Yayın Geliş Tarihi (Submitted): **05/10/2023**

Yayın Kabul Tarihi (Accepted): **27/05/2024**

Makele Türü (Paper Type): **Araştırma Makalesi – Research Paper**

Please Cite As/Atıf için:
Diler, S. ve Demir, Y. (2024), Çoklu doğrusal bağlantı olması durumunda veri madenciliği algoritmaları performanslarının karşılaştırılması, *Nicel Bilimler Dergisi*, 6(1), 40-67. *doi: 10.51541/nicel.1371834*

ÇOKLU DOĞRUSAL BAĞLANTI OLMASI DURUMUNDA VERİ MADENCİLİĞİ ALGORİTMALARI PERFORMANSLARININ KARŞILAŞTIRILMASI

Saygın Diler[1] ve Yıldırım Demir[2]

ÖZ

Bilgisayar teknolojilerindeki gelişmelere paralel olarak veri madenciliği algoritmaları ile yapılan çalışmalarda artış yaşanmaktadır. Sınıflandırma algoritmalar ile yapılan çalışmalarda veri kalitesinin bozulması algoritmaların performansında önemli rol oynamaktadır. Bu çalışmada veri kalitesini bozan etmenlerden birisi olan çoklu doğrusal bağlantının veri setinde bulunması durumunda sınıflandırma algoritmalarının performansının nasıl etkilendiği incelenmiştir. Çoklu doğrusal bağlantının varlığını tespit etmek için veri setlerine ait korelasyon grafikleri incelenmiş daha sonrasında ise koşul endeksi ile çoklu doğrusal bağlantının derecesi belirlenmiştir. Sınıflandırma algoritmalarından olan Naive Bayes (NB), Lojistik Regresyon (LR) ve K-En Yakın Komşu Algoritması (kNN), Destek Vektör Makineleri (SVM) ve Aşırı Gradyan Arttırma Algoritması (XGBoost) ile uygulamalar gerçekleştirilmiştir. Yöntemlerin performanslarının incelenmesi için simülasyon çalışması ve gerçek veri setleri ile uygulamalar yapılmış, sonuçlar tablolar halinde sunulmuştur. Analiz sonuçlarına göre, çoklu doğrusal bağlantı varlığında büyük örneklem hacimli veri setlerinde doğruluk ve F-ölçütü metriklerine göre XGBoost algoritmasının diğer algoritmalardan dikkate değer performans farklılığı gösterdiği belirlenmiştir. Çoklu doğrusal bağlantından performansı en olumsuz etkilenen algoritmanın ise Naive Bayes olduğu gözlenmiştir.

[1] Sorumlu yazar, Türkiye İstatistik Kurumu, Ankara, Türkiye. ORCID ID: https://orcid.org/0000-0002-9056-412X
[2] Doç. Dr., İktisadi ve İdari Bilimler Fakültesi, Van Yüzüncü Yıl Üniversitesi, Van, Türkiye. ORCID ID: https://orcid.org/0000-0002-6350-8122

Anahtar Sözcükler: Çoklu doğrusal bağlantı, Sınıflandırma, Veri madenciliği

COMPARISON OF DATA MINING ALGORITHMS PERFORMANCES IN CASE OF MULTICOLLINEARITY

ABSTRACT

As advancements in computer technologies progress, there has been an increase in research utilizing data mining algorithms. In studies involving classification algorithms, the degradation of data quality plays a significant role in algorithm performance. This study investigates the impact of multicollinearity, one of the factors that compromise data quality, on the performance of classification algorithms. To identify the presence of multicollinearity, correlation graphs of the datasets were examined, followed by the determination of the degree of multicollinearity using the condition index. The classification algorithms, namely Naive Bayes (NB), Logistic Regression (LR), k-Nearest Neighbors (kNN), Support Vector Machines (SVM), and Extreme Gradient Boosting (XGBoost), were implemented for the analysis. Simulation studies and real dataset analyses were conducted to assess the performance of these methods, and the results were presented in tabular form. According to the analysis results, it has been determined that XGBoost algorithm shows a notable performance difference compared to other algorithms in terms of accuracy and F-measure metrics in the presence of multicollinearity in large sample-sized datasets. On the other hand, Naive Bayes was observed to be the algorithm most adversely affected by multicollinearity, showing diminished performance.

Keywords: Classification, data mining, multicollinearty.

1. GİRİŞ

Teknolojik gelişmelerin günümüzde hızla ilerlemesine bağlı olarak büyük miktarda ve çeşitlilikte veri üretilmektedir. Her geçen gün daha da artarak veri tabanlarında biriken karmaşık ve büyük veri setlerinden değerli bilgilerin ortaya çıkarılması için geliştirilmiş farklı yöntemler bulunmaktadır. Veri madenciliği yöntemi büyük veri setlerinden bilgilerin keşfedilmesi için geliştirilmiş yaklaşımlar arasında yer almaktadır (Han vd., 2012). Veri setinin kalitesi istatistiksel analiz süreçlerinde ve veri madenciliği sınıflandırma algoritmaları

ile yapılan çalışmalarda büyük bir öneme sahiptir. Çünkü bu algoritmalar ile yapılan analizler veri setinin kalitesinden etkilenmekte ve dolayısıyla algoritmaların etkin performans göstermesi veri kalitesiyle ilişkili olmaktadır (Batista ve Monard, 2002). Veri kalitesini bozan çeşitli etmenler bulunmakta ve çoklu doğrusal bağlantı bu etmenler arasında yer almaktadır. Veri madenciliği yöntemleri ile yapılan analizlerin başarısı, büyük ölçüde algoritma ve veri kalitesi ile ilişkilidir. Verilerin analize uygunluğu veri madenciliği çalışmalarının varsayımları arasında bulunmaktadır. Veri yapısı ve kalitesini etkiyen sorunlarla karşılaşmak istatistik biliminin doğal bir parçası olarak görülebilir. Bu tür sorunlar, veri madenciliği çalışmalarında sonuçları etkileyebilir. Veri kalitesi düşük olduğunda klasik veri madenciliği yöntemleri etkisiz hale gelebilmekte ve bu, algoritma performanslarını etkilemektedir (Zhu vd., 2018).

Burada yöntem performanslarının kötü etkilenmesi iki şekilde ele alınabilir: birincisi veriyi temsil edememek ve düşük doğuruluk, keskinlik gibi değerlere sebep olmak, ikincisi makine öğrenmesi yöntemlerinde sıkça karşılaşılan aşırı uyum problemi ki bu da performans kriterlerinin yüksek çıkmasına sebep olmaktadır. Bu çalışmada söz konusu sorun, birinciden ziyade ikinci sorun olan aşırı uyum problemidir. Bu durum klasik regresyon modellerinde daha çok çoklu doğrusal bağlantıdan kaynaklanmaktadır. Aşırı uyum sorununun çözümü, farklı makine öğrenmesi yöntemleriyle birçok farklı alanda araştırılmıştır (Roelofs vd., 2019; Ying, 2019). Bu araştırmanın temel motivasyonlarından birisi, çözümden bağımsız şekilde bu sorunun farklı sınıflandırma algoritmalarını nasıl etkilediğinin ortaya çıkarılması ve karşılaştırma yapılarak yöntemlerin dayanaklılığı veya avantaj ve dezavantajları hakkında çıkarsamalar yapmaktır.

Çoklu doğrusal bağlantı sorunu için istatistik alanında oldukça fazla literatür bulunmasına rağmen veri madenciliği, makine öğrenmesi gibi alanlarda çok az literatür bulunmaktadır (Garg ve Tai, 2013). Bu bağlamda veri kalitesini bozan etmenler arasında yer alan çoklu doğrusal bağlantının veri setinde olması durumunda sınıflandırma algoritmalarına ait performansların incelenmesi çalışmanın temel motivasyonunu oluşturmaktadır.

Literatürde, çoklu doğrusal bağlantı olması durumunda veri madenciliği yöntemlerini kullanan çalışmalardan bazılarına kısaca değinecek olunursa; Garg ve Tai (2013) vücuttaki yağ içeriğinin tahmin edilmesi için çoklu doğrusal bağlantıya sahip veri setinde uygulama gerçekleştirmiştir. Çalışmada makine öğrenmesi algoritmaları kullanılmış ve Yapay Sinir Ağır (YSA) ile genetik programlamanın başarılı sonuçlar verdiği gözlenmiştir. Blommaert vd. (2014) tarafından yapılan çalışmada, cezalandırılmış genelleştirilmiş tahmin denklemleri kullanılarak çoklu doğrusallık ve zamana bağımlılık altında uzunlamasına veriler için veri

madenciliği algoritmaları ile uygulama yapılmıştır. Dumancas ve Bello (2015) büyük veri analitiği ve yüksek performanslı veri madenciliğinde çoklu doğrusallığı ele almada makine öğrenmesi tekniklerinin karşılaştırılması üzerine araştırma gerçekleştirmiştir. Senawi vd. (2017), çoklu doğrusallığı azaltmak için özellik seçimi yöntemi olarak maksimum ilişki-minimum çoklu doğrusallık (MRmMC) yöntemi kullanmış ve sınıflandırma algoritmalarında iyi performans elde etmiştir. Obite vd. (2020) çoklu doğrusal bağlantıya sahip gerçek ve simülasyon verileri ile uygulama gerçekleştirmiştir. Uygulamada, Yapay Sinir Ağır (YSA) modelinin, Sıradan En Küçük Kareler Regresyonundan daha iyi bir uyum ve tahmin elde ettiği belirtilmiştir. Zhang vd. (2021), çoklu doğrusallığa sahip hisse senedi fiyat hareketlerini tahmin etmek derin çarpanlara ayırma makinesi ve dikkat mekanizmasına dayalı (FA-CNN) sinir ağı modeli üzerine odaklanmıştır. Çalışma, girdi özellikleri arasındaki gün içi etkileşimleri ve yan sanayi endeksi gibi ek bilgilerin tahmin doğruluğunu arttırdığını göstermişlerdir. Rahman vd. (2021), meme kanserinin erken teşhisindeki performansı artırmak amacıyla çoklu doğrusallık analizi ve makine öğrenimi modelleri kullanarak meme kanseri belirteçlerini ve teşhis sistemini geliştirmeyi hedeflemektedir. Çalışmada destek vektör makineleri algoritması en başarılı algoritma olarak belirlenmiştir. Urooj vd. (2022), Android işletim sistemi uygulamalarındaki kötü amaçlı yazılımları tespit etmek için makine öğrenimi sınıflandırma algoritmalarını kullanarak çoklu doğrusallığa sahip verileri ele almış ve önerilen modelin %96.24 doğrulukla kötü amaçlı yazılımları tespitte etkili olduğunu göstermişlerdir. Chan vd. (2022), çoklu doğrusallığı azaltmak için değişken seçimi ve değiştirilmiş tahmin yöntemlerinin kullanıldığına değinmiş, son araştırmalarda makine öğrenimi ile optimizasyon yaklaşımının çoklu doğrusallığı daha iyi ele aldığını belirtmişlerdir. Derraz vd. (2023), pirinç biyokütlesinin tahmininde temel ve topluluk makine öğrenmesi algoritmaları kullanmışlardır. Çalışmada çoklu doğrusal bağlantı olan ve çoklu doğrusal bağlantı olamayan veriler ile pirinç biyokütle tahmini gerçekleştirilmiştir. Çalışmada, topluluk makine öğrenmesi algoritmalarının temel makine öğrenmesi algoritmalarına göre daha başarılı performans gösterdiği belirlenmiştir.

Bu çalışmada, çoklu doğrusal bağlantı olması durumunda veri madenciliği sınıflandırma algoritmalarına ait sınıflandırma performansları incelenmektedir. Analiz öncesinde çoklu doğrusal bağlantının varlığını incelemek için çoklu korelasyon ısı grafikleri ve daha sonrasında ise çoklu doğrusal bağlantının derecesini belirlemek için koşul indeksi incelenmiştir. Çoklu doğrusal bağlantıya sahip iki gerçek veri seti ve simülasyon verileri ile uygulamalar gerçekleştirilmiş ve algoritmaların performansları karşılaştırılmıştır.

2. MATERYAL VE YÖNTEM

Veri setinde çoklu doğrusal bağlantı olduğunda sınıflandırma algoritmalarının performanslarını karşılaştırılmak amacıyla ilk önce iki gerçek veri setiyle ve daha sonra ise simülasyon çalışmasıyla uygulamalar gerçekleştirilmiştir. Veri setleri eğitim (%75) ve test (%25) verisi olacak şekilde ikiye ayrılmıştır. Eğitim veri setiyle modeller oluşturulmuş ve daha sonra test veri setiyle bu modellerin performansları ölçülmüştür. Model performanslarını karşılaştırmada ise kesinlik, seçicilik, duyarlılık, doğruluk ve F-ölçüt kriterleri kullanılmıştır. Çalışmada, Destek Vektör Makineleri, Lojistik Regresyon, Naïve Bayes, k-En Yakın Komşu (kNN) ve Aşırı Gradyan Arttırma algoritmaları ile uygulamalar gerçekleştirilmiş ve R programlama dili kullanılmıştır.

2.1. Sınıflandırma Algoritmaları

Sınıflandırma yöntemlerinde amaç bağımsız değişkenler yardımıyla genellikle kategorik verilerden oluşan bağımlı değişkeni tahmin etmektir. Sınıflandırma için çalışma yapısı ve varsayımları birbirinden farklılık gösteren birçok algoritma geliştirilmiştir (Davidson ve Tayi, 2009).

2.1.1. K-En Yakın Komşu Algoritması

kNN algoritması, veri madenciliğinde en çok kullanılan algoritmalardan birisidir. Algoritma, sınıfları bilinen veri setinde sınıfları bilinmeyen yeni verileri en yakın komşularına atama mantığı ile çalışmaktadır (Mucherino vd., 2009). Bu çalışmada, kNN algoritması için "k" değeri doğruluk ölçütünü maksimum yapabilecek şekilde çapraz geçerlilik yöntemiyle 1 ile 20 arasında bir değer seçilerek optimize edilmiştir. Zira kNN algoritmasının doğru sınıflandırma yapabilmesi büyük ölçüde komşu sayısının (k değeri) uygun seçimiyle ilgilidir.

2.1.2. Naive Bayes Algoritması

Bayesci sınıflama yöntemi istatistik tabanlı algoritmalar arasında yer almakta ve algoritma gözlemelere dayalı olasılıklar ile olasılık dağılımındaki parametreleri hesaplamaktadır (McNamara vd., 2006). Yöntem, veri setinde sınıfları belli olan verileri kullanarak yeni bir verinin mevcut sınıflardan birine girme olasılığını belirlemektedir (Silahtaroğlu, 2013).

Sınıfları belirlemede algoritma koşullu olasılıkları kullanmakta ve olasılıklar,

$$P(C_i|X) = \frac{P(X|C_i)P(C_i)}{P(X)} \quad (1)$$

olarak hesaplanmaktadır. Burada $C_1, C_1, \dots, C_m$ sınıf değerlerini ve $X = \{x_1, x_2, \dots, x_n\}$ nitelik değerlerinden oluşan ve sınıfı bilinmeyen veri örneğini göstermektedir.

Bilinmeyen bir örneği sınıflandırmak amacıyla maksimum değer Eşitlik (1) ile hesaplanmakta ve bilinmeyen X örneğinin bu sınıfta olabileceğine karar verilmektedir. Eşitlik (2) sonlu olasılıkları kullanmakta ve en büyük sonlu sınıflandırma yöntemi olarak bilinmektedir. Bayes sınıflayıcısı Eşitlik (3)'ü kullanmaktadır (Özkan, 2008).

$$arg\max_{Ci}\{P(X|C_i)P(C_i)\} \quad (2)$$

$$C_{MAP} = arg\max_{Ci}\prod_{k=1}^{n} P(X_k|C_i) \quad (3)$$

2.1.3. Lojistik Regresyon Analizi

Lojistik regresyonda bağımlı değişkenler kategorik verilerden oluştuğu için doğrusal regresyon analizinin özel bir hali olarak ifade edilmektedir (Lewis, 2017). Lojistik regresyon analizi, veri madenciliğinde sınıflandırma amacı ile kullanılmaktadır. Bağımsız değişken sayısı p olduğunda lojistik fonksiyon Eşitlik (4) ile gösterilmekte ve değişken sayısı 1 olduğunda bu fonksiyon sigmoid fonksiyon olarak adlandırılmaktadır (Harrington, 2012).

$$P(Y = 1|\, x) = \frac{e^{\beta_0+\beta_1X_1+\cdots+\beta_pX_p}}{1 + e^{\beta_0+\beta_1X_1+\cdots+\beta_pX_p}} = \frac{1}{1 + e^{-(\beta_0+\beta_1X_1+\cdots+\beta_pX_p)}} \quad (4)$$

Y genellikle 0 ve 1 değerlerini alan kategorik bağımlı değişkeni, x açıklayıcı değişkenleri ve β model parametrelerini ifade etmektedir. Bağımsız değişkenlerin değerleri bilindiğinde bağımlı değişkenin olasılığını $P(Y|\, x)$ göstermektedir. Logaritmik dönüşüm uygulanarak bu ilişki Eşitlik (5)'deki gibi doğrusal bir şekilde incelenebilir.

$$ln\left(\frac{P(Y = 1|\, x)}{1 - P(Y = 1|\, x)}\right) = \beta_0 + \beta_1X_1 + \cdots + \beta_pX_p \quad (5)$$

$P(Y = 1|\, x)/(1\text{-}\, P(Y = 1|\, x))$, odds oranı olup bir olayın gerçekleşme olasılığının gerçekleşmeme olasılığına oranını ifade etmektedir (Hosmer vd., 2013).

2.1.4. Destek Vektör Makineleri

Destek vektör makinaları (Support vector machine-SVM) el yazılarını, yüz, nesne, örüntü tanıma, zaman serileri, konuşmacı tanıma metin kategorizasyonu, gibi çok çeşitli alanlarda hem sınıflandırma hem de regresyon amacı ile kullanılmaktadır (Burges, 1998; Campbell ve Yiming, 2011). Destek vektör makinaları doğrusal ve doğrusal olmayan sınıflandırma olmak üzere iki kısımda incelenmektedir (Han vd., 2012).

Doğrusal destek vektör makinelerinde sınıflandırma yapılırken girdi uzayındaki örnekler dikkate alınır ve sınıflar arasında çizilebilecek sonsuz sayıdaki hiperdüzlemler arasından sınıfları birbirinden olabildiğince uzak sınıflandıran hiperdüzlem optimum seçim kabul edilir (Cervantes vd., 2020).

Hard-Marjin Destek Vektör Makineleri: D boyutlu (nitelik) örnek sayısı L ve x_i (i=1, …, L) girdileri yalnızca iki sınıfa ait olsun. Verilerin birbirlerinden doğrusal olarak ayrıldığı varsayımına dayanarak D=2 olduğu durumda iki sınıf birbirinden bir doğru ile ayrılırken, D>2 olduğu durumda ise iki sınıf bir hiperdüzlem ile birbirlerinden ayrılabilir. w hiperdüzleminin ağırlık vektörü, ‖w‖ ise w'nin öklit normunu gösterirse b/‖w‖ hiperdüzlemden orijine dik uzaklık olmak üzere hiperdüzleme ait denklem Eşitlik (6)'daki gibi yazılabilir (Burges, 1998).

$$w * x + b = 0 \tag{6}$$

Destek vektörleri hiperdüzleme en yakın noktalar olup amacı her iki sınıfın birbirlerine yakın üyelerine hiperdüzlemi en uzak biçimde yönlendirmektir. Destek vektör makinaları w ve b'nin seçimi olarak da ifade edilebilir (Kartal ve Balaban, 2019). Bu durumda eğitim veri seti Eşitlik (7) ile yazılabilir (Cristianini ve Taylor, 2000).

$$\forall_i \; için \; y_i(\, x_i * w + b) - 1 \geq 0 \tag{7}$$

Ayırıcı hiperdüzlemin konumu da destek vektörleri ile belirlenmekte ve iki destek düzlemi arasında herhangi bir veri bulunmamaktadır. Destek düzlemleri arasındaki uzaklık $\frac{1}{||w||}$'ye eşit olup marjin olarak ifade edilmektedir. Ayırıcı optimal hiperdüzlemin destek vektörden mümkün oldukça uzak olması için marjinin maksimize edilmesi gerekmektedir (Kartal ve Balaban, 2019). $\frac{1}{||w||}$ ifadesinin maksimizasyonu ||w|| normunun minimizasyonu ile mümkün olmaktadır (Uğuz, 2019). Bu minimizasyon yönteminin çözümünde Lagrange çarpanları kullanılmakta ve $\forall_i$ için Lagrange çarpanları $\alpha_i \geq 0$ olur (Burges, 1998).

$\forall_i \; için \; y_i(\, x_i * w + b) - 1 \geq 0$ olmak üzere; $\quad min \frac{1}{2} \|w\|^2$ (8)

olur. Eşitlik (8)'i minimize eden w ve b değerleri ile aynı eşitliği maksimize eden α değerinin bulunması istenmektedir $(\alpha_i \geq 0, \forall_i)$. Bunun için; Karush-Kush-Tucker koşullarına bağlı kalarak sırasıyla w'ye ve b'ye göre Eşitlik (9)'un kısmı türevi alınıp sıfıra eşitlenmektedir.

$$L_P(w, b, \alpha) = \frac{1}{2}\|w\|^2 - \sum_{i=1}^{L} \alpha_i \, y_i(\, x_i * w + b) + \sum_{i=1}^{L} \alpha_i \quad (9)$$

Soft-Marjin Destek Vektör Makineleri: Hard-Marjin destek vektör makineleri gibi sıfır hatalı sınıflandırma yapmak yerine bazı verilerin sınırın karşı sınıfında yer almasına izin vermektedir. Bu sayede daha esnek, aşırıyı öğrenmeye (overfitting) duyarlı ve iyi bir genelleme yeteneğine sahip bir model elde edilmektedir (Uğuz, 2019).

Soft-Marjin destek vektör makinelerinde ξ_i, negatif olmayan aylak (slack) değişkenini tanımlamaktadır (Cortes ve Vapnik, 1995);

$\forall_i \; için \; y_i(\, x_i * w + b) - 1 + \; \xi_i \geq 0$ (10)

Bu yöntemde, ξ_i değerlerinin toplamının az olması istenmektedir (Uğuz, 2019). $\xi_i = 0$ ise örneklem doğru sınıflandırılmış, $0 < \xi_i < 1$ ise örneklem sınırın içerisinde ve doğru sınıflandırılmış, $\xi_i \geq 1$ ise örneklem yanlış sınıflandırılmıştır. Böylece ξ_i'nin 1'den büyük olduğu örneklerin sayısı yanlış sınıflandırılan örnekleri vermektedir (Öz, 2019). Aylak değişken (ξ_i) ile marjin arasında bir denge sağlamak için bir C parametresi kullanılmaktadır. C'nin büyük olması; daha az sayıda hatalı sınıflandırılan örnekler sağlamasına ve marjinin küçük olmasına yol açmaktadır. C'nin küçük olması ise tersini vermektedir. En uygun C parametresi, çapraz doğrulama ve ızgara arama yöntemleriyle belirlenmektedir (Uğuz, 2019). Hard-Marjinde tanımlanan ilk problem, Soft-Marjinde Eşitlik (11)'deki gibi ifade edilebilir (Kartal ve Balaban, 2019). $\forall_i \; için \; y_i(\, x_i * w + b) - 1 + \; \xi_i \geq 0$ ve $\xi_i \geq 0$ olmak üzere;

$$min \frac{1}{2}\|w\|^2 + C \sum_{i=1}^{L} \xi_i \quad (11)$$

olur.

Bazı durumlarda veri setleri bir düzlem ile birbirlerinden doğrusal olarak ayrılamamaktadır. Bu tarz veriler doğrusal olarak birbirinden ayrılamadığından, iki boyutlu bir uzaydan üç boyutlu bir uzaya taşınarak bir düzlem yardımıyla birbirlerinden ayrılmaktadır

(Uğuz, 2019). Verilerin daha yüksek boyutlu bir uzaya dönüştürülmesi çekirdek fonksiyonlarıyla gerçekleşmekte ve $\Phi: R^n \rightarrow H$ biçiminde haritalanmaktadır. Böylece veriler Hilbert uzayı denilen daha yüksek boyutlu uzaya yerleştirilir (Öz, 2019). $\Phi(x)$'in uygulanmasından sonra veriler, daha yüksek boyutlu uzayda doğrusal olarak ayrılabilmekte ve doğrusal bir karar sınırı bulunabilmektedir (Kartal ve Balaban, 2019). Farklı bir uzaydaki vektörlerin iç çarpımı sonuçlarını döndüren ve öznitelik uzayına taşınan vektörlerin iç çarpımını ifade eden $K(x, y)$ fonksiyonuna çekirdek fonksiyon (kernel) denilmektedir (Stoean ve Stoean, 2014).

$$K\left(x_i, x_j\right) = \Phi(x_i).\Phi(x_j) \tag{12}$$

Bir doğrusal sınıflandırma probleminin amaç fonksiyonunda $\langle x_i, x_j \rangle$ şeklinde vektörlerin bir iç çarpımı yer alıyorsa, bu iç çarpım yerine uygun bir $K(x_i, x_j)$ çekirdek fonksiyonu yazılabilir. Bu durumda dual optimizasyon problemi Eşitlik (13)'deki gibi güncellenebilir (Uğuz, 2019).

$$\max_a \quad L(a) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j \, y_i y_j K(x_i, x_j)$$
$$Kısıtlar; \sum_{i=1}^{n} \alpha_i \, y_i = 0, \qquad 0 \leq \alpha_i \leq c, \qquad \mathrm{i} = 1, \ldots, \mathrm{n} \tag{13}$$

Sınıflandırma performansını doğrudan etkilediği için doğrusal olmayan SVM'de çekirdek fonksiyonun seçimi çok önemlidir (Öz, 2019). Bu çalışmada gerçek veri ve simülasyon uygulamalarında, SVM algoritması için Radyal çekirdek kullanılmış ve cost hiperparametresi (0.1, 0.5, 1, 2) aralığında, gamma hiperparametresi ise (0.1, 0.5, 1, 1.5, 2) aralığında doğruluk kriterini maksimize edecek şekilde çapraz geçerlilikle optimize edilmiştir.

2.1.5. XGBoost Algoritması

Chen ve Guestrin (2016) tarafından literatüre kazandırılan Aşırı Gradyan Arttırma algoritması (Extreme Gradient Boosting) XGBoost olarak adlandırılmaktadır. Temelde Karar Ağaçları ve Gradyan Arttırma yöntemine dayanan algoritma, büyük ölçekli veri setlerinde daha yüksek sınıflandırma başarısı elde etmek için yaygın olarak kullanılmaktadır. Boosting yöntemi ile paralel işlem yaparak büyük veri setlerini hızlı bir şekilde işleyebildiği için

XGBoost, sınıflandırma problemlerinde yüksek sınıflandırma performansı sağlayabilir (Singh vd., 2022).

XGBoost, birden çok ağacın tahmin sonuçlarını toplayarak nihai bir model oluşturur. Tahmin edilen değer ile mevcut değer arasındaki hataları (artık) sürekli olarak iyileştirilmesi ve güçlü bir model elde edilmesi algoritmanın çalışma mantığını oluşturur (Dong vd., 2023). Algoritma, model performansını iyileştirmek için kayıp fonksiyonun çözümünde Newton yöntemini kullanır, kayıp fonksiyonu Taylor serisini ikinci mertebeden genişletir ve genişletilmiş seriyi minimuma götüren bir dizi iterasyon izler. Ek olarak, kayıp fonksiyonuna düzenlileştirme terimleri eklenir. Eğitim sırasında amaç fonksiyonu, gradyan kaldırma algoritmasının kaybı ve düzenlileştirme terimi olmak üzere iki kısımdan oluşmaktadır (Yan vd., 2022).

Veri seti: $D = (x_i, y_i),\ i = 1, \dots, n,\ x_i \in \mathbb{R}^m, y_i \in \mathbb{R}$ varsayımına sahip ve n gözlem ile m değişken olursa, tahmin edilen $\hat{y}_i$ değeri Eşitlik (14) ile tanımlanır (Asselman vd., 2021).

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i), f_k \in K \qquad (14)$$

Chen ve Guestrin (2016), XGBoost algoritmasını Eşitlik (15) ile tanımlamaktadır. Gradyan arttırma algoritmasının her bir iterasyonunda artık değerler önceki tahmin edicinin düzeltilmesi için manipüle edilerek kayıp fonksiyonu optimize edilmektedir. $f_k(x_i)$, k ağacın i. örneği için tahmin değerini belirtmektedir. f_k fonksiyon seti amaç fonksiyonu minimize edilerek öğrenilebilir (Chen ve Guestrin, 2016; Asselman vd., 2021).

$$Obj = \sum_{i=1}^{n} l(\hat{y}_i, y_i) + \sum_{k=1}^{K} \Omega(f_k) \qquad (15)$$

Eşitlik (15)'de l ile belirtilen ilk terim kayıp fonksiyonu temsil etmekte ve tahmin edilen $\hat{y}_i$ değeri ile gerçek y_i değeri arasındaki farkı ölçmektedir. Ω ise regülasyon terimini temsil etmekte ve f_k ağacının karmaşıklığını ölçmektedir.

$$\Omega(f_k) = \gamma T + \frac{1}{2}\lambda\omega^2 \qquad (16)$$

Burada γ ve λ parametreleri regülasyon derecelerini, T yaprak düğümleri sayısını ve ω ise her yapraktaki puanı belirtir.

XGBoost algoritmasına özgü birçok hiperparametre bulunmakta (Asselman vd., 2021) ve bunların değerleri veri setinin yapısına göre değişkenlik göstermektedir. Bu çalışmanın gerçek veri ve simülasyon uygulamalarında, XGBoost algoritması için nround hiperparametresi 100 ila 1000 aralığında 100 artacak şekilde, eta hiperparametresi (0.01, 0.05, 0.1, 0.3) aralığında çapraz geçerlilik yöntemiyle doğruluk kriterini maksimize edecek şekilde optimize edilmiştir. Diğer hiperparametreler için ise varsayılan değerler kullanılmıştır.

2.2. Çoklu Doğrusal Bağlantı

Çoklu doğrusal bağlantı (multicollinearty) çoklu doğrusal regresyon problemlerinde karşılaşılan önemli sorunlardan birisidir. Çoklu doğrusal bağlantı, bağımsız değişkenlerin kendi aralarında tam ya da güçlü ilişki yapısını ifade etmektedir (Demir, 2020). Ancak, regresyon uygulamaları bağımsız değişkenler arasında ilişki olmaması varsayımına dayalıdır. Uygulamalarda genellikle bu varsayım ihmal edilmekte ve bağımsız değişkenler kendi aralarında ilişkili olabilmektedir (Alpar, 2013). Veri setinde çoklu doğrusal bağlantı hatalı katsayılar ve bağımsız değişkenler arasında yüksek korelasyon gibi çeşitli problemlere yol açabilir. Hatalı regresyon katsayıları büyük standart hatalara neden olmakta, dolayısıyla model güvenirliliğini ve tahminleri olumsuz etkilemektedir. Bağımsız değişkenler arasındaki yüksek korelasyon ise modelde R^2 değerinin büyümesine yol açmaktadır (Chan vd., 2022).

Yapay zeka, veri madenciliği ya da makine öğrenmesi alanlarındaki çalışmalarda da çoklu doğrusal bağlantı sorunuyla karşılaşılabilmekte ve bu bağlantının belirlenmesi için çeşitli yöntemler bulunmaktadır. İlk yöntem, korelasyon matrisinde değişkenler arasındaki ikili korelasyon değerlerine bakmaktır. İkili değişkenler arasındaki korelasyonun 0,8'den büyük olması çoklu doğrusal bağlantı göstergesi olabilir (Mason ve Perreault, 1991).

Korelasyon matrisinin özdeğer ($\lambda_1, \lambda_2, \ldots, \lambda_p$) ve özvektörleri ($v_1, v_2, \ldots, v_p$) çoklu doğrusal bağlantıyı belirlemede kullanılan bir diğer yöntemdir. Özdeğerlerin sıfıra yakın olması çoklu doğrusal bağlantı hakkında fikir vermekte, ancak çoklu doğrusal bağlantının derecesi hakkında fikir edinmek için Eşitlik (17)'de verilen ve koşul indeksinde yer alan maksimum değeri ifade eden koşul numarasına bakılmaktadır (Alin, 2010).

$$K = \sqrt{\frac{\lambda_{\max}}{\lambda_{\min}}} \qquad (17)$$

Koşul endeksinin 10 ila 30 arasında olması orta düzeyde çoklu doğrusal bağlantıya, 30'dan büyük olması ise ciddi derecede çoklu doğrusal bağlantıya işaret etmektedir (Chan vd., 2022).

2.3. Sınıflandırma Performanslarının Değerlendirilmesi

Tablo 1'de verilen gerçek ve tahmin değerlerinden elde edilen Eşitlik (18) ile (22) arasındaki 5 metrik kullanılarak model başarıları değerlendirilebilir (Mulla vd., 2021).

Tablo 1. Karmaşıklık matrisi

		Tahmin Edilen Sınıf	
		Pozitif	**Negatif**
Gerçek Sınıf	**Pozitif**	DP (Doğru Pozitif)	YN (Yanlış Negatif)
	Negatif	YP (Yanlış Pozitif)	DN (Doğru Negatif)

$$Doğruluk = \frac{DP + YN}{DP + DN + YP + YN} \quad (18)$$

$$Duyarlılık = \frac{DP}{DP + YN} \quad (19)$$

$$Seçicilik = \frac{DN}{DN + YP} \quad (20)$$

$$Kesinlik = \frac{DP}{DP + YP} \quad (21)$$

$$F\ Ölçütü = 2\ x\ \frac{Duyarlılık\ x\ Kesinlik}{Duyarlılık + \ Kesinlik} \quad (22)$$

3. UYGULAMA

İki gerçek veri seti ve bir simülasyon uygulaması ile çalışma gerçekleştirilmiştir. Lojistik regresyon algoritması parametrik bir model olup sıkı varsayımları bulunmakta ve bunlar arasında çoklu doğrusal bağlantı olmaması da yer almaktadır. Gerçek veri ve simülasyon çalışmalarının sonuçları yorumlanırken bu varsayım üzerinde ayrıca durulmuştur.

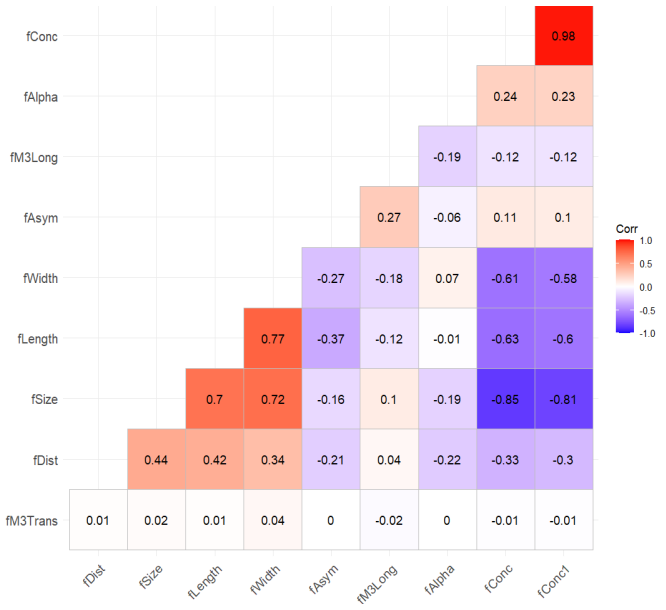
3.1. MAGIC Gamma Teleskop Veri Seti

19020 gözlem, 10 bağımsız ve 1 bağımlı değişkenden oluşmakta olan MAGIC gamma teleskop (Telescope) veri seti, https://archive.ics.uci.edu (UCI Machine Learning Repository) adresinden temin edilmiştir. Veri setine ait değişkenler ve özet istatistikler Tablo 2'de verilmiştir.

Tablo 2. MAGIC gamma teleskop veri setine ait değişkenler

Öznitelik Adı	**Veri Türü**	**Nümerik=**Min, Max **Kategorik=** Değerler	**Nümerik=**Ort.±s.s **Kategorik=** n(%)
fOdak uzunluğu	Nümerik	4.2-334.1	53.3±42.4
fOdak genişliği	Nümerik	0.0-256.3	22.2±18.3
fAlan boyutu	Nümerik	1.9-5.32	2.83±0.47
fÖzellik yoğunluğu1	Nümerik	0.01-0.89	0.38±0.18
fÖzellik yoğunluğu2	Nümerik	0.00-0.67	0.21±0.11
fAsimetri	Nümerik	-457.9-575.2	-4.33±59.2
fM3Uzunluk	Nümerik	-331.7-238.3	10.5±51.0
fM3Yatay	Nümerik	-205.8-179.8	0.25±20.8
fAlfa	Nümerik	0.-90	27.6±26.1
fMesafe	Nümerik	1.2-495.5	193.8±74.7
Bağımlı Değişken (Sınıf)	Kategorik	0, 1	**0:** n=12332 (%64.8) **1:** n= 6688 (%35.2)

Veri setinde yer alan bağımsız değişkenlere ait korelasyon grafiği Şekil 1'de verilmiştir.



Şekil 1. MAGIC gamma teleskop veri seti korelasyon ısı grafiği

Şekil 1 incelendiğinde bağımsız değişkenler arasında pozitif ve negatif yönlerde yüksek ilişkilerin varlığı gözlenmektedir. Ayrıca çoklu doğrusal bağlantının tespiti için koşul endeksine bakılmış, en büyük özdeğer 173.588 ve en küçük özdeğer 1.683 olmak üzere koşul endeksi yaklaşık 103 olarak belirlenmiştir. Bu değer 30'dan çok büyük olduğu için veri setinde yüksek düzeyde çoklu doğrusal bağlantının olduğu söylenebilir.

MAGIC Gamma Teleskop veri seti için sınıflandırma algoritmalarına ait performans sonuçları Tablo 3'de verilmiş ve değerlendirmede beş sınıflandırma performans kriteri kullanılmıştır.

Tablo 3. MAGIC gamma teleskop veri setine ait sınıflandırma performansları

Ölçüt / Algoritma	**kNN**	**Naive Bayes**	**Lojistik Reg.**	**SVM**	**XGBoost**
Doğruluk	0.837	0.725	0.790	0.871	**0.886**
Duyarlılık	0.956	0.919	0.799	**0.958**	0.942
Seçicilik	0.617	0.367	0.765	0.710	**0.772**
Kesinlik	0.821	0.728	**0.903**	0.859	0.884
F-Ölçütü	0.883	0.812	0.848	0.906	**0.909**

Çoklu doğrusal bağlantıya sahip MAGIC gamma teleskop veri seti ile yapılan analizde en yüksek sınıflandırma performansı %88.6 doğrulukla XGBoost algoritmasına, en düşük performans ise %72.5 doğrulukla Naive Bayes algoritmasına ait olduğu görülmektedir. Ayrıca seçicilik ve F ölçütüne göre de en yüksek performansın XGBoost ve en düşük performansın ise Naive Bayes tarafından sergilendiği gözlenmiştir. Duyarlılıkta en yüksek değer SVM ve

yakın bir değer kNN ile en düşük değer ise Lojistik Regresyon ile elde edilmiştir. Kesinlik ölçütüne göre en yüksek performans Lojistik Regresyon ve en düşük performans ise Naive Bayes ile elde edilmiştir.

MAGIC gamma teleskop veri seti için bir değerlendirme yapıldığında XGBoost algoritması ile SVM algoritmasının başarılı performans sergilediği ve Naive Bayes algoritmasının düşük performansa sahip olduğu ve iyi performans sergilemediği söylenebilir.

3.2. Kredi Kartı Müşterilerinin Temerrüdü Veri Seti

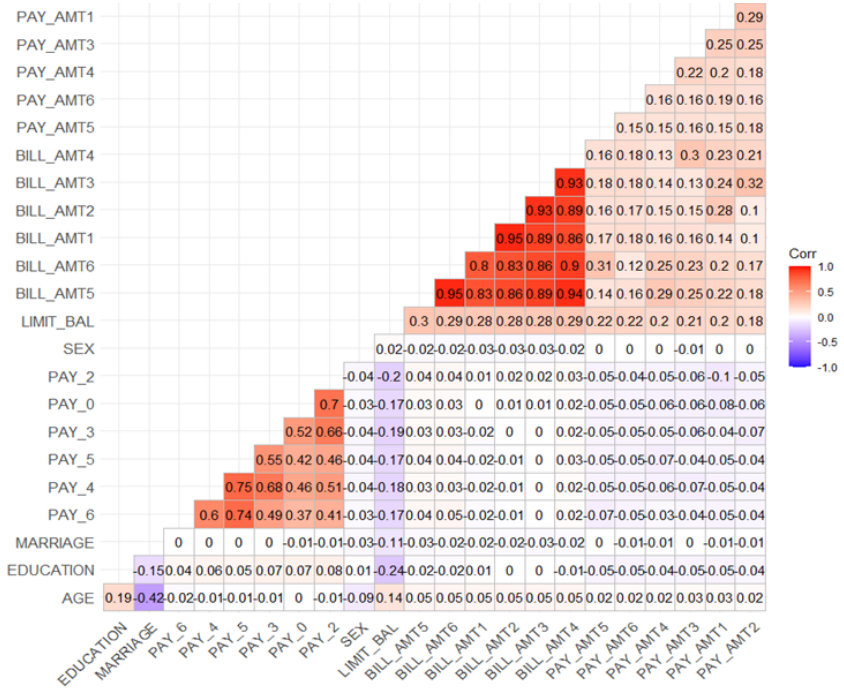
29602 gözlem, 23 bağımsız ve 1 bağımlı değişkenden oluşan Kredi Kartı Müşterilerinin Temerrüdü (Default of Credit Card Clients) veri seti, https://archive.ics.uci.edu (UCI Machine Learning Repository) adresinden elde edilmiştir. Veri setinde yer alan değişkenler ve bu değişkenlere ait özellikler Tablo 4'de verilmiştir.

Tablo 4. Kredi kartı veri setinde yer alan değişkenler

Öznitelik Adı	**Veri Türü**	**Nümerik**=Min, Max **Kategorik=** Değerler	**Nümerik**=Ort.±s.s **Kategorik=** n(%)
Limit Bakiyesi*	Nümerik	10.000-100.000	167.551±129.944
Fatura Tutarı (1, 2, 3, 4, 5, 6)	Nümerik	-69.777-1.664.089	44.820±66.655
Ödeme Tutarı (1, 2, 3, 4,5, 6)	Nümerik	0-896.040	5.258±17.835
Ödeme (0,1,2,3,4,5,6)[1]	Kategorik	0, 1, 2, 3, 4, 5, 6, 7, 8	**0:** n=152.720 (%86.0) **1:** n=3.696 (%2.1) **2:** n=18.861 (%10.6) **3:** n=1.423 (%0.8) **4:** n=451 (%0.3) **5:** n=135 (%0.1) **6:** n=74 (%0.04) **7:** n=218 (%0.1) **8:** n=28 (%0.02)
Yaş	Nümerik	21-79	35.5± 9.21
Cinsiyet	Kategorik	1, 2	**1:** n=11.746 (%39.7) **2:** n=17.855 (%60.3)
Eğitim[1]	Kategorik	1, 2, 3, 4	**1:** n=10.581 (%35.7) **2:** n=14.024 (%4.4) **3:** n=4.873 (%16.5) **4:** n=123 (%0.4)
Evlilik[1]	Kategorik	1, 2, 3	**1:** n=13.477 (%45.5) **2:** n=15.806 (%53.4) **3:** n=318 (%1.1)
Bağımlı Değişken (Ödeme durumu)	Kategorik	0, 1	**0:** n=22.996 (%77.7) **1:** n=6.605 (%22.3)

Kredi kartı veri setinde yer alan bağımsız değişkenlere ait korelasyon grafiği Şekil 2'de verilmiştir.

Şekil 2 incelendiğinde bazı bağımsız değişkenler arasında pozitif yönde yüksek derecede ilişkilerin olduğu gözlenmektedir. Ayrıca bu veri seti için koşul endeksi yaklaşık 893 olarak hesaplanmış ve böylece veri setinde çok yüksek düzeyde çoklu doğrusal bağlantı olduğu söylenebilmektedir.



Şekil 2. Kredi kartı veri setinde bağımsız değişkenlere ait korelasyon ısı grafiği

Kredi kartı veri seti için sınıflandırma algoritmalarına ait performans sonuçları Tablo 5'de verilmiş ve değerlendirmede beş sınıflandırma performans değerlendirme kriteri kullanılmıştır.

Tablo 5. Kredi kartı veri setine ait sınıflandırma performansları

Ölçüt / Algoritma	**kNN**	**Naive Bayes**	**Lojistik Reg.**	**SVM**	**XGBoost**
Doğruluk	0.816	0.771	0.823	0.822	**0.826**
Duyarlılık	0.945	0.828	0.837	**0.955**	0.952
Seçicilik	0.368	0.572	**0.707**	0.360	0.389
Kesinlik	0.839	0.871	**0.958**	0.839	0.844
F-Ölçütü	0.889	0.849	0.894	0.893	**0.895**

Tablo 5 incelendiğinde, en yüksek sınıflandırma performansının %82.6 doğrulukla XGBoost algoritmasına ve en düşük performansın ise %77.1 doğrulukla Naive Bayes algoritmasına ait olduğu görülmektedir. Duyarlılığa göre en yüksek değer SVM ve en düşük değer ise Naive Bayes ile sağlanmıştır. Seçicilik ve kesinlik ölçütlerine göre en yüksek değer Lojistik Regresyon ile elde edilirken, seçicilik için en düşük değer SVM ile kesinlik için ise SVM ve kNN ile elde edilmiştir. F Ölçütüne göre algoritmaların birbirine çok yakın değerler verdiği ve en yüksek performans %89.5 değeriyle XGBoost algoritmasından, en düşük performans ise %84.9 değeriyle Naive Bayes algoritmasından elde edilmiştir.

Kredi kartı veri seti için bir değerlendirme yapıldığında XGBoost, SVM ve Lojistik Regresyon algoritmalarının başarılı performans sergilediği söylenebilir. Naive Bayes algoritması bu veri seti için de düşük performans gösterdiği söylenebilir.

3.3. Simülasyon Çalışması

Simülasyon çalışması için veri üretim mekanizması Tablo 6'da sunulmuştur. Her bir simülasyon senaryosu için tekrar sayısı 500 olarak kararlaştırılmıştır. Simülasyon çalışmalarında bağımlı değişkenin iki sınıf içerdiği varsayılmıştır. Simülasyon kurgusu göz önüne alındığında, sınıflandırma algoritmalarının performansları için üç farklı çoklu doğrusal bağlantı oranı (ÇDB), iki farklı örneklem büyüklüğü ve iki farklı bağımsız değişken sayısı dikkate alınmıştır. Eğitim ve test veri setleri gerçek veri setlerinde olduğu gibi sırasıyla %75 ve %25 olarak alınmıştır. Sonuçlar, gerçek veri sonuçlarıyla karşılaştırılarak uyum ve uyumsuzluklar tartışılmıştır.

Tablo 6. Simülasyon kurgusu ve veri üretimi

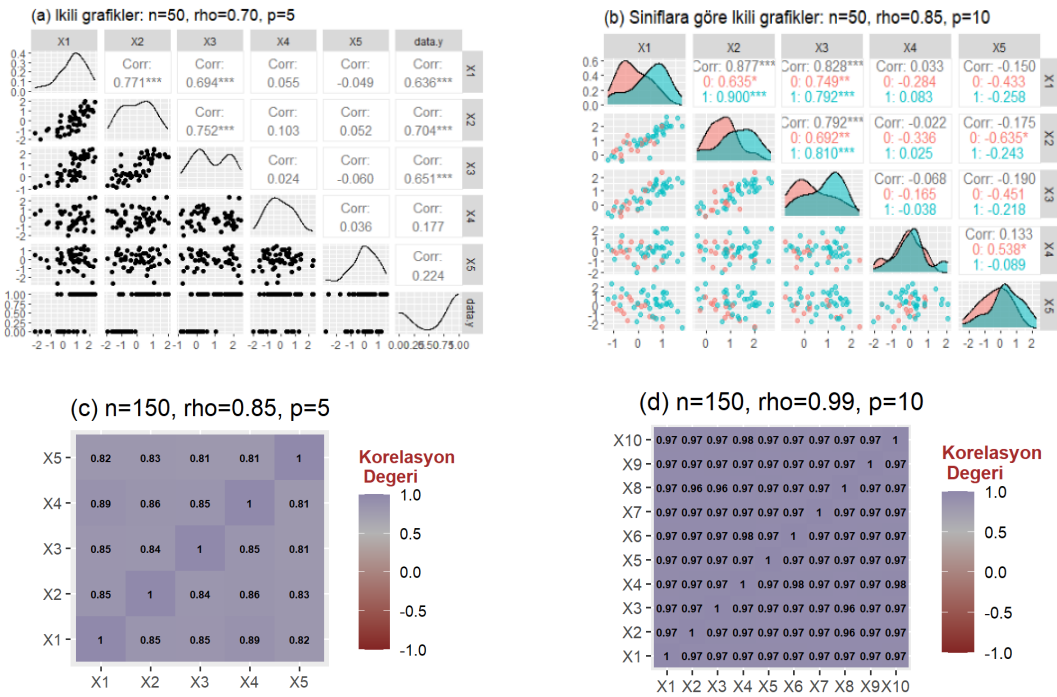
Simülasyon kurgusu		
Örneklem büyüklüğü	Çoklu doğrusal bağlantı oranı	Açıklayıcı değişken sayısı
$n = 50, 150$	$\rho = (0.70, 0.85, 0.99)$	$p = (5, 10)$
Verilerin üretilmesi		
Bağımsız değişkenler	Olasılıklar üretilmesi (sınıflar için)	Kategorik bağımlı değişken
$X \sim MV[\mu_x, \Sigma_x] \in \mathbb{R}^{n \times p}$	$z = 1 + \theta_p X,$ $P_y = \frac{1}{(1 + e^{-z})}$	$y = binom(n, P_y)$

Üretilen verilerin, çoklu doğrusal bağlantı durumlarını gözlemleyebilmek için Şekil 3 verilmiştir. Şekil 3, her biri farklı simülasyon konfigürasyonunu temsil eden dört panelden

oluşmaktadır. Ayrıca çoklu doğrusal bağlantı içeren açıklayıcı değişkenlerin üretimi aşağıdaki gibi ifade edilebilir.

$$X \sim MV(\mu_x, \Sigma_x)$$

Burada, $\{\mu_i \sim U[0,1]\}_{i=1}^{p}$ ve $\Sigma_x = \begin{bmatrix} 1 & \rho\Sigma_{ij} \\ \rho\Sigma_{ij} & 1 \end{bmatrix}$ şeklinde olup $MV(\mu_x, \Sigma_x)$, μ_x ortalama vektörüne ve Σ_x varyans kovaryans matrisine sahip çok değişkenli normal dağılıma sahip olduğunu gösteriyor. Çoklu doğrusal bağlantı üretilirken Σ_x köşegenleri bir ve kovaryansları korelasyon değerleri ile orantılı olacak şekilde dizayn edilmektedir. Çoklu doğrusal bağlantı içermeyen veri üretilirken ρ değeri sıfır veya çok küçük bir değer verilerek elde edilir ve veri üretimi aynı kalır. Şekil 3 incelendiğinde, veri setinin tamamı için ikili korelasyon değerleri ve ilişkilerin yoğunlukları Panel (a) ile ve bağımlı değişkene göre sınıfları içeren ikili grafikler ise Panel (b) ile gösterilmiştir. Panel (c), 0.85 korelasyon ile çoklu doğrusal bağlantı seviyeleri ve panel (d) 0.99 korelasyon ile çoklu doğrusal bağlantı seviyeleri için sırasıyla küçük ve büyük örneklem hacimlerine göre veri dağılımlarını göstermektedir. Bağımsız değişkenler arasında çoklu doğrusal bağlantı bulunmakta ve her bir değişkenin aynı dağılımdan üretildiği düşüncesiyle, temsilen sadece biri için saçılım grafiği verilmiştir. Grafikler incelendiğinde bağımsız değişkenler arasında güçlü bir ilişkinin olduğu görülmektedir.



Şekil 3. Çoklu doğrusal bağlantı için üretilen verilere ait tanımlayıcı grafikler

Tablo 7'de simülasyon çalışması için üretilen çoklu doğrusal bağlantılı veri setlerine ait koşul endeks değerleri yer almaktadır.

Tablo 7. Çoklu doğrusal bağlantılı üretilen veriler için koşul endeks değerleri

Simülasyon seti			**Koşul Endeksi**
n	**ρ**	**p**	
50	0.70	5	18.847
50	0.85	5	32.185
50	0.99	5	50.517
50	0.70	10	20.884
50	0.85	10	46.131
50	0.99	10	80.801
150	0.70	5	13.583
150	0.85	5	28.646
150	0.99	5	44.457
150	0.70	10	20.929
150	0.85	10	50.207
150	0.99	10	65.683

Tablo 7 incelendiğinde, veri setlerinde çoklu doğrusal bağlantı olduğu gözlenmektedir. Veri setindeki değişkenler arasındaki ilişki arttıkça çoklu doğrusal bağlantıda doğru orantılı olarak artmaktadır.

Sınıflandırma algoritmalarının genel performanslarına ait olası tüm simülasyon sonuçları Tablo 8'de verilmektedir.

Tablo 8. Çoklu doğrusal bağlantı simülasyon sonuçları

n	**p**	**Ölçüt / Alg.**	**kNN**	**Naive Bayes**	**Lojistik Reg.**	**SVM**	**XGB**	**kNN**	**Naive Bayes**	**Lojistik Reg.**	**SVM**	**XGB**
50	**5**		*ÇDB yok*					$\rho = 0.70$				
		Doğruluk	0.611	0.653	**0.858**	0.650	0.635	0.573	0.635	**0.777**	0.510	0.610
		Duyarlılık	0.936	0.913	0.876	0.938	**0.944**	0.913	0.878	0.892	**0.972**	0.930
		Seçicilik	0.735	0.848	0.681	**0.887**	0.833	0.714	0.608	0.693	**0.857**	0.750
		Kesinlik	0.812	0.689	0.722	**0.876**	0.804	0.800	0.666	0.697	**0.857**	0.782
		F-Ölçütü	0.774	0.783	**0.867**	0.624	0.789	0.743	0.756	**0.835**	0.741	0.720
			$\rho = 0.85$					$\rho = 0.99$				

		Doğruluk	0.760	0.855	**0.891**	0.673	0.851	0.891	0.900	**0.921**	0.735	0.771
		Duyarlılık	0.944	0.860	**0.977**	0.962	0.896	0.936	0.866	**0.977**	0.976	0.896
		Seçicilik	0.925	0.726	**0.946**	0.829	0.831	0.900	0.817	**0.950**	**0.950**	0.816
		Kesinlik	0.810	0.694	**0.905**	0.804	0.823	0.874	0.835	**0.926**	0.803	0.774
		F-Ölçütü	0.852	0.857	**0.934**	0.817	0.873	0.913	0.883	**0.949**	0.855	0.834
	10		*ÇDB yok*					$\rho = 0.70$				
		Doğruluk	0.646	0.596	**0.745**	0.583	0.540	0.496	0.641	**0.718**	0.606	0.676
		Duyarlılık	0.744	0.770	**0.938**	0.900	0.716	0.882	0.884	**0.926**	0.914	0.817
		Seçicilik	0.721	0.628	**0.888**	0.854	0.503	0.819	0.824	**0.845**	0.827	0.654
		Kesinlik	0.687	0.606	0.787	**0.790**	0.508	0.786	**0.803**	0.758	0.781	0.650
		F-Ölçütü	0.695	0.683	**0.842**	0.691	0.628	0.689	0.762	**0.822**	0.781	0.746
			$\rho = 0.85$					$\rho = 0.99$				
		Doğruluk	0.618	0.751	**0.855**	0.526	0.678	0.415	0.618	**0.791**	0.520	0.470
		Duyarlılık	**0.966**	0.888	0.940	0.903	0.862	0.880	0.892	0.868	0.805	**0.894**
		Seçicilik	**0.902**	0.815	0.866	0.869	0.725	**0.871**	0.833	0.701	0.806	0.723
		Kesinlik	0.820	0.675	**0.848**	0.815	0.675	**0.793**	0.654	0.720	0.789	0.719
		F-Ölçütü	0.792	0.820	**0.897**	0.642	0.770	0.647	0.755	**0.830**	0.604	0.682
150	**5**		*ÇDB yok*					$\rho = 0.70$				
		Doğruluk	0.827	0.754	**0.893**	0.712	0.785	0.749	**0.851**	0.838	0.726	0.721
		Duyarlılık	0.926	**0.956**	0.951	0.928	0.922	0.938	0.893	**0.962**	0.943	0.932
		Seçicilik	0.781	**0.869**	0.830	0.736	0.773	0.846	0.781	**0.918**	0.850	0.838
		Kesinlik	0.734	0.706	**0.849**	0.640	0.709	0.778	0.808	**0.874**	0.772	0.820
		F-Ölçütü	0.826	0.805	**0.922**	0.770	0.804	0.844	0.872	**0.900**	0.834	0.876
			$\rho = 0.85$					$\rho = 0.99$				
		Doğruluk	0.900	**0.934**	0.866	0.820	0.798	0.791	0.888	**0.898**	0.851	0.820
		Duyarlılık	**0.952**	0.906	0.962	0.963	0.933	0.956	0.859	0.953	**0.960**	0.940
		Seçicilik	0.894	0.812	**0.923**	0.919	0.843	0.898	0.746	0.888	**0.916**	0.881
		Kesinlik	**0.892**	0.867	0.890	0.854	0.814	0.837	0.804	**0.889**	0.878	0.843
		F-Ölçütü	**0.926**	0.920	0.914	0.892	0.865	0.873	0.873	**0.925**	0.906	0.880
	10		*ÇDB yok*					$\rho = 0.70$				
		Doğruluk	0.523	0.623	**0.893**	0.587	0.615	0.690	0.741	**0.898**	0.517	0.701
		Duyarlılık	**0.960**	0.948	0.949	0.851	0.874	0.935	0.903	**0.953**	0.868	0.927
		Seçicilik	**0.916**	0.892	0.874	0.795	0.701	0.848	0.815	**0.909**	0.776	0.851

		Kesinlik	0.627	0.700	**0.881**	0.760	0.634	0.758	0.769	**0.902**	0.753	0.749
		F-Ölçütü	0.742	0.785	**0.921**	0.674	0.744	0.812	0.822	**0.925**	0.705	0.814
			$\rho = 0.85$					$\rho = 0.99$				
		Doğruluk	0.647	0.844	**0.892**	0.671	0.713	0.727	0.834	**0.884**	0.642	0.837
		Duyarlılık	0.941	0.867	0.944	**0.991**	0.901	0.898	0.834	**0.943**	0.939	0.866
		Seçicilik	0.868	0.743	**0.879**	0.824	0.763	0.842	0.761	**0.910**	0.907	0.811
		Kesinlik	0.729	0.783	**0.881**	0.786	0.734	0.772	0.792	**0.894**	0.558	0.820
		F-Ölçütü	0.794	0.856	**0.918**	0.731	0.807	0.813	0.834	**0.913**	0.696	0.852

Tablo 8'deki simülasyon sonuçları n=50 ve p=5 çoklu doğrusal bağlantı olmayan durum için incelendiğinde; en yüksek doğruluk ve F-Ölçütü değerlerine sahip algoritma Lojistik Regresyon algoritmasıdır. Seçicilik ve kesinlik açısından SVM, duyarlılık açısından ise XGB en yüksek değere sahiptir. Çoklu doğrusal bağlantı varlığında ise; Lojistik Regresyon, Naive Bayes ve SVM algoritmalarının performansı çoklu doğrusal bağlantının artmasıyla birlikte genellikle arttığı gözlenmiştir. kNN ise çoklu doğrusal bağlantının artmasıyla birlikte bazı metriklerde performansını artırabilirken, bazılarında ise performanslarını düşürebildiği gözlenmiştir. Ayrıca ρ değeri 0.7'den 0.85'e çıkarılırken XGB performansı duyarlılık hariç diğer tüm metriklerde artarken 0.99'a çıkarıldığında ise duyarlılık için değişmezken diğer tüm metrikler için performanslarda düşüş gözlemlenmiştir. Böylece, çoklu doğrusal bağlantının bazı algoritmaların performansını olumlu ve bazılarını ise olumsuz etkilediği gözlenmiştir.

n=50 ve p=10 çoklu doğrusal bağlantı olmayan durum için incelendiğinde; tüm metriklerde en yüksek değere sahip algoritmanın Lojistik Regresyon olduğu ve en düşük değere sahip algoritmanın XGBoost algoritması olduğu söylenebilir. Çoklu Doğrusal Bağlantı varlığında ise: Lojistik Regresyon algoritması, çoğu metrikte en yüksek değeri vermektedir. Çoklu doğrusal bağlantı oranını gösteren ρ değeri 0.7'den 0.85'e çıkarıldığında kNN, Lojistik Regreyson ve XGBoost performanslarının tüm metriklere göre artığı, Naive Bayes performansının ise Seçicilik ve Kesinlik metriğine göre düştüğü geri kalan diğer üç metriğe göre de artığı görülmüştür. SVM algoritması ise Naive Bayes'in tam tersine bir davranış sergilemiştir. Ancak ρ değeri 0.85'ten 0.99'a çıkarıldığında duyarlılık ve seçicilik metriklerine göre Naive Bayes algoritmasının, duyarlılık ve kesinlik metriklerine göre ise XGBoost algoritmasının performansında hafif bir artış gözlenirken, geriye kalan metriklere göre ise bu iki algoritmanın performansında ciddi bir düşüş gözlenmiştir. Ayrıca kNN,

Lojistik Regresyon ve SVM algoritmaları tüm metriklerde farklı düzeylerde performans kaybına uğramıştır. Hem bağımsız değişken sayısındaki (p) artış hem de çoklu doğrusal bağlantı oranındaki (ρ) artış birlikte değerlendirildiğinde bu artışların algoritmaların performansı üzerinde genel olarak olumsuz etkiye sahip olduğu söylenebilir.

n=150 ve p=5 çoklu doğrusal bağlantı olmayan durum için incelendiğinde; Lojistik Regresyon Doğruluk, Kesinlik ve F-Ölçütü metriklerinde ve Naive Bayes algoritması ise Duyarlılık ve Seçicilik metriklerde en yüksek değerlere sahip algoritmalardır. Çoklu Doğrusal Bağlantı varlığında ise: Lojistik Regresyon birçok metrikte en yüksek değerlere sahip algoritmadır. Zira ρ, 0.70 olduğunda sadece Doğruluk metriğine göre Naive Bayes; 0.85 olduğunda Duyarlılık, Kesinlik ve F-Ölçütüne göre kNN, Doğruluk metriğine göre ise Naive Bayes; 0.99 olduğunda Duyarlılık ve Seçicilik metriklerine göre SVM en yüksek değerlere sahipken, geriye kalan tüm durumlarda Lojistik Regresyon en yüksek değerlere sahiptir. Çoklu doğrusal bağlantı arttıkça çoğu metriğe göre SVM algoritması performans artışı göstermektedir. ρ değeri 0.7'den 0.85'e çıkarıldığında Naive Bayes algoritması tüm metriklere göre performans artışı gösterirken 0.85'den 0.99'a çıkarıldığında ise tüm metriklere göre performans düşüşü göstermektedir. Naive Bayes algoritmasına çok yakın performans değişimi kNN algoritması içinde gözlenmiştir. Çoklu doğrusal bağlantı arttıkça XGBoost algoritması neredeyse tüm metriklerde daha kararlı bir duruş sergilemişi ve performansında önemli değişkenlik gözlemlenmemiştir.

n=150 ve p=10 çoklu doğrusal bağlantı olmayan durum için incelendiğinde; Lojistik Regresyon doğruluk, kesinlik ve F-Ölçütü metriklerinde yüksek performans göstermiştir. kNN ise seçicilik ve duyarlılık metriklerinde en yüksek performansı göstermiştir. Çoklu doğrusal bağlantı varlığında ise: ρ, 0.85 olduğu durumda sadece Duyarlılık metriğine göre SVM, hem geriye kalan diğer metriklere göre hem de ρ, 0.70 ve 0.99 olduğu durumlarda tüm metriklere göre Lojistik Regresyon en yüksek değerlere sahip algoritma olduğu gözlemlenmiştir. Çoklu doğrusal bağlantı arttıkça kNN ve Naive Bayes algoritmalarının performanslarında düzensiz dalgalanmalar yaşanmış ve bu dalgalanmalar metrikten metriğe farklılık göstermiştir. ρ değeri 0.7'den 0.85'e çıkarıldığında SVM algoritması tüm metriklere göre performans artışı gösterirken, 0.85'den 0.99'a çıkarıldığında ise seçicilik hariç geriye kalan diğer tüm metriklere göre performans düşüşü göstermiştir. Çoklu doğrusal bağlantı arttıkça XGBoost algoritmasının; doğruluk metriğine göre performans artışı, duyarlılık metriğine göre performans düşüşü gösterdiği ve diğer üç metriğe göre ise performans artış ve düşümlerinde bir kararsızlığın olduğu belirlenmiştir.

Simülasyon sonuçları için algoritmalar özelinde genel bir değerlendirme yapıldığında: Lojistik Regresyon, tüm n ve p değerleri için genel olarak en yüksek performans gösteren algoritma olmuştur. Özellikle doğruluk ve F-Ölçütü açısından oldukça başarılı sonuçlar vermiştir. XGBoost algoritmasının büyük örneklem hacminde genel olarak daha iyi sonuçlar verdiği gözlenmiştir. XGBoost algoritmasına ait bu sonuç, Tablo 3 ve Tablo 5'te verilen sonuçlarla genel olarak uyumsuzluk göstermektedir. kNN, genel olarak daha düşük performans göstermiş ve özellikle örneklem hacmi artırıldığında (p=10 olduğunda) performansı oldukça düşmüştür. Naive Bayes, çoklu doğrusal bağlantı varlığında özellikle doğruluk ve F-Ölçütü bakımından kNN'e göre genellikle daha iyi performans sergilemektedir. SVM algoritmasının performansı genel olarak Lojistik Regresyon ve XGBoost algoritmaları performansına yetişmezse de seçicilik açısından bu algoritmalardan aşağı olmadığı söylenebilir.

4. SONUÇ

Çoklu doğrusal bağlantıya sahip iki gerçek veri seti ve simülasyon çalışmaları ile gerçekleştirilen bu çalışmadan elde edilen sonuç ve öneriler aşağıda verilmiştir.

MAGIC gamma teleskop veri kümesi için, XGBoost algoritması ile SVM algoritmasının başarılı performans sergilediği, en düşük performansın ise Naive Bayes algoritması tarafından gösterildiği söylenebilir.

Kredi kartı veri setine ait sonuçlar incelendiğinde; XGBoost, SVM ve Lojistik Regresyon algoritmalarının başarılı performans gösterdiği, Naive Bayes algoritmasının düşük performans gösterdiği söylenebilir. Bu sonucun, MAGIC gamma teleskop veri seti ile uyumlu olduğu söylenebilir.

Simülasyon çalışmaları, kNN ve Lojistik Regresyon algoritmalarının çoklu doğrusal bağlantı karşısında farklı tepkiler verdiğini göstermiştir. Küçük örneklemli veri setlerinde ve yüksek bağımsız değişken sayısında kNN algoritmasının performansı olumsuz etkilenirken, Lojistik Regresyon algoritması bu duruma karşı daha dirençlidir ve artan çoklu doğrusal bağlantı oranına bağlı olarak birçok metrikte iyileşme gösterir. Bu durum, Lojistik Regresyonun varsayımları arasında yer alan çoklu doğrusal bağlantı olmaması varsayımının ne kadar önemli olduğunu ve ihmal edilmemesi gerektiğini açıkça göstermektedir. Aksi takdirde yanıltıcı sonuçlar elde edilebilir. Örneklem hacmi artmaya başladığında ise XGBoost algoritması artan

çoklu doğrusal bağlantı ile birlikte başarılı sonuçlar vermektedir. Bu durum, XGBoost algoritması için gerçek verilerden elde edilen bulgularla uyumlu olduğunu göstermektedir.

Sonuç olarak, örneklem büyüklüğü yeterli seviyelere ulaştığında çoklu doğrusal bağlantıya sahip veri setlerinde sınıflandırma için XGBoost algoritmasının iyi bir alternatif olduğu söylenebilir. Derraz vd. (2023), çoklu doğrusal bağlantı durumunda topluluk öğrenme algoritmalarının geleneksel makine öğrenmesi yöntemlerinden daha başarılı performans sergilediğini belirtmişler. Georganos vd. (2018) ile McNamara vd. (2022), örneklem hacminin artmasıyla XGBoost algoritması performansının da artığını ifade etmişlerdir. Çalışma ilgili çalışmalara ait sonuçlar ile benzerlik göstermektir. Naive Bayes ile kNN algoritmasını çoklu doğrusal bağlantıdan olumsuz etkilendiği söylenebilir. Bu çalışmadan elde edilen sonuçlara benzer şekilde Dumancas ve Bello (2015) tarafından, çoklu doğrusal bağlantıdan en çok etkilen algoritmanın Naive Bayes olduğu gösterilmiştir.

Gelecekteki çalışmalarda çoklu doğrusal bağlantı çözümü daha kapsamlı bir şekilde ele alınmalıdır. Bu çalışmanın sonuçları, kullanılan veri setleri ve algoritmalar ile sınırlandırılabilir. Daha genel sonuçlara varabilmek için farklı veri setleri ve algoritmalar üzerinde daha fazla araştırma yapılması gerekmektedir.

ETİK BEYAN

"Çoklu Doğrusal Bağlantı Olması Durumunda Veri Madenciliği Algoritmaları Performanslarının Karşılaştırılması" başlıklı çalışmanın yazım sürecinde bilimsel, etik ve alıntı kurallarına uyulmuş, toplanan veriler üzerinde herhangi bir tahrifat yapılmamış ve bu çalışma herhangi başka bir akademik yayın ortamına değerlendirme için gönderilmemiştir.

KAYNAKÇA

Alin, A. (2010), Multicollinearity, *Wiley Interdisciplinary Reviews Computational Statistics*, 2(3), 370–374.

Alpar, R. (2013), *Çok değişkenli istatistiksel yöntemler*, Detay Yayıncılık: Ankara, Türkiye.

Asselman, A., Khaldi, M. and Aammou, S. (2021), Enhancing the prediction of student performance based on the machine learning xgboost algorithm, *Interactive Learning Environments*, 1–20.

Batista, G. E. A. P. A. and Monard, M. C. (2002), A study of k-nearest neighbour as an imputation method. In Abraham, A., Solar, J.R., Köppen, M. (Ed.), *Frontiers in artificial intelligence and applications*, 87, 251–260, IOS Press.

Blommaert, A., Hens, N. and Beutels, P. (2014), Data mining for longitudinal data under multicollinearity and time dependence using penalized generalized estimating equations, *Computational Statistics & Data Analysis*, 71(0), 667–680.

Burges, C. J. (1998), A tutorial on support vector machines for pattern recognation, *Data Mining and Knowledge Discovery*, 2(2), 121–167.

Cervantes, J., Garcia-Lamont, F., Rodríguez-Mazahua, L. and Lopez, A. (2020). A comprehensive survey on support vector machine classification: Applications, challenges and trends. *Neurocomputing*, *408*, 189–215.

Chan, J.-L., Leow, S., Bea, K., Cheng, W., Phoong, S., Hong, Z.-W. and Chen, Y. L. (2022), Mitigating the multicollinearity problem and its machine learning approach: A review, *Mathematics*, 10(8), 1283.

Chen, T. and Guestrin, C. (2016), XGBoost: A scalable tree boosting system, KDD ’16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, USA.

Cortes, C. and Vapnik, V. N. (1995), Support vector networks, *Machine Learning*, 20, 273–297.

Cristianini, N. and Taylor, J. S. (2000), *An introduction to support vector machines and other kernel-based learning methods*, Cambridge University Press: Cambridge, UK.

Davidson, I. and Tayi, G. (2009), Data preparation using data quality matrices for classification mining, *European Journal of Operational Research,* 197(2), 764-772.

Demir, Y. (2020), Çoklu doğrusal regresyon ve bazı cezalı tahmin yöntemlerinin incelenmesi. In S. Öztürk (Ed.), *Sosyal ve beşeri bilimlerde teori ve araştırmalar II,* 2, 261-276, Gece Akademi: Ankara.

Derraz, R., Melissa Muharam, F., Nurulhuda, K., Ahmad Jaafar, N. and Keng Yap, N. (2023), Ensemble and single algorithm models to handle multicollinearity of UAV vegetation indices for predicting rice biomass, *Computers and Electronics in Agriculture*, *205*, 107621.

Dong, Z., Li, X., Luan, F., Ding, J. and Zhang, D. (2023), Point and interval prediction of the effective length of hot-rolled plates based on IBES-XGBoost, *Measurement*, *214*(0), 112857.

Dumancas, G. and Bello, G. (2015), Comparison of machine-learning techniques for handling multicollinearity in big data analytics and high performance data mining, The International Conference for High Performance Computing, Networking, Storage, and Analysis, Texas, USA.

Garg, A. and Tai, K. (2013), Comparison of statistical and machine learning methods in modelling of data with multicollinearity, *International Journal of Modelling, Identification and Control*, 18(4), 295–312.

Georganos, S., Grippa, T., Vanhuysse, S., Lennert. M., Shimoni, M. and Wolff, E. (2018), Very high resolution object-based land use–land cover urban classification using extreme gradient boosting, *IEEE Geoscience and Remote Sensing*, 15(4), 607-611.

Han, J., Kamber, M. and Pei, J. (2012), *Data mining concepts and techniques* (Third Edition). Morgan Kaufman Publishers: Massachusetts, USA.

Harrington, P. (2012), *Machine learning in action*, Manning Publications: New York, USA.

Hosmer, D. W., Lemeshov, S. and Sturdivant, R. X. (2013), *Applied logistic regression* (Third Edition). John Wiley & Sons, Inc: New Jersey, USA.

Kartal, E. and Balaban, M. E. (2019), Destek vektör makineleri: teori ve R dili ile bir uygulama. In M. E. Balaban, E. Kartal (Eds.), *Veri madenciliği ve makine öğrenmesi temel kavramlar, algoritmalar, uygulmalar* (207-241), Çağlayan Kitapevi: İstanbul.

Lewis, N. D. (2017), *Machine learning made easy with R: An intuitive step by step blueprint for beginners*, CreateSpace Independent Publishing Platform: Carolina, USA.

Mason, C. H. and Perreault, W. D. (1991), Collinearity, power, and interpretation of multiple regression analysis, *Journal of Marketing Research*, 28(3), 268–280.

McNamara, J. M., Green, R. F. and Olsson, O. (2006). Bayes' Theorem and ıts applications in animal behaviour, *Oikos*, 112(2), 243–251.

McNamara, M. E., Zisser, M., Beevers, C. G. and Shumake, J. (2022), Not just "big" data: Importance of sample size, measurement error, and uninformative predictors for developing prognostic models for digital interventions, *Behaviour Research and*

Therapy, 153(0), 1-12.

Mucherino, A., Papajorgji, P. J. and Paradalos, P. M. (2009), *Data mining in agriculture*, Springer: Dordrecht, Hollanda.

Mulla, G. A. A., Demir, Y. and Hassan, M. (2021), Combination of PCA with SMOTE Oversampling for Classification of High-Dimensional Imbalanced Data, *Bitlis Eren Üniversitesi Fen Bilimleri Dergisi*, *10*(3), 858–869.

Obite, C. P., Olewuezi, N. P., Ugwuanyim, G. U. and Bartholomew, D. C. (2020), Multicollinearity effect in regression analysis: A feed forward artificial neural network approach*, Asian Journal of Probability and Statistics*, 6(1), 22-33.

Öz, E. (2019), Destek vektör makineleri. In S. Alp, E. Öz (Ed.), *Makine öğreniminde sınıflandırma yöntemleri ve R uygulamaları* (67-189), Nobel Akademik Yayıncılık: Ankara.

Rahman, M. M., Ghasemi, Y., Suley, E., Zhou, Y., Wang, S. and Rogers, J. (2021), Machine learning based computer aided diagnosis of breast cancer utilizing anthropometric and clinical features, *IRBM*, 42(4), 215-226.

Roelofs, R., Shankar, V., Recht, B., Fridovich-Keil, S., Hardt, M., Miller, J. and Schmidt, L. (2019), A meta-analysis of overfitting in machine learning. Advances in Neural Information Processing Systems, 32.

Senawi, A., Wei, H.-L. and Billings, S. A. (2017), A new maximum relevance-minimum multicollinearity (MRmMC) method for feature selection and ranking, *Pattern Recognition*, 67, 47-61.

Silahtaroğlu, G. (2013), *Veri Madenciliği Kavram ve Algoritmaları*, Papatya Yayınevi: İstanbul.

Singh, R., Biswas, M. and Pal, M. (2022), Cloud detection using sentinel 2 imageries: A comparison of XGBoost, RF, SVM, and CNN algorithms. *Geocarto International*, 0(0), 1–32.

Stoean, C., Stoean, R. (2014), *Evolutionary support vector machines and their application for classification*, Springer International Publishing: New York, USA.

Uğuz, S. (2019), *Makine öğrenmesi teorik yönleri ve python uygulamaları* (1. Basım). Nobel Akademik Yayıncılık: Ankara.

Urooj, B., Shah, M. A., Maple, C., Abbasi, M. K., Riasat, S. (2022), Malware detection: a framework for reverse engineered android applications through machine learning algorithms*, IEEE Access*, 10(6), 89031-89050.

Yan, Z., Chen, H., Dong, X., Zhou, K. and Xu, Z. (2022), Research on prediction of multi-class theft crimes by an optimized decomposition and fusion method based on XGBoost, *Expert Systems with Applications*, 207, 117943.

Ying, X. (2019), An overview of overfitting and its solutions, In Journal of physics: Conference series, 1168, 022022, IOP Publishing.

Zhang, X., Liu, S. and Zheng, X. (2021), Stock Price Movement Prediction Based on a Deep Factorization Machine and the Attention Mechanism, *Mathematics*, 9(8), 800.

Zhu, J., Ge, Z., Song, Z. and Gao, F. (2018), Review and big data perspectives on robust data mining approaches for industrial process modeling with outliers and missing data, *Annual Reviews in Control,* 46(1), 107–133.