticaret işlemleri kaynak ve yüz yüze işlemler sırasıyla kaynak ve hedef alan olarak ele alınmıştır. İstatistiksel testlere dayanarak önerilen modeller daha önceden önerilen tüm diğer modellerden daha iyi performans göstermiştir.

Wang ve Zhu, [16]'da e-ticaret işlemlerinden kaynaklı artan çevrim içi ödeme dolandırıcılık tespitine çözüm olarak, işlem özelliklerinin birlikte oluşma ilişkilerini modelleyerek etkili bir veri geliştirme planı önermişlerdir. Düşük kaliteli davranışsal verileri kullanarak yüksek çözünürlüklü davranış modelleri oluşturmak büyük bir zorluktur. Bu çalışmada bu sorun esas alınarak bir bilgi grafiği kullanılmış ve işlem niteliklerinin birliktelik ilişkileri kapsamlı bir şekilde belirtilmiştir. Önerilen yöntemin performansı ticari bir bankadan gerçek veri seti üzerinde yapılan deneylerle doğrulanmıştır. Bu çalışma öznitelik düzeyinden ağ yerleştirme algoritmaları çeşitlendirilmiş davranış modelleri için veri gerçekleştirmeye yönelik ilk çalışmadır. Önerilen yöntemin son teknoloji sınıflandırıcılardan daha iyi performans gösterdiği gözlemlenmiştir.

Shiguihara ve ark., [17]'de dolandırıcılık tespiti için denetimli makine öğrenme tekniklerini kullanmıştır. Alanla ilgili kısıtlamalar kullanılarak dolandırıcılık tespiti için olasılıklı bir grafik model oluşturulması amacıyla bir yöntem önerilmiştir. Çalışmada BankSim veri seti üzerinde K2 arama, Hill-Climbing, Simulated Annealing gibi Bayes Ağları algoritmaları kullanılmıştır.

Lopez ve Axelsson, [18]'de Banksim veri seti üzerinde k-NN, XGB, RF makine öğrenimi algoritmaları kullanarak dolandırıcılık tespiti gerçekleştirmişlerdir. Bu çalışmada dengesiz veri seti undersampling yöntemiyle dengeli hale getirilmiştir. K-NN precision (hassasiyet) değeri 0.83, recall (duyarlılık) değeri 0.61, f1 skor değeri 0.70 olarak bulunmuştur. Aynı değerler sırasıyla XGB için; 0.89, 0.76, 0.82 olarak bulunurken, RF için aynı değerler sırasıyla; 0.24, 0.98, 0.82 olarak bulunmuştur.

Islam, [19]'da kredi ödemelerindeki sahtecilik tespiti için en etkili yöntemleri ve makine öğrenimi algoritmalarını ele almıştır. Banksim veri seti üzerinde undersampling yöntemi ile veri setini dengeledikten sonra k-NN için kesinlik değerini 0.80, SVM için kesinlik değerini 0.77, RF için kesinlik değerini 0.93 olarak elde etmiştir. K-NN için precision, recall, f skor değerleri sırasıyla; 0.80, 0.80, 079 olarak, SVM için aynı değerler sırasıyla; 0.76, 0.77, 0.76 olarak, RF için ayı değerler sırasıyla; 0.93, 0.93 ve 0.93 olarak elde edilmiştir.

Tablo 1.'de incelenen literatür çalışmalarının bir karşılaştırılması sunulmaktadır. Olasılıklı grafik modellerin diğer temel tekniklere göre %99.272 doğrulukla daha iyi performans gösterdiği gözlemlenmiştir.

Tablo 1. Literatür Taraması.

Referans	Methodlar	Algoritmalar	Veri Seti	Metrikler
Lopez ve Axelsson, 2014 [18]	Makine Öğrenmesi	RF, XGB, k-NN	Banksim	Precision, Recall, F1 Skor
Yang ve ark., 2015 [10]	TADW (Metin Tabanlı Derin Yürüyüş)	Derin Yürüş PLSA Transdüktif SVM SVM	Cora, Wiki, Citeseer	Matrix Çarpanlara Ayırma
Xie ve ark., 2016 [11]	POI (Grafik tabanlı ilgi noktası iyileştirme)	İki parçalı grafik gömme optimizasyon ortak yerleştirme öğrenim	Foursquare, Gowalla	Precision Recall Accuracy Rank
Roy ve ark., 2017 [12]	Derin Öğrenme	Yapay Sinir Ağları Özyinelemeli Sinir Ağları Uzun-kısa vadeli bellek Geçişli Özyinelemeli Birim	Parakende bankacılık yapan bir firmanın hesap ve işlem ayrıntıları	Precision Recall Accuracy
Shiguihara ve ark., 2018 [17]	Makine Öğrenmesi	Naive Bayes (NB) Hill-Climbing Simulated Annealing K2 Arama	Banksim	Precision Recall Accuracy
Islam., 2018 [19]	Makine Öğrenmesi	k-NN, SVM, RF	Banksim	Precision Recall F skor Accuracy
Lebichot ve ark., 2019 [15]	Derin transfer öğrenme	Derin sinir ağı algoritmaları	Bir kuruluş tarafından elde edilmiş 5 aylık e-ticaret ve yüz yüze işlem bilgileri	Precision Recall AUC-PR
Wang ve Zhu., 2020 [16]	İşlem kayıtlarının grafik gösterimi Network embedding Dolandırıcılık tespit modelleri	Çok-ajanlı davranışsal modeller Tek-ajanlı davranışsal model Lojistik regresyon (LR) RF, XGB, CNN, NB	Çin'deki ticari bankalar tarafından elde edilmiş 3 aylık B2C ve C2C işlem kayıtları	Precision Recall AUC-ROC
Cheng ve ark., 2020 [9]	Mekansal-zamansal dikkat tabanlı bir grafik ağı (STAGN)	LR, GradientBoosting, MLP, AdaBoost, STAGN	1 Ocak-31 Aralık 2016 tarihleri arasında toplanan gerçek kelime kredi kartı işlem kayıtları	AUC, Recall

İncelenen literatür çalışmaları göz önüne alındığında çizge tabanlı yöntemlerin makine öğrenimi yöntemlerine kıyasla daha yüksek doğruluk oranı sağladığı ve daha hızlı geri dönüşler gerçekleştirdiği gözlenmiştir. Bu çalışmada ise literatürdeki çalışmalardan farklı olarak, makine öğrenimi algoritmaları çizge algoritmaları ile birlikte kullanılarak performans sonuçları optimize edilmiştir. Bu çalışma makine öğrenimi algoritmalarının birlikte kullanılarak daha iyi sonuçlar elde edebileceğini gösterme konusunda literatüre katkı sağlamaktadır.

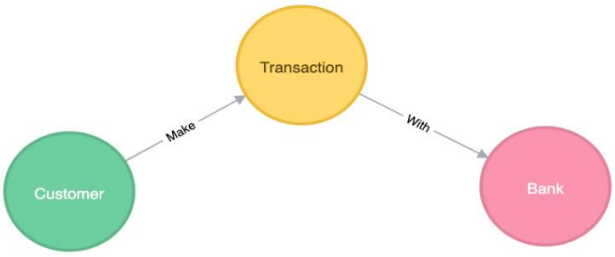
Metodoloji

Grafik veri bilimi, aramaları, sorguları ve çizge algoritmalarını kullanarak ağ yapılarının keşfedilmesini ve analiz edilmesini sağlar. Ayrıca ayrık matematiğin bir alt alanı olan grafik teorisinden yararlanarak, dolandırıcılık tahmin doğruluğunu arttırır ve finans hizmetleri firmalarına küçük yüzdelik oranlarına sahip doğruluk artışı sağlasa bile firmalar milyonlarca dolar tasarruf elde edebilir [20].

Dolandırıcılık tespiti için kullanılacak veriler çeşitli yerlerden çeşitli formatlarla toplansa bile kendi başlarına bir değer sağlamazlar. Verileri anlamlandırmak için veriler arası iletişim kurularak ve veriler düzenlenerek bilgi üretimi sağlanmalıdır. Bu bilgiler ve aralarındaki ilişkiler oldukça karmaşıktır ve dönüşüm işlemi oldukça çaba gerektirir. Uzun bir dönüşüm işleminin sonunda veri gerçek değerini bulur ve bu bilgiden değer elde edilerek eyleme dönüştürülebilen bir öğrenme elde etmek zekâ gerektirir. Bu zekâ makine öğrenimidir [21]. Grafik modeli iyi analiz edilmiş verinin yönetimine yardımcı olmaktadır. İlişkisel bir veritabanı sorgulamak veya NoSQL veri tabanındaki bir değerden anahtar çıkarmak karmaşık bir iştir ve grafiklerin sorgulanması bu açıdan kolaydır, birden fazla kaynaktan gelen verileri birleştirebilir ve eğitimde kullanılacak değişken listesinin bulunmasını ve çıkarılmasını kolaylaştırır. Model oluşturma sürecini hızlandırır, verileri dış bilgi kaynaklarıyla kolaylıkla birleştirir ve veriler istenilen formatta dışarı aktarılabilir. Tüm bu avantajlarından dolayı bu çalışmada dolandırıcılık analizi yapılırken makine öğrenimi yöntemleri çizge madenciliği yöntemleriyle birleştirilerek analizin doğruluğu arttırılmıştır. Grafik veri tabanı ve çizge madenciliği için Neo4j aracı kullanılmıştır. Neo4j'nin en büyük grafik topluluğu olması, yüksek performanslı okuma ve yazma ölçeklenebilirliği sağlaması, grafik depolama ve işlemede yüksek performans sağlaması, öğrenmesi ve kullanmasının kolay olması ve güvenilir olması bu çalışmada Neo4j'nin seçilmesine sebep olmuştur [22]. Makine öğrenimi yöntemleri CRISP-DM modelinin 6 aşamasını içermektedir. Bunlar; iş anlayışı, verileri anlama, veri hazırlama, modelleme, değerlendirme ve dağıtımdır [23]. Dolandırıcılık tespiti için makine öğrenimi sınıflandırma algoritmalarından SVM, RF, XGB, k-NN algoritmaları kullanılmıştır ve 5-katlamalı çapraz doğrulama kullanılarak performans değerlendirmeleri gerçekleştirilmiştir.

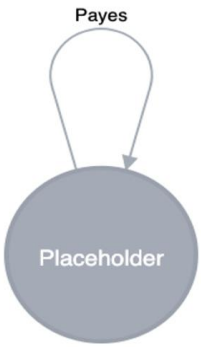
Veri seti ve neo4j

Bu çalışmada İspanya'daki bir bankaya ait toplu banka ödemelerinin bir örneğine dayanan, BankSim veri seti kullanılmıştır [24]. Veri setinin temel amacı, dolandırıcılık tespiti araştırmalarında kullanılabilecek sentetik verilerin üretilmesidir. Modeli geliştirmek ve kalibre etmek için satıcı ve müşteriler arasındaki ilişkilerin istatiksel ve sosyal ağ analizi kullanılmıştır. Veri setinde normal ödemeler ve dolandırıcılık verileri mevcuttur. BankSim yaklaşık altı ay boyunca 180 adımda çalıştırılmış ve test için güvenilir bir dağıtım elde etmek için parametreler kalibre edilmiştir. 587443 tane normal ödemelerin ve 7200 hileli ödemelerin olduğu toplam 594643 kayıt üretilmiştir [25]. Neo4j, günümüzde finansal hizmetler, enerji, yönetim, teknoloji, parakende ve imalat dahil olmak üzere neredeyse tüm sektörlerde ürün ve hizmetleri geliştirmek amacıyla binlerce şirket ve kuruluş tarafından kullanılan bir grafik veri tabanı aracıdır [26]. Bu çalışmada ön işleme aşamasından geçmiş veri setini görselleştirmek amacıyla Neo4j veritabanı ve CypherQL dili kullanılmıştır. Görselleştirilen veriseti python 3 ile Neo4j veri tabanına bağlanılarak CypherQL sorgu dili ile veritabanından çekilmiştir. CypherQL'le görselleştirilmiştir grafik veri setinin temel yapısı Şekil 1.'de gösterildiği gibidir.



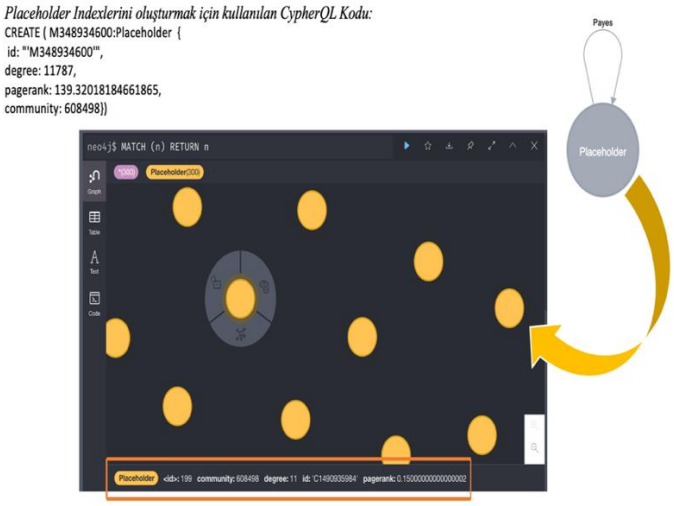
Şekil 1. Neo4j çizge veri seti yemel yapısı.

Şekil 1'de CypherQL'le oluşturulan grafik veri seti daha sonra Placeholder düğümüne bağlı müşteri ve banka indeksleri eklenerek genişletilmiştir. Burada Customer ve Bank, Constraint (ana düğümler) olarak etiketlenmiştir. İkisi arasındaki bağlantı ise, transaction düğümü ile gerçekleştirilmiştir. PageRank, Community ve Degree gibi ağda yüksek bir değere sahip düğümleri belirleyen çizge algoritmaları Placeholder olarak etiketlenen müşteri ve banka indeksleri üzerinde gerçekleştirilmiştir. Bu işlemin amacı bir düğüm bozulmak istendiğinde ağın ne kadarını etkileyeceğini belirlemektir. Örneğin bir ağda yalnızca en büyük değere sahip son müşterilerden birine uygun bir biçimde dönüştürülen bir dönüşüm süreci olduğu varsayılırsa süreçten geçen çok fazla yol yoktur. Bu nedenle düğümün ara merkezliği oldukça düşüktür ama düğümün değeri zincirdeki önemli bir öğeyi etkilediği için yüksektir. Placeholder düğümü Şekil 2.'de görüldüğü gibidir.



Şekil 2. Placeholder düğümü.

Placeholder düğümüne bağlı müşteri ve banka indekslerinin her biri Community, Degree, PageRank değerlerini içerir. Çizge algoritmaları Placeholder etiketi üzerinde uygulanmış ve Placeholder'ın bağlantısı kendisine ödemeler linkiyle bağlanacak şekilde gerçekleştirilmiştir. CypherQL ile genişletilen veri seti indeksleri ve CypherQL kodu Şekil 3.'de görüldüğü gibidir.



Şekil 3. Placeholder düğümü genişletme.

Veri ön işleme

Veri ön işleme aşamasında öncelikle her bir sütundaki boş değer sayıları tespit edilmiştir. Daha sonra veri çerçevesindensın ıf öznitelikleri alınmıştır. Zipcodeori ve zipMerchant sütunları tüm satırlar için aynı değere sahip olduğundan bu sütunlar silinmiştir. Veri seti özellik ağında yüksek değere sahip düğümler belirlendikten sonra step, age, gender, customer, fraud sütunları silinerek One Hot Encoding ile veri kategorilendirilmesi işlemi gerçekleştirilmiştir. Daha sonra özellik standartlaştırılması işlemi yapılmıştır. İç özellikler ve çizge tabanlı özellikler kullanılarak denetimli öğrenme modelleri eğitilmiştir. Bu işlemlerden sonra sahte düğümler ve gerçek düğümler arasında çok fark olduğundan veri setini dengelemek amacıyla veri setine undersampling yöntemi uygulanmıştır. Undersampling işlemi sonrasında sahte ve gerçek düğüm sayısı 7200 olarak belirlenmiştir. Ön işleme aşamasında boyut küçültme işlemi PCA (Principal Component Analyses) kullanılarak ve bileşen sayısını varyansın %95'i açıklanacak şekilde sınırlandırılarak gerçekleştirilmiştir.

Makine öğrenimi ve çizge algoritmaları

Destek vektör makinesi (SVM), günümüzde makine öğrenimi ve desen sınıflandırması alanında önem kazanmıştır [27]. SVM, girdi uzayında doğrusal veya doğrusal olmayan bir ayırma yüzeyi gerçekleştirilerek elde edilir. Destek vektörü sınıflandırmasında ayırma işlemi destek vektörleri ile ilişkili çekirdeklerin doğrusal bir kombinasyonu olarak ifade edilebilir. Bu çalışmada Lineer SVM algoritması kullanıldı. Lineer SVM destek vektör kümesini aşamalı olarak oluşturan sezgisel olarak tahmin yapan bir algoritmadır [28]. Son zamanlarda, karşı sınıfa en yakın nokta çiftinin her zaman Destek Vektörleri olduğu kanıt lanmıştır. Lineer SVM, aday destek vektörü kümesindeki bu n okta çifti ile başlar. Her yineleme sırasında maksimum ihlal e denin bir destek vektörü olduğu varsayılmıştır [29]. Algoritm a her yineleme sırasında maksimum ihlal edeni bulur ve maks imum ihlal edeni bir destek vektörü yapmak için aday destek düzlemini döndürür. Uzayın boyutunun aşılması veya tüm ver i noktalarının yakınsama olmaksızın kullanılması durumunda, algoritma, zıt sınıflardan bir sonraki en yakın nokta çifti ile ye niden başlatılır. Lineer SVM geometrik olarak tanımlanmıştır ve anlaşılması kolaydır.

Random Forest algoritması ağaç tahmin edicilerinin bir kolek siyonudur. H (x; θk), k=1,…,K ifadesinde x, ilişkili rasgele v ektörü, θk ve k bağımsız ve özdeş olarak dağıtılmış (id) rasgel e vektörleri temsil eder. Algoritmada sayısal bir sonuca sahip olunduğunda regresyon ayarına odaklanılır, ancak sınıflandırma (kategorik sonuç) problemleriyle bazı özel temas noktaları kurulur [30].

XGBoost algoritması Gradient Boosting algoritmasının çeşitli düzenlemeler sonucu optimize edilmiş halidir. Yüksek tahmin gücü elde edebilmesi, aşırı öğrenmeyi önlemesi, boş verileri yönetmesi özellikleri ile iyi bir performans gösterir [31]. XGBoost algoritmasında ilk adım olarak ilk tahmin gerçekleştirildi. Bu değer 0,5 olarak varsayıldı. Gözlemlenen değerden tahmin edilen değerlerin çıkarılmasıyla hatalar elde edildi. Hatalar öğrenilerek doğru tahmine yaklaşma amaçlandı.

K-Nearest-Neighbors algoritması parametrik olmayan (non-parametric) bir makine öğrenimi algoritmasıdır. Eğitim seti kullanan algoritmalara kıyasla k-Neighbors eğitim veri seti kullanarak eğitim verilerini öğrenmek yerine ezberler ve tahmin yapmak için veri seti içerisindeki en yakın komşuları arar. İlk aşamada bir k değeri belirlenir. Bu k değeri kadar elemana bakılır. Gelen değer ile en yakın komşu eleman arasındaki uzaklık öklid fonksiyonu ile hesaplanır. Öklid'e alternatif olarak Manhattan, Minkowski ve Hamming fonksiyonları da kullanılabilir. Uzaklık hesaplamasından sonra sıralama yapılır ve gelen değer uygun sınıfa atanır [32].

PangeRank algoritması düğümlerin geçiş etkisini veya bağlanabilirliğini ölçen en popüler çizge algoritmasıdır ve adını Larry Page'den almıştır [33]. Bu algoritma Neo4j'de düğümlerin öncelikli olmasına bağlı olan hesaplamalarda kullanılmıştır. PageRank hesaplaması Denklem 1.'de gösterildiği gibidir.

$$PR(A) = (1 - d) + d\left(\frac{PR(T_1)}{C(T_1)} + \cdots + \frac{PR(T_n)}{C(T_n)}\right) \qquad (1)$$

Burada bir A sayfasının T_1 ile T_n sayfalarına sahip olduğu varsayılır. Buradaki d, 0 (dahil) ile 1(hariç) arasında ayarlanabilen bir sönümleme faktörüdür ve genellikle 0.85'e ayarlanır. C (A), A sayfasından çıkan bağlantıların sayısıdır.

Degree algoritması bir düğüme bağlı ilişkilerin sayısını ölçmektedir [34]. Çalışmada bu algoritma hangi düğümün en çok alt düğüme sahip olduğunu belirlemek için kullanılmıştır.

Community algoritması, modülerlik puanını en üst düzeye çıkarmaya dayalı olarak ağlardaki toplulukları algılamaktadır [35]. Çalışmada bu algoritma grafikte alt toplulukları bulmak için kullanılmıştır.

Deney sonuçları

Neo4j'de oluşturulan çizge veri setinin üzerinde işlem yapmadan önce klasik makine öğrenimi yöntemleriyle SVM, XGB, kNN ve RF algoritmalarının k-katlamalı çapraz doğrulamasına göre performans karşılaştırması yapılmıştır. Performans değerlendirmeleri için karışıklık matrisi kullanılmıştır.

Karışıklık Matrisi: Gerçek değerlerin bilindiği bir test verisi üzerinde herhangi bir sınıflandırma modelinin performans tanımlaması için sıklıkla kullanılan bir tablodur.

Bu çalışmada SVM, XGB, RF, k-NN sınıflandırıcı modellerinin karışıklık matrisinden elde edilen değerler Tablo 2.'de görüldüğü gibidir. Bu tabloya göre XGB, RF, SVM ve k_NN algoritmalarının kesinlik ve F1 skor değerleri oldukça benzer çıkmıştır. Bu sonuçlar dengesiz veri seti üzerinde yani; undersampling yöntemi uygulanmadan önce elde edilen sonuçlardır. Undersampling yönteminin uygulanması sonucunda RF, XGB, SVM ve k-NN algoritması sınıflandırıcıları için elde edilen sonuçlar K-katlamalı çapraz doğrulama ile karşılaştırılmıştır. Elde edilen sonuçlar Tablo 3.'de görüldüğü gibidir.

K-katmanlı çapraz doğrulama bir makine öğrenmesi modelinde gerçekleştirilen testin hatasını daha yüksek doğrulukla tahmin etmek için kullanılan bir tekniktir. Bu yöntemde temel mantık eğitim veri setinden doğrulama kümeleri olarak da bilinen örnek gözlem bölümleri oluşturmaktır. Veri kümesi k alt kümeye bölünür ve k kez tekrarlanır. Her defasında k alt kümelerinden biri test kümesi olarak kullanılırken, diğer k1 alt kümeleri bir eğitim kümesi oluşturacak şekilde birleştirilir. Bundan sonra her bir k denemesi için ortalama hata değeri hesaplanır. Bu yöntemin avantajı verilerin nasıl bölündüğünü odak noktasına almamasıdır. Her veri noktası k kere test setinde yer alırken k-1 kere eğitim setine girer. K arttıkça sonuç tahmin varyansı da azalır. Bu yöntem k kere tekrarlanma zorunluluğu içerse de her bir test kümesinin büyüklüğünün anlaşılması ve kaç tane bağımsız deneme olduğunun belirlenmesi açısından avantajlıdır. Tablo 3'de XGB, RF, SVM ve k-NN algoritmalarının veri setinde undersampling uygulanması sonucunda veri seti dengelenerek kesinlik ve F1 skor metrikleri için daha tutarlı değerler elde edilmiştir.

Python üzerinden Neo4j veri tabanına bağlanıldıktan sonra CypherQL ile veri tabanındaki veri düğümleri çekilmiştir. Bu veri üzerinde daha önceki bölümlerde gösterilmiş olan ön işleme aşamaları gerçekleştirilmiştir. Ön işlemeden sonra veri seti üzerindeki PageRank, Community, Degree gibi düğümler hakkında ana merkez ve derece yüksekliklerini belirleyen çizge algoritmaları uygulanmıştır. Son olarak uygulanan XGB, RF, SVM ve k-NN yakın komşu algoritmaları ve genişletilmiş çizge veri seti üzerinde 5-katlamalı çapraz doğrulama kullanılarak performanslar karşılaştırılmıştır. Sonuçlar Tablo 4.'de görüldüğü gibidir. RF ve k-NN algoritmalarının düşük de olsa F1 skor ve kesinlik değerleri optimize edilmiştir. XGB algoritması için duyarlılık (recall) değerinde küçük de olsa bir performans iyileştirilmesi gözlemlenmiştir.

Ayrıca RF için klasik makine öğrenimi yöntemi ve çizge algoritmalarıyla birlikte çizge veri setine uygulanan yöntemin eğitim ve tahmin süreleri karşılaştırılmıştır. Sonuç Şekil 4'de görüldüğü gibidir.

```
---------- Time Evaluation ----------
training_time_std: 50.490270488262176
training_time_enh 49.02800601005558:
pred_time_std 1.3038832473754878:
pred_time_enh 1.2147937798500064:
```

Şekil 4. Tahmin ve eğitim süresi.

Training_time_std ve pred_time_std standart makine öğrenimi algoritmalarıyla yapılan eğitim ve tahmin sürelerini ifade etmektedir. Training_time_enh ve pred_time_enh ise genişletilmiş çizge analizi algoritmaları ve standart makine öğrenimi algoritmalarının birlikte kullanıldığı eğitim ve tahmin sürelerini ifade etmektedir. Küçük de olsa genişletilmiş çizge analizi ve klasik makine öğrenimi yöntemlerinin birlikte kullanılması eğitim ve tahmin sürelerinin kısalmasını sağlamıştır.

Tablo 2. Performans sonuçları ve sınıflandırıcı algoritmaları

Sınıflandırıcı Algoritmaları ve Performans Sonuçları				
Sınıflandırıcı Modeli	**Kesinlik (Accuracy)**	**Duyarlılık (Recall)**	**Hassasiyet (Precision)**	**F1 Score**
XGBoost	1	0.76	0.90	0.82
Random Forest	0.99	0.74	0.91	0.82
SVM	0.99	0.64	0.88	0.74
3-En Yakın Komşu	1	0.74	0.85	0.79

Tablo 3. Performans sonuçları ve sınıflandırıcı algoritmaları

Sınıflandırıcı Algoritmaları ve Performans Sonuçları				
Sınıflandırıcı Modeli	**Kesinlik (Accuracy)**	**Duyarlılık (Recall)**	**Hassasiyet (Precision)**	**F1 Score**
XGBoost	0.89	0.86	0.92	0.89
Random Forest	0.86	0.84	0.88	0.86
SVM	0.84	0.76	0.91	0.83
3-En Yakın Komşu	0.89	0.85	0.93	0.89

Tablo 4. Çizge madenciliği ve ML performans sonuçları ve sınıflandırıcı algoritmaları

Sınıflandırıcı Algoritmaları ve Performans Sonuçları				
Sınıflandırıcı Modeli	**Kesinlik (Accuracy)**	**Duyarlılık (Recall)**	**Hassasiyet (Precision)**	**F1 Score**
XGBoost	0.89	0.87	0.92	0.89
Random Forest	0.88	0.83	0.92	0.87
SVM	0.84	0.76	0.91	0.83
3-En Yakın Komşu	0.90	0.85	0.93	0.89

Tablo 2 dengesiz veri setinden yalnızca makine öğrenimi yöntemleri kullanılarak elde edilen sonuçları, Tablo 3 veri seti undersampling yöntemi ile dengeli hale getirildikten sonra yalnızca makine öğrenimi yöntemleri kullanılarak elde edilen sonuçları göstermektedir. Tablo 4, Neo4j aracı ile görselleştirilen veri setinin düğümleri üzerinde Community, PageRank, Degree gibi çizge algoritmaları uygulanarak düğüm derecelerinin belirlenmesinden sonra uygulanan makine öğrenimi yöntemlerinin sonuçlarını göstermektedir. Tablo 3 ve Tablo 4 karşılaştırıldığında RF ve k-NN kesinlik değerlerinde küçük de olsa bir iyileşme gözlemlenmiştir. Ayrıca XGB duyarlılık değeri ve RF hassasiyet ve f1 skoru değerleri de küçük de olsa iyileşme göstermiştir. Kurumsal firmalar açısından çok küçük iyileştirmelerin bile firmaya kazanç sağlama konusunda önemsendiği bilinmektedir [20]. Bu nedenle küçük de olsa yapılan iyileştirmeler kurumsal firmalar tarafından önem arz etmektedir.

Yapılan iyileştirmeleri daha net gözlemleyebilmek amacıyla Tablo 5.'de literatür taraması kısmında incelenen [18] ve [19]'daki çalışma sonuçlarıyla elde ettiğimiz sonuçlar kıyaslanmıştır. Her bir algoritma için sırasıyla kesinlik, hassasiyet, duyarlılık ve f1 skor değerleri belirtilmiştir.

Tablo 5. Diğer çalışmalarla elde edilen sonuçların kıyaslanması.

Referans	K-NN Algoritması	RF Algoritması	XGB Algoritması	SVM Algoritması	Metrikler Precision (P) Recall (R) Accuracy (A) F1 Score (F1)
Lopez ve Axelsson, 2014 [18]	√	√	√	χ	k-NN P: 0.83 k-NN R:0.61 k-NN F1: 0.70 XGB P: 0.89 XGB R: 0.76 XGB F1: 0.82 RF P: 0.24 RF R: 0.98 RF F1: 0.82
Islam., 2018 [19]	√	√	χ	√	k-NN P: 0.80 k-NN R: 0.80 k-NN F1: 0.79 k-NN A: 0.80 RF P: 0.93 RF R: 0.93 RF F1: 0.93 RF A: 0.93 SVM P: 0.76 SVM R: 0.77 SVM F1: 0.76 SVM A: 0.77
ML ve Çizge Algoritmasının birlikte kullanılmasıyla elde edilen sonuçlar	√	√	√	√	k-NN P: 0.93 k-NN R:0.85 k-NN F1: 0.89 k-NN A: 0.90 RF P: 0.92 RF R: 0.83 RF F1: 0.87 RF A: 0.88 SVM P: 0.91 SVM R: 0.76 SVM F1: 0.83 SVM A: 0.84 XGB P: 0.92 XGB R: 0.87 XGB F1: 0.89 XGB A: 0.89

Tablo 5.'de bu çalışmada elde edilen sonuçlarla literatürde aynı veri setini, aynı ön işleme yöntemini ve yalnızca makine öğrenimi algoritmalarını kullanan çalışmaların sonuçları listelenmiştir. Her bir algoritma için değerlendirme metrikleri Precision (P), Recall (R), Accuracy (A), F1 Score (F1) olacak şekilde kısaltılarak değerleri verilmiştir. K-NN, XGB ve SVM için çizge algoritmalarıyla birlikte kullanılan makine öğrenimi algoritmaları yönteminin daha iyi sonuçlar elde ettiği gözlemlenmiştir.

Sonuç

Bu çalışmada öncelikle BankSim veri seti üzerinde veri ön işleme aşaması gerçekleştirilerek makine öğrenimi modelleri için hazır hale getirilmiştir. Daha sonra XGB, RF, SVM, k-NN algoritmalarıyla veri setinin sınıflandırılması gerçekleştirilmiştir. Veri setinin dengesizliğinden kaynaklanan ve tutarsız olan kesinlik ve F1 skor değerleri nedeniyle veri seti üzerinde undersampling uygulanarak daha tutarlı kesinlik ve F1 skor değerleri elde edilmiştir. Neo4j veri tabanında veri seti grafiği oluşturulduktan sonra Community, PangeRank ve Degree algoritmalarıyla birlikte standart makine öğrenimi algoritmaları uygulanmıştır. Çizge algoritmalarıyla XGB, RF, k-NN algoritmalarının sonuçlarının optimize edildiği görülmüştür. Makine öğrenimi algoritmalarının ve çizge algoritmalarının birlikte kullanılmasıyla daha iyi sonuçlar elde etmek ve daha kolay işlemler gerçekleştirmek mümkündür. Neo4j aracı bu işlemleri gerçekleştirmede ve çizge algoritmalarını kullanmada oldukça kullanışlıdır. Birden çok kaynaktan toplanan verilerin kullanılmasına imkân vermesiyle birlikte kendi içinde de eğitim amaçlı çeşitli veri setleri barındırmaktadır ve kaynak sayısı oldukça fazladır. Sürekli yeni algoritmaların çıkarılması ve optimize edilmesi sonucunda çalışma gelecekte farklı makine öğrenimi algoritmalarıyla ve çizge algoritmalarının birleştirilmesiyle gerçekleştirilebilir. Çalışmada sadece CypherQL kullanılarak makine öğrenimi algoritmaları olmadan Neo4j üzerinde script'ler kullanarak dolandırıcılık risk tahminlerinin yapılması mümkündür. Burada kullanılan scriptler bir özellik olarak belirtilebilir. Bu çalışmada dolandırıcılık tespitinde çizge algoritmalarının ve makine öğrenimi algoritmalarının birlikte kullanılmasının zaman ve performans açısından sonuçlarda iyileştirme meydana getirdiği gösterilmiştir.

Kaynaklar

[1] G. Sadowski, & P. Rathle. Fraud detection: Discovering connections with graph databases. *White Paper-Neo Technology-Graphs are Everywhere*, *13*, 2014.

[2] K. Julisch. Risk-based payment fraud detection. Research Report, IBM Research, Zurich, (2010).

[3] S. Rehman, U. Khan, A. U., S. Fong. Graph mining: A survey of graph mining techniques. In Seventh

International Conference on Digital Information Management (ICDIM 2012) (pp. 88-92), IEEE, (2012).

[4] D. Koutra, C. Faloutsos. Individual and collective graph mining: principles, algorithms, and applications. Synthesis Lectures on Data Mining and Knowledge Discovery, 9(2), 1-206, (2017).

[5] C. Jiang, F. Coenen, M. Zito. A survey of frequent subgraph mining algorithms. The Knowledge Engineering Review, 28(1), 75-105i (2013).

[6] S. Suthaharan. Big data classification: Problems and challenges in network intrusion prediction with machine learning. ACM SIGMETRICS Performance Evaluation Review, 41(4), 70-73, (2014).

[7] J. Qiu, Wu, Ding Q., G., Xu, Y., S. Feng.A survey of machine learning for big data processing. EURASIP Journal on Advances in Signal Processing, 2016(1), 1-16,(2016).

[8] E. Kurshan, H. Shen, & H. Yu. Financial Crime & Fraud Detection Using Graph Computing: Application Considerations & Outlook. In *2020 Second International Conference on Transdisciplinary AI (TransAI)* (pp. 125-130). IEEE, September,2020.

[9] D. Cheng, X. Wang, Y. Zhang, & L. Zhang. Graph Neural Network for Fraud Detection via Spatial-temporal Attention. *IEEE Transactions on Knowledge and Data Engineering*, 2020

[10] C. Yang, Z. Liu, D. Zhao, Sun, M., & E. Y. Chang. Network representation learning with rich text information. In *IJCAI* (Vol. 2015, pp. 2111-2117), July, 2015.

[11] M. Xie, H. Yin, H. Wang, F., Xu, W. Chen, & S. Wang. Learning graph-based poi embedding for location-based recommendation. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management* (pp. 15-24), October, 2016.

[12] A. Roy, , J. Sun, R. Mahoney, L. Alonzi, S. Adams, & P. Beling. Deep learning detecting fraud in credit card transactions. In *2018 Systems and Information Engineering Design Symposium (SIEDS)* (pp. 129-134). IEEE, April, 2018.

[13] H. M. Vidanelage, T. Tasnavijitvong, , P. Suwimonsatein & P. Meesad. Study on machine learning techniques with conventional tools for payment fraud detection. In *2019 11th International Conference on Information Technology and Electrical Engineering (ICITEE)* (pp. 1-5). IEEE, October, 2019.

[14] F. Carcillo, Y. A. Le Borgne, O. Caelen, Y. Kessaci, F. Oblé, & G. Bontempi. Combining unsupervised and supervised learning in credit card fraud detection. *Information Sciences*, 2019

[15] B. Lebichot, Y. A. Le Borgne, L. He-Guelton, F. Oblé, & G. Bontempi, Deep-learning domain adaptation techniques for credit cards fraud detection. In *INNS Big Data and Deep Learning conference* (pp. 78-88). Springer, Cham, April, 2019

[16] C. Wang, & H. Zhu. Representing Fine-Grained Co-Occurrences for Behavior-Based Fraud Detection in Online Payment Services. *IEEE Transactions on Dependable and Secure Computing*, 2020.

[17] P. Shiguihara-Juárez, & N. Murrugarra-Llerena. A Bayesian Classifier Based on Constraints of Ordering of Variables for Fraud Detection. In *2018 Congreso Internacional de Innovación y Tendencias en Ingeniería (CONIITI)* (pp. 1-6). IEEE, October, 2018.

[18] E. A. Lopez-Rojas,, S. Axelsson. Banksim: A bank payments simulator for fraud detection research Inproceedings. In 26th EuropeanModeling and Simulation Symposium, EMSS, (2014).

[19] S. R. Islam. An efficient technique for mining bad credit accounts from both olap and oltp (Doctoral dissertation, Tennessee Technological University), (2018).

[20] S. Even. *Graph algorithms*. Cambridge University Press, 2011.

[21] A. Castelltort. Review of Graph-Powered Machine Learning, Alessandro Negro, Manning Publication, 2020.

[22] Packpub, URL: https://hub.packtpub.com/neo4j-most-popular-graph-database/, Varangaonkar, A. Why Neo4j is the most popular Graph database. (Erişim zamanı: 2021)

[23] R. Wirth, & J. Hipp. CRISP-DM: Towards a standard process model for data mining. In *Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining* (Vol. 1). London, UK: Springer-Verlag, April 2000.

[24] E. A. Lopes-Rojas, & S. Axelsson. Banksim: A bank Payment Simulation for Fraud Detection Research, 2014.

[25] Neo4j, URL: https://neo4j.com/developer/graph-database/, (Erişim zamanı: 2021)

[26] Kaggle, URL: https://www.kaggle.com/ntnu-testimon/banksim1, E. Alonso, Axelsson, Stefan. Banksim: A bank payments simulator for fraud detection research Inproceedings, (Erişim zamanı: 2021)

[27] D. Roobaert. DirectSVM: A simple support vector machine perceptron. *Journal of VLSI signal processing systems for signal, image and video technology*, *32*(1), 147-156, 2002.

[28] D. Roobaert. *Pedagogical support vector learning: A pure learning approach to object recognition* (Doctoral dissertation, Numerisk analys och datalogi), 2001.

[29] V. N. Vapnik. Introduction: Four periods in the research of the learning problem. In *The nature of statistical learning theory* (pp. 1-15). Springer, New York, NY, 2000.

[30] M. R. Segal. Machine learning benchmarks and random forest regression, 2004.

[31] R. Mitchell & E. Frank. Accelerating the XGBoost algorithm using GPU computing. *PeerJ Computer Science*, *3*, e127, 2017.

[32] M. Sarkar, & T. Y. Leong. Application of K-nearest neighbors algorithm on breast cancer diagnosis problem. In *Proceedings of the AMIA Symposium* (p. 759). American Medical Informatics Association, 2000.

[33] Neo4j, URL: https://neo4j.com/docs/graph-data-science/current/algorithms/page-rank/, (Erişim zamanı: 2021)

[34] Neo4j, URL: https://neo4j.com/docs/graph-algorithms/current/labs-algorithms/degree-centrality/, (Erişim zamanı: 2021).

[35] Neo4j, URL: https://neo4j.com/docs/graph-data-science/current/algorithms/community/, (Erişim zamanı: 2021).