www.dergipark.gov.tr
ISSN:2148-3736
El-Cezerî Fen ve Mühendislik Dergisi
Cilt: 8, No: 2, 2021 (1007-1023)

El-Cezerî Journal of Science and Engineering
Vol: 8, No: 2, 2021 (1007-1023)
DOI : 10.31202/ecjse.908260

ECJSE

Makale / Research Paper

Yapay Zeka Destekli ÇOKS Yöntemi ile Kredi Kartı Sahtekarlığının Tespiti

Mustafa Furkan KESKENLER[1a*], Deniz DAL[1b], Tolga AYDIN[1c]

[1]Atatürk Üniversitesi, Mühendislik Fakültesi, Bilgisayar Mühendisliği Bölümü. Erzurum/TÜRKİYE
mfkeskenler@gmail.com

Received/Geliş: 02.04.2021 **Accepted/Kabul:** 02.05.2021

Öz: Ödeme ve bankacılık sistemleri yeni teknolojik imkânlarla her geçen gün bir değişime ve gelişime uğramaktadır. Bu kapsamda kredi kartı teknolojisi de barındırdığı çeşitli avantajlar dolayısıyla kullanımı hızla artan bir ödeme seçeneği olarak karşımıza çıkmaktadır. Diğer taraftan kredi kartları en yaygın ödeme şekli haline geldikçe, sanal ortamdaki dolandırıcılık oranının da bu duruma paralel bir biçimde artma eğilimi gösterdiği bildirilmektedir. Hem yasal hem de sahtekârlığa yönelik işlemlerin benzer davranış eğilimine sahip olduğu gerçeği kredi kartı sahteciliğinin sanal ortamda tespitini oldukça zorlaştırmaktadır. Literatür incelendiğinde kredi kartı sahteciliğini tespite yönelik araştırmalarda çoğunlukla makine öğrenmesi algoritmalarından faydalanıldığı ve bu çalışmalar kapsamında farklı sınıflandırma algoritmalarının bireysel olarak dikkate alındığı görülmektedir. Öte yandan literatürde sınıflandırma işlemi için makine öğrenmesi algoritmalarının birlikte kullanıldıkları yöntemlere de rastlanıldığı ve bu sayede son derece hassas sınıflandırıcılara ulaşılabildiği rapor edilmektedir. Buna rağmen kredi kartı sahteciliğini tespit etmek amacıyla karar ağacı, k en yakın komşu ve näive bayes sınıflandırıcıların bir arada kullanıldığı bir çalışma literatürde mevcut değildir. Bu gözlemden hareketle bu çalışma kapsamında eldeki problemin çözümüne yönelik karar ağacı, k en yakın komşu ve näive bayes makine öğrenmesi algoritmalarından yararlanan ve Çoğunluk Oyu ile Karar Verme Sistemi (ÇOKS) olarak adlandırılan yeni bir sezgisel algoritma geliştirilmiştir. Geliştirilen yöntem ile literatürdeki çalışmalarda elde edilen başarının üzerine çıkıldığı saptanmıştır. Bu algoritmanın ortak karar verme mekanizması için de bir sayısal devre tasarımı lojik fonksiyonu olan çoğunluk fonksiyonundan faydalanılmıştır. Bu sayede ilgili algoritmaların güçlü yönlerinin stratejik bir şekilde birleştirilmesi amaçlanmıştır. ÇOKS'nin etkinliği her biri 30 farklı özniteliğe sahip 284,807 kredi kartı işleminin yer aldığı bir veri kümesi üzerinde test edilmiştir. Yürütülen testler finansal güvenliği hedefleyen bu yeni yöntemin %99,93 doğruluk oranı, %95,60 kesinlik oranı ve %80,0 ROC AUC değeri ile veri kümesindeki bir işlemi "sahte" veya "yasal" işlem olarak sınıflandırabilmeyi başardığını göstermiştir. Literatürdeki benzer çalışmalarla yapılan kıyaslamalar doğruluk oranıyla birlikte ÇOKS'nin özellikle kesinlik ve ROC AUC performans ölçütleri açısından yüksek bir başarı gösterdiğini ortaya koymuştur.

Anahtar Kelimeler: Kredi kartı; sahtekârlık tespiti; veri madenciliği; makine öğrenmesi; çoğunluk oyu ile karar verme.

Detection of Credit Card Fraud Using Artificial Intelligence Supported ÇOKS Method

Abstract: Payment and banking systems are changing and developing day by day with the new technological advances. In this context, the use of credit card as the preferred payment method is rapidly increasing due to its various advantages. On the other hand, as credit cards become the most common form of payment, the rate of fraud in the online world tends to increase in parallel with this trend. The fact that both legal and fraudulent transactions have similar behavioral tendency makes it difficult to detect the credit card fraud in the online world. When the literature is analyzed, it is seen that machine learning algorithms are mostly utilized to detect the credit card fraud and different classification algorithms are individually taken into account within the scope of the associated studies. On the other hand, the existence of the methods that employ multiple machine learning algorithms to obtain highly sensitive classifiers is also reported in the literature. However, there exists no study

Bu makaleye atıf yapmak için
Keskenler, M.F., Dal, D., Aydın, T., *"Yapay Zeka Destekli ÇOKS Yöntemi ile Kredi Kartı Sahtekarlığının Tespiti"* El-Cezerî Fen ve Mühendislik Dergisi 2021, 8(2); 1007-1023.

How to cite this article
Keskenler, M.F., Dal, D., Aydın, T., "*Detection of Credit Card Fraud Using Artificial Intelligence Supported ÇOKS Method*" El-Cezerî Journal of Science and Engineering, 2021, 8(2); 1007-1023.
ORCID ID: [a] 0000-0002-7604-4179; [b]0000-0003-0120-4315; [c]0000-0002-8971-3255

in the literature that uses decision tree, k nearest neighbor and näive bayes classifiers together to detect the credit card fraud. Based on this observation, a new heuristic algorithm called ÇOKS, that employs decision tree, k nearest neighbor and näive bayes machine learning algorithms, has been developed towards the solution of the problem at hand. It has been determined that the success achieved in the studies in the literature has been exceeded with the developed method. The majority function, which is a logic function of digital circuit design, is also used for the common decision-making mechanism of this algorithm. In this way, it is aimed to strategically combine the strengths of the related algorithms. The effectiveness of ÇOKS was tested on a data set containing 284,807 credit card transactions, each with 30 different features. The tests conducted have shown that this new method, which aims financial security, has been able to classify a transaction in the dataset as " fraud" or "legal" with 99.93% accuracy rate, 95.60% precision rate and 80.0% ROC AUC value. The comparisons with the similar studies in the literature revealed that ÇOKS has shown a high success rate especially in terms of precision and ROC AUC along with the accuracy.

Keywords: Credit card, fraud detection, data mining, machine learning, decision-making with majority voting.

1. Giriş

Ödeme ve bankacılık sistemleri yeni teknolojik imkânlarla her geçen gün bir değişime ve gelişime uğramaktadır. Bu kapsamda kredi kartı teknolojisi de barındırdığı çeşitli avantajlar dolayısıyla kullanımı hızla artan bir ödeme seçeneği olarak karşımıza çıkmaktadır. Örneğin günümüzde tüm dünyayı etkileyen Covid-19 salgını nedeniyle insanların alışverişlerini kâğıt ve madeni paralar gibi değiş tokuş esaslı bir ödeme yerine daha hijyenik olan temassız kredi kartları ile yaptıkları bilinmektedir. Öte yandan insan hayatını kolaylaştırmayı hedef alan bu gelişmeler yeni güvenlik sorunlarını da beraberinde getirmektedir. Kredi kartları en yaygın ödeme şekli haline geldikçe, dolandırıcılık oranının da bu duruma paralel bir biçimde artma eğilimi gösterdiği rapor edilmektedir [1]. Bu nedenle kredi kartlarında artan dolandırıcılığa karşı bir takım güvenlik önlemleri alınmaktadır. Çip&Pin teknolojisine geçiş bu tedbirlerden birisidir ve fiziksel ortamda gerçekleşen kredi kartı sahteciliğindeki risk etmenini nispeten bertaraf etmektedir. Bununla birlikte sanal ortamda gerçekleştirilen kart sahteciliğinde bir artış olduğu gözlemlenmektedir [2]. Sanal ortamlardaki sahteciliği ve hileli işlemleri geleneksel manuel algılama tekniklerini kullanarak tespit etmek zaman alıcıdır ve yanıltıcı sonuçlar doğurmaktadır [3, 4]. Sektördeki finansal kurumlar bu nedenle yapay zekâya dayalı akıllı tekniklere yönelmişlerdir.

Günümüzde kredi kartları, ön ödemeli kartlar ve debit kartlarının kullanımının oldukça yaygınlaştığı şu istatistiklerden anlaşılmaktadır. Bankalararası Kart Merkezi (BKM) tarafından yayınlanan ilgili istatistikler incelendiğinde, Türkiye'de Şubat 2021'de 77.254.183 adet kredi kartı, 141.270.606 adet debit kartı ve 45.241.461 adet ön ödemeli kartın mevcut olduğu görülmektedir. Türkiye'de 2021 yılında internet üzerinden kart kullanılarak toplam 1.005.243.176 işlemin gerçekleştirildiği belirtilen raporlarda sunulmaktadır. Kart kullanımının yaygınlaşması kredi kartı dolandırıcılığındaki artışı da beraberinde getirmektedir. Dünyanın dört bir yanındaki bankaların ve işletmelerin toplam dolandırıcılık kayıplarının miktarının 2014 yılında bir önceki yıla kıyasla yaklaşık 2,5 milyar dolarlık bir artışla 16 milyar doların üzerine çıktığı rapor edilmiştir [5]. Türkiye özelinde ise sanal ortamda gerçekleşen kart sahteciliklerinin en önemli kısmını özellikle kart verilerinin çalınması, sızdırılması, kopyalanması ve kötü niyetli kişiler tarafından ele geçirilmesi oluşturmaktadır [2]. Ayrıca, Türkiye'de 2016 yılında en üst seviyeleri gören kredi kartı dolandırıcılığı 2017 yılında da azalmadan devam etmiştir. 2017 yılında gerçekleşen kart dolandırıcılığındaki sanal ortam faktörü %85 ila %90 oranlarına ulaşmış ve bu konuda ülkemizin dünyada ilk 10'a giren ülkelerden birisi olmasına neden olmuştur [32].

Kredi kartı sahteciliğinin sanal ortamda tespiti oldukça zordur zira hem yasal hem de sahtekârlığa yönelik işlemlerin benzer davranış eğilimine sahip olduğu gözlemlenmektedir [1]. Bu nedenledir ki sahtekârlık tespiti, istenmeyen davranışları tahmin ve tespit etmek veya önlemek için kullanıcıların davranışlarının izlenmesini içeren zor bir süreci kapsamaktadır.

Bu çalışmada, araştırma kapsamında geliştirdiğimiz ÇOKS'nin etkinliği her biri 30 farklı özniteliğe sahip 284,807 kredi kartı işleminin yer aldığı bir veri kümesi üzerinde test edilmiştir. Yürütülen testler finansal güvenliği hedefleyen bu yeni yöntemin %99,93 doğruluk oranı, %95,60 kesinlik oranı ve %80,0 ROC AUC değeri ile veri kümesindeki bir işlemi "sahte" veya "yasal" işlem olarak sınıflandırabilmeyi başardığını göstermiştir. Literatürdeki benzer çalışmalarla yapılan kıyaslamalar doğruluk oranıyla birlikte ÇOKS'nin özellikle kesinlik ve ROC AUC performans ölçütleri açısından yüksek bir başarı gösterdiğini ortaya koymuştur.

Bu çalışmanın geri kalan bölümü şu şekilde organize edilmiştir. 2. Bölümde çalışma kapsamında literatürde yer alan çalışmalardan ve araştırma problemine yönelik son durumlardan bahsedilmiştir. 3. Bölümde materyal ve önerilen yöntem detaylandırılmıştır. 4. bölümde ise bulgular analiz edilmiş ve araştırma sonuçları benzer çalışmaların sonuçları ile karşılaştırılmıştır. Son olarak 5. bölümde gelecek çalışmalara ışık tutacak bir değerlendirme ile makale sonuçlandırılmıştır.

2. Literatür Özeti

Literatürdeki çalışmalar incelendiğinde kredi kartı sahtekârlığını tespit algoritmalarının çoğunlukla veri madenciliğine dayalı olduğu anlaşılmaktadır. Örneğin kredi kartı sahtekârlığında, şaibeli olan işlemler meşru (yasal) ve hileli işlemler olmak üzere iki sınıfa ayrılarak tespit edilebilmektedir [6, 7]. Ayrıca karar ağacı, yapay sinir ağı, destek vektör makinesi (DVM), genetik algoritma, näive bayes algoritması, bulanık mantık ve lojistik regresyon gibi algoritmalardan sahtekârlık tespit işlemlerinde sıklıkla faydalanılmaktadır [8-13].

Kredi kartı sahtekârlığının son yıllarda giderek artması nedeniyle literatürde bu konuya yönelik yapılan çalışmalarda da bir artış olduğu gözlenmektedir. Bu alandaki çalışmalardan ikisinde [14, 15], lojistik regresyon ve näive bayes algoritmalarının performansları karşılaştırılmıştır. Analiz sonuçları lojistik regresyon'un performansının näive bayes'in performansından daha düşük olduğuna dair birkaç vakaya işaret etmekte, fakat bunun özellikle küçük veri kümelerinde meydana geldiği rapor edilmektedir. Näive bayes ve yapay sinir ağı kullanılarak kredi kartı sahtekârlığı tespiti üzerine yapılan karşılaştırmalı iki farklı çalışma [16, 17], näive bayes yönteminin kredi kartı sahtekârlığını tespit etmede yapay sinir ağından daha iyi performans gösterdiğini bildirmektedir.

Bir diğer çalışmada, üç farklı sınıflandırma yöntemi, (karar ağacı, yapay sinir ağı ve lojistik regresyon) sahtekârlık tespitinde uygulanabilirlikleri açısından test edilmiştir [3]. Sonuçlar, yapay sinir ağları ve lojistik regresyon yaklaşımlarının ilgili problemin çözümünde karar ağacından daha iyi performans sergilediğini ortaya koymuştur.

Karar ağaçları ve destek vektör makineleri kullanılarak kredi kartı sahtekârlığının tespitinin araştırıldığı bir başka çalışma, karar ağacı yaklaşımının araştırılan sorunun çözümünde DVM yaklaşımından daha iyi bir performans sergilediğine vurgu yapmaktadır [18]. Eğitim verileri sınanırken, DVM tabanlı modelin karar ağacı tabanlı modelin başarısına yakın olduğu belirtilmektedir. Ancak bu algoritmanın gerçek zamanlı sahtekârlık tespitinde yetersiz kaldığı rapor edilmektedir.

Konuyla ilgili bir başka çalışmada, kredi kartı sahtekârlığının tespitinde DVM'nin, rastgele orman'ın ve lojistik regresyon'un performansı karşılaştırılmaktadır [19]. Bu çalışma, lojistik regresyon'un farklı düşük örnekleme seviyeleri ile rastgele orman algoritmasıyla benzer bir performansa sahip olduğunu, DVM'nin ise eğitim verilerinde daha düşük dolandırıcılık oranına rağmen nispeten iyi sonuçlar verdiğini göstermiştir. Lojistik regresyon'un genellikle DVM modelinin performansını aşan kayda değer bir performans sergilediği de belirtilmiştir.

Yukarıdaki paragraflarda da detaylandırıldığı üzere, kredi kartı sahteciliği alanındaki araştırmalarda çoğunlukla makine öğrenmesi algoritmalarından faydalanılmıştır ve bu çalışmalar kapsamında farklı sınıflandırma algoritmaları bireysel olarak dikkate alınmıştır. Öte yandan literatürde sınıflandırma işlemi için makine öğrenmesi algoritmalarının birlikte kullanıldıkları yöntemlere de rastlanıldığı ve bu sayede son derece hassas sınıflandırıcılara ulaşılabildiği rapor edilmektedir [30]. Söz konusu bu yöntemlerin ortak karar verme mekanizması için ise birçok matematiksel, istatistiksel ve mantıksal metottan yararlanılmaktadır [31].

Makine öğrenmesi algoritmalarının kredi kartı sahtekarlığının tespiti üzerindeki etkinliğinin irdelendiği ve bu algoritmaların bir arada kullanılarak çoğunluk oyuyla karar verme sistemlerinin ele alındığı son yapılan literatür çalışmalarının birinde gerçek dünya örneklerinden oluşan ve %30'a varan gürültülü veri içeren veri kümelerinde dahi çoğunluk oyu ile karar vermenin pozitif etkisinin gözlemlendiği bildirilmiştir. İlgili çalışmada yüksek gürültülü veri oranına sahip veri kümeleri üzerinde makine öğrenmesi algoritmalarıyla gerçekleştirilen karar sistemlerinde dolandırıcılık tespit oranları ve doğruluk oranlarının beklenmedik şekilde bozulduğu söylenmektedir. Bu problemin birlikte karar veren ikili makine öğrenmesi algoritmalarıyla iyileştirildiği ve %30'lara varan gürültü oranlarına rağmen çoğunluk oyuyla karar verme sistemlerinde doğruluk oranlarının çeşitli veri kümelerinde %90'ları aşan doğruluklara ulaştığı açıklanmıştır. Çalışmada örnek gösterilen ve çoğunluk oyuyla karar verme işlemlerinde kullanılan ikili makine öğrenmesi yöntemleri, karar ağacı (KA) + gradyan artırma ağacı (GBT), näive bayes (NB) + KA ve NB + GBT olarak sıralanmaktadır [33]. Öte yandan, kredi kartı sahteciliğini tespit etmek amacıyla karar ağacı, k en yakın komşu ve näive bayes sınıflandırıcıların bir arada kullanıldığı bir çalışma literatürde mevcut değildir. Bu gözlemden hareketle bu çalışma kapsamında eldeki problemin çözümüne yönelik karar ağacı, k en yakın komşu ve näive bayes makine öğrenmesi algoritmalarından yararlanan ve Çoğunluk Oyu ile Karar Verme Sistemi (ÇOKS) olarak adlandırılan yeni bir sezgisel algoritma geliştirilmiştir. Bu algoritmanın ortak karar verme mekanizması için de bir sayısal devre tasarımı lojik fonksiyonu olan çoğunluk fonksiyonundan faydalanılmıştır. Bu sayede ilgili algoritmaların güçlü yönlerinin stratejik bir şekilde birleştirilmesi suretiyle hassasiyeti ve doğruluk oranı yüksek sonuçlara ulaşılmıştır.

3. Materyal ve Metot

Bu bölümün alt bölümlerinde, bu çalışma kapsamında geliştirilen ÇOKS, bu yöntem kapsamında kullanılan veri madenciliği algoritmaları ve test için kullanılan veri kümesi detaylandırılmıştır.

3.1. Materyal

Bu çalışmada Worldline ve ULB Machine Learning Group tarafından oluşturulmuş veri kümesi kullanılmıştır [20]. Bu veri kümesi, Avrupa'da yaşayan kredi kartı sahipleri tarafından Eylül 2013'te yapılan kredi kartı işlemlerini içermektedir ve iki gün içerisinde gerçekleştirilen toplam 284,807 ödeme işleminden oluşmaktadır. Veri kümesindeki her bir veriye ait bir etiket sütunu mevcuttur. Bu sütunda yer alan 1 değeri ilgili işlemin sahtekârlık içeren bir işlem (pozitif), 0 değeri ise yasal bir işlem (negatif) olarak sınıflandırıldığına işaret etmektedir. Pozitif işlemler verilerin sadece %0.172'sini oluşturmaktadır. Dolayısıyla veri kümesi oldukça dengesizdir ve bu durumun muhtemel etkileri bulgular ve tartışma bölümünde ayrıntılı olarak ele alınmıştır.

Veri kümesi içerisindeki 28 sütun/öznitelik (V1-V28) kredi kartı sahiplerine ait bazı kişisel bilgileri içermektedir ve kişisel verilerin korunması kapsamında doğrudan paylaşılmamıştır. Bu nedenle veri kümesi paylaşılmadan önce araştırmacılar tarafından öncelikle söz konusu 28 özniteliğin her biri TBA (Temel Bileşen Analizi) algoritması kullanılarak sayısal bir veriye dönüştürülmüştür. Bu işlem, öznitelik çıkarma önişlem adımını da bir nebze kolaylaştırmıştır. Ödeme işlemine ait zaman ve işlem tutarı bilgileri ise veri kümesinde gizlenmeden doğrudan kullanılmıştır. Dolayısıyla

toplamda 30 adet öznitelik mevcuttur. Bu çalışmada veri kümesindeki işlemlerin %70'i eğitim verisi (199364 adet), %30'u ise test verisi (85443 adet) olarak değerlendirmeye alınmıştır.

3.2. Metot

3.2.1. Makine Öğrenmesi Algoritmaları

Bu çalışma kapsamında yürütülen literatür taraması esnasında kredi kartı sahtekârlığının tespitinde başarılı sonuçlarıyla öne çıkan makine öğrenmesi algoritmalarının karar ağacı (C4.5), k en yakın komşu ve näive bayes olduğu anlaşılmıştır [1, 10, 27]. Bu nedenle araştırmanın ilk bölümünde bu üç algoritmanın her biri için en iyi parametre değerlerini elde etmek amacıyla denemeler yapılmıştır. Sonrasında önerilen ÇOKS algoritması bu üç algoritmanın tespit ettiği kararları birleştirerek nihai sınıflandırmayı tamamlamıştır. Çalışmanın tamamı Python programlama dili ve scikit-learn kütüphanesi kullanılarak gerçekleştirilmiştir.

3.2.1.1. Karar Ağacı (KA) Algoritması

Karar ağacı algoritması, yaygın olarak kullanılan denetimli öğrenme algoritmalarından birisidir. Diğer denetimli öğrenme algoritmalarının aksine, bu algoritma hem regresyon hem de sınıflandırma problemlerinin çözümü için tercih edilmektedir [18].

Karar Ağaçlarında, bir verinin sınıf etiketini tahmin etmek için ağacın kök değerinin bulunmasıyla işleme başlanır. En yüksek kazanç sağlayan öznitelik ağacın kök düğümü olarak kullanılır. Seçilen özniteliğin almış olduğu değerlere göre dallanma işlemi yapılır. Ardından her bir dala göre yaprak düğümler yine kazanç hesabı yapılarak kalan öznitelikler arasından seçilir. Böylece sınıflandırma modeli eğitim süresince oluşturulur. Entropi, Information Gain, Gini Index, Gain Ratio, Reduction in Variance ve Chi-Square karar ağacı algoritmasının öznitelikler arasında seçim yapmak için kullandığı kazanç hesaplama yöntemlerinden bazılarıdır. Verilerin makine öğrenimine göre sınıflandırma prosedürü esas olarak iki aşamalı sürece ayrılmıştır. İlk aşamada, öğrenme süreci alınan bilgilerden bir model oluşturur. İşveren öğrenme aşamasında sınıflandırma bilgisi sağlarsa, buna denetimli öğrenme denir; aksine, sınıfların önceden sınıflandırma bilgisi olmaksızın bir veri setinden türetildiği gözetimsiz öğrenme olarak adlandırılır. Karar ağaçlarını (DT) kullanmanın belirgin bir avantajı, işverene hem denetimli hem de denetimsiz öğrenmeyi yürütme olanağı sağlamasıdır. Bu nedenle, genellikle bilgi keşfi için kullanılırlar [34].

Bu çalışmada karar ağacı algoritması olarak C4.5 ve kazanç hesabı için Entropi tercih edilmiştir. Ayrıca karar ağacı algoritması için Python'a ait sklearn kütüphanesinde tanımlı DecisionTreeRegressor fonksiyonundan yararlanılmıştır.

3.2.1.2. K En Yakın Komşu (KNN) Algoritması

KNN algoritması Cover ve Hart tarafından 1967 yılında sınıflandırma problemleri için ilk olarak geliştirilen yöntemlerden biri olup, etkili ve basit olmasından dolayı da yaygın kullanılan bir makine öğrenmesi algoritması olarak nitelendirilebilir. Makine öğrenmesi algoritmalarında yönteme dair parametreler sınıflandırma performansında oldukça önem arz etmektedir ve bu parametrelerin optimizasyonu yapılan çalışmalarda dikkat gerektirmektedir. KNN algoritmasında ise k parametresi optimizasyona dahil edilerek daha başarılı sonuçlarla veri kümesindeki özniteliklere ait örneklerin dahil olacağı sınıf belirlenebilmektedir. KNN algoritması, prensipte veri kümesindeki örneklerin birbirine olan uzaklıklarının Manhattan, Euclidean ve Minkowski gibi uzaklık ölçüm yöntemleri ile hesaplanması temeline dayanır. Hesaplama sonucunda k parametresine göre birbirine mesafe olarak yakın örnekler aynı sınıf olarak değerlendirilirler [35]. Aynı zamanda KNN Algoritması, mevcut

tüm verileri saklayan, yeni verileri benzerlik ölçüsüne göre belirtildiği gibi uzaklık formülleri kullanarak sınıflandırma işlemini gerçekleştiren ve oldukça yaygın kullanılan öğrenme algoritmalarından birisidir. Ayrıca KNN'den, 1970'lerin başında istatistiksel tahmin ve örüntü tanımada parametrik olmayan bir teknik olarak faydalanılmıştır [21].

KNN algoritmasında bir veri kendisine en yakın K tane komşusunun çoğunluk oyu ile sınıflandırılır. Yani veri, bir mesafe fonksiyonu ile ölçülen ve kendisine yakın K komşu arasında en yakın olan sınıfa atanır. KNN algoritmasının kullandığı bazı uzaklık mesafesi ölçüm denklemleri Eşitlik (1), (2) ve (3) ile verilmiştir [22].

$$Öklid = \sqrt{\sum_{i=1}^{k}(xi - yi)^2} \tag{1}$$

$$Manhattan = \sum_{i=1}^{k} |xi - yi| \tag{2}$$

$$Minkowski = \left(\sum_{i=1}^{k}(|xi - yi|)\text{^}q \right)^{\frac{1}{q}} \tag{3}$$

Bu çalışmada, veri kümesinde deneysel sonuçlarda en iyi başarıyı sağlayan Minkowski denklemi, yani Eşitlik (3) kullanılarak KNN algoritması gerçeklenmiştir. Ayrıca sırasıyla denenerek K değerinin 11 olduğu durumlarda başarının en üst seviyeye ulaştığı saptanmıştır. KNN Algoritması için Python'a ait sklearn kütüphanesinde tanımlı KNeighborsClassifier fonksiyonundan faydalanılmıştır.

3.2.1.3. Näive Bayes (NB) Algoritması

Bu algoritma öznitelikler arasında bağımsızlık varsayımına ve Bayes Teoremi'ne dayanan bir sınıflandırma tekniğidir. Sade bir ifadeyle, näive bayes sınıflandırıcısı, bir sınıftaki belirli bir özelliğin varlığının, başka herhangi bir özelliğin varlığıyla ilgisiz olduğunu varsaymaktadır.

Naive Bayes algoritması, bir veri setindeki değerlerin frekans ve kombinasyonlarını sayarak olasılık setini hesaplayan bir olasılık sınıflandırma algoritmasıdır [36]. Algoritma, sınıflandırılacak verinin her bir durumunun olasılığını hesaplar, ardından hesaplanan en yüksek olasılık değerine göre sınıflandırmayı gerçekleştirir. X girdi değişkeninin m öznitelikleri X = (x1, x2,…, xm) ile sunulur. Naive Bayes (NB) sınıflandırıcı Bayes kuralına dayanmasına rağmen, durumların gerçek olasılıklarını hesaplamaz. NB, sınıfın maksimum olasılıkla belirlenmesine odaklanır [37].

Näive bayes modeli özellikle çok büyük veri kümeleri için kullanışlıdır. Sınıflandırma sürecinde sistem belirli bir oranda etiketi belli olan veri ile beslenmektedir. Bu yöntemde eğitim için kullanılan verilerin kesinlikle bir sınıfı bulunmalıdır. Eğitilmiş veriler üzerinde yapılan olasılık işlemleri sonucu veri kümesinde yer alan test verileri işleme dâhil edilerek sınıflandırma sonuçları elde edilmektedir.

Bayes Teoremi P(c), P(x), P(x|c) ve P(c|x) parametrelerinin kullanıldığı Eşitlik (4)'deki denklem ile ifade edilmektedir [1].

$$P(c|x) = \frac{P(x \mid c) \cdot P(c)}{P(x)} \tag{4}$$

Bu denklemde yer alan P(c) c olayının gerçekleşme olasılığını, P(x) x olayının gerçekleşme olasılığını, P(c|x) x olayı gerçekleştiğinde c olayının gerçekleşme olasılığını, P(x|c) ise c olayı gerçekleştiğinde x olayının gerçekleşme olasılığını ifade etmektedir. NB Algoritması için Python'a ait sklearn kütüphanesinde tanımlı GaussianNB fonksiyonundan yararlanılmıştır.

3.2.2. Çoğunluk Fonksiyonu

Boolean cebrinde girişlerinin en az yarısı doğru iken doğru değer üreten fonksiyona çoğunluk fonksiyonu (majority function) adı verilmektedir. Şekil 1'de 3 değişkenli çoğunluk fonksiyonuna ait doğruluk tablosu yer almaktadır. Görüldüğü üzere F fonksiyonu, 2 veya 3 değişkenin değeri doğru (1) olduğunda çıkışında doğru değeri üretmektedir.

F fonksiyonunun minterimlerin toplamı şeklinde ifade edildiği denklem Eşitlik (5) ile verilmiştir. Eşitlik (6) ise bu fonksiyonun Karnaugh haritası kullanılarak sadeleştirilmiş halini ifade etmektedir.

$$F(x,y,z) = \bar{x}yz + x\bar{y}z + xy\bar{z} + xyz \tag{5}$$

$$F(x,y,z) = xy + yz + xz \tag{6}$$

x	y	z	F
0	0	0	0
0	0	1	0
0	1	0	0
0	1	1	1
1	0	0	0
1	0	1	1
1	1	0	1
1	1	1	1

Şekil 1. Üç boolean değişkenli çoğunluk fonksiyonunun doğruluk tablosu

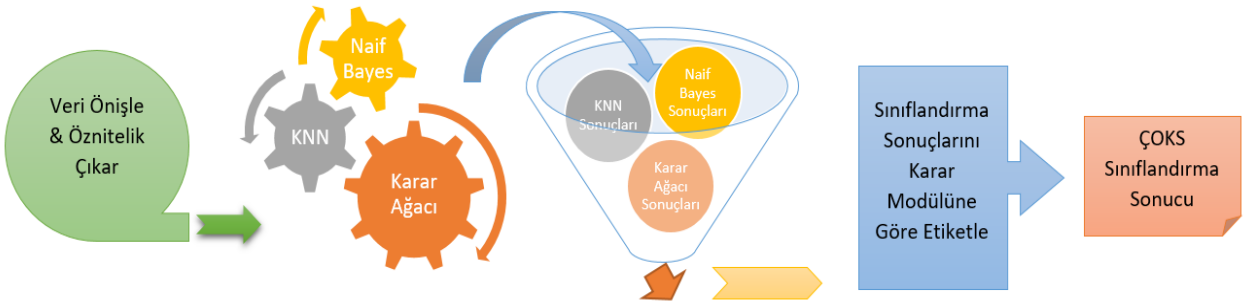
3.2.3. Çoğunluk Oyu ile Karar Verme Sistemi (ÇOKS)

Literatürdeki çalışmalar makine öğrenmesi algoritmalarının her problem türü için veri kümesine bağlı olarak farklı başarı sonuçları ürettiğini göstermektedir [1,16, 27]. Başka bir deyişle bu çalışmalarda öğrenme algoritmalarının performanslarının veri kümesi ile doğrudan ilişkili olduğu gerçeğine vurgu yapılmaktadır. Kredi kartı sahtekârlığı tespiti için de literatürde var olan veri kümeleri incelendiğinde yasal işlemlere kıyasla sahtekârlık şeklinde etiketlenmiş verilerin sayısının veri kümesi miktarına oranla oldukça düşük seviyede seyrettiği anlaşılmaktadır. Dolayısıyla bu veri kümelerinin entropisi bayağı düşük durumdadır. Bu nedenledir ki bu türden problemlerde başarı ölçütü olarak doğruluk değerinin yanında kesinlik ve ROC AUC değerlerinin dikkate alınması gerekmektedir.

Konuyla ilgili literatürdeki diğer çalışmalar incelendiğinde başarı ölçütü olarak özellikle doğruluğa odaklanıldığı, kesinlik ve ROC AUC değerlerinin genelde düşük oranlarda elde edildiği ve hatta bazı çalışmalarda bu ölçütlerin hiç dikkate alınmadığı görülmektedir. Bu noktadan hareketle kesinliğin ve ROC AUC'nin de doğruluk ile birlikte makul seviyelerde bir başarı oranına ulaştırılabilmesi amacıyla bu çalışmada Çoğunluk Oyu ile Karar Verme Sistemi (ÇOKS) adı verilen yeni bir sezgisel algoritma önerilmiştir. ÇOKS, kredi kartı sahtekârlığının tespitinde doğruluk, kesinlik ve ROC AUC ölçütleri için kısmi olarak iyi sonuçlar ürettikleri bilinen karar ağacı, k en yakın komşu ve näive bayes algoritmalarının karar verme güçlerinin birleştirilmesi mantığıyla çalışmaktadır. Bu ortak karar verme mekanizması sayesinde kesinlik için iyi sonuçlar veren bir algoritmanın düşük ROC AUC değeri bir başka algoritma tarafından iyileştirilebilmektedir. Şekil 2'de de görüldüğü üzere ÇOKS sayesinde üç farklı sınıflandırma algoritmasının aldığı kararlar üç başarı ölçütünü de üst seviyelere çıkarabilmek amacıyla stratejik bir şekilde birleştirilmektedir.

Şekil 3'te fonksiyon isimlendirmelerinde Paskal notasyonundan ve değişken isimlendirmelerinde Deve notasyonundan faydalanılan ÇOKS algoritmasının sözde kodu yer almaktadır. Algoritmanın

girdileri veri kümesi, test verisi oranı ve KNN algoritmasında kullanılacak K değeridir. Çıktıları ise sınıflandırma ve öznitelik katkı vektörleridir. Algoritma 1'de kırmızı çerçeve (Eşit kesikli çizgilerle gösterilen, Satır 1-6) içerisine alınan ilk kısım yöntemin ön işleme adımlarını oluşturmaktadır. Sırasıyla yeşil çerçeve (Bir uzun ve bir kısa çizgi ile gösterilen, Satır 7-9) içerisinde karar ağacı (C4.5), mavi çerçeve (Noktalı çerçeve ile gösterilen, Satır 10-12) içerisinde k en yakın komşu ve sarı çerçeve (Bir uzun ve iki kısa çizgi ile gösterilen, Satır 13-14) içerisinde de näive bayes algoritması gerçeklenmiştir. Son olarak turuncu çerçeveli (Düz çizgi ile gösterilen, Satır 15) bölüm ise çoğunluk fonksiyonu ile 3 algoritmanın sonuçlarının birleştirildiği bölümdür.



Şekil 2. Çoğunluk oyu ile karar verme sisteminin mimari yapısı

Algoritma 1: ÇOKS Algoritmasının Sözde Kodu

Girdi: *Veri Seti, Test Verisi Oranı, K Değeri*
Çı***kt***ı: *Sınıflandırma Vektörü*, *Öznitelik Katkı Vektörü*

1: $veriMatrisi \leftarrow VeriSetiniOku(veriSeti)$
2: $öznitelikKatkıVektörü \leftarrow ÖznitelikSeç(veriMatrisi, ExtraTreesClassifier)$
3: $[öznitelikler Matrisi, etiketlerVektörü] \leftarrow VeriMatrisiniBöl(veriMatrisi)$
4: $medyanVektörü \leftarrow OrtancaDeğerleriHesapla(öznitelikler Matrisi)$
5: $öznitelikler Matrisi \leftarrow EksikVerileriTamamla(öznitelikler Matrisi, medyanVektörü)$
6: $[eğitimMatrisi, testMatrisi] \leftarrow EğitimVeTestVerisiniAyır(öznitelikler Matrisi, etiketlerVektörü, testVerisiOranı)$

7: $entropiVektörü \leftarrow EntropiDeğerleriniHesapla(öznitelikler Matrisi)$
8: $KANesnesi \leftarrow KararAğacıAlgoritmasıylaEğit(eğitimMatrisi, entropiVektörü)$
9: $KASonuçVektörü \leftarrow KANesnesi.Uygula(testMatrisi)$

10: $minkowskiVektörü \leftarrow MinkowskiDeğerleriniHesapla(öznitelikler Matrisi)$
11: $KNNNesnesi \leftarrow KNNAlgoritmasınıOlustur(eğitimMatrisi, minkowskiVektörü, kDeğeri)$
12: $KNNSonuçVektörü \leftarrow kNNNesnesi.Uygula(testMatrisi)$

13: $NBNesnesi \leftarrow NaifBayesAlgoritmasıylaEğit(eğitimMatrisi)$
14: $NBSonuçVektörü \leftarrow NBNesnesi.Uygula(testMatrisi)$

15: $sınıflandırmaVektörü \leftarrow KararFonksiyonu(KASonuçVektörü, KNNSonuçVektörü, NBSonuçVektörü)$

16: ***return*** $sınıflandırmaVektörü, öznitelikKatkıVektörü$

Şekil 3. Çoğunluk oyuyla karar verme sisteminin sözde kodu

Sözde kodun 1. satırında veri kümesi girdisi hafızadan okunarak veriMatrisi isimli değişkene atanmaktadır. 2. satırda veriMatrisi içerisindeki özniteliklerin sınıflandırma başarısına olan katkıları

hesaplanmaktadır. Bu hesaplama, ÇOKS yönteminin sonucuna doğrudan bir etkiye sahip değildir; sınıflandırma işlemi için en çok öneme sahip özniteliğin belirlenmesinde araştırmacılara bir fikir vermesi açısından faydalı görülmüştür. İlgili hesaplamada literatürde sıklıkla kullanılan ve Python kütüphanesinde bulunan ExtraTreesClassifier fonksiyonundan yararlanılmıştır. 3. satırda veriMatrisi içerisindeki veriler, öznitelikler (matris) ve etiketler (vektör) olarak ikiye ayrılmaktadır. 4. satırda her bir öznitelik kolonuna ait medyan değeri hesaplanmaktadır. 5. satırda özniteliklerMatrisi içerisindeki her bir eksik alan tespit edilmekte ve medyanVektörü değerleri ile doldurulmaktadır. 6. satırda eğitim ve test verisi ayrıştırılmaktadır. (Bu çalışmada test verisi %30 oranıyla bölünmüştür.) 7. satırda KA algoritmasının ihtiyaç duyduğu entropi değerleri hesaplanmakta ve 8. satırda KA algoritmasıyla eğitim işlemi gerçekleştirilmektedir. 9. satırda karar ağacı algoritmasına ait test sonuçları elde edilmektedir. 11. satırda, 10. satırda hesaplanan minkowski değerleri kullanılarak KNN algoritması hesaplamaları gerçekleştirilmektedir. (Bu çalışmada K değerini tespit etmek amacıyla 0-50 arası ardışık tam sayılar sırasıyla denenmiştir ve en iyi sonucu veren değerin 11 olduğu belirlenmiştir.) 12. satırda KNN algoritmasına ait test sonuçları alınmaktadır. Benzer şekilde 13. satırda näive bayes algoritmasıyla eğitim süreci gerçekleştirilmekte ve 14. satırda ilgili test sonuçları elde edilmektedir. 15. satırda üç sınıflandırıcının sonuçları Karar Fonksiyonuna gönderilmekte ve bu fonksiyon sayesinde birleştirilen sonuçlar geri döndürülerek ÇOKS algoritması sonlandırılmaktadır.

Nihai sonuçlar sözde kodu Şekil 4 ile verilen karar fonksiyonuna göre belirlenmektedir. Bu algoritma kendisine girdi olarak üç sınıflandırma algoritmasının sonuçlarını almakta ve Eşitlik (6) ile ifade edilen çoğunluk fonksiyonunu kullanarak elde ettiği sınıflandırmaVektörü'nü çıktı olarak geriye döndürmektedir. Algoritma kapsamında öncelikle üç sonuç vektörünün boyutunda ve içi sıfırlarla dolu bir vektör oluşturulmaktadır. Daha sonra her bir algoritma ile elde edilen her bir test verisi için çoğunluk oyu ile bir sınıflandırma yapılmaktadır. Sözde kod içerisindeki KASonuçVektörü Eşitlik (6)'daki x değişkenini, KNNSonuçVektörü y değişkenini ve NBSonuçVektörü de z değişkenini temsil etmektedir. Karar fonksiyonunun işlevine denk gelen ve bu işlevi görselleştiren mantıksal devre ise Şekil 5'te görülmektedir.

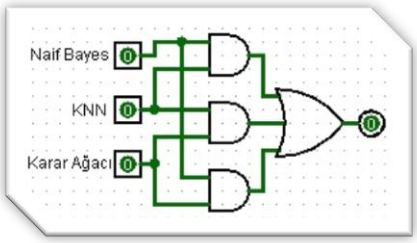
Algoritma 2: Karar Fonksiyonunun Sözde Kodu

Girdi: *KA Sonuç Vektörü*, *KNN Sonuç Vektörü*, *NB Sonuç Vektörü*
Çı***ktı***: *Sınıflandırma Vektörü*

```
1:  boyut ← VektörBoyutunuBul(KASonuçVektörü)
2:  sınıflandırmaVektörü ← BoşVektörOluştur(boyut, 0)
3:  for i ← 1 to boyut by 1
4:      if (KASonuçVektörü[i] VE KNNSonuçVektörü[i])
5:          sınıflandırmaVektörü[i] ← 1
6:      elseif (KNNSonuçVektörü[i] VE NBSonuçVektörü[i])
7:          sınıflandırmaVektörü[i] ← 1
8:      elseif (KASonuçVektörü[i] VE NBSonuçVektörü[i])
9:          sınıflandırmaVektörü[i] ← 1
10:     End if
11: End for
12: return sınıflandırmaVektörü
```

Şekil 4. ÇOKS'nin faydalandığı karar fonksiyonunun sözde kodu

ÇOKS, üç makine öğrenmesi algoritmasını eşit ağırlıkla karar sürecine dâhil etmiştir. Fakat bu tasarım her bir algoritma için değiştirilebilir bir esnekliğe sahiptir. Örneğin daha fazla sayıda makine öğrenmesi algoritmasının kullanılacağı bir sistemde hangi ölçütte (kesinlik, doğruluk, hassasiyet gibi) daha yüksek başarı elde edilmesi isteniyorsa o ölçütte daha iyi sonuçlar üreten algoritmanın karara olan ağırlığının yükseltilebilmesi mümkündür



Şekil 5. Karar modülüne ait mantıksal devre

3.2.4. Başarı Değerlendirme Ölçütleri

Bu çalışmada literatürde başarı ölçütü hesaplamalarında yaygın olarak kullanılan metriklerden üç tanesi seçilerek işlemler gerçekleştirilmiştir. Bu metrikler Doğruluk, Kesinlik ve ROC AUC değeridir ve Karmaşıklık Matrisi yardımıyla hesaplanmaktadır [23].

Karmaşıklık matrisi içerisinde başarı ölçüm metriklerinin hesaplanmasında Doğru Pozitif (True Positive, TP), Yanlış Pozitif (False Positive, FP), Yanlış Negatif (False Negative, FN) ve Doğru Negatif (True Negative, TN) değerleri kullanılmaktadır [24].

3.2.4.1. Doğruluk (Accuracy) değeri

Sınıflandırılan durumların ne kadarının doğru sınıflandırma sonucu ürettiğinin bir ölçüsüdür. Doğru sınıflandırılan verilerin tüm veri kümesi sayısına oranına doğruluk değeri denilmektedir ve bu değer Eşitlik (7) ile hesaplanmaktadır [25]. Bu çalışmada veri kümesinin entropisinin düşük olmasından dolayı başarı konusunda doğruluk değeri önem arz etse de gerçek sahtekarlık oranının tespitinde başarı ölçütü olarak yetersiz kalmaktadır.

$$Doğruluk = \frac{TP+TN}{TP+FP+FN+TN} \tag{7}$$

3.2.4.2. Kesinlik (Precision) değeri

Veri kümesi içerisinde doğru olarak etiketlenen durumların ne kadarının doğru pozitif etiketlendiğinin bir ölçüsüdür. Kısaca Kesinlik değeri, Positive olarak tahmin edilen değerlerin gerçekten kaç adedinin Positive olduğunu göstermektedir. Bu ölçüt ne kadar hassas bir sınıflandırma yapıldığı bilgisini vermektedir ve ilgili değer Eşitlik (8) ile hesaplanmaktadır [22].

$$Kesinlik = \frac{TP}{TP+FP} \tag{8}$$

3.2.4.3. ROC Eğrisi (ROC AUC) Değeri

Dengeli bir veri değeri dağılımına sahip olmayan veri kümeleri için önemli ve kullanılabilecek olan performans değerlendirme ölçütlerinin başında ROC eğrisi gelmektedir [38]. ROC Curve (Receiver Operating Characteristic Curve, Alıcı İşletim Karakteristik Eğrisi) değeri, doğru pozitif oranının (TPR), yanlış pozitif oranına (FPR) bölümü ile hesaplanan bir sınıflandırma başarısı ölçütüdür. TPR ve FPR değerlerinin hesaplanışları sırasıyla Eşitlik (9) ve (10) ile verilmiştir [26].

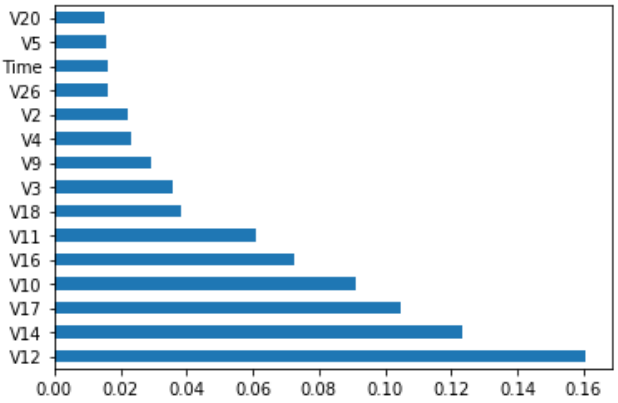
$$TPR = \frac{TP}{TP+FN} \tag{9}$$

$$FPR = \frac{FP}{FP+TN} \tag{10}$$

Eşitlik (9) ve (10) değerlerinin oranı ROC AUC değerini vermektedir. Aynı zamanda iki boyutlu koordinat ekseninde bu iki değeri baz alarak çizilen eğrinin altında kalan alan ne kadar fazla ise sınıflandırmanın başarısı o denli yüksek olarak değerlendirilmektedir.

4. Bulgular ve Tartışma

Deneysel sonuçlar, 10 GB RAM'e ve çift çekirdekli, 4 iş parçacıklı ve 2.4 GHz saat frekansında İntel işlemciye sahip bir bilgisayar kullanılarak yaklaşık 1 dakikalık sürede (algoritmaların eğitim süreleri de dâhil olmak üzere) elde edilmiştir. Yazılımda Python programlama dili kullanılmış ve Python 3.7 sürümünden faydalanılmıştır.



Şekil 6. Sınıflandırma işleminde başarı için en önemli etkiyi sağlayan ilk 15 öznitelik ve bunların katkı oranları (%)

Worldline ve ULB Machine Learning Group'un araştırma ve akademik çalışmalar yapmak için toplamış olduğu veri kümesinde bazı verilerin/özniteliklerin (V1-V28) gizliliği gözetilmiş ve bu verilere TBA algoritması uygulandıktan sonra veri kümesi erişime açılmıştır. Veri kümesindeki ödeme miktarı ve zaman öznitelikleri ise TBA uygulanmadan doğrudan kullanılmıştır. Şekil 6'da bu veri kümesindeki toplam 30 adet öznitelikten kredi kartı sahtekârlığı tespitinde sınıflandırma işlemine en çok katkı sağlayanlar fayda oranlarıyla birlikte listelenmiştir. Bu işlem için Extra Trees Classifier algoritmasından yararlanılmıştır ve ilgili problem için en değerli beş öznitelik sırasıyla V12, V14, V17, V10 ve V16 olarak belirlenmiştir.

ÇOKS yönteminde öncelikle veri kümesi %30 test (85443 adet) ve %70 (199364 adet) eğitim verisi olarak ikiye ayrılmıştır. Sonrasında karar ağacı, KNN ve näive bayes algoritmaları ilgili veri kümesinde eğitilmiştir. Ardından bu algoritmalar test işlemine tabi tutulmuş ve Tablo 1'de verilen doğruluk, kesinlik ve ROC AUC değerleri hesaplanmıştır. Daha sonra adı geçen 3 algoritmanın kararları birleştirilmiş ve ÇOKS yöntemi için test işlemi yapılarak ilgili parametreler tekrar hesaplanmıştır. Deneysel sonuçlar üç kez tekrar edilerek ortalama değerler kullanılmıştır.

Tablo 1 ile verilen bu çalışmadaki dört algoritmaya ait sonuçlar incelendiğinde tüm yöntemlerin doğruluk oranlarının çok yüksek (yaklaşık %99) ve birbirlerine yakın olduğu, en yüksek doğruluk değerinin ise ÇOKS ile elde edildiği anlaşılmaktadır. Bu değerler tüm algoritmaların TP ve TN değerleri yüksek sınıflandırmalar yaptığını göstermektedir.

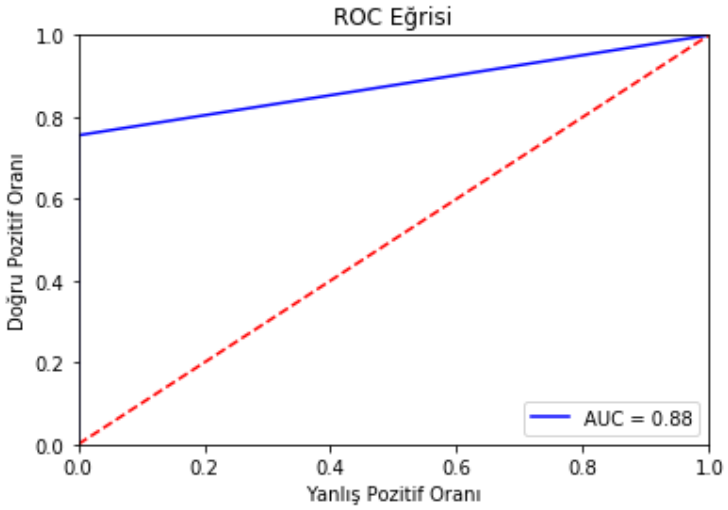
Tablo 1'deki kesinlik değerleri incelendiğinde KNN algoritmasının %100 değeri ile en iyi sonucu, näive bayes algoritmasının da %14,89 ile en kötü sonucu ürettiği görülmektedir.

Tablo 1. Çalışmada kullanılan yöntemlere ait başarı oranları

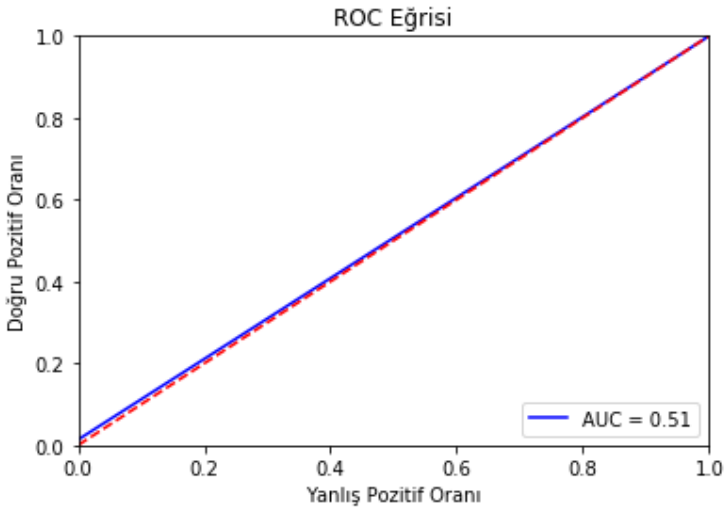
Algoritma	Doğruluk (%)	Kesinlik (%)	ROC AUC (%)
Karar Ağacı	99,92	81,02	88,09
KNN	99,83	100,0	51,80
Näive Bayes	99,30	14,89	82,04
ÇOKS	**99,93**	**95,60**	**80,00**

Tablo 1'deki ROC eğrisi değerleri dikkate alındığında en iyi sonucun karar ağacı algoritması ile elde edildiği anlaşılmaktadır. Öte yandan ÇOKS'de algoritmaların birlikte karar verebileceği bir yapı kurgulandığı için hem doğruluk hem kesinlik ve hem de ROC AUC değerlerinde bir iyileşme gözlemlenmektedir. Her ne kadar ÇOKS ile üç başarı ölçütü için her zaman en iyi değerler elde edilememiş olsa da toplu bir değerlendirme yapıldığında ÇOKS'nin üstünlüğü ortaya çıkmaktadır. Örneğin näive bayes algoritması bireysel olarak %14,89 kesinlik değeri ile kötü bir tablo çizmiştir. Algoritmanın bu eksik yönü ÇOKS içerisinde diğer algoritmalar tarafından kapatılmıştır. Benzer bir durum KNN algoritmasının ROC AUC ve KA'nın da kesinlik değeri için geçerlidir. Böylece ÇOKS'nin amacına ulaştığı deneysel sonuçlarla ortaya konulmuştur.

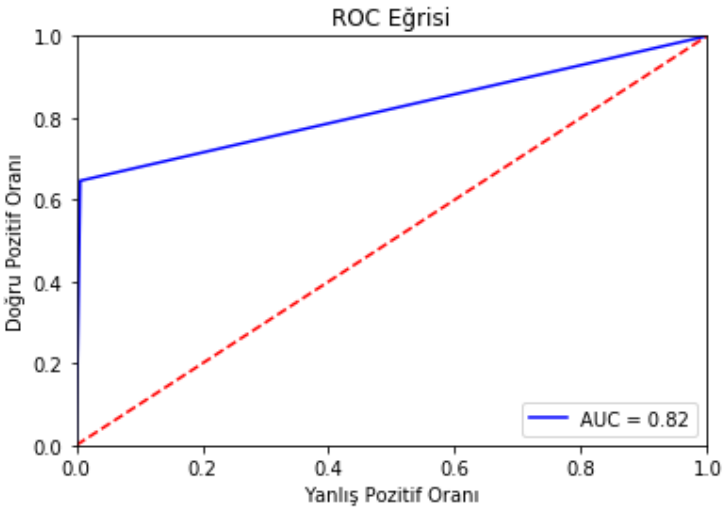
Şekil 7'de verilen karar ağacı algoritmasına ait ROC Eğrisi grafiği incelendiğinde mavi çizginin altında kalan alanın yaklaşık olarak tüm alanın %88'ine tekabül ettiği görülmektedir. Oldukça başarılı bir sınıflandırma işlemi gerçekleştirildiği anlamına gelen bu değerin, literatürdeki diğer çalışmalar incelendiğinde önemi daha iyi anlaşılmaktadır. Şekil 8'deki KNN algoritmasına ait ROC eğrisini gösteren grafik incelendiğinde KNN algoritmasının çalışmada kullanılan veri kümesinde başarılı bir değer ortaya koyamadığı anlaşılmaktadır. İki farklı sınıflandırma etiketi kullanılan bir çalışmada olasılıksal olarak %50 ihtimal bulunmaktadır ve Şekil 8'de kırmızı kesik çizgilerle gösterilen eğri bunu ifade etmektedir. KNN algoritmasının da %51 değeriyle bu orana çok yakın olduğu gözlemlenmektedir. Şekil 9 ve 10 da sırasıyla näive bayes ve ÇOKS algoritmalarının ROC eğrilerine ait grafikler gösterilmektedir. ÇOKS yönteminde diğer üç algoritmanın başarısı bir araya getirilerek %80 değerinde bir başarı elde edildiği görülmektedir.



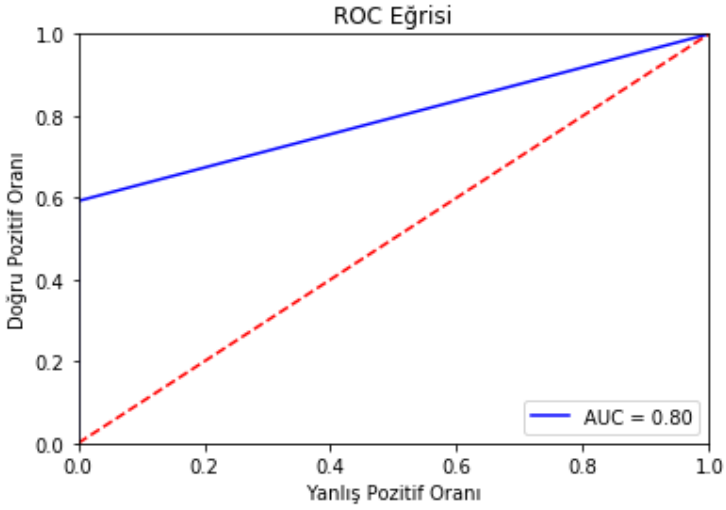
Şekil 7. KA için ROC eğrisi ve AUC değeri



Şekil 8. KNN için ROC eğrisi ve AUC değeri



Şekil 9. NB için ROC eğrisi ve AUC değeri



Şekil 10. ÇOKS için ROC eğrisi ve AUC değeri

Tablo 2, literatürdeki benzer araştırmalar ile ÇOKS'nin üç başarı ölçütüyle karşılaştırıldığı tablodur. Bu tablo incelendiğinde dördü hariç diğer algoritmaların kesinlik değerlerinin oldukça düşük

olduğu gözlemlenmektedir. Kesinlik değeri, tespit edilen dolandırıcılık işlemlerinin ne kadarının gerçekte dolandırıcılık işlemi olduğunu gösteren önemli bir ölçüttür. Veri kümesi incelendiğinde sahtekârlık işlemlerinin tüm işlemlerin sadece %0,172'sini oluşturduğu görülmektedir. Bu nedenledir ki yasal işlemlerin ezici çoğunluğu doğruluğu çok büyük ölçüde etkilemektedir. Yani algoritma, sahtekârlık işlemleri de dâhil tüm işlemleri sınıflandırırken %100'lük bir yasallık tespit ederse, başka bir deyişle hiç sahtekârlık tespit edemez ise doğruluk oranı %99,822 olacaktır. Dolayısıyla bu veri kümesi açısından doğruluğun tek başına bir ölçüt olarak kullanılması uygun değildir. Ayrıca eldeki problem açısından yasal işlemlerin tespitindeki başarı yerine sahtekârlık işlemlerindeki başarı dikkate alınmalıdır.

Tablo 2. ÇOKS ile literatürde ilgili alanda yapılmış diğer üç çalışmadaki algoritmaların başarı oranları

Algoritma	Çalışma	Doğruluk (%)	Kesinlik (%)	ROC AUC (%)
Lojistik Regresyon	[1]	36,39	16,78	50,47
Näive Bayes		97,52	5,46	89,75
Yapay Sinir Ağı	[16]	60	-	-
Bayes İnanç Ağları		68	-	-
Yerel Anormallik Faktörü	[27]	89,9	00,38	-
İzolasyon Ormanı		90,1	01,47	-
Destek Vektör Makinesi		99,8	76,81	-
Lojistik Regresyon		99,9	87,50	-
Karar Ağacı		99,9	88,54	-
Rastgele Orman		99,9	93,10	-
Derin Öğrenme Modeli	[40]	99,9	83,30	81,0
Rastgele Orman		99,9	94,10	84,9
Sınıflandırıcı Yığını Modeli		99,9	81,20	81,7
ÇOKS (Geliştirilen Yöntem)		**99,93**	**95,60**	**80,00**

Bu çalışmada kullanılan veri kümesi gibi entropisi düşük olan veri kümelerinde yukarıda bahsedilen nedenden dolayı kesinlik değeri ve ROC AUC değeri gerçek başarı tespitinde oldukça önem kazanmaktadır. Bu nedenle bu çalışmada, doğruluk değerinin yanında bu iki değeri de yükseltecek ÇOKS yöntemi önerilmiştir.

Tablo 2'de ÇOKS, aynı veri kümesinin kullanıldığı dört farklı çalışma kapsamında gerçeklenen algoritmalarla karşılaştırılmıştır. ÇOKS ile mukayese edilen birinci çalışmada ([1]) lojistik regresyon ve näive bayes algoritmaları gerçeklenmiştir. Kıyaslamanın yapıldığı ikinci çalışmada ([16]) yapay sinir ağı ve bayes inanç ağları yöntemleri kullanılarak ölçümler yapılmıştır. Üçüncü çalışmada ([27]) 6 farklı algoritma ile deneysel sonuçlar elde edilmiştir. Dördüncü çalışmada [40] ise optimizasyona önem verilerek üç farklı algoritma kullanılarak problem giderilmeye çalışılmıştır. Tüm bu değerler incelendiğinde üç ölçüt için de makul başarıda sonuçların ÇOKS algoritmasıyla elde edildiği anlaşılmaktadır. Öte yandan en yüksek doğruluk ve kesinlik değerlerini veren algoritmanın da ÇOKS olduğu görülmektedir.

5. Sonuç ve Öneriler

Yapay zeka çalışma alanı olarak birçok teori, metot ve teknolojiyi içermektedir ve insanoğlunun karşılaştığı bir çok güçlüğe çözüm üretmektedir [39]. Kredi kartı sahtekârlığının yapay zekâsız tespitinde çekilen güçlükler ve yetersizlikler ile bu alanda son yıllarda artan dolandırıcılıklar araştırmacıları sorunun çözümüne yönelik yeni çalışmalar yapmaya ve algoritmalar geliştirmeye

zorlamaktadır. Bu motivasyonla bu çalışma kapsamında bankacılık ve ödeme sistemlerinin güvenliğine katkı sunacak yeni bir sezgisel bir algoritma literatüre kazandırılmıştır. ÇOKS olarak adlandırılan bu algoritma bu alanda yürütülen çalışmalarda sıklıkla kullanılan ve 284,807 adet işlem içeren büyük bir veri kümesi ile eğitilmiştir ve sonrasında elde edilen model ile testler yapılmıştır. Literatürdeki benzer çalışmaların sonuçları ile yapılan karşılaştırmalar önerilen bu yeni yöntemin %99,93 doğruluk, %95,60 kesinlik ve %80,0 ROC AUC değerleri ile önemli bir başarıya imza attığını göstermiştir. ÇOKS'nin eğitim ve sınıflandırma sürecinde çok uzun bir çalışma zamanı maliyeti oluşturmaması ise bir başka avantaj olarak ortaya çıkmıştır.

n adet Boolean değişken ile 2^(2^n) farklı lojik fonksiyon gerçeklenebilmektedir. Örneğin n=3 iken 256 farklı lojik fonksiyonun varlığı söz konusudur. Bu fonksiyonlardan üçü F=0, F=1 ve çoğunluk fonksiyonudur. Dolayısıyla bu üç fonksiyon devre dışı bırakıldığında ÇOKS algoritmasının karar verme modülü bünyesinde değerlendirilebilecek 253 farklı lojik fonksiyon mevcuttur. Bu fonksiyonların her birinin ÇOKS'ye entegre edilmesi ve ilgili sonuçların çoğunluk fonksiyonu ile elde edilen sonuçlarla karşılaştırılması ise gelecekte yapılması planlanan bir çalışmadır.

Alınan güvenlik tedbirlerine rağmen gün geçtikçe farklı yöntemler geliştiren dolandırıcılara karşı ilgili çalışmaların hız kesmeden devam ettirilmesi ve yeni koşullara uygun tekniklerin geliştirilmesi elzem görünmektedir. Bu nedenle gelecek çalışmalarda öncelikle daha dengeli işlemler içeren veri kümelerinin oluşturulmasının ve eldeki problem için daha iyi sonuçlar üretebilecek çok katmanlı yapay sinir ağı ve derin öğrenme yöntemlerinden faydalanılması gerektiği değerlendirilmektedir. Buna ek olarak eğitim ve test sürelerinin uzaması ihtimaline karşı çeşitli optimizasyon ve paralel programlama tekniklerinin de [28, 29] değerlendirmeye alınabileceği düşünülmektedir.

Teşekkür

Bu çalışma TÜBITAK 2211-A programı tarafından desteklenmiştir.

Yazar katkı oranları

MFK araştırma için fikir ya da hipotezin oluşturulmasında ve sonuçlara ulaşmak için yöntemlerin planlanması aşamasında görev almıştır. Çalışma ve yazının organizasyonu ve seyrinin gözetimi ve sorumluluğunda DD ve TA iş birliği yapmıştır. MFK ve TA deneylerin yapılması, takibi, verilerin düzenlenmesi ve bildirilmesi için sorumluluk almıştır. MFK ve DD bulguların mantıklı açıklaması ve sunumu için sorumluluk almıştır ve her iki yazar makale metni yazma sürecinin tümünde görev almıştır. DD yazıyı teslim etmeden önce sadece imla ve dil bilgisi açısından değil, aynı zamanda entelektüel içerik açısından yeniden çalışma yapmıştır.

Her üç yazar da son olarak makaleyi okumuş ve onaylamıştır.

Çıkar çatışması

Makalenin hazırlanması, uygulanması veya değerlendirilmesiyle ilgili olarak herhangi bir çıkar çatışması bulunmadığını yazarlar beyan etmektedir.

Kaynaklar

[1]. Awoyemi J. O., Adetunmbi A. O., Oluwadare S. A., Credit card fraud detection using machine learning techniques: A comparative analysis, 2017 International Conference on Computing Networking and Informatics (ICCNI), 1-9, (2017).
[2]. Kaya D. F., Türkiye'de Kredi Kartı Uygulaması, İstanbul: Türkiye Bankalar Birliği, (2009).

[3]. Shen A., Tong R., Deng Y., Application of classification models on credit card fraud detection, In Service Systems and Service Management, 2007, 1-4.
[4]. Chaudhary K., Mallick B., Credit Card Fraud: The study of its impact and detection techniques, International Journal of Computer Science and Network (IJCSN), 2012, 1 (4): 31-35.
[5]. Robertson D., U.S. Credit & Debit Cards 2015, The Nilson Report, (2015).
[6]. Maes S., Tuyls K., Vanschoenwinkel B., Manderick B., Credit card fraud detection using Bayesian and neural networks, Proceeding International NAISO Congress on Neuro Fuzzy Technologies, (2002).
[7]. Kundu A., Panigrahi S., Sural S., Majumdar A. K., Credit card fraud detection: A fusion approach using Dempster–Shafer theory and Bayesian learning, Special Issue on Information Fusion in Computer Security, 2009, 10 (4): 354-363.
[8]. RamaKalyani K., UmaDevi D., Fraud Detection of Credit Card Payment System by Genetic Algorithm, International Journal of Scientific & Engineering Research, 2012, 3 (7): 1-6.
[9]. Meshram P. L., Bhanarkar P., Credit and ATM Card Fraud Detection Using Genetic Approach, International Journal of Engineering Research & Technology (IJERT), 2012, 1 (10): 1-5.
[10]. Maes S., Tuyls K., Vanschoenwinkel B., Manderick B., Credit card fraud detection using Bayesian and neural networks, Interactive image-guided neurosurgery, 1993, 261-270.
[11]. Haykin S., Neural Networks: A Comprehensive Foundation (2nd Edition), (1999).
[12]. Chiu A., Tsai C., A Web Services-Based Collaborative Scheme for Credit Card Fraud Detection, Proceedings of the IEEE International Conference on e-Technology, e-Commerce and e-Service, 177-181, (2004).
[13]. Brause R., Langsdorf T., Hepp M., Neural Data Mining for Credit Card Fraud Detection, International Conference on Tools with Artificial Intelligence, 103-106, (1999).
[14]. Ng A. Y., Jordan M. I., On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes, Advances in neural information processing systems, 2002, 2: 841-848.
[15]. Raj S. B. E., Portia A. A., Analysis on Credit Card Fraud Detection Methods, International Conference on Computer, Communication and Electrical Technology, 152-156, (2011).
[16]. Maes S., Tuyls K., Vanschoenwinkel B., Manderick B., Credit card fraud detection using Bayesian and neural networks, In Proceedings of the 1st international naiso congress on neuro fuzzy technologies, 261-270, (2002).
[17]. Soylu K., Kredi Kartı Sahte İşlem Tespiti, Master, Bilgisayar Mühendisliği, Ankara Üniversitesi, Ankara, (2018).
[18]. Sahin Y., Duman E., Detecting Credit Card Fraud by Decision Trees and Support Vector Machines, Proceedings of International Multi-Conference of Engineers and Computer Scientists, 1: 1-6, (2011).
[19]. Bhattacharyya S., Jha S., Tharakunnel K., Westland J. C., Data mining for credit card fraud: A comparative study, Decision Support Systems, 2011, 50 (3): 602-613.
[20]. Fabrizio C., Andrea D. P., Yann-Aël L. B., Olivier C., Yannis M., Gianluca B., Scarff: a scalable framework for streaming credit card fraud detection with Spark, Information fusion Elsevier, 2018, 41: 182-194.
[21]. Zhang S., KNN-CF Approach: Incorporating Certainty Factor to kNN Classification, IEEE Intelligent Informatics Bulletin, 2010, 11 (1).
[22]. Keskenler M. F., Mikroskobik Görüntülerde Sperm Yoğunluk Tespiti, Yüksek Lisans, Fen Bilimleri Enstitüsü Bilgisayar Mühendisliği Anabilim Dalı, Atatürk Üniversitesi, Erzurum, (2019).
[23]. Keskenler M. F., Haşıloğlu A., Özyer G. T., Özyer B., Şimşek E., Sperm Detection and Analysis Using Feature Description Algorithms, Signal Processing and Communications Applications Conference (IEEE), 1-4, (2019).

[24]. W. Y., L. J., Y. M. H., Online Object Tracking: A Benchmark, Computer Vision Foundation, 2411- 2418.
[25]. Thompson T., Lloyd A., Joseph A., Weiss M., The Weiss Functional Impairment Rating Scale-Parent Form for assessing ADHD: evaluating diagnostic accuracy and determining optimal thresholds using ROC analysis, Quality of Life Research, 2017, 26 (7): 1879-1885.
[26]. Pinchi V., Pradella F., Vitale G., Rugo D., Nieri M., Norelli G.-A., Comparison of the diagnostic accuracy, sensitivity and specificity of four odontological methods for age evaluation in Italian children at the age threshold of 14 years using ROC curves, 2016, 56 (1): 13-18.
[27]. Vaishnavi Nath D., Geetha S., Credit Card Fraud Detection using Machine Learning Algorithms, Procedia Computer Science, 2019, 165: 631-641.
[28]. Keskenler M. F., Keskenler E. F., Yoğun İşlem Yüküne Sahip Matris Çarpımı Hesaplama Sürelerinin Önbellek Kullanım Optimizasyonu ve Paralel Programlama Teknikleri Kullanılarak İyileştirilmesi, Muş Alparslan Üniversitesi Fen Bilimleri Dergisi, 2018, 6 (2): 545-551.
[29]. Lotfollahi M., Jafari Siavoshani M., Shirali Hossein Zade R., Saberian M., Deep packet: a novel approach for encrypted traffic classification using deep learning, Soft Computing, 2020, 24 (3): 1999-2012.
[30]. Dietterich T. G, Ensemble Methods in Machine Learning, Oregon State University USA, (2001).
[31]. Matteo R., Giorgio V., Ensemble methods: A review, (2012).
[32]. Elçiboğa İ. K., Türkiye'de Kart Dolandırıcılık Trendlerinin Değişimi, Fintechtime, (2019).
[33]. Fathima N., Shoukat S., Mohiuddin S., Afzal M., Implementation Of Majority Voting And Adaboost For Credit Card Fraud Detection, International Journal Of Merging Technology And Advanced Research In Computing, 2020, 8 (29): 1-8.
[34]. Yıldız B., Applying Decision Tree Techniques to Classify European Football Teams, Journal of Soft Computing and Artificial Intelligence, 2020, 1 (2): 86-91.
[35]. Uğuz S., Oral O. Ç., PV Güç Santrallerinden Elde Edilecek Enerjinin Makine Öğrenmesi Metotları Kullanılarak Tahmin Edilmesi, International Journal of Engineering Research and Development, 2019, Aralık 2019-Özel Sayı: 769-779.
[36]. Patil T.R., Sherekar S.S., Performance analysis of Naive Bayes and J48 classification algorithm for data classification, International journal of computer science and applications, 2013, 6(2): 256-261.
[37]. Uludağ O., Gürsoy A., On the Financial Situation Analysis with KNN and Naive Bayes Classification Algorithms, Journal of the Institute of Science and Technology, 2020, 10 (4): 2881-2888.
[38]. Uğuz S., Makine öğrenmesi : teorik yönleri ve Python uygulamaları ile bir yapay zeka ekolü, Nobel Akademik Yayıncılık, Ankara, (2019).
[39]. Aylak B., Oral O., Yazıcı K., Yapay Zeka ve Makine Öğrenmesi Tekniklerinin Lojistik Sektöründe Kullanımı, El-Cezeri Journal of Science and Engineering, 2021, 8 (1): 74-93.
[40]. Soylu K., Kredi Kartı Sahte İşlem Tespiti, Yüksek Lisans, Ankara Üniversitesi Fen Bilimleri Enstitüsü, (2018).