İstatistik Araştırma Dergisi
Journal of Statistical Research

ISSN: 2791-7614

(2023), 13 (1)
Başvuru/Received: 27/10/2022
Kabul/Accepted: 17 /01/2023

https://journal.tuik.gov.tr/

ARAŞTIRMA MAKALESİ **RESEARCH ARTICLE**

Sağdan Sansürlü Veriler için Veri Madenciliği Algoritmaları Performanslarının Karşılaştırılması

Saygın DİLER
Van Yüzüncü Yıl Üniversitesi / Doktora Öğrencisi
saygin.diler@tuik.gov.tr
Orcid No: 0000-0002-9056-412X

Yıldırım DEMİR
Van Yüzüncü Yıl Üniversitesi / Dr. Öğr. Üyesi
ydemir@yyu.edu.tr
Orcid No: 0000-0002-6350-8122

Özet

Veri madenciliği algoritmaları ile gerçekleştirilen modelleme çalışmaları bilgisayar teknolojisinin gelişmesiyle birlikte artış göstermiştir. Ancak bu algoritmalar ile yapılan çalışmalarda veri kalitesinin bozulması elde edilecek sınıflandırma performanslarında önemli rol oynamaktadır. Bu çalışmada veri madenciliği sınıflandırma algoritmalarının performanslarının veri kalitesini bozan etmenlerden biri olan sansürlü verinin veri setinde yer alması durumunda nasıl etkilendiği incelenmiştir. Sansürlü verilerinin etkisini veri setinde gösterilebilmesi amacı ile K en yakın komşu algoritması (kNN) imputasyon yöntemi kullanılmıştır. Daha sonra sınıflandırma algoritmalarından olan Naive Bayes (NB), Lojistik Regresyon (LR) ve K en yakın komşu algoritması (kNN) ile uygulamalar gerçekleştirilmiştir. Yöntemlerin performanslarının incelenmesi için simülasyon çalışması ve gerçek veri seti çalışmaları yapılmış, sonuçlar sunulmuştur. Analiz sonuçlarına göre, yüksek sansür seviyesinde ve düşük sansür seviyesinde Lojistik Regresyon algoritmasının sansür ile baş etmede dikkate değer performans gösterdiği belirlenmiştir. Ayrıca örneklem büyüklüğü arttıkça genel olarak algoritmaların doğru sınıflama performanslarının arttığı gözlenmiştir. Özetle büyük örneklemeli veri setlerinde Lojistik Regresyon algoritmasının doğru sınıflandırma oranı ile başarılı sınıflandırma performansı gösterdiği söylenebilir.

Anahtar Sözcükler: Sağdan Sansürlü Veri, Sınıflandırma, Veri Madenciliği

Sorumlu Yazar / Corresponding Author: 1-Saygın DİLER, Van Yüzüncü Yıl Üniversitesi, Fen Bilimleri Enstitüsü İstatistik

2-Yıldırım DEMİR, Van Yüzüncü Yıl Üniversitesi, İktisadi ve İdari Bilimler Fakültesi, Ekonometri Bölümü

Atıf / Citation: DİLER S., DEMİR Y. (2023). Sağdan Sansürlü Veriler için Veri Madenciliği Algoritmaları Performanslarının Karşılaştırılması. İstatistik Araştırma Dergisi, 13 (1), 34-47.

Bu makale Saygın DİLER tarafından yazılmakta olan "Veri Kalitesinin Bozulduğu Durumlarda Veri Madenciliği Sınıflandırma Algoritmalarının Performanslarının Karşılaştırılması" isimli doktora tezinden üretilmiştir.

Atıf / Citation: DİLER S., DEMİR Y. (2023). Sağdan Sansürlü Veriler için Veri Madenciliği Algoritmaları Performanslarının Karşılaştırılması. İstatistik Araştırma Dergisi, 13 (1), 34-47.

Comparison of Data Mining Algorithms Performances for Right-Censored Data

Abstract

Modeling studies performed with data mining algorithms have increased with the development of computer technology. However, the deterioration of data quality in studies with these algorithms plays an important role in the classification performances to be obtained. In this study, it has been examined how the performance of data mining classification algorithms is affected when censored data, which is one of the factors that deteriorates data quality, is included in the data set. In order to show the effect of the censored data in the data set, the K nearest neighbor algorithm (KNN) imputation method was used. Then, applications were carried out with Naive Bayes (NB), Logistic Regression (LR) and K nearest neighbor algorithm (KNN), which are among the classification algorithms. To inspect the performance of the mentioned methods, simulation study and real data example are carried out. According to the results of the analysis, it was determined that Logistic Regression algorithm at high and low censorship level showed remarkable performance in dealing with censorship. In addition, it was observed that the correct classification performance of the algorithms increased as the sample size increased. In summary, it can be said that the correct classification success of Logistic Regression algorithm in data sets with large samples show successful classification performance with values.

Keywords: Classification, Data Mining, Right-Censored Data

1. Giriş

Günümüzde teknolojik gelişmeler çok hızlı ilerleme göstermektedir. Bu hızlı gelişmelerle birlikte teknolojinin aktif olarak kullanılması sonucunda veriler üretilmektedir. Üretilen veri çeşitliliği, miktarı ve hacmi her geçen gün daha da artmaktadır. Bu artış ile veri tabanlarında analiz edilmeyi bekleyen sayısızca veri bulunmaktadır. Bu verilerin analiz edilmesi için çeşitli yöntemler geliştirilmiş ve geliştirilmeye de devam edilmektedir. Yaygın olarak kullanılan bu yöntemlerden birisi de veri madenciliği algoritmaları yaklaşımıdır. Veri madenciliği algoritmaları sağlık, güvenlik, finans, lojistik, ekonomi gibi birçok farklı bilim alanında sıklıkla kullanılmaktadır (Han ve ark., 2012). Ancak veri madenciliği için geliştirilen algoritmaların sınıflandırma performanslarını etkileyen önemli etkenlerden biri de veri setinin kalitesidir. Zira algoritmalar veri setini kullandığı ve çıkarımlar veri setinden yapıldığı için algoritma performansları veri kalitesinden etkilenmektedir (Batista & Monard, 2002). Veri kalitesini bozan durumlara; eksik veri, sansürlü veri, aykırı değer, çoklu doğrusal bağlantı sorunu örnek gösterilebilir.

Başarılı bir şekilde veri madenciliği uygulamasının yapılabilmesi, büyük ölçüde veri kalitesi ile doğru orantılıdır. Veri bütünlüğü ve verilerin analize uygun olması veri madenciliği çalışmalarının ön koşulları arasında yer almaktadır. Birçok bilim alanında toplanan verilerde, veri yapısı ve kalitesini etkileyen sorunlarla karşılaşmak mümkündür. Söz konusu sorunların varlığında veri madenciliği uygulanmalarından veya istatistik analizlerden güvenilir tahminler yapmak zorlaşabilir. Bu nedenle, veri kalitesinin bozuk olduğu durumlarda klasik veri madenciliği yöntemleri hantal hale gelmektedir. Özellikle yüksek boyutlu ve çeşitli süreç verilerinde varsayım ve hipotezlerin derinlemesine incelenmesi araştırmacıları zorlamaktadır (Zhu ve ark., 2018). Veri madenciliği çalışmalarında, veri yapısı ve kalitesini etkileyen önemli faktörlerden birisi de sansürlü verilerdir. Sansürlü veriler üzerine yapılan çalışmalarda, genellikle sağdan sansürlü veriler kullanılmaktadır (Eröz & Tutkun, 2020). Bu çalışmada da sağdan sansürlü verilere odaklanılmıştır.

Veri madenciliğinde sınıflandırma amacı ile kullanılan algoritmalar olay bilgilerinin bütün veriler için bilindiği varsayımı ile çalışmaktalar. Sansürlü veri içeren veri setlerinde bilgileri takip edilemeyen denekler için belirsiz bir durum bulunmaktadır (Vock ve ark., 2016). Sınıflandırma algoritmaları gerçek veri setleri üzerinde başarılı performanslar göstermesine rağmen, literatürde sansürlenmiş verilere bu algoritmaların uyguladığı çok az çalışma bulunmaktadır (Goldberg ve Kosorok, 2012).

Literatürde, sansürlü verilerde makine öğrenmesi yöntemlerini kullanan çalışmalardan bazıları şu şekildedir. Shivaswamy ve ark. (2007) ile Khan ve Zubek (2008), sansürlü veriyi hesaba katmak için kayıp fonksiyonunu değiştirerek destek vektör makinelerini uyarlamayı önermişlerdir. Her iki çalışmada da sansürlü veri setlerine uyarlanmış destek vektör makinelerinin geleneksel yöntemlere göre daha başarılı sonuçlar verdiği gözlenmiştir. Ishwaran ve ark. (2008), sağdan sansürlenmiş hayatta kalma verilerinin analizi için rasgele orman algoritmasının bir versiyonu olan rastgele hayatta kalma orman algoritmasını tanıtmışlar. Hayatta kalma dağılımını sınıflandırma ağaçları ile göstermişler. Štajduhar ve ark. (2009), sansürlü verilerde bayes ağları ile oluşturulan modellerin yaşam sürelerinin hesaplanmasındaki etkisini incelemişler. Hafif, orta ve ağır sansür altında sentetik veriler üzerinde bir simülasyon çalışması kullanarak sansürlemenin etkisini ve sansürün bayes ağlarının öğrenme olasılığını nasıl etkilediğini belirlemeye çalışmışlar. Bandyopadhyay ve ark. (2015), sansürlü elektronik sağlık verileriyle kardiyovasküler riskleri

tahmin etmek amacıyla bayes ağı modeli önermişlerdir. Sağdan sansürlü verilerde bayes ağı modelinin Cox orantılı risk analizine göre daha başarılı tahmin etme performansı gösterdiğini belirtmişlerdir. Vock ve ark. (2016), sağdan sansürlü elektronik sağlık verilerine; sansürleme ağırlığının ters olasılığını hesaba katarak makine öğrenmesi/veri madenciliği algoritmalarını uygulamışlar. İleri sürdükleri yaklaşımın daha başarılı sonuçlar verdiğini göstermişlerdir.

Sağdan sansürlü veriler olması durumunda, literatürde yer alan çalışmalarda izlenen yollardan birisi sağdan sansürlü veriler hariç tutularak analizler gerçekleştirilmektedir. Bir diğer yol ise veri ön işleme adımlarında sansürlü verilerin tahminleme yöntemi ile tamamlanmasıdır. Bu çalışmalar genel olarak değerlendirildiğinde şu sonuçlar elde edilmektedir; sansürlü veriler ile ilgili yapılan çalışmalar tek bir sınıflandırma algoritmasına özgüdür ve genel olarak uygulanabilirler ancak analizler sonucunda elde edilen tahminler hata payı içermektedir (Vock ve ark., 2016).

Bu çalışmada, sağdan-sansürlü verilerde veri madenciliği sınıflandırma algoritmalarının sınıflandırma performansları üzerinde durulmuştur. Sansürlenmiş gözlemleri tahmin etmeye yarayan yerine koyma yöntemlerinden biri olan KNN imputasyon yöntemi ile sansürlü veriler tahmin edilerek analize uygun hale getirilmiş ve daha sonra uygulamalar, sansürlü veri setleri üzerinde gerçekleştirilmiş ve çeşitli sansür seviyelerine göre algoritma başarıları karşılaştırılmıştır.

Çalışmanın temel amacı, sağdan sansürlü verilerde makine öğrenmesi algoritmalarının performanslarını ve bu tür verilerde mevcut kullanılan algoritmalardan hangisinin daha iyi olduğunu belirleyerek literatüre katkı sağlamaktır.

2. Materyal ve Yöntem

2.1 Sınıflandırma Algoritmaları

Sınıflandırma yöntemlerinde amaç bir bağımlı ve birden fazla bağımsız değişkenden oluşan veri setlerinde bağımsız değişkenlerden bağımlı değişkene eşleyen bir model oluşturmaktır. Bağımlı değişkenler kategorik verilerden oluşmaktadır. Veri madenciliği alanında sınıflandırma amacı için geliştirilen algoritmalar istatistik, yapay zeka, makine öğrenmesi gibi bilim alanlarından faydalanmaktadır. Her algoritma için geçerli varsayımlar ve sınıflandırma görevleri, kullanılan yönteme göre birbirinden farklılık göstermektedir (Davidson ve Tayi, 2009).

Sınıflandırma yöntemlerinde genellikle kullanılan başlıca teknikler arasında; K-En Yakın Komşu, Naïve Bayes, Lojistik Regresyon, Destek Vektör Makineleri, Karar Ağaçları, Yapay Sinir Ağları ve Genetik Algoritmalar yer almaktadır (Silahtaroğlu, 2013). Bu çalışmada literatürde fazlaca uygulama yapılan algoritmalar arasında yer alan KNN, Naïve Bayes ve Lojistik Regresyon algoritmaları ile uygulama gerçekleştirilmiştir ve performansları karşılaştırılmıştır.

2.1.1 K-En Yakın Komşu Algoritması

K-en yakın komşu (KNN) algoritması veri madenciliğinde en çok kullanılan algoritmalar arasında yer almaktadır. Parametrik olmayan yöntemler arasında yer alan algoritma (Bishop, 2006), sınıfları belli olan veri setinden sınıfı bilinmeyen yeni bir veriyi en yakın komşusuna atama mantığına dayanmaktadır (Mucherino ve ark., 2009).

KNN algoritmasının doğru sınıflandırma başarısı büyük ölçüde k değerinin (komşu sayısı) optimum seçimine bağlıdır. k değerini seçmenin bir çok yolu bulunmakta ve en basit seçim, farklı k değerleri kullanılarak en iyi performans sergileyen k değerini belirleme yöntemidir (Guo ve ark., 2003). Ayrıca bu değer, örnek sayısı ve öznitelikler göz önünde bulundurularak da seçilebilir. Örnek sayısı n ise k değeri; $k = \sqrt{n}$ şeklinde seçilebilir (Balaban ve Kartal, 2015). k en yakın komşu algoritmasının doğru sınıflama başarısında k değerinin seçimi kadar en yakın komşuya olan uzaklık (benzerlik) ölçüsü de büyük önem arz etmektedir. Öznitelikler arasındaki mesafeyi ölçmenin farklı yöntemleri bulunmakta ve bu çalışmada uzaklık ölçüsü olarak Öklid kullanılmıştır. Öklid ölçüsü;

$$d(x,y) = \sqrt{\sum_{j=1}^{N}\left(x_j - y_j\right)^2} \qquad (1)$$

olarak gösterilebilir (Bramer, 2007).

K-en yakın komşu algoritması ile veriler basit ve etkili bir şekilde sınıflandırılabilir. Ancak algoritma örnek tabanlı olduğundan uygulama yapabilmek için eğitim verilerinin mevcut olması gerekmekte ve bu nedenle büyük veri kümeleri için büyük miktarda depolama alanına ihtiyaç duyulmaktadır. Verilerin yapısı hakkında bilgi vermemesi (model oluşturmaz) KNN'nin bir diğer dezavantajıdır (Harrington, 2012).

2.1.2 Naive Bayes Algoritması

Naive bayes algoritması gözlemelere dayalı olasılıkları ve olasılık dağılımındaki parametreleri hesaplayan bir yöntemdir (McNamara ve ark., 2006). İstatistik tabanlı algoritmalar arasında yer alan Bayesci sınıflama tekniği (Bramer, 2007), veri seti içerisinde sınıfı bilinen verileri kullanarak yeni bir verinin mevcut sınıflardan herhangi birine girme olasılığını belirlemektedir (Silahtaroğlu, 2013).

Naive Bayes, makine öğrenmesi ve veri madenciliği için en verimli ve etkili tümevarımsal öğrenme algoritmalarından biridir. Bu algoritma basit olmasına rağmen, çeşitli öğrenme problemlerinde iyi performans sergilemektedir (Frank ve ark., 2003). Naive Bayes sınıflandırıcı, test örneklerinin sınıfını doğru tahmin eden ve sınıf bilgisi taşıyan eğitim örnekleri gibi danışmanlı tümevarım işlemlerinde kullanılmak üzere tasarlanmıştır (Balaban & Kartal, 2015).

Naive Bayes sınıflandırıcı ile başarılı bir sınıflandırma modeli oluşturmak için eğitim veri setinin büyük olmasına gerek yoktur. Konu ile ilgili olmayan niteliklere karşı güçlü olup bu sınıflandırıcı için örneklem boyutu arttıkça ilgisiz veriler önemsiz hale gelmekte ve gerçek zamanlı çevrimiçi sistemlere kolayca entegre edilebilmektedir (Lewis, 2017).

Algoritma, sınıfları belirlemek için koşullu olasılıkları kullanmaktadır. Burada $X = \{x_1, x_2, \dots, x_n\}$ nitelik değerlerinden oluşan ve sınıf üyeliği bilinmeyen veri örneğini, $C_1, C_1, \dots, C_m$ m sınıfın sınıf değerlerini göstermektedir. Sınıfı belirleyecek olan örneğe ilişkin olasılıklar,

$$P(C_i|X) = \frac{P(X|C_i)P(C_i)}{P(X)} \tag{2}$$

olarak hesaplanır. $P(X|C_i)$ olasılığı basitleştirilerek hesaplamadaki işlem yükü azaltılabilir. Bunun için, x_k değerlerinin birbirinden bağımsız olduğu kabul edilmekte ve Eşitlik (3) kullanılmaktadır.

$$P(X|C_i) = \prod_{k=1}^{n} P(X_k|C_i) \tag{3}$$

Bilinmeyen X örneğini sınıflandırmak için eşitlik (2) deki en büyük değer belirlenmekte ve bilinmeyen örneğin bu sınıfa ait olduğuna karar verilmektedir.

$$arg\max_{Ci}\{P(X|C_i)P(C_i)\} \tag{4}$$

Sonlu olasılıkları kullanan Eşitlik (4), en büyük sonlu sınıflandırma yöntemi (Maximum A Posteriori classification=MAP) olarak da bilinir ve Bayes sınıflayıcısı Eşitlik (5)'i kullanmaktadır (Özkan, 2008):

$$C_{MAP} = arg\max_{Ci} \prod_{k=1}^{n} P(X_k|C_i) \tag{5}$$

2.1.3 Lojistik Regresyon Analizi

Lojistik regresyon modeli doğrusal regresyon modelinin özel bir hali olup bu modelde bağımlı değişken iki sınıflı kategorik bir değişkendir (Lewis, 2017). Lineer regresyonda aranan varsayımların aranmaması, değişken tipi ve dağılımı ile ilgili varsayımların az olması nedeniyle lojistik regresyon araştırmacıların ilgisini çekmektedir (Balaban & Kartal, 2015). Lojistik regresyon analizinde amaç en az değişken ile regresyon katsayılarını optimize ederek model oluşturmaktır (Akpınar, 2014). Veri madenciliği çalışmalarında ise sınıflandırma amacı ile kullanılmaktadır. Lojistik regresyon analizinde sigmoid fonksiyon olarak adlandırılan lojistik fonksiyon aşağıdaki gibi yazılabilir (Harrington, 2012):

$$P(Y=1)=\frac{e^{\beta_0+\beta_1X_1}}{1+e^{\beta_0+\beta_1X_1}}=\frac{1}{1+e^{-(\beta_0+\beta_1X_1)}} \quad (6)$$

Burada β'lar model parametrelerini, x'ler açıklayıcı değişkenleri, Y ise genellikle 0 ve 1 değerlerinden oluşan kategorik bağımlı değişkenini ifade etmektedir. Logaritmik dönüşüm uygulanarak bu ilişki doğrusal bir şekilde incelenebilir. Bu dönüşüm lojit dönüşüm olduğundan yöntem lojistik regresyon olarak adlandırılır (Gamgam & Altunkaynak, 2017). Bağımsız değişken sayısı p olduğu zaman regresyon modeli:

$$P(Y=1|\,x)=\frac{e^{\beta_0+\beta_1X_1+\cdots+\beta_pX_p}}{1+e^{\beta_0+\beta_1X_1+\cdots+\beta_pX_p}}=\frac{1}{1+e^{-(\beta_0+\beta_1X_1+\cdots+\beta_pX_p)}} \quad (7)$$

olur. Burada $P(Y|\,x)$, bağımsız değişkenlerin değeri bilindiğinde bağımlı değişkenin olasılığını göstermektedir. Bu model Eşitlik (8)'deki gibi düzenlenebilir.

$$ln\left(\frac{P(Y=1|\,x)}{1-P(Y=1|\,x)}\right)=\beta_0+\beta_1X_1+\cdots+\beta_pX_p \quad (8)$$

Bir olayın gerçekleşme olasılığının gerçekleşmeme olasılığına oranı olan $P(Y=1|\,x)$/(1- $P(Y=1|\,x)$) ifadesi odds oranını göstermektedir (Hosmer ve ark., 2013).

Bağımlı değişkenin iki kategoriden oluşması ikili (Binary), kategorilerin ikiden fazla ve sıralı olması sıralı (Ordinal), ikiden fazla ve sıralı olmaması durumunda ise nominal lojistik regresyon analizi geçerlidir (Özdamar, 2019).

2.2. Sağdan-sansürlü veri

Sansürlü veriler; sağdan sansürlü (right censored), soldan sansürlü (left censored) ve aralıklı sansürlü (interval censored) veriler olmak üzeri üç ana başlık altında incelenmektedir (Bardakçı & Kartal, 2018). Bu çalışmada sağdan sansürlü veriler olması durumunda sınıflandırma algoritmalarının performansları karşılaştırıldığı için sağdan sansürlü veri tanıtıldıktan sonra, sansürün çözümlenmesi için kullanılan K en yakın komşu algoritması (kNN) imputasyon yöntemi belirtilmiştir.

Bir araştırma için klinik bir deney yapıldığı ve çalışmaya katılan n adet hastanın yaşam sürelerinin takip edildiği düşünüldüğünde, gözlem süresince maalesef çalışmadaki tüm (n) deneklerin hayatta kalma süreleri gözlemlenemez (sansürlenir). Bu tür durumlar aşağıdaki sebeplerden dolayı yaşanmaktadır;

- **Çalışmanın bitmesi**: Bazı bireyler için çalışma süresince ilgilenilen olay henüz gerçekleşmemiştir (ölüm, hastalık belirtisi vb.)
- **Bırakma**: Bireyin artık çalışmaya katılmak istememesi, bırakıp gitmesi.
- **Takipte kayıp**: Çalışmadaki kişi artık çalışmada görünmez, kişinin izi kaybedilmiştir.

Böyle durumlarda çalışmada bulunan bazı bireyler için kısmi bilgi elde edilir. "c" çalışmanın gerçek bitiş süresini gösterdiğinde bazı bireyler için çalışma sonuna kadar ilgilenilen durum gerçekleşmediğinde en kötü ilgili bireyin hayatta kalma süresi c_i olacaktır. Böylece i'inci birey için kısmi bilgiye sahip olunur ve y_i hayatta kalma gözleminin c_i tarafından sansürlendiği söylenebilir. Dolayısıyla her bir özne veya nesne için c_i ve y_i değerleri arasından en küçük olanının seçilmesiyle yaşam sürelerini sansür durumlarıyla birlikte sansür bilgisini içeren yeni bir yanıt değişkeni elde edilir. Gözlemler;

$$Z_i=\min(y_i,c_i)\ ve\ \delta_i=I(y_i\le c_i) \quad (9)$$

şeklinde olur. Böylece sansürü taşıyan gözlem çiftleri elde edilmekte ve burada δ_i'ler sansür bilgisini taşıyan gösterge fonksiyonunu vermektedir. Sansür varsa "0" yoksa "1" değerini göstermektedir (Gijbels, 2010; Yılmaz & Aydın, 2019).

Sağdan sansürlü verilerin tahmin edilmesi ile ilgili çalışmalarda önemli bir varsayım olan bağımsızlık varsayımı: y_i hayatta kalma süresinin tüm c_i sansürleme süresinden bağımsız olduğudur (Gijbels, 2010). Bu sansür mekanizması, sağ rastgele sansür olarak anılmaktadır. İkinci varsayım ise $P(y_i\le c_i|y_i,x_i)=P(y_i\le c_i|y_i)$'dir. Yani olay (ölüm, belirti vb.) zamanı için açıklayıcı değişkenler veri sansürlü olsa da olmasa da elde edilenden daha fazla bilgi sağlamamaktadır (Yılmaz & Aydın, 2019).

2.2.1. k-NN yerine koyma yöntemi ile sansürün çözümlenmesi

kNN yerine koyma yöntemi, sağdan sansürlü veri noktalarını tahmin etmek için kullanılabilir. Bu yöntem dağılımdan tamamen bağımsız bir yöntem olup hiçbir varsayım kullanmadığı için parametrik olmayan bir yöntemdir. Kesikli ve sürekli veri yapısındaki değişkenler için kullanılabilir. Yerine koymak için tahmin edilen değerler gerçek verilerden elde edildiği için açıklayıcı değişkenler hakkında daha fazla bilgi içermektedir. Bunlar yöntemin önemli avantajları arasında yer almaktadır. Bu, veri noktaları arasındaki mesafelere bağlı olarak benzerliğe dayalı çalışan bir yöntemdir. Genellikle bu mesafe, Eşitlik (1)'de verilmiş olan Öklid uzaklıkları ile ölçülmektedir. k-NN yerine koyma algoritmasının uygulama adımları Tablo 1'de verilmiştir (Ahmed ve ark., 2020);

Tablo 1. k-NN için algoritma adımları

Algoritma: $Sağdan-sansürlü\ veriler\ için\ kNN\ yerin\ koyma\ yöntemi$
Girdi: Sağdan sansürlü veriseti z_i
Sansür işaretçisi δ_i
En yakın komşu sayısı k
Açıklayıcı değişken değerleri x_i
Çıktı: Sansürlü verilerin tahminlerini içeren yeni değişken $\mathbf{y}^{knn} = \left(y_1^{knn}, \dots, y_n^{knn}\right)^T$
1 **başla**
2 **for** $(i = 1$ to $n)$
3 **Eğer** $(\delta_i = 0)$ **yap** (eğer veri noktası sansürlü ise)
4 **for**$(j = 1$ to $n)$
5Her bir sansürlü gözlem için x_j ve x_i için eşitlik $(1)'$te verilen öklit uzaklıkları bulunur
6 Mesafeler küçükten büyüğe sıralanır
7 **for** $(j = 1$ to $k)$
8 Sıralanan mesafelerle ilişkili ilk k adet sansürlü olmayan gözlem alınır.
9 i′nci sansürlü veri tahmini $\left(y_i^{knn}\right)$ en yakın k adet y_i değerinin ortalaması alınarak hesaplanır.
10 Tahmin edilen gözlemler $\left(y_i^{knn}\right)$ sansürlü gözlemler $(z_i, \delta_i = 0)$ile yer değiştirilir.
11 $\mathbf{y}^{knn} = \left(y_1^{knn}, \dots, y_n^{knn}\right)^T$ oluşturulur.
12 **Son**

Böylece veri setinde sansürün etkisi modele dahil edilerek yaşam sürelerini temsil eden veriler yerine $\mathbf{y}^{knn}$ kullanılabilir.

2.3. Sınıflandırma Performanslarının Değerlendirilmesi

Sınıflandırma algoritmaları ile kurulan modellerin başarısını değerlendirmek için Doğruluk, Duyarlılık, Seçicilik, Kesinlik ve F-Ölçütü gibi metrikler bulunmaktadır. Hata matrisi tablosunda yer alan gerçek ve tahmin değerlerinden elde edilen bazı metriklerle sınıflandırma model performansları değerlendirilmektedir (Balaban ve Kartal, 2015).

Tablo 2. Hata matrisi (Confusion matrix)

		Tahmin Edilen Sınıf	
		Pozitif	**Negatif**
Gerçek Sınıf	**Pozitif**	TP (True Positive)	FN (False Negative)
	Negatif	FP (False Positive)	TN (True Negative)

Doğruluk (Accuracy): Modelin ortalama performansını ve sınıflandırma başarısını gösteren en basit ve en sık kullanılan ölçüttür. Doğru sınıflandırılmış değerlerin toplam örnek sayısına bölünmesi ile hesaplanmaktadır.

$$Doğruluk = \frac{TP + TN}{TP + TN + FP + FN} \tag{10}$$

Duyarlılık (Sensitivity): Doğru sınıflandırılmış pozitif değerlerin toplam gerçek pozitif örnek sayısına bölünmesi ile hesaplanmaktadır.

$$Duyarlılık = \frac{TP}{TP + FN} \tag{11}$$

Seçicilik (Spesificity): Doğru sınıflandırılmış negatif örneklerin toplam negatif tahmin edilen örneklere bölünmesi ile hesaplanmaktadır.

$$Seçicilik = \frac{TN}{TN + FP} \tag{12}$$

Kesinlik (Precision): Doğru sınıflandırılmış pozitif örneklerin toplam pozitif tahmin edilen örneklere bölünmesi ile hesaplanmaktadır

$$Kesinlik = \frac{TP}{TP + FP} \tag{13}$$

F-Ölçütü (F-Measure): Kesinlik ve Duyarlılık metriklerinin harmonik ortalaması alınarak her iki ölçüyü beraber değerlendirme imkânı vermektedir (Mulla ve ark., 2021).

$$F\ Ölçütü = 2\ x\ \frac{Duyarlılık\ x\ Kesinlik}{Duyarlılık +\ Kesinlik} \tag{14}$$

3. Uygulama

Sağdan sansürlü veriler bulunması durumunda sınıflandırma algoritma performanslarının karşılaştırılması amacı ile iki gerçek veri seti ve daha sonra ise simülasyon çalışması ile uygulama gerçekleştirilmiştir. Uygulamalarda R programlama dili kullanılmıştır.

%75'i eğitim ve %25'i test olacak şekilde veri setleri iki gruba ayrılmıştır. Modeller eğitim veri seti ile oluşturulmuş daha sonra model performansları test veri setiyle ölçülmüştür. Model performanslarını karşılaştırmak üzere doğruluk, duyarlılık, seçicilik, kesinlik ve F-ölçütü değerleri kullanılmıştır.

3.1. Rektum Kanseri veri seti

Rektum veri setinde (Aydin ve Yilmaz, 2018) yer alan değişkenler Tablo 3'de gösterilmiştir. Veri seti 97 gözlemden oluşmakta ve bu gözlemlerin 32'si sağdan sansürlüdür. Böylece veri seti %33 orta düzey sansür içermektedir. Sınıflandırılan veri setinde bir bağımlı ve dört bağımsız değişken bulunmaktadır.

Tablo 3. Rektum veri setinde yer alan değişkenler.

Öznitelik Adı	Veri Türü	**Nümerik**=Min, Max **Kategorik**= Değerler	**Nümerik**=Ort.±s.s **Kategorik**= n (%)
Yaş	Nümerik	18- 85	56.6±14.3
Preop.Alb<3gr:1 >3gr:2	Nümerik	1, 2	1: n=15 (%15) 2: n=82 (%85)
Toplam kan tx	Nümerik	0- 17	3.4±3.5
Operasyon süresi (dk.)	Nümerik	60- 330	205.8±69.0
Sansür Durumu	Kategorik	0=sansürlü, 1=sansürsüz	0: n=32 (%33) 1: n=65 (%67)
Yaşam Süresi*	Nümerik	1- 94	25.9±23.3
İmputed Data (Yaşam Süresi)	Nümerik	1- 98.2	31.0±27.2
Tümörün evresi**	Kategorik	1, 2, 3, 4	1: n=3 (%3) 3: n=20 (%21) 2: n=28 (%29) 4: n=46 (%47)
Bağımlı Değişken (Tümörün evresi)	Kategorik	1: Başlangıç safhası 2: İleri safha	1: n=23 (%24) 2: n=74 (%76)

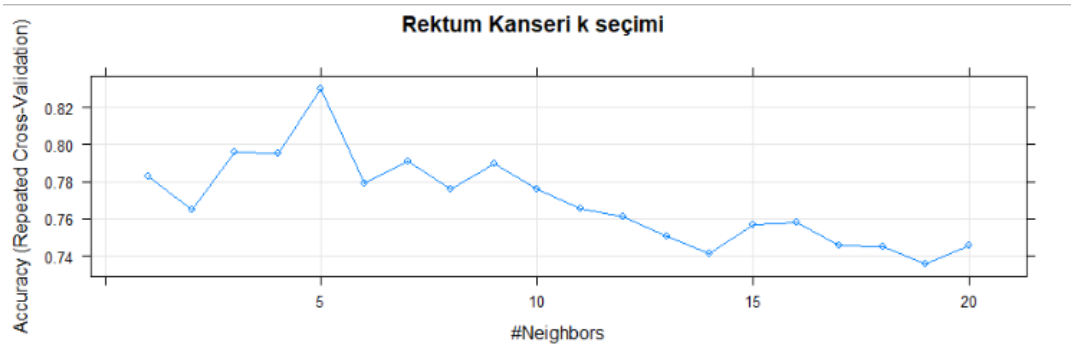
*yaşam süresi değişkeni sağdan sansürlü verileri temsil etmekte ve modelde bu değişken yerine İmputed data kullanılmıştır.
**tümörün evresi değişkeni yeniden sınıflandırılmış ve veri setinde bağımlı değişken olarak kullanılmıştır.

Rektum veri seti için beş sınıflandırma performans değerlendirme kriteri dikkate alınarak sınıflandırma algoritmalarının sınıflandırma performans sonuçları Tablo 4'de verilmiştir.

Tablo 4. Rektum veri setinde için sınıflandırma performansları.

Ölçütler / Algoritma	Doğruluk	Duyarlılık	Seçicilik	Kesinlik	F-Ölçütü
K-NN	0.78	0.60	0.83	0.50	0.54
Naive Bayes	0.70	**0.60**	0.72	0.38	0.46
Lojistik Regresyon	**0.83**	**0.60**	**0.89**	**0.60**	**0.60**

Sağdan sansürlü rektum veri setinde oluşturulan modellerden en iyi performans Lojistik Regresyon algoritması tarafından %83 doğrulukla elde edilmiştir. En düşük performans ise %70 doğrulukla Naïve Bayes algoritmasıyla elde edilmiştir. Duyarlılık, Seçicilik, Kesinlik ve F-Ölçütü metriklerine göre incelendiğinde en yüksek değerler Lojistik Regresyon algoritması ile elde edildiği gözlenmiştir.



Şekil 1. Rektum veri seti için k değerinin seçimi (k=5)

3.2. Hepatosellüler veri seti

Hepatosellüler veri seti 26 Mayıs 2022 tarihinde (https://rdrr.io/cran/asaur/man/hepatoCellular.html) adresinden alınmış ve veri setinde yer alan değişkenler Tablo 5'de verilmiştir. Veri seti 227 gözlemden oluşmakta ve bu gözlemlerin 84'ü sağdan sansürlüdür. Böylece veri seti %37 orta düzey sansür içermektedir. Sınıflandırma yapılan veri seti, 1 bağımlı ve

11 bağımsız değişkenden oluşmaktadır.

Tablo 5. Hepatosellüler veri setinde yer alan değişkenler.

Öznitelik Adı	**Veri Türü**	**Nümerik**=Min, Max **Kategorik**= Değerler	**Nümerik**=Ort.±s.s **Kategorik**= n(%)
Yaş	Nümerik	13- 79	49.8 - 12.4
RFS	Nümerik	1- 81	26.3 - 23.2
CXCL17T	Nümerik	0- 1184.4	99.3 - 181.4
CXCL17P	Nümerik	2,0- 1016.0	100.5 - 130
CXCL17N	Nümerik	0- 1171.1	100 - 181.4
Tümör boyutu	Kategorik	1,2	1: n=96 (%42.3) 2: n=131 (%57.7)
Cinsiyet	Kategorik	0,1	0: n=30 (%13.2) 1: n=197 (%86.8)
HBsAg	Kategorik	0,1	0: n=18 (%7.9) 1: n=209 (%92.1)
Sansür Durumu	Kategorik	0=sansürlü, 1=sansürsüz	0: n=84 (%37.0) 1: n=143 (%63.0)
Tumormultiplicity	Kategorik	0,1	0: n=170 (%74.9) 1: n=57 (%25.1)
OS*	Nümerik	2- 83	36.5- 22.2
İmputed Data (OS)	Nümerik	2- 94.9	42- 26.6
Bağımlı Değişken (Vascularinvasion)	Kategorik	0, 1	0: n=186 (%81.9) 1: n=41 (%18.1)

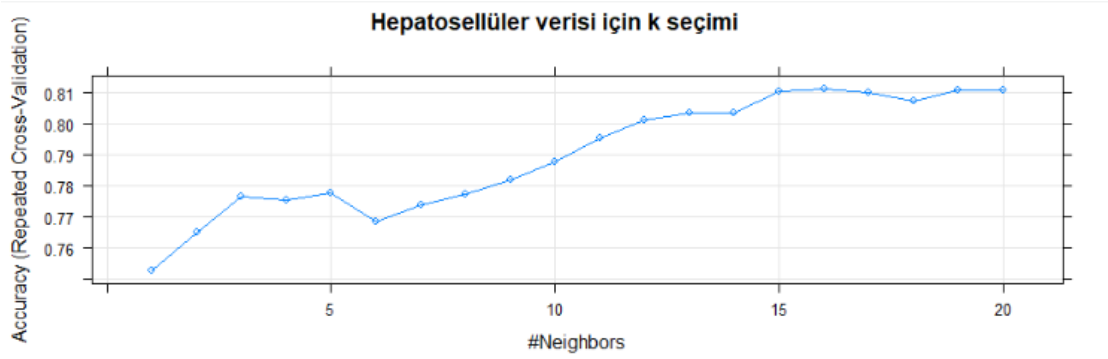
*OS değişkeni sağdan sansürlü verileri temsil etmektedir ve modelde bu değişken yerine impüstasyon ile elde edilen veri kullanılmıştır.

Beş sınıflandırma performans değerlendirme kriteri dikkate alınarak hepatosellüler veri seti için sınıflandırma algoritmalarının sınıflandırma performans sonuçları Tablo 6'da verilmiştir.

Tablo 6. Hepatosellüler veri setinde için sınıflandırma performansları.

Ölçütler / **Algoritma**	**Doğruluk**	**Duyarlılık**	**Seçicilik**	**Kesinlik**	**F-Ölçütü**
K-NN	**0.82**	1.00	0.00	0.82	**0.90**
Naive Bayes	0.77	**0.87**	**0.30**	0.85	0.86
Lojistik Regresyon	0.79	0.83	0.25	**0.93**	0.88

Sağdan sansürlü hepatosellüler veri setinde oluşturulan modellerden en iyi performans %82 doğrulukla K-NN algoritmasıyla ve en düşük performans ise %77 doğrulukla Naive Bayes algoritmasıyla elde edilmiştir. Duyarlılık ve Seçicilik için en yüksek değerler Naïve Bayes algoritmasıyla, kesinlik için ise en yüksek değer %93 ile Lojistik Regresyon algoritması ile elde edilmiştir.



Şekil 2. Hepatosellüler veri seti için k değerinin seçimi (k=16)

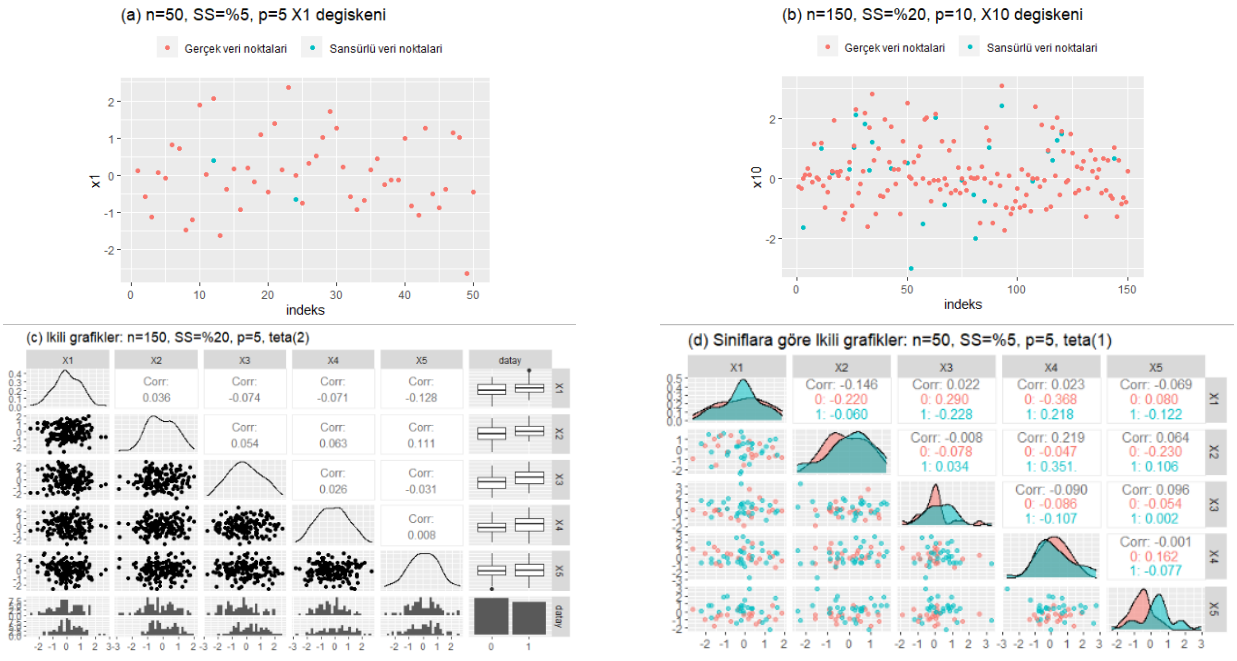
3.3. Simülasyon Çalışması

Simülasyon çalışması için veri üretimi ve simülasyon kurgusu Tablo 7'de verilmiştir. Her bir simülasyon 1000 tekrarlı olup sonuçlar tablolar halinde sunulmuştur. Simülasyon çalışmalarında bağımlı değişkenin içerdiği sınıf sayısı her zaman iki olarak alınmıştır. Simülasyon kurgusu göz önüne alındığında, sınıflandırma algoritmalarının performanslarına etkileri bakımından hem sansür seviyesinin hem örneklem büyüklüğünün hem de bağımsız değişken sayısının performansa nasıl etki yaptıkları gözlemlenebilmektedir. Eğitim ve test veri setleri gerçek veride olduğu gibi ayarlanmıştır. Ayrıca, elde edilen sonuçlar, gerçek veri çalışmaları ile de karşılaştırılmış ve uyum ya da uyumsuzluklar tartışılmıştır.

Tablo 7. Simülasyon kurgusu ve veri üretimi

Simülasyon kurgusu		
Örneklem büyüklüğü	Sansür seviyesi	Açıklayıcı değişken sayısı
$n = 50, 150$	$SS = (\%5, \%20)$	$p = (5, 10)$
Verilerin üretilmesi		
Bağımsız değişkenler	Olasılıklar üretilmesi (sınıflar için)	Kategorik bağımlı değişken
$X \sim U[0,1] \in \mathbb{R}^{n \times p}$	$z = 1 + \theta_p X,$ $P_y = \frac{1}{(1 + e^{-z})}$	$y = binom(n, P_y)$

Burada θ_p, $(p \times 1)$ bir vektörü temsil eder ve açıklayıcı değişkenin katsayılarını ifade eder. Veri üretiminde iki farklı katsayı vektörü ele alındı. Bunlar; $\theta_p^{(1)} = (0.5, 0.5, \ldots, 0.5)$ ve $\theta_p^{(2)} = (3, 3, \ldots, 3)$. Böylece, açıklayıcı değişkenlerin, sınıf değişkeni üzerindeki farklı seviyedeki etkilerinin incelenmesi planlanmıştır.



Şekil 3. Simülasyonda belirli konfigürasyonlar için üretilen verilere ait tanımlayıcı grafikler

Şekil 1, üretilen verileri, sansürlülük durumlarını ve değişkenlerin birbirleriyle ilişkilerini gözlemleyebilmek amacıyla elde edilmiştir. Dört panelden oluşmakta ve her bir panel, farklı simülasyon konfigürasyonlarını temsil etmektedir. Panel (a), düşük sansür ve panel (b) yüksek sansür seviyeleri için sırasıyla küçük ve büyük örneklem büyüklüklerine göre verilerin dağılımını göstermektedir. Burada sansürlenen değişkenler, açıklayıcı değişkenlerdir. Her bir değişken aynı dağılımdan üretildiğinden, temsilen bir tanesi için saçılım grafiği elde edilmiştir. Panel (c), verilerin tamamı için ikili ilişkilerin yoğunluklarını ve korelasyon değerlerini gösterirken, Panel (d) bağımlı değişkene göre sınıfları içeren ikili grafikleri göstermektedir. Grafiklere göre bağımsız değişkenler arasında güçlü bir ilişkinin olmadığı görülmüş ki bu, lojistik regresyon modelinin doğru çalışabilmesi için gerekli bir varsayımdır. Burada belirtmek gerekir ki

korelasyon değerleri olarak Spearman korelasyon katsayıları hesaplanmıştır. Aksi halde, tanımsızlıkların elde edilmesi mümkündür. Diğer makine öğrenmesi yöntemlerinin sıkı varsayımlar gerektirmemesi nedeniyle veri üretiminde başka bir varsayım ele alınmamıştır. Sonuçlardan önce belirtmelidir ki, kNN için "k" gerçek veri çalışmalarında olduğu gibi doğruluk kriterini maksimum yapacak şekilde çapraz geçerlilik (cross-validation) ile optimize edilerek seçilmiştir.

Sınıflandırma algoritmalarının genel performanslarına ait bütün muhtemel simülasyon senaryoları Tablo 8 ve Tablo 9'da verilmiş ve Tablo 8, bağımsız değişken sayısı $p = 5$ olduğu durum içindir. En iyi performans gösteren yönteme ait değerler tablolarda kalın şeklide vurgulanmıştır.

Tablo 8. $p = 5$ olduğunda örnek büyüklüğü ve sansür seviyesine göre sonuçlar

		%5						%20					
n	Algoritma	K-NN		Naive Bayes		Lojistik Reg.		K-NN		Naive Bayes		Lojistik Reg.	
		$\theta_p^{(1)}$	$\theta_p^{(2)}$	$\theta_p^{(1)}$	$\theta_p^{(2)}$	$\theta_p^{(1)}$	$\theta_p^{(2)}$	$\theta_p^{(1)}$	$\theta_p^{(2)}$	$\theta_p^{(1)}$	$\theta_p^{(2)}$	$\theta_p^{(1)}$	$\theta_p^{(2)}$
50	**Duyarlılık**	0.652	0.900	**0.840**	0.843	**0.840**	**0.942**	0.608	0.850	**0.804**	0.810	0.768	**0.864**
	Seçicilik	0.917	**0.886**	0.787	**0.886**	**0.920**	0.880	0.847	0.840	0.747	0.780	**0.876**	**0.850**
	Kesinlik	0.786	0.876	0.722	0.881	**0.840**	**0.843**	0.700	**0.888**	0.751	0.764	**0.864**	0.810
	F-Ölçütü	0.687	**0.880**	0.762	0.858	**0.832**	0.878	0.604	**0.842**	0.774	0.779	**0.809**	0.816
	Doğruluk	0.784	0.893	0.813	0.865	**0.880**	**0.911**	0.777	0.845	0.725	0.795	**0.822**	**0.857**
150	**Duyarlılık**	0.711	0.792	**0.937**	0.855	0.809	**0.931**	0.605	0.779	**0.871**	0.769	0.793	**0.857**
	Seçicilik	0.777	**0.920**	0.878	0.907	**0.928**	0.898	0.810	0.899	**0.909**	**0.907**	0.839	0.829
	Kesinlik	0.733	**0.906**	0.792	0.901	**0.885**	0.870	0.744	0.838	**0.844**	**0.898**	0.789	0.892
	F-Ölçütü	0.712	0.842	0.754	0.871	**0.833**	**0.898**	0.754	0.804	**0.854**	0.824	0.797	**0.872**
	Doğruluk	0.744	0.856	0.808	0.892	**0.861**	**0.915**	0.790	0.839	**0.890**	0.838	0.824	**0.893**

Tablo 8 incelendiğinde, genel çerçevede lojistik regresyon yönteminin diğer üç yönteme göre üstün olduğu görülmektedir. Ayrıca, ölçüt değerlerindeki düşüş ile sansürün performans üzerinde negatif etkiye sahip olduğu gözlemlenebilir. Fakat bu genel bir doğru olarak sunulmamalıdır. Zira hem sansürlülük çözümü için kullanılan k-NN imputasyon yönteminin, hem sınıflandırma algoritmalarının sıkı varsayımlara dayanamaması nedeniyle sansür seviyesinin kesinlikle negatif etki oluşturacağı söylenemez. Bu durum hem Tablo 8 hem de Tablo 9'da pek çok kez gözlenmiştir. Aynı şekilde, örneklem büyüklüğü arttığında performansların artması beklenir. Fakat aynı sebepten bu beklenti gerçekleşmeyebilir. Bu bağlamda, yöntemlerin performansı incelendiğinde, örneklem büyüklüğü artıkça performansın arttığı gözlenmiştir. Diğer yandan, sansür seviyesi arttıkça hem Tablo 8 hem de Tablo 9'da Naive-Bayes algoritmasının ağır sansür altında ve n=150 olduğunda lojistik regresyon ile birlikte performansının kNN modeline göre daha iyi sınıflandırma performansı gösterdiği açıktır. Ek olarak kNN yönteminin her iki örneklem büyüklüğü için de seçicilik ve kesinlik kriterleri özelinde fakat veri üretiminde $\theta_p^{(2)}$ kullanıldığında iyi sonuçlar gösterdiği saptanmıştır. Burada belirtmek gerekir ki katsayı vektörü θ_p tahmin performanslarını önemli ölçüde etkilemiştir. $\theta_p^{(1)}$ için elde edilen performanslar $\theta_p^{(2)}$'ye göre çok daha düşük elde edilmiştir ki bu beklenen bir sonuç olmasına rağmen, çalışmanın önemli sonuçlarından sayılabilir. Çünkü açıklayıcı değişkenlerin sınıf değişkeni üzerindeki etkisi arttıkça sınıflandırma performansının iyileşeceği öngörülebilir ve bu öngörü verilen Tablo 8 ve Tablo 9'daki değerlerle ispatlanmıştır.

Tablo 9. $p = 10$ olduğunda örnek büyüklüğü ve sansür seviyesine göre sonuçlar

		%5						%20					
n	Algoritma	K-NN		Naive Bayes		Lojistik Reg.		K-NN		Naive Bayes		Lojistik Reg.	
		$\theta_p^{(1)}$	$\theta_p^{(2)}$	$\theta_p^{(1)}$	$\theta_p^{(2)}$	$\theta_p^{(1)}$	$\theta_p^{(2)}$	$\theta_p^{(1)}$	$\theta_p^{(2)}$	$\theta_p^{(1)}$	$\theta_p^{(2)}$	$\theta_p^{(1)}$	$\theta_p^{(2)}$
50	**Duyarlılık**	0.600	**0.853**	**0.800**	0.767	0.753	0.785	0.590	0.826	0.640	**0.900**	**0.653**	0.800
	Seçicilik	0.826	0.876	**0.892**	0.590	0.740	**0.893**	**0.735**	0.666	0.726	0.790	0.718	**0.966**
	Kesinlik	0.733	0.820	**0.820**	0.680	0.580	**0.933**	0.614	0.761	**0.708**	0.736	0.616	**0.950**
	F-Ölçütü	0.545	0.820	**0.803**	0.719	0.613	**0.847**	0.522	0.778	0.647	0.801	**0.658**	**0.863**
	Doğruluk	0.658	**0.864**	**0.846**	0.678	0.747	0.839	0.667	0.746	**0.683**	0.845	0.676	**0.883**
150	**Duyarlılık**	0.734	0.788	**0.936**	0.844	0.840	**0.920**	0.774	0.753	0.766	**0.963**	**0.822**	0.910
	Seçicilik	0.735	0.831	0.833	0.834	**0.955**	**0.889**	0.841	0.841	0.774	0.832	**0.842**	**0.929**
	Kesinlik	0.733	0.814	0.803	0.824	**0.941**	**0.872**	0.782	0.796	0.746	0.809	**0.825**	**0.895**
	F-Ölçütü	0.721	0.792	0.864	0.829	**0.885**	**0.889**	0.775	0.767	0.753	0.876	**0.823**	**0.899**
	Doğruluk	0.729	0.810	0.885	0.839	**0.897**	**0.905**	0.807	0.797	0.770	0.898	**0.832**	**0.919**

Tablo 9, bağımsız değişken sayısı $p = 10$ olduğunda muhtemel tüm simülasyon senaryolarını içermektedir. Tablo incelendiğinde, $n = 50$ ve CL=5% olduğunda Tablo 8'deki sonuçlardan farklı olarak Naive-Bayes yönteminin $\theta_p^{(1)}$ vektörü için diğer iki yöntemden daha iyi performans gösterdiği görülmüştür. Buna göre Naive-Bayes yönteminin açıklayıcı değişken sayısı arttığında ve bu değişkenlerin katkılarının küçük olduğu durumda kullanılması önerilebilir ki bu sonuç, çalışmanın katkılarından biri olarak sunulabilir. Diğer yandan, genel tablo incelendiğinde, lojistik regresyon modelinin diğer iki yönteme göre çok daha iyi sonuçlar sunduğu ve Naive-Bayes'in lojistik regresyon yakın değerler verdiği görülmüştür. kNN sınıflandırma yöntemi her ne kadar $n = 150$ olduğunda ve $\theta_p^{(2)}$ katsayıları kullanıldığında tatmin edici sonuçlar verse de, diğer iki yöntem kadar iyi bir performans gösterememiştir. Bu bağlamda, $p = 10$ özelinde, yüksek sayıda açıklayıcı değişken içeren veri setlerinde kNN yerine lojistik regresyon veya uygun konfigürasyon altında ($n = 50, CL = 5\%$) Naive-Bayes kullanılması önerilebilir. Burada ayrıca belirtmek gerekir ki $n = 150$ olduğunda performanslar dikkate değer şekilde artmıştır. Ayrıca duyarlılık ölçütü bakımından en iyi performansı lojistik regresyon ve Naive Bayes yöntemleri göstermiştir. Sansür seviyesi $SS = \%20$ olduğunda Naive-Bayes yönteminin özellikle $\theta_p^{(1)}$ için ayırt edici şekilde sansürle baş etmede lojistik regresyon ile kNN algoritmasından daha iyi sonuçlar verdiği söylenebilir.

Gerçek verilerle simülasyon çalışmalarının uyumlu sonuçlar verdiği belirlenmiştir. Hem Rektum kanseri hem de Hepatosellüler veri setinde lojistik regresyon yönteminin performans ölçütleri bağlamında iyi sonuçlar göstermesi, simülasyon sonuçlarının gerçek veri ile uyumlu olduğu şeklinde yorumlanabilir.

4. Tartışma ve Sonuç

Bu çalışmada sansürlü veri setleri üzerinde üç farklı sınıflandırma yöntemi performanslarının karşılaştırılması hedeflenmiştir. Bu bağlamda sağdan sansürlü veriler için k-NN imputasyon yöntemi ile sağdan-sansürlü veriler tamamlanmış ve sınıflandırma yöntemleri elde edilen yeni veriler kullanılarak bağımlı değişken için sınıflandırma modelleri kurulmuş ve performansları ölçülmüştür. Sonuçları gözlemlemek için Rektum kanseri veri seti ve Hepatosellüler veri seti ile yapılan uygulamalara ek olarak farklı senaryolardan yöntemlerin davranışlarını gözlemlemek amacıyla simülasyon çalışması gerçekleştirilmiş ve sonuçlar sunulmuştur. Elde edilen sonuç ve öneriler aşağıda sıralanmıştır.

- Rektum kanseri veri seti sonuçları incelendiğinde, en iyi modellerin lojistik regresyon yöntemi ile elde edildiği görülmüştür. Sansürlü veri bağlamında oldukça yüksek sayılabilecek bir sansür seviyesi için 0.83 doğruluk değeriyle Lojistik regresyon algoritması tatmin edici sonuçlar vermiştir. Böylece sansürle baş etmede bu yöntemin dikkate değer performans gösterdiği görülmüştür.
- Hepatosellüler veri seti sonuçları incelendiğinde, hemen hemen aynı sansür seviyesi için (SS=%37) her yöntem farklı performans kriteri için iyi sonuçlar vermiştir. Bu durum, $n = 227$ geniş örneklem büyüklüğü söz konusu olduğunda elde edilmiş ve $n = 150$ olduğunda simülasyon sonuçları bunu doğrular niteliktedir. Doğruluk kriterinde en iyi sonuçlar K-NN yöntemiyle sağlanmıştır.
- Simülasyon sonuçlarına göre ise, sansüre karşı en dayanıklı iki yöntemin lojistik regresyon ve Naive-Bayes olduğu görülmüştür. Ayrıca lojistik regresyon yönteminin büyük örneklemlerde iyi sonuçlar verdiği gözlemlenmiş ve bu durum Hepatosellüler veri seti sonuçlarıyla uyumluluk göstermektedir. K-NN simülasyon çalışmasında her ne kadar diğer yöntemlere yakın performanslar gösterse de dikkate değer bir farklılık ortaya koyamamıştır.

Bu bağlamda gerçek veri setleri ile simülasyon çalışmaları beraber değerlendirildiğinde doğruluk ölçütüne göre; yüksek ve düşük sansür seviyesinde Lojistik Regresyon algoritmasının sansür ile baş etmede dikkate değer performans gösterdiği söylenebilir. Ayrıca örneklem büyüklüğü arttıkça genel olarak algoritmaların doğru sınıflama performanslarının arttığı gözlenmiştir. Özetle, büyük örneklemli veri setlerinde Lojistik Regresyon algoritmasının doğru sınıflandırma oranı ile başarılı sınıflandırma performansı gösterdiği söylenebilir.

Kaynakça

Ahmed, S. E., Aydin, D., & Yılmaz, E. (2020). Nonparametric regression estimates based on imputation techniques for right-censored data. *Advances in Intelligent Systems and Computing, 1001*, 109–120. https://doi.org/10.1007/978-3-030-21248-3_8

Akpınar, H. (2014). *Data : Veri Madenciliği Veri Analizi* (Genişletil). Papatya Bilim Yayınevi.

Aydin, D., & Yilmaz, E. (2018). Modified spline regression based on randomly right-censored data: A comparative study. *Communications in Statistics: Simulation and Computation, 47*(9), 2587–2611. https://doi.org/10.1080/03610918.2017.1353615

Balaban, M. E., & Kartal, E. (2015). *Veri Madenciliği ve Makine Öğrenmesi Temel Algoritmaları ve R Dili İle Uygulamaları* (Birinci Ba). Çağlayan Kitapevi.

Bandyopadhyay, S., Wolfson, J., Vock, D. M., Vazquez-Benitez, G., Adomavicius, G., Elidrisi, M., Johnson, P. E., & O'Connor, P. J. (2015). Data mining for censored time-to-event data: a Bayesian network model for predicting cardiovascular risk from electronic health record data. In *Data Mining and Knowledge Discovery* (Vol. 29, Issue 4, pp. 1033–1069). https://doi.org/10.1007/s10618-014-0386-6

Batista, G. E. A. P. A., & Monard, M. C. (2002). A study of k-nearest neighbour as an imputation method. In *Frontiers in Artificial Intelligence and Applications* (Vol. 87, pp. 251–260).

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer.

Bramer, M. (2007). *Principles of Data Mining. Undergraduate Topics in Computer Science*. Springer Verlag.

Davidson, I., & Tayi, G. (2009). Data preparation using data quality matrices for classification mining. In *European Journal of Operational Research* (Vol. 197, Issue 2, pp. 764–772). https://doi.org/10.1016/j.ejor.2008.07.019

Eröz, İ., & Tutkun, N. A. (2020). Aralıklı Sansürlü Veriler için Sağkalım Modelleri. *Süleyman Demirel Üniversitesi Fen Bilimleri Enstitüsü Dergisi*, *24*(2), 267–280. https://doi.org/DOI: 10.19113/sdufenbed.652776

Frank, E., Hall, B., & Pfahringer, B. (2003). Locally weighted naive bayes. *Proceedings of Conference on Uncertainty in Artificial Intelligence*, 249–256.

Gamgam, H., & Altunkaynak, B. (2017). *SPSS Uygulamalı Regresyon Analizi* (2. Basım). Seçkin Kitapevi.

Gijbels, I. (2010). Censored data. *Wiley Interdisciplinary Reviews: Computational Statistics*, *2*(2), 178–188. https://doi.org/10.1002/wics.80

Goldberg, Y., & Kosorok, M. R. (2012). Q-learning with censored data. *Annals of Statistics*, *40*(1), 529–560.

Guo, G., Wang, H., Bell, D., Bi, Y., & Greer, K. (2003). KNN Model-Based Approach in Classification. In R. Meersman, Z. Tari, & D. C. Schmidt (Eds.), *On The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE* (pp. 986–996). Springer Berlin Heidelberg.

Han, J., Kamber, M., & Pei, J. (2012). *Data Mining Concepts and Techniques* (Third Edit). Morgan Kaufman Publishers.

Harrington, P. (2012). *Machine Learning In Action*. Manning Publications.

Hosmer, D. W., Lemeshov, S., & Sturdivant, R. X. (2013). *Applied Logistic Regression* (Third Edit). John Wiley & Sons, Inc.

Ishwaran, H., Kogalur, U. B., Blackstone, E. H., & Lauer, M. S. (2008). Random survival forests. *Ann. Appl. Stat.*, *2*(3), 841–860.

Khan, F. M., & Zubek, V. B. (2008). Support vector regression for censored data (SVRc): A novel tool for survival analysis. *Proceedings - IEEE International Conference on Data Mining, ICDM*, 863–868. https://doi.org/10.1109/ICDM.2008.50

Lewis, N. D. (2017). *Machine Learning Made Easy with R: An Intuitive Step by Step Blueprint for Beginners*. CreateSpace Independent Publishing Platform.

McNamara, J. M., Green, R. F., & Olsson, O. (2006). Bayes' Theorem and Its Applications in Animal Behaviour. *Oikos*, *112*(2), 243–251. http://www.jstor.org/stable/3548663

Mucherino, A., Papajorgji, P. J., & Paradalos, P. M. (2009). *Data Mining In Agriculture*. Springer.

Mulla, G. A. A., Demir, Y., & 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. https://doi.org/10.17798/bitlisfen.939733

Özdamar, K. (2019). *Paket Programları İle İstatistiksel Veri Analizi-1* (11. Baskı). Nisan Kitapevi.

Özkan, Y. (2008). *Veri Madenciliği Yöntemleri*. Papatya Yayınevi.

Shivaswamy, P. K., Chu, W., & Jansche, M. (2007). A support vector approach to censored targets. *Proceedings - IEEE International Conference on Data Mining, ICDM*, 655–660. https://doi.org/10.1109/ICDM.2007.93

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

Štajduhar, I., Dalbelo-Bašić, B., & Bogunović, N. (2009). Impact of censoring on learning Bayesian networks in survival modelling. In *Artificial Intelligence in Medicine* (Vol. 47, Issue 3, pp. 199–217).

https://doi.org/10.1016/j.artmed.2009.08.001

Vock, D. M., Wolfson, J., Bandyopadhyay, S., Adomavicius, G., Johnson, P. E., Vazquez-Benitez, G., & O'Connor, P. J. (2016). Adapting machine learning techniques to censored time-to-event health record data: A general-purpose approach using inverse probability of censoring weighting. *Journal of Biomedical Informatics*, *61*, 119-131. https://doi.org/https://doi.org/10.1016/j.jbi.2016.03.009

Yılmaz, E., & Aydın, D. (2019). Regresyon Analizinde Sağdan Sansürlü Veriler İçin Önerilen Çözüm Yöntemleri Üzerine Bir İnceleme. *Turkiye Klinikleri Journal of Biostatistics*, *11*(3), 224–238. https://doi.org/10.5336/biostatic.2019-66838

Zhu, J., Ge, Z., Song, Z., & 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. https://doi.org/https://doi.org/10.1016/j.arcontrol.2018.09.003