Bilgisayar Bilimleri ve Mühendisliği Dergisi

Cilt-16
Sayı-1
2023

Sağlık Verilerinin Analizinde Veri Ön işleme Adımlarının Makine Öğrenmesi Yöntemlerinin Performansına Etkisi

The Effect of Data Preprocessing Steps on the Performance of Machine Learning Methods in the Analysis of Health Data

Hatice NİZAM ÖZOĞUR
İstanbul Üniversitesi-Cerrahpaşa
Bilgisayar Mühendisliği Bölümü
İstanbul, Türkiye
haticenizam@outlook.com
ORCID: 0000-0002-9722-4355

Zeynep ORMAN
İstanbul Üniversitesi-Cerrahpaşa
Bilgisayar Mühendisliği Bölümü
İstanbul, Türkiye
ormanz@istanbul.edu.tr
ORCID: 0000-0002-0205-4198

Öz

Günümüzde verilerin hızla artmasıyla makine öğrenmesi yöntemleri ile veri analizi birçok alanda popüler hale gelmiştir. Gerçek dünya veri kümelerinde eksik değerler ve dengesiz sınıf verileri sıklıkla karşılaşılan sorunlardır. Bu sorunlar, makine öğrenmesi yöntemlerinin başarımlarını olumsuz yönde etkilemekte ve modelin hatalı veya yanlış sonuçlar elde etmesine neden olmaktadır. Verilerdeki eksik değerlerin doldurulması ve sınıf dengesizliğinin ortadan kaldırılması veri ön işleme aşamasında önem arz etmektedir. Özellikle, sağlık verilerinde sınıfların dengesi verilerin doğruluğu ve eksiksizliği makine öğrenmesi yöntemlerinin performansını etkilediğinden çok önemlidir. Bu makalede, makine öğrenmesinde eksik değerlere sahip dengesiz veri sınıflandırması ile ilgili sorunları araştırmak için literatürde başarılı olan yöntemlerin karşılaştırmalı bir çalışması PIMA diyabet veri kümesi kullanılarak yapılmıştır. Elde edilen sonuçlara göre, sınıf dengesizliğinde eksik ve aşırı örnekleme yöntemlerinin birleştirildiği SMOTEENN algoritması ile eksik değerlerde zincirleme denklemlerle çoklu atama yönteminin kullanılması hasta ve sağlıklı bireylerin sınıflandırılmasında %91 F-skor değeri ile diğer en iyi yöntemlerden yaklaşık %9 oranında daha iyi performans göstermiştir.

Gönderme, düzeltme ve kabul tarihi: 26.08.2022 - 10.10.2022 – 17..06.2022
Makale türü: Araştırma

Anahtar sözcükler: Eksik değer, Dengesiz veri kümesi, Makine öğrenmesi, Sağlık veri kümesi

Abstract

Today, with the rapid increase in data, data analysis with machine learning methods has become popular in many areas. Missing values and imbalanced class data are common problems in real-world datasets. These problems negatively affect the performance of machine learning methods and cause the model to obtain erroneous or incorrect results. The missing values imputation and eliminating the class imbalance are important in the data preprocessing stage. In particular, the balance of classes in health data is very important as the accuracy and completeness of the data affect the performance of machine learning methods. In this article, a comparative study of successful methods in the literature for investigating problems with imbalanced data classification with missing values in machine learning was conducted using the PIMA diabetes dataset. According to the results, the SMOTEENN algorithm, which combines undersampling and oversampling methods in class imbalance, and the use of multiple imputation with chained equations for missing values, were showed an F-score value of 91%, approximately 9% better than the other best methods in classifying patients and healthy individuals.

Keywords: Missing value, Imbalanced dataset, Machine learning, Health dataset

1. Giriş

Makine öğrenmesi uygulamalarında veri kümelerindeki eksik veriler ve dengesiz veri sınıfları karşılaşılan en büyük sorunlardır. Geleneksel makine öğrenmesi algoritmaları eksiksiz veriler ve dengeli sınıf dağılımlara sahip veri kümelerine göre tasarlandığından bu tip problemlerde iyi performans gösteremezler. Gerçek dünya veri kümeleri, genellikle belirli bir sınıfın örneklerinin diğer sınıflara oranla daha az temsil edildiği dengesiz sınıflardan, eksik veya geçersiz verilerden meydana gelmektedir. Bu sorunlar ile başa çıkabilmek için araştırmacılar çeşitli alanlarda çalışmalar yapmışlardır. Eksik veri probleminde; kardiyak aritmi sınıflandırma [1], finansal dolandırıcılık tahmini [2], mikro dizi ve RNA dizileme verilerinin tahmini [3], ürün hatası tahmini [4], tıbbi teşhis [5], sınıf dengesizliği sorununda; duygu analizi [6], spam tespiti [7], hastalıkların tespiti [8, 9, 10, 11, 12, 13, 14], kredi kartı dolandırıcılık tespiti [15, 16], siber güvenlik [17] gibi çeşitli alanlarda literatürde yapılan çalışmalar örnek verilebilir.

Gerçek dünya verilerinden özellikle sağlık veri kümeleri, genellikle eksik değerlere sahip verilerden ve dengesiz veri sınıflarından oluşmaktadır [18]. Dengesiz veri kümesi sorunu, hedef hastalığa sahip bireylerin sayısı popülasyona göre çok az olduğunda ortaya çıkmaktadır. Veri kümelerindeki sınıfların dengesizliği nedeniyle makine öğrenmesi yöntemleri azınlık sınıfını göz ardı ederek çoğunluk sınıfına ayrılma eğilimi gösterdiğinden bu yöntemlerin performansını olumsuz yönde etkilemektedir. Veri kümesindeki kayıp değer sorunu genellikle bireylerin sağlık verilerinin toplanmasında talep edilen tüm bilgileri sağlamaması sonucu ortaya çıkmaktadır. Birçok veri kümesinde bazı eksik değerler bu şekilde açıkça temsil edilemeyebilir ve potansiyel olarak geçerli veri değerleri olarak görülebilir. Bu sorun, makine öğrenmesi yöntemlerinde yeterince yanlılığa ve yanıltıcı sonuçlara neden olacağından veri analizinin doğruluğunu ciddi şekilde etkilemektedir. Araştırmacıların ve hekimlerin dengesiz ve eksik değerlere sahip veri kümeleri için analiz yöntem ve araçlarının geliştirilmesine yönelik taleplerinin olduğu açıktır [18]. Literatürde eksik değer veya dengesiz veri kümeleri problemlerine yönelik birçok çalışma geliştirilmiştir. Kim ve Chung, kişisel sağlık kayıtlarında veri toplama ve işleme aşamalarında meydana gelen eksik verileri tahmin etmek için yığınlanmış bir gürültü giderme otomatik kodlayıcı yöntemini önermişlerdir [19]. Le, Beuran ve Tan, sağlık verilerinde eksik değerlerin doldurulmasında Düzenlileştirilmiş Beklenti-Maksimizasyon (EM), Çoklu Atama (MI), k-En Yakın Komşu Atama (kNNI) ve Ortalama Atama eksik veri atama algoritmalarının performansını iki gerçek sağlık hizmeti veri kümesinde karşılaştırmalı çalışmasını yapmışlardır [20]. Azimi ve arkadaşları, veri kümesindeki eksik değerleri tahmin etmek için IoT tabanlı sağlık hizmeti sistemlerde çeşitli veri kaynaklarıyla güçlendirilmiş Çoklu Atama yöntemi kullanılarak kişiselleştirilmiş bir eksik veri esnek karar verme yaklaşımı önermişlerdir [21]. Phung, Kumar ve Kim, sağlık verilerindeki eksik değerleri doldurmak için verilerin gizli temsillerini öğrenebilen bir derin öğrenme yöntemi önermişlerdir [22]. Xu ve arkadaşları, sağlık verilerinde eksik değer tahmini için verilerdeki çok yönlü bilgileri kullanan ağaç tabanlı bir yöntem önermişlerdir [23]. Zhao, Wong ve Tsui, dengesiz dağılıma sahip sağlık verilerinin analizi için sentetik azınlık aşırı örnekleme tekniği (SMOTE) ile lojistik regresyonu birleştiren bir öğrenme yöntemi önermişlerdir. Bu öğrenme sistemi temel sınıflandırıcıların seçilmesi ve yeniden dengeleme stratejilerinin değerlendirilmesi olmak üzere iki aşamadan oluşmaktadır. İlk aşamada, her bir sınıflandırıcının performans ölçütlerini değerlendirerek bir dizi aday arasından bir temel sınıflandırıcı seçilmiştir. İkinci aşamada ise yeniden dengeleme stratejileri temel sınıflandırıcı ile birleştirilmiştir [24]. Farhadi ve arkadaşları, meme kanseri verilerinde dengesiz veri sorunu ile başa çıkabilmek için bir derin aktarım öğrenme yöntemi önermişlerdir [25]. Tran, Le ve Shi, makine öğrenimi modellerinin tahmine dayalı performansını iyileştirmek ve meme kanserine ait dengesiz verileri işlemek için tasarlanmış bir yukarı örnekleme yöntemi (ENUS) önermişlerdir [26]. Lan ve arkadaşları, dengesiz verilerin işlenmesi için sentetik azınlık aşırı örnekleme üreten çekişmeli üretici ağları (SMOGANs) yöntemini önermişlerdir [27]. Sağlık veri kümelerinde dengesiz sınıfların dengelenmesi ve eksik değerlerin doldurulmasına yönelik incelenen çalışmaların özeti Çizelge-1'de verilmiştir.

Çizelge-1: Sağlık veri kümelerindeki çalışmaların özeti

Çalışma	Problem	Yöntem	Başarı
Kim ve Chung [19]	Eksik değer	Çok Modlu Yığılmış Gürültü Giderme Otomatik Kodlayıcı yönteminin önerilmesi	%93 Doğruluk
Le, Beuran ve Tan [20]	Eksik değer	Düzenlileştirilmiş EM, MI, kNNI ve Ortalama Atama algoritmalarının karşılaştırılması	EM yöntemi RMSE=3
Azimi ve arkadaşları [21]	Eksik değer	Kişiselleştirme yoluyla sağlık hizmeti IoT'si için eksik veriye dayanıklı karar verme sistemi	5,6 RMSE
Phung, Kumar ve Kim [22]	Eksik değer	Aşırı Tamamlanmış Gürültü Giderme Otomatik Kodlayıcı Tasarımı (ODAE)	0,00988 MSE %70,37 F1
Xu ve arkadaşları [23]	Eksik değer	MD-MTS	0,1717 RMSE
Zhao, Wong ve Tsui [24]	Dengesiz veri kümesi	Yeniden dengeleme sistemi	%75,7 Duyarlılık
Farhadi ve arkadaşları [25]	Dengesiz veri kümesi	Derin Transfer Öğrenimi (DTL)	0,81 AUC
Tran, Le ve Shi [26]	Dengesiz veri kümesi	ENUS	%97,5 Doğruluk %97,88 Kesinlik
Lan ve arkadaşları [27]	Dengesiz veri kümesi	SMOGAN	WBC veri kümesinde %94 Doğruluk

Literatürde yapılan çalışmalar genellikle verilerde eksik değerleri doldurmak ya da dengesiz sınıfları dengelemek için önerilmiştir. Ancak, eksik değerlerle birlikte sınıf dengesizliği problemi makine öğrenmesi yönteminin performansı üzerinde önemli bir etkiye sahiptir. Bu iki problemin birlikte değerlendirilmesi ve çözüme uygun tekniğin seçilmesi veri analizinde oldukça önemlidir. Bu makale, veri kümesinin sınıf dengesizliği ve eksik değer varlığında hem eksik değer teknikleri hem de yeniden örnekleme tekniklerinin farklı kombinasyonlarının makine öğrenmesi yöntemleri ile elde edilmiş farklı değerlendirme ölçütlerine dayalı karşılaştırmalı analizini içermektedir. Sağlık verilerinin analizinde ortaya çıkan eksik veriler ile dengesiz sınıf sorununu kapsamlı bir şekilde incelemek ve bu sorunlara yönelik literatürde sıklıkla kullanılan yöntemlere genel bakış sağlamak amacıyla bu çalışmanın literatüre katkı sağlaması amaçlanmaktadır.

Makalenin geri kalan kısmı şu şekilde organize edilmiştir: birinci bölümde, çalışmaya ve arkasındaki motivasyona bir giriş sunulmuştur. İkinci bölümde, veri kümesi tanıtımı, sınıf dağılımlarının dengesizliği, veri kümesini dengelemek için yeniden örnekleme yöntemleri, verilerdeki eksik değer sorunu, eksik değerleri doldurma yöntemleri ve sınıflandırma açıklanmaktadır. Üçüncü bölümde, deneysel sonuçlara yer verilmiş ve son bölümde, gelecekteki kapsamı ile makale sonlandırılmıştır.

2. Yöntem ve Materyal

Bu bölüm esas olarak dört bölüme ayrılmıştır. İlk olarak, bu çalışmada kullanılan veri kümesi tanıtılmıştır. Ardından, veri kümelerinde sınıf dengesizliği ve eksik değer sorunlarından bahsedilerek sınıfları dengelemek ve eksik değerleri doldurmak için kullanılan popüler yöntemler açıklanmıştır. Son olarak, veri analizinde sınıflandırma aşamasından bahsedilmiştir.

2.1 Veri Kümesi Tanıtımı

PIMA diyabet veri kümesi, 21 yaş ve üzeri PIMA Kızılderili kadın nüfusundan alınan sağlık bilgilerini içeren bir veri kümesidir. Bu veri kümesi Kaggle sitesinden indirilmiştir [28]. Veri kümesi 268 örneği diyabetik ve 500 örneği diyabetik olmayan toplam 768 vakaya sahip bir ikili sınıflandırma problemine aittir. Diyabet tahmininde bulunan toplam 8 özellik şu şekildedir: gebe olma sayısı, Vücut Kitle İndeksi (BMI), insülin seviyesi, yaş, kan basıncı, deri kalınlığı, glikoz, Diyabet Pedigre işlevi ve etiket sonucudur. Çizelge-2'de veri kümesi özellikleri ve açıklamaları verilmiştir.

Çizelge-2: PIMA diyabet veri kümesine genel bakış

Özellik	Açıklama	Veri Tipi	Aralık
Gebelik durumu	Hamile kalma sayısı	Sayısal	[0-17]
Kan şekeri	Oral glikoz tolerans testinde (GTIT) 2 saatte plazma glikoz konsantrasyonu	Sayısal	[0-199]
BP	Diyastolik Kan Basıncı (mm Hg)	Sayısal	[0-122]
Deri kalınlığı	Triseps deri kıvrım kalınlığı (mm)	Sayısal	[0-99]
İnsülin	2 Saatlik Serum insülini (µh/ml)	Sayısal	[0-846]
BMI	Vücut kitle indeksi (kg cinsinden ağırlık/m cinsinden boy)	Sayısal	[0-67,10]
Diyabet Pedigre İşlevi	Diyabet soyağacı işlevi	Sayısal	[0,078-2,42]
Yaş	Yaş	Sayısal	[21-81]
Etiket	Diyabetik durumunu gösteren ikili değer	Sayısal	[0-1]

2.2 Sınıf Dengesizliği

Sınıf dengesizliği, veri kümesindeki sınıfların eşit olmayan bir gözlem dağılımına sahip olduğunda ortaya çıkan bir durumdur. Bu durumun makine öğrenmesi yöntemlerinin genelleme performansını önemli ölçüde engellediği bilinmektedir. Sınıf dengesizliği sınıf dengesizlik oranı (DO) ile gösterilen bir veri kümesinde çoğunluk sınıf örnekleri (m) ile azınlık sınıf örnekleri (n) arasındaki oran ile temsil edilir (1).

$$\mathrm{DO} = \frac{m}{n} \tag{1}$$

Azınlık sınıfındaki gözlem sayısı azaldıkça sınıf dengesizlik oranı arttığından makine öğrenmesi yöntemlerinde düşük model başarımı, aşırı öğrenme veya eksik öğrenme gibi problemler meydana gelmektedir. Veri kümelerinde sınıfları dengelemek için literatürde eksik yeniden örnekleme, aşırı yeniden örnekleme, sentetik yeniden örnekleme, melez yeniden örnekleme gibi çeşitli yöntemler önerilmiştir [29, 30]. Eksik yeniden örnekleme yaklaşımında, çoğunluk sınıfına ait rastgele seçilen örnekler çıkarılarak sınıflardaki örnek sayıları dengelenmektedir. Aşırı yeniden örnekleme yaklaşımında ise, azınlık sınıfından rastgele örnekler seçilerek yeniden örnekleme ile sınıflardaki örnek sayıları dengelenmektedir. Sentetik yeniden örnekleme yaklaşımında mevcut veriler kullanılarak yeni örnekler üretilmektedir. Bu yaklaşımlardan eksik yeniden örnekleme yaklaşımlarına rastgele eksik örnekleme (RUS), Tomek bağlantıları ve NearMiss, aşırı yeniden örnekleme yaklaşımlarına rastgele aşırı örnekleme (ROS), sentetik yeniden örnekleme yaklaşımlarına SMOTE, Borderline-SMOTE, SVM-SMOTE, K-ortalama-SMOTE ve ADASYN, aşırı ve eksik yeniden örnekleme yöntemlerini birleştiren melez yeniden örnekleme yaklaşımlara ise SMOTEENN ve SMOTETomek yöntemleri örnek verilebilir.

2.2.1 Eksik Yeniden Örnekleme Yöntemleri

Eksik yeniden örnekleme yöntemleri, çoğunluk sınıfının bir alt kümesini kullanarak eğitim kümesini dengeler ve eğitim sürecini kısaltır. Eksik yeniden örnekleme algoritmalarından RUS yöntemi, çoğunluk sınıfının örneklerini veri kümesinden rastgele çıkararak sınıfları dengelemeyi amaçlar. Bu durumda çoğunluk sınıfından göz ardı edilen örneklerde yer alan faydalı bilgiler ihmal edilebilir ve sınıflandırıcının performansını olumsuz yönde etkileyebilir. Bu olumsuz durumu ortadan kaldırmak için Tomek bağlantıları gibi gelişmiş yöntemler kullanılabilir. Tomek bağlantıları, çoğunluk sınıfındaki gürültülü veriler olarak kabul edilen sınır örneklerini silerek sınıflar arası dengesizliği ortadan kaldırmaya yönelik geliştirilmiş bir eksik örnekleme yöntemidir [31]. NearMiss

yöntemi, veri kümesindeki sınıf dengesizliğini ortadan kaldırmak için kullanılan bir diğer eksik örnekleme yöntemidir. Bu yöntem, iki farklı sınıfa ait iki nokta birbirine çok yakın olduğunda çoğunluk sınıfındaki örnekleri rastgele silerek sınıf dağılımını dengelemeye çalışır.

2.2.2 Aşırı Yeniden Örnekleme Yöntemleri

Aşırı yeniden örnekleme yöntemleri, çoğunluk ve azınlık sınıfı örneklerinin sayıları arasında denge kurmak için azınlık sınıfını çoğaltır. En basit aşırı yeniden örnekleme yöntemi ROS yöntemidir. Bu yöntemde azınlık sınıfı örnekleri rastgele kopyalanarak çoğunluk sınıfı örneklerinin sayısı ile eşitlenir. Bu yöntemin dezavantajı çoğunluk sınıfının karar bölgelerini daha küçük ve özel hale getirebileceğinden sınıflandırıcının aşırı öğrenmesine neden olabilir [32].

2.2.3 Sentetik Yeniden Örnekleme Yöntemleri

Sentetik yeniden örnekleme yöntemleri, sınıflar arası dengesizlik sorununu çözmek için sentetik örneklerin oluşturulduğu bir aşırı yeniden örnekleme yaklaşımıdır. Aşırı yeniden örnekleme yöntemlerinin çoğu, azınlık sınıfının örneklerini kopyaladığından aşırı öğrenmeye neden olabilir ve dengesiz dağılıma sahip büyük veri kümelerinde hesaplama açısından çok maliyetli olabilir. Sentetik yeniden örnekleme yaklaşımları ise azınlık sınıfı verilerini kopyalamak yerine incelenen örneklerde keşifsel yöntemler kullanarak yapay örnekler üretir. Böylece aşırı öğrenme olasılığı daha düşük olmaktadır. Bu yaklaşımlara örnek olarak SMOTE ve türevleri verilebilir. SMOTE yönteminde, veri kümesindeki örneklerin tamamı girdi olarak alınır ve azınlık sınıfının en yakın komşuları belirlenir. Azınlık sınıfındaki örneklerin ve en yakın komşularını birleştiren doğru parçası boyunca yeni sentetik örnekler üretilir. Çoğunluk sınıfının örnek sayısı değiştirilmez ve sentetik örnekler sınıflandırıcının daha küçük ve daha özel bölgeler yerine daha büyük ve daha az özel karar bölgeleri oluşturmasını sağlar [33]. SMOTE algoritması, azınlık sınıfındaki her örnek için sentetik örnekler oluşturduğundan aşırı genellemeye yol açabilir. Bu sorunu çözmek için Borderline-SMOTE yöntemi önerilmiştir. Bu yöntem, karar sınırına yakın azınlık örnekleri aşırı yeniden örneklemede kullanarak sınıf dağılımı dengesizlik sorunu çözmeyi amaçlar [34]. Borderline-SMOTE algoritmasının bir alternatifi olan SVM-SMOTE algoritması k-en yakın komşu algoritması yerine SVM algoritmasını kullanarak karar sınırını belirler. Bu karar sınırı boyunca azınlık sınıfından rastgele yeni örnekler oluşturulur ve sınıflar arası dengesizlik ortadan kaldırılır [35]. Veri kümesindeki dengesizliği yeniden dengelemek için kullanılan bir diğer yaklaşım ise k-ortalama kümeleme algoritması ve SMOTE yöntemini birleştiren k-ortalama-SMOTE yöntemidir. Bu yöntemde gürültü oluşumu engellenerek sınıflar arası ve sınıf içi dengesizlik problemi önlenir [36]. Sentetik örneklerin oluşturulmasında çakışma sorununu çözmek için veri kümesindeki sınıflar arası dengesizliğin giderilmesinde uyarlamalı sentetik örnekleme olan ADASYN yöntemi kullanılabilir. Bu yöntemde, sentetik örneklerin oluşturulmasında azınlık sınıfındaki örneklerin tümünü kullanmak yerine bu örneklere önem derecesine göre ağırlıklandırma yapılır. Yüksek ağırlıklandırılmış örneklerden daha çok sentetik örnekler üretilerek sınıflar arası denge sağlanmış olur [34].

2.2.4 Melez Yeniden Örnekleme Yöntemleri

Melez yeniden örnekleme yöntemleri aşırı ve eksik örnekleme yöntemlerinin birleşimini içerir. Bu yöntemlerden SMOTEENN ve SMOTETomek yöntemlerinde veri kümesindeki sınıf dengesizliği problemi aşırı örnekleme yöntemi olan SMOTE ile çözülmektedir. SMOTE yöntemi ile oluşan örtüşmeyi azaltmak için ENN algoritması veya Tomek bağlantıları kullanılarak veri temizleme işlemi yapılmaktadır. SMOTEENN yönteminde her bir örneğin k-en yakın komşusu hesaplanır ve bu komşular çoğunluk sınıfı örneği sınıfından farklı ise bu örnekler ve k-en yakın komşular veri kümesinden çıkarılır [37]. SMOTETomek yönteminde, farklı sınıflardan ve birbirinin komşularından örnek olan bir çift gözlem seçilir [37]. Seçilen bu gözlemler Tomek bağlantıları olarak adlandırılır ve azınlık sınıfının gürültülü ve sınır bilgilerinin ortadan kaldırılmasında bu bağlantılar kullanılarak veri temizleme işlemi yapılır [38].

2.3 Veri Kümelerinde Eksik Değer Problemi

Bir veri kümesindeki eksik değerler, öznitelik değerlerinde çeşitli nedenlerden dolayı meydana gelen eksiklik anlamına gelmektedir. Eksik veriler çoğu araştırma alanında yaygın bir sorundur ve makine öğrenmesi yöntemlerinin başarımını etkilemektedir [39]. Modelin değerlendirilmesinden önce verilerdeki eksik değerlerin doldurulması oldukça önemlidir. Eksik veri kümesi sorununun en basit çözümü, bir veya daha fazla eksik değere sahip verilerin doğrudan kaldırıldığı silme yöntemidir. Bu yöntem yalnızca veri kümesi çok az miktarda eksik veri içerdiğinde uygulanabilir [40]. Ancak gerçek dünya verilerine ait veri kümeleri daha fazla eksik değerlere sahip olabilir ve bu durum veri kaybına neden olarak makine öğrenmesi yöntemlerinin genelleme başarımını etkileyebilir. Bu nedenle, literatürde önerilen eksik değer atama yöntemleri kullanılabilir. Bu yöntemlere ortalama, mod, medyan, regresyon gibi istatistiksel yöntemler, k-en yakın komşu gibi makine öğrenmesi yöntemleri ve zincirleme denklemlerle çoklu atama gibi karmaşık yöntemler örnek verilebilir. Eksik değerleri doldurma yöntemlerinden ortalama yaklaşımında, eksik değerler gözlenen tüm verilerde bu özelliğin ortalama değeri ile doldurulurken, mod yaklaşımında, eksik değerler tüm gözlenen verilerdeki en fazla frekansa sahip olan değer ile doldurulur [41]. Medyan yaklaşımında, eksik değerler gözlenen tüm verilerde bu özelliğin medyan değeri ile doldurulur [42]. Regresyon tabanlı atama yöntemleri için, öznitelikler arasındaki ilişkiler tahmin edilir ve ardından eksik öznitelik değerlerini tahmin etmek için regresyon katsayıları kullanılır [41]. Makine öğrenmesi yaklaşımlarından k-en yakın komşu yönteminde, verilerdeki bir eksik değerin doldurulması için kendisine en yakın k değişkenini arar ve tanımlanan komşularının gözlenen değerlerinin ağırlıklı bir ortalamasını hesaplayarak değer ataması yapılır [43]. Zincirleme denklemlerle çoklu atama yöntemi (MICE) başarılı bir şekilde kullanılan bir diğer eksik değer atama yöntemidir. Bu yöntemde veri kümesindeki eksik değerler yinelemeli bir dizi tahmine dayalı model aracılığıyla doldurulur. Her iterasyonda eksik değere sahip bir değişkenin değerini tahmin etmek için diğer değişkenlerin eksik olmayan

ve doldurulan değerlerini kullanarak atanır. Bu işlem tüm eksik değere sahip değişkenlerin değerini tahmin etmek için tekrarlanır [44].

Veri kümesindeki eksik değeri doldurmak ve dengesiz dağılıma sahip sınıfları dengelemek için kullanılan algoritmaların özeti Çizelge-3'te verilmiştir.

Çizelge-3: Veri ön işleme adımında uygulanan yöntemler

Veri Ön İşleme Adımları	Yöntemler	Örnek
Sınıf Dengeleme Algoritmaları	Eksik Yeniden Örnekleme Yöntemleri	RUS Tomek bağlantıları NearMiss
	Aşırı Yeniden Örnekleme Yöntemleri	ROS
	Sentetik Yeniden Örnekleme Yöntemleri	SMOTE Borderline-SMOTE SVM-SMOTE K-ortalama-SMOTE ADASYN
	Melez Yeniden Örnekleme Yöntemleri	SMOTEENN SMOTETomek
Eksik Değer Doldurma Algoritmaları	Basit Yöntemler	Satır silme algoritması
	İstatistiksel Yöntemler	Ortalama Mod Medyan Regresyon
	Makine Öğrenmesi Yöntemleri	kNN algoritması
	Karmaşık Yöntemler	Zincirleme denklemlerle çoklu atama yöntemi (MICE)

2.4 Sınıflandırma

Sınıflandırma, problem alanından örneklere bir sınıf etiketi atamayı öğrenen makine öğrenmesinin önemli bir parçasıdır. Sınıflandırmada örnekler eğitim kümesi ve test kümesi olmak üzere iki parçaya veya geçerleme kümesinin de dahil olduğu üç parçaya ayrılır. Eğitim kümesinde model kurma işlemi yapıldıktan sonra çapraz geçerleme gibi teknikler ile model genellemesi gerçekleştirilir ve test kümesi üzerinde modelin performansı ölçülür. Modelin performansının ölçülmesinde doğruluk, kesinlik, duyarlılık ve F1-skoru gibi metrikler kullanılabilir. Makine öğrenmesi sınıflandırma yöntemlerine Lojistik Regresyon (LR), Rastgele Orman (RF), Destek Vektör Makinesi (SVM), Karar Ağaçları (DTs) ve Naive Bayes (NB) gibi literatürde sıklıkla kullanılan algoritmalar örnek verilebilir.

2. 3. Deneysel Sonuçlar ve Tartışma

Bu çalışmada, Çizelge-2'deki PIMA diyabet veri kümesi kullanılarak veri kümesindeki sınıf dengesizliği ve eksik değer sorunları ele alınmıştır. Bu kapsamda yapılan ilk analizde sınıf dengesizliğini dengelemek ve eksik değerleri doldurmak için herhangi bir yöntem uygulanmadan veri kümesi incelenmiştir. İkinci analizde, veri kümesinde sınıf dengesizliği problemi için literatürde sıklıkla kullanılan yöntemler kullanılarak veri kümesi incelenmiştir. Üçüncü analizde, veri kümesindeki sınıf dengesizliğini dengelemek ve eksik değerleri doldurmak için literatürde sıklıkla kullanılan yöntemler bir arada kullanılarak veri kümesi incelenmiştir. Son analizde ise sınıf dengesizliği için hem aşırı hem de eksik örnekleme yaklaşımlarını birleştiren melez yöntemler ile veri kümesi incelenmiştir. Bu kapsamda tasarlanan tüm deneylerde veri kümesi eğitim ve test kümesi (test boyutu=0,33) olarak ikiye ayrılmıştır. Bu çalışmada kullanılan model, veri ön işleme aşamasında normalizasyon işlemi ve sınıflandırma aşamasında sınıflandırıcıların uygulanması için bir veri hattı olarak tasarlanmıştır. Normalizasyon aşamasında verileri 0 ve 1 arasında yeniden ölçekleyen Minimum-Maksimum normalizasyonu (MinMaxScaler), veri kümesindeki tüm değişkenlerin ortalamasının sıfıra ve standart sapmasının ise bire eşitlendiği Standart ölçekleme (StandartScaler), maksimum mutlak değer 1 olacak şekilde giriş değerlerini ölçekleyen Maksimum mutlak ölçekleme (MaxAbsScaler) ve en az bir sıfır olmayan bileşene sahip her örnek normu (l1 veya l2) bire eşit olacak şekilde diğer örneklerden bağımsız olarak yeniden ölçeklendirildiği Normalleştirici (Normalizer) yöntemleri seçilmiştir. Veri kümesindeki bireylerin hasta veya sağlıklı olarak sınıflandırılmasında LR, RF, SVM, DTs ve NB algoritmaları sınıflandırıcı olarak kullanılmıştır. Deneylerde en iyi performansı gösteren veri normalizasyon fonksiyonunun ve sınıflandırıcının seçilmesi için parametre uzayı ile çapraz doğrulama yöntemi kullanılmıştır. Bu yöntemde modelde denenmesi istenen normalizasyon fonksiyonları ile sınıflandırıcıların hiper-parametrelerin değerlerinin bütün kombinasyonları için ayrı ayrı model kurulur ve belirtilen model değerlendirme metriğine göre en başarılı hiper-parametre kümesi belirlenir. Tüm deneylerde sınıflandırıcılar için kullanılan parametre uzayı Çizelge-4' te verilmiştir. Parametre optimizasyonunda sınıf dengesini alt kümelerde de görmek için çapraz doğrulama yönteminden katmanlı alt örnekleme yöntemi (StratifiedKFold) kullanılmıştır. Katmanlı çapraz doğrulama yöntemindeki katman değeri 5 olarak belirlenerek eğitim kümesi 5 bölüme ayrılmış ve her bir bölümde sınıflandırıcıların doğruluk metriğine göre performansları ölçülmüştür. En başarılı olan bölümün parametre değerleri tüm eğitim kümesine uygulanarak nihai eğitim modelinin başarımı elde edilmiştir. Test kümesi üzerinde bu parametre değerleri ile sınıflandırıcıların doğruluk, kesinlik, duyarlılık ve F1-skoru metriklerinin sonuçları elde edilerek başarımları karşılaştırılmıştır.

Bu çalışmada, tüm deneyler AMD Ryzen 7 5700U CPU @ 1.80 GHz (8 çekirdek), 8 GB RAM ve 64-bit Windows 10 Pro işletim sistemine sahip dizüstü bilgisayarda gerçekleştirilmiştir. Deneylerin gerçekleştirilmesinde Python dili ve Scikit-learn kütüphanesinden yararlanılmıştır. Veri analizinin ön işleme adımında kullanılan normalizasyon fonksiyonunun seçilmesinde ve sınıflandırma aşamasında sınıflandırıcıların parametre optimizasyonunda en iyi parametrelerinin belirlenmesinde bu kütüphanede yer alan GridSearchCV fonksiyonu kullanılmıştır.

Çizelge-4: Sınıflandırıcıların Parametre Uzayı

Sınıflandıcı	Parametreler
LR	penalty: ['l1', 'l2', 'elasticnet', 'none'] C : [0.001, 0.01, 0.1, 1.0, 10, 100] class_weight): ['dict','balanced', 'None'] random_state: [42, 50, 100] solver :['newton-cg', 'lbfgs', 'liblinear', 'sag', 'saga2'] max_iter: [10, 50, 100]
RF	n_estimators: [10, 30, 50, 100] criterion: ['gini', 'entropy'] max_features: ['auto', 'sqrt', 'log2'] random_state: [42, 50, 100] class_weight: ['balanced', 'balanced_subsample', 'dict', 'None'] bootstrap: [0,1] max_depth:[2,3]
SVM	C: [0.1, 1, 10, 100, 1000] gamma: ['scale', 'auto',1, 0.1, 0.01, 0.001, 0.0001] random_state: [42, 50, 100] kernel: ['linear','rbf'] class_weight: ['dict' , 'balanced', 'None'] decision_function_shape:['ovo', 'ovr'] degree: [2, 3, 5]
DTs	criterion: ['gini', 'entropy'] splitter: ['best', 'random'] max_features: ['auto', 'sqrt', 'log2', 'None'] random_state:[42, 50, 100] class_weight:['dict', 'balanced', 'None'] ccp_alpha:[0.0, 0.005, 0.010, 0.015, 0.030, 0.035] max_depth:[1, 2, 3, 4, 5, 6, 7, 'None']
NB	GaussianNB() Varsayılan parametreler kullanılmıştır.

3.1 Veri Kümesinin Analizi

Bu bölümde, veri analizinde önemli bir aşama olan veri ön işleme adımına ait sınıf dengesizliğinin ortadan kaldırılması ve eksik değerlerin doldurulmasına yönelik hiçbir yöntem uygulanmadan bireylerin hasta veya sağlıklı olarak sınıflandırılması gerçekleştirilmiştir. Sınıflandırıcılar için en iyi performansı gösteren parametre değerleri Çizelge-4'te verilen parametre uzayı kullanılarak belirlenmiştir. Test kümesi üzerinde sınıflandırıcıların seçilen en iyi parametre değerleri ile doğruluk, kesinlik, duyarlılık ve F1-skoru metriklerinin sonuçları Çizelge-5'te verilmiştir.

Çizelge-5: Orijinal veri kümesi üzerinde modellerin performansları

Algoritma	Doğruluk	Kesinlik	Duyarlılık	F1-skor
LR	%74	%62	%59	%61
RF	**%74**	**%59**	**%77**	**%67**
SVM	%71	%58	%52	%55
DTs	%71	%55	%78	%64
NB	%74	%61	%63	%62

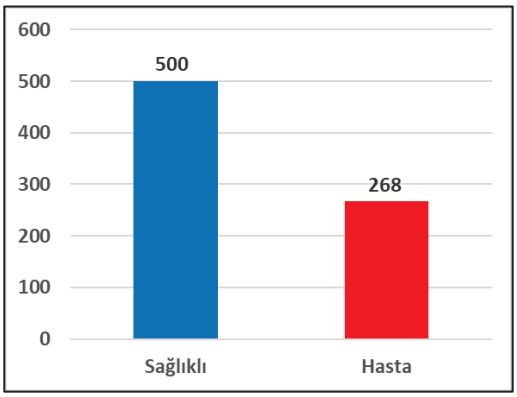
Hastalıkların tespitinde sınıflandırma algoritmalarının amacı hasta sınıfını ayırt etmektir. Veri kümesinde dengesiz bir dağılım bulunduğunda sınıflandırıcı genel doğruluğu maksimize etmek için veri sayısının çoğunlukta olduğu sınıfı daha iyi öğrenirken azınlık sınıfını iyi öğrenemez. Deneysel sonuçlara göre, veri kümesinin ön işleme adımında Standart ölçekleme fonksiyonu ile sınıflandırma adımında Rastgele Orman algoritmasının kullanılması en iyi modeli oluşturmuştur. Parametre optimizasyonu sonucu RF sınıflandırıcısının en iyi parametre değerleri şu şekildedir:

Rastgele örneklem seçimi (bootstrap): 1,
Sınıf ağırlığı (class_weight): balanced_subsample,
Bölünmenin kalitesini ölçme işlevi kriteri (criterion): gini,
Ağacın maksimum derinliği (max_depth): 3,
Maksimum özellik sayısı (max_features):log2,
Ormandaki ağaç sayısı (n_estimators) : 100,
Örneklerin önyükleme rastgelelik değeri (random_state): 42

Bu modelde kesinlik değerinin (%59) düşük olması modelin hasta olan bireyleri sağlıklı olarak tahmin ettiğini belirtmektedir. Bu durum hastaların tedavisini olumsuz yönde etkilemekte ve hastalığın ilerlemesine neden olmaktadır. Duyarlılık değerinin (%77) yüksek çıkması sağlıklı bireyleri hasta olarak tahminin az olduğunu göstermektedir. Doğruluk oranı (%74) yüksek olmasına rağmen kesinlik ve duyarlılık metriklerinin harmonik ortalaması olan F1-skoru (%67) oldukça düşüktür. Veri kümesindeki sınıf dengesizlik oranı 1,87 olduğundan veri kümesindeki sınıflar dengesiz bir dağılıma sahiptir. Bu durumda, F1-skor değerini değerlendirmek modelde uç değerleri göz ardı etmemizi engeller ve dengesiz dağılıma sahip bir veri kümesinde doğru bir model seçimi yapmamızı sağlamaktadır. Modelde doğruluk başarımının iyi ve diğer metrik sonuçlarının düşük olması sınıflandırıcıların azınlık sınıfını iyi öğrenemediğini göstermektedir. Ayrıca, veri kümesinde eksik değerlerin yer alması sınıflandırıcının genelleme performansını olumsuz yönde etkilediği elde edilen sonuçlar arasındadır.

3.2 Sınıf Dengesizliği Probleminin Analizi

Bu bölümde, veri kümesindeki sınıf dengesizliği sorunu ele alınmıştır. PIMA diyabet veri kümesinin sınıf dağılımı Şekil-1'de gösterilmiştir.



Şekil-1: Veri kümesinin sınıf dağılımı

Veri kümesindeki eksik değerler için herhangi bir yöntem uygulanmadan sadece sınıf dengesizliğini dengelemek için literatürde sıklıkla kullanılan yöntemler PIMA diyabet veri kümesi üzerinde uygulanarak bireylerin hasta veya sağlıklı olarak sınıflandırılması gerçekleştirilmiştir. Sınıflandırıcılar için en iyi performansı gösteren parametre değerleri Çizelge-4'te

verilen parametre uzayı kullanılarak belirlenmiştir. Test kümesi üzerinde sınıflandırıcıların seçilen en iyi parametre değerleri ile doğruluk, kesinlik, duyarlılık ve F1-skoru metriklerinin sonuçları Çizelge-6 - Çizelge-14'te verilmiştir.

Çizelge-6: Rastgele eksik örnekleme dengeleme algoritmasının analizi

Algoritma	Doğruluk	Kesinlik	Duyarlılık	F1-skor
LR	%75	%77	%72	%74
RF	**%74**	**%71**	**%82**	**%76**
SVM	%72	%72	%71	%72
DTs	%66	%63	%76	%69
NB	%71	%74	%66	%70

Çizelge-7: Eksik örnekleme Tomek bağlantıları dengeleme algoritmasının analizi

Algoritma	Doğruluk	Kesinlik	Duyarlılık	F1-skor
LR	%73	%60	%80	%69
RF	**%76**	**%62**	**%88**	**%73**
SVM	%73	%61	%77	%68
DTs	%75	%66	%66	%66
NB	%72	%61	%66	%63

Çizelge-8: NearMiss dengeleme algoritmasının analizi

Algoritma	Doğruluk	Kesinlik	Duyarlılık	F1-skor
LR	%72	%74	%69	%71
RF	**%72**	**%73**	**%69**	**%71**
SVM	%68	%70	%65	%67
DTs	%64	%67	%58	%62
NB	%72	%74	%67	%71

Çizelge-9: Rastgele aşırı örnekleme dengeleme algoritmasının analizi

Algoritma	Doğruluk	Kesinlik	Duyarlılık	F1-skor
LR	%75	%75	%71	%73
RF	**%75**	**%70**	**%83**	**%76**
SVM	%76	%74	%79	%76
DTs	%73	%69	%82	%75
NB	%74	%73	%73	%73

Çizelge-10: ADASYN dengeleme algoritmasının analizi

Algoritma	Doğruluk	Kesinlik	Duyarlılık	F1-skor
LR	**%74**	**%67**	**%87**	**%76**
RF	**%**71	**%**67	**%**77	**%**72
SVM	**%**76	**%**73	**%**77	**%**75
DTs	**%**69	**%**65	**%**73	**%**69
NB	**%**73	**%**74	**%**69	**%**71

Çizelge-11: SMOTE dengeleme algoritmasının analizi

Algoritma	Doğruluk	Kesinlik	Duyarlılık	F1-skor
LR	%76	%75	%76	%76
RF	%73	%70	%76	%73
SVM	**%78**	**%75**	**%81**	**%78**
DTs	%71	%67	%78	%72
NB	%75	%76	%72	%74

Çizelge-12: Borderline-SMOTE dengeleme algoritmasının analizi

Algoritma	Doğruluk	Kesinlik	Duyarlılık	F1-skor
LR	%71	%68	%76	%72
RF	**%72**	**%70**	**%74**	**%72**
SVM	%72	%69	%75	%72
DTs	%70	%66	%78	%72
NB	%71	%70	%69	%70

Çizelge-13: SVM-SMOTE dengeleme algoritmasının analizi

Algoritma	Doğruluk	Kesinlik	Duyarlılık	F1-skor
LR	%77	%77	%74	%76
RF	%75	%70	%85	%77
SVM	**%77**	**%73**	**%84**	**%78**
DTs	%71	%65	%86	%74
NB	%75	%75	%74	%74

Çizelge-14: k-ortalama-SMOTE dengeleme algoritmasının analizi

Algoritma	Doğruluk	Kesinlik	Duyarlılık	F1-skor
LR	%79	%78	%79	%78
RF	**%80**	**%79**	**%81**	**%80**
SVM	%79	%79	%77	%78
DTs	%79	%77	%81	%79
NB	%79	%77	%80	%79

Veri kümesindeki sınıfları dengelemek için uygulanan yöntemler Çizelge-6 ila Çizelge-14 incelendiğinde bir önceki bölümde analiz edilen sonuçlara göre tüm sınıflandırıcıların F1-skoru %70'in üzerine çıktığı ve veri kümesindeki sınıfların dengelenmesinin veri analizinde önemli bir adım olduğu gözlemlenmektedir. Çizelge-6, Çizelge-7 ve Çizelge-8'de uygulanan eksik yeniden örnekleme yöntemlerinin sonuçları değerlendirildiğinde kesinlik, duyarlılık ve F1-skorunun düşük olması ve dolayısıyla modellerin sınıf etiketlerini yeterince doğru tahmin edemediği elde edilen sonuçlar arasındadır. Bu yöntemler ile sınıflar dengelediğinde çoğunluk sınıfından faydalı olabilecek örneklerin sınıftan çıkarılabilme olasılığı olduğundan model öğrenmesi zayıf kalabilmektedir. Veri kümesinin küçük boyutlu olduğu durumlarda önerilmiş olan aşırı yeniden örnekleme yöntemlerinden Çizelge-9'da verilen ROS'un performansı analiz edildiğinde kesinlik, duyarlılık ve F1-skoru sonuçlarının eksik yeniden örnekleme yöntemlerinden daha başarılı olduğu görülmüştür. Bu yöntemde, azınlık sınıfı örnekleri çoğunluk sınıfı örnekleri ile eşit sayıda olana kadar örnekler rastgele kopyalandığından sınıf içinde çok sayıda aynı tip örnekler yer alabilmektedir ve modelin genelleme performansı olumsuz yönde etkilenmektedir. Bu yöntemin dezavantajın giderilmesine yönelik sentetik yeniden örnekleme yöntemleri önerilmiştir. Çizelge-10 ila Çizelge-14'te bu yöntemlerin performansları sınıflandırıcılar üzerinde ölçülmüştür. Elde edilen sonuçlara göre sentetik yeniden örnekleme yöntemi SMOTE ve türevlerinin klasik yöntemlere göre daha başarılı sonuçlar verdiği gözlemlenmiştir. Veri analizinin ön işleme adımında Standart ölçekleme ve sınıflandırma adımında Rastgele Orman algoritmasının kullanılması ile Çizelge-14'te verilen k-ortalama-SMOTE yöntemi en yüksek model değerlendirme sonuçlarını Kesinlik (%80), Duyarlılık (%81) ve F1-skor (%80) olarak elde etmiştir. Parametre optimizasyonu sonucu RF sınıflandırıcısının en iyi parametre değerleri şu şekildedir:

Rastgele örneklem seçimi (bootstrap): 1,
Sınıf ağırlığı (class_weight): balanced_subsample,
Bölünmenin kalitesini ölçme işlevi kriteri (criterion): entropy,

Ağacın maksimum derinliği (max_depth): 3,
Maksimum özellik sayısı (max_features): auto,
Ormandaki ağaç sayısı (n_estimators) : 100,
Örneklerin önyükleme rastgelelik değeri (random_state): 42

Değerlendirme ölçütlerine göre sonuçlar değerlendirildiğinde modelin hasta ve sağlıklı bireyleri sınıflandırmasında oldukça başarılı olduğu görülmektedir.

3.3 Verilerdeki Eksik Değer Probleminin Analizi

Bu bölümde, veri analizinin önemli bir aşaması olan veri temizleme işleminde veri kümesindeki eksik değerler tespit edilmiş ve Çizelge-15'te gösterilmiştir.

Çizelge-15: Veri kümesindeki özelliklerin eksik değer miktarı

Özellik	Eksik Değer Miktarı
Gebelik durumu	0
Kan şekeri	5
BP	35
Deri kalınlığı	227
İnsülin	374
BMI	11
DiyabetPedigre İşlevi	0
Yaş	0
Etiket	0

PIMA diyabet veri kümesindeki eksik değerlerin doldurulmasında literatürde sıklıkla kullanılan eksik değerlere sahip satırları silme, ortalama değer ile doldurma, Zincirleme Denklemlerle Çoklu Atama (MICE) ve k-en yakın komşu (kNN) algoritmaları kullanılmıştır. Veri kümesindeki sınıf dengesizliğini dengelemek için SMOTE algoritması ile veri kümesi dengelenmiştir. Sınıflandırıcılar için en iyi performansı gösteren parametre değerlerinin belirlenmesinde Çizelge-4'te verilen parametre uzayı kullanılmıştır. Test kümesi üzerinde sınıflandırıcıların seçilen en iyi parametre değerleri ile doğruluk, kesinlik, duyarlılık ve F1-skoru metriklerinin sonuçları Çizelge-16-Çizelge-19'da verilmiştir.

Çizelge-16: SMOTE dengeleme ve satır silme algoritmalarının analizi

Algoritma	Doğruluk	Kesinlik	Duyarlılık	F1-skor
LR	%73	%79	%69	%74
RF	**%82**	**%89**	**%77**	**%82**
SVM	%80	%86	%76	%80
DTs	%75	%82	%69	%75
NB	%71	%81	%62	%70

Çizelge-17: SMOTE dengeleme ve ortalama değer ile doldurma algoritmalarının analizi

Algoritma	Doğruluk	Kesinlik	Duyarlılık	F1-skor
LR	%75	%72	%80	%76
RF	**%76**	**%73**	**%82**	**%77**
SVM	%77	%74	%79	%77
DTs	%69	%67	%72	%69
NB	%71	%71	%66	%68

Çizelge-18: SMOTE dengeleme ve MICE algoritmalarının analizi

Algoritma	Doğruluk	Kesinlik	Duyarlılık	F1-skor
LR	**%78**	**%72**	**%91**	**%80**
RF	%77	%72	%84	%78
SVM	%76	%73	%81	%77
DTs	%71	%72	%68	%69
NB	%76	%76	%76	%76

Çizelge-19: SMOTE dengeleme ve kNN algoritmalarının analizi

Algoritma	Doğruluk	Kesinlik	Duyarlılık	F1-skor
LR	**%75**	**%69**	**%89**	**%78**
RF	%75	%70	%82	%76
SVM	%76	%72	%83	%77
DTs	%72	%69	%75	%72
NB	%74	%73	%74	%74

Veri kümesindeki sınıfların dengelenmesi ve eksik değerleri doldurmak için uygulanan yöntemlerin performansları Çizelge-16 ila Çizelge-19 incelendiğinde veri kümesindeki eksik değerlerin doldurulması sonucu bir önceki bölümdeki sınıflandırıcı performanslarında artışlar olduğu gözlemlenmiştir. Çizelge-16'da uygulanan satır silme yöntemi ile eksik değerlerin giderilmesi Rastgele Orman sınıflandırıcısının hasta olan bireyleri doğru bir şekilde tahmin ederek kesinlik (%89) sonucunu artırırken, duyarlılık (%77) sonucu ile sağlıklı olan bireylerin tahmininde azalmalar görülmüştür. Bu yöntem, uygulaması basit olmasına karşın veri kümesi boyutunu azalttığı ve standart hatayı artırdığından modelin tahmininde yanlı sonuçlara neden olabilmektedir. Veri kaybı probleminin giderilmesi için Çizelge-17'de ortalama değer ile eksik değerlerin doldurulması gerçekleştirilmiştir ve modelin kesinlik (%73) değeri düşerek hasta bireylerin tahmini azalmıştır. Bu yöntemde, eksik değerlere sabit bir değerin eklenmesi değişkenin varyans değerini düşürdüğünden modelin genelleme performansını etkilemektedir. Eksik değerlerin doldurulmasında eksik değere sahip değişkene en yakın k adet değişkenin mesafesine göre atama yapan kNN yönteminin değerlendirme ölçütlerine göre sonuçları Çizelge-19'da verilmiştir. Bu yöntem, kesinlik (%69) ile hasta bireylerin tahmininde hatayı artırmış ve modelin başarımını azaltmıştır. Veri kümesindeki gözlemlenen tüm değerleri kullanarak eksik değerlerin doldurulması yöntemi olan MICE algoritması ile model değerlendirme metriklerinin başarımlarında artışlar olduğu Çizelge-18'de gözlemlenmiştir. Veri analizinin ön işleme adımında Min-Max ölçekleme ve sınıflandırma aşamasında LR algoritmasının kullanılması sonucunda Duyarlılık (%91) oranı ile sağlıklı bireylerin tahmininde artışlar gözlemlenmiştir. Parametre optimizasyonu sonucu LR sınıflandırıcısının en iyi parametre değerleri şu şekildedir:

Düzenlileştirme değeri (C): 0.01,
Sınıf ağırlığı (class_weight): dict,
Maksimum iterasyon sayısı (max_iter): 10,
Ceza norm terimi (penalty): l2,
Verileri karıştırmak için rastgele değer (random_state): 100,
Optimizasyon algoritması (solver): sag

Bu yöntem ile silme ve tekli atama yöntemlerinin getirdiği dezavantajlar ortadan kaldırılmıştır. Ayrıca, eksik değerlere atama yaparken veri kümesindeki eksik veriden kaynaklanan belirsizliği de dikkate aldığı için diğer yöntemlere göre avantajlı olduğu gözlemlenmiştir.

3.4 Melez Yaklaşımların Performans Analizi

Bu bölümde, PIMA diyabet veri kümesindeki sınıfları dengelemek için aşırı örnekleme ve eksik örnekleme yöntemlerinin kombinasyonlarından SMOTEENN ve SMOTETomek algoritmalarının performansı ölçülmüştür. Veri kümesindeki eksik değerler önceki analizde başarılı olan MICE algoritması ile doldurulmuştur. Sınıflandırıcılar için en iyi performansı gösteren parametre değerlerinin belirlenmesinde Çizelge-4'te verilen parametre uzayı kullanılmıştır. Test kümesi üzerinde sınıflandırıcıların seçilen en iyi parametre değerleri ile doğruluk, kesinlik, duyarlılık ve F1-skoru metriklerinin sonuçları Çizelge-20 ve Çizelge-21'de verilmiştir.

Çizelge-20: SMOTEENN dengeleme ve MICE algoritmalarının analizi

Algoritma	Doğruluk	Kesinlik	Duyarlılık	F1-skor
LR	%86	%87	%89	%88
RF	%89	%90	%90	%90
SVM	**%90**	**%91**	**%91**	**%91**
DTs	%81	%78	%89	%83
NB	%89	%90	%81	%85

Çizelge-21: SMOTETomek dengeleme ve MICE algoritmalarının analizi

Algoritma	Doğruluk	Kesinlik	Duyarlılık	F1-skor
LR	%77	%76	%79	%78
RF	**%81**	**%79**	**%87**	**%82**
SVM	%80	%78	%84	%81
DTs	%77	%76	%80	%78
NB	%73	%75	%71	%73

Deneysel sonuçlar incelendiğinde azınlık sınıfı için sentetik veri üretilmesi ve çoğunluk sınıfı örneklerinin azaltılması yöntemine dayanan melez yöntemler ile hasta ve sağlıklı bireylerin sınıflandırılmasında daha başarılı model değerlendirme metriklerinin elde edildiği görülmektedir. Elde edilen sonuçlara göre veri analizinin ön işleme adımında Maksimum mutlak ölçekleme ve sınıflandırma aşamasında SVM algoritmasının kullanılması ile SMOTEENN yönteminin en başarılı değerlendirme metrik sonuçlarını elde ettiği gözlemlenmiştir. Parametre optimizasyonu sonucu SVM sınıflandırıcısının en iyi parametre değerleri şu şekildedir:

Düzenlileştirme parametresi (C): 1000,
Sınıf ağırlığı (class_weight): balanced,
Karar fonksiyon yapısı (decision_function_shape): ovo,
'poly' fonksiyonunun derecesi (degree): 2,
'rbf', 'poly' ve 'sigmoid' için çekirdek katsayısı (gamma): 0,1,
Çekirdek türü (kernel): rbf,
Verileri karıştırmak için rastgele sayı (random_state): 42

Model değerlendirme ölçütlerinin sonuçları bu algoritmanın azınlık sınıfını iyi öğrendiğini ve azınlık sınıfının etiketini doğru bir şekilde tahmin ettiğini göstermektedir. Böylelikle, farklı veri üretme yöntemleri ile azınlık sınıfı için sentetik veri üretilmesi sağlanarak çeşitli veri azaltma yöntemleri ile gürültülü verilerin veri kümesinden çıkartılması ile her iki yöntemin avantajlarından faydalanılabileceği elde edilen sonuçlar arasındadır.

3. Sonuç

Bu çalışmada, eksik değerlere ve sınıf dengesizliğine sahip PIMA diyabet veri kümesi kullanılarak veri kümesinin dengelenmesi ve eksik değerlerin doldurulması için literatürde sıklıkla kullanılan yöntemlerin sonuçları ile orijinal veri kümesinin sonuçlarının karşılaştırmalı bir analizi yapılmıştır. Veri kümesinin dengelenmesi için rastgele eksik örnekleme, rastgele aşırı örnekleme, Tomek bağlantıları, NearMiss, ADASYN, SMOTE, Borderline- SMOTE, SVM-SMOTE ve k-ortalama-SMOTE yöntemleri kullanılmıştır. Veri kümesindeki eksik değerlerin doldurulması için satır silme, ortalama değer ile doldurma, MICE ve kNN yöntemleri kullanılmıştır. Veri kümesindeki bireylerin hasta veya sağlıklı olarak sınıflandırılması için Lojistik Regresyon, Rastgele Orman, SVM, Karar Ağaçları ve Naive Bayes algoritmaları seçilmiştir.

Deneysel çalışmalarda, orijinal ve yeniden örneklenen veri kümesinin performans değerleri karşılaştırılmış ve yeniden örneklenen veri kümesi hemen hemen tüm model değerlendirme metriklerinde daha başarılı olmuştur. Ayrıca, SMOTE ve türevleri ile yeniden örneklenen veri kümesi diğer klasik örnekleme yöntemlerine göre daha yüksek F1-skoru ile azınlık sınıfındaki örnekleri ayırt etmede oldukça başarılı olduğu görülmüştür. Sınıf dengesizliğinin giderilmesinde hem eksik hem de aşırı örnekleme yaklaşımlarını birleştiren melez yöntemlerden SMOTEENN ve SMOTETomek yöntemlerinin sınıflandırma başarımını artırdığı elde edilen sonuçlar arasındadır. Veri kümesindeki eksik değerlerin doldurulmasında çoklu veri atama prensipleri ile çalışan gelişmiş bir veri atama yöntemi olan MICE'nin diğer atama yöntemlerine göre daha başarılı olduğu tespit edilmiştir. SMOTEENN ve MICE yöntemlerin sağlık veri kümelerinin ön işleme adımında kullanımı ile daha tutarlı makine öğrenimi modellerinin elde edileceği gözlemlenmiştir. Gelecek çalışma olarak veri kümelerindeki dengesiz sınıfları dengelemek için SMOTE yönteminin geliştirilmesi hedeflenmektedir.

4. Kaynakça

[1] Fei Y., Jiazhi D., Jiying L., Weigang L., Lei Liu, Changlong Jin, and Qinma Kang. Missing value estimation methods research for arrhythmia classification using the modified kernel difference-weighted knn algorithms. BioMed research internati- onal, 2020, 2020.

[2] Ching-Hsue C., Yung-Fu K., ve Hsien-Ping L.. A financial statement fraud model based on synthesized attribute selection and a dataset with missing values and imbalanced classes. Applied Soft Computing, 108:107487, 2021.

[3] Saskya M. S., Titin S., Yoel F., Devvi S., Her-ley Shaori A., Sarah S., ve Noval S., Iterative bicluster-based bayesian principal component analysis and least squares for missing-value imputation in microarray and rna-sequencing data. Mathematical Biosciences and Engineering, 19(9):8741–8759, 2022.

[4] Seokho K. Product failure prediction with missing data using graph neural net- works. Neural computing and applications, 33(12):7225–7234, 2021.

[5] Mingjing W. ve Huiling C. Chaotic multi-swarm whale optimizer boosted support vector machine for medical diagnosis. Applied Soft Computing, 88:105946, 2020.

[6] Nizam H. Ve Saliha S. A.. Sosyal medyada makine öğrenmesi ile duygu analizinde dengeli ve dengesiz veri setlerinin performanslarının karşılaştırılması. XIX. Türkiye'de İnternet Konferansı, 1(6), 2014.

[7] Chaoliang L. and Shigang L.. A comparative study of the class imbalance problem in twitter spam detection. Concurrency and Computation: Practice and Experience, 30(5):e4281, 2018.

[8] Jinyan L., Lian-sheng L., Simon F., Raymond K W., Sabah M., Jinan F., Yunsick S., ve Kelvin KL W., Adaptive swarm balancing algorithms for rare-event prediction in imbalanced healthcare data. PloS one, 12(7):e0180830, 2017.

[9] Koichi F., Yukun H., Kentaro H., Kenichi N., Masao K., Mai K., ve Manabu K.. Over-and under-sampling approach for extremely imbalanced and small minority data problem in health record analysis. Frontiers in Public Health, 8, 2020.

[10] Vanaja R. ve Saswati M., An effective clinical decision support system using swarm intelligence. The Journal of Supercomputing, 76(9):6599–6618, 2020.

[11] Tince E. T. ve Aina M.. The implementation of genetic algorithm in smote (synthetic minority oversampling technique) for handling imbalanced dataset problem. In 2018 4th international conference on science and technology (ICST), pages 1–4. IEEE, 2018.

[12] Apurva S., Ruhi P., ve Nitin P., A new approach for handling imba- lanced dataset using ann and genetic algorithm. In 2016 International Conference on Communication and Signal Processing (ICCSP), pages 1987–1990. IEEE, 2016.

[13] Everlandio RQ F., Carvalho A., ve Xin Y.. Ensemble of classifiers based on multiobjective genetic sampling for imbalanced data. IEEE Tran- sactions on Knowledge and Data Engineering, 32(6):1104–1115, 2019.

[14] Chakraborty A., Kushal K. G., Rajonya De, E. C., ve Ram S.- kar. Learning automata based particle swarm optimization for solving class imbalance problem. Applied Soft Computing, page 107959, 2021.

[15] Wei W., Jinjiu L., Longbing C., Yuming O., ve Jiahang C., Effective detection of sophisticated online banking fraud on extremely imbalanced data. World Wide Web, 16(4):449–475, 2013.

[16] Dal Pozzolo A, Caelen O., Borgne Y. L, Waterschoot S., ve Bontempi G., Learned lessons in credit card fraud detection from a practitioner perspective. Expert systems with applications, 41(10):4915–4928, 2014.

[17] Sikha B. ve Kunqi L., Resampling imbalanced data for network intrusion detec- tion datasets. Journal of Big Data, 8(1):1–41, 2021.

[18] Nizam Ozogur H. and Orman Z., The effect of heuristic methods toward performance of health data analysis. Next Generation Healthcare Informatics, page 147.

[19] Joo-Chang K. ve Kyungyong C., Multi-modal stacked denoising autoencoder for handling missing data in healthcare big data. IEEE Access, 8:104933–104943, 2020.

[20] Tan D. L., Razvan B., ve Yasuo T., Comparison of the most influential mis- sing data imputation algorithms for healthcare. In 2018 10th International Conference on Knowledge and Systems Engineering (KSE), pages 247–251. IEEE, 2018.

[21] Iman A., Tapio P., Amir M R., Hannakaisa N.V., Anna A.L.,ve Pasi L., Missing data resilient decision-making for healthcare iot thro- ugh personalization: A case study on maternal health. Future Generation Computer Systems, 96:297–308, 2019.

[22] Son P., Ashnil K., ve Jinman K., A deep learning technique for imputing missing healthcare data. In 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pages 6513–6516. IEEE, 2019.

[23] Xiao X., Xiaoshuang L., Yanni K., Xian X., Junmei W., Yuyao S., Quanhe C., Xiaoyu J., Xinyue M., Xiaoyan M., ve ark. A multi-directional approach for missing value estimation in multivariate time series clinical data. Journal of Healthcare Informatics Research, 4(4):365–382, 2020.

[24] Yang Z., Zoie S.-Y. W., ve Kwok L. T., A framework of rebalancing imbalanced healthcare data for rare events' classification: a case of look-alike sound- alike mix-up incident detection. Journal of healthcare engineering, 2018, 2018.

[25] Akram F., David C., Rozalina M., Christopher S., John A M., Celine M V., ve Che N., Breast cancer classification using deep transfer learning on structured healthcare data. In 2019 IEEE International Conference on Data Science and Advanced Analytics (DSAA), pages 277–286. IEEE, 2019.

[26] Tran, T., Le, U., & Shi, Y. (2022). An effective up-sampling approach for breast cancer prediction with imbalanced data: A machine learning model-based comparative analysis. Plos one, 17(5), e0269135.

[27] Zi-Ching L., Guan-Yu H., Yun-Pei L., Seungmin R., S V., ve Bo-Wei C.. Conquering insufficient/imbalanced data learning for the internet of medical things. Neural Computing and Applications, pages 1–10, 2022.

[28] Pima indians diabetes dataset. "https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database" , [Ziyaret tarihi: 29 Haziran 2022].

[29] Ganganwar, V. (2012). An overview of classification algorithms for imbalanced datasets. International Journal of Emerging Technology and Advanced Engineering, 2(4), 42-47.

[30] Khushi, M., Shaukat, K., Alam, T. M., Hameed, I. A., Uddin, S., Luo, S., ... & Reyes, M. C. (2021). A comparative performance analysis of data resampling methods on imbalance medical data. IEEE Access, 9, 109960-109975.

[31] Ivan T.. Two modifications of cnn. 1976.

[32] Fan, X., Tang, K., & Weise, T. (2011, May). Margin-based over-sampling method for learning from imbalanced datasets. In Pacific-Asia Conference on Knowledge Discovery and Data Mining (pp. 309-320). Springer, Berlin, Heidelberg.

[33] Nitesh V C., Kevin W B., Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. Journal of artificial intelligence research, 16:321–357, 2002.

[34] Varsha S B. ve Roshani A., A review on imbalanced learning methods. Int. J. Comput. Appl, 975:23–27, 2015.

[35] Nguyen H. M, Cooper E. W, ve Kamei K., Borderline over-sampling for imbalanced data classification. In Proceedings: Fifth International Workshop on Computational Intelligence & Applications, volume 2009, pages 24–29. IEEE SMC Hiroshima Chapter, 2009.

[36] Last F., Douzas G., ve Bacao F., Oversampling for imbalanced learning based on k-means and smote. arXiv preprint arXiv:1711.00837, 2017.

[37] Alisha B., Ravinder A. ve Sharma S. C., Accurate detection of electricity theft using classification algorithms and internet of things in smart grid. Arabian Journal for Science and Engineering, pages 1–17, 2021.

[38] Kumar T. R, Linesh Raja, Kumar A., Dadheech P., Kumar A.,ve Nachappa MN. A cluster based classification for imbalanced data using smote. In IOP Conference Series: Materials Science and Engineering, volume 1099, page 012080. IOP Publishing, 2021.

[39] Gordana I., Tome E., ve Koroušić Seljak B. Evaluating missing value imputation methods for food composition databases. Food and Chemical Toxi- cology, 141:111368, 2020.

[40] Wei-Chao L., Chih-Fong T., ve Zhong J. R., Deep learning for missing value imputation of continuous data and the effect of data discretization. Knowledge-Based Systems, 239:108079.

[41] Lin, W. C., & Tsai, C. F. (2020). Missing value imputation: a review and analysis of the literature (2006–2017). Artificial Intelligence Review, 53(2), 1487-1509.

[42] Hunt, L. A. (2017). Missing data imputation and its effect on the accuracy of classification. In Data Science (pp. 3-14). Springer, Cham.

[43] Liao, S. G., Lin, Y., Kang, D. D., Chandra, D., Bon, J., Kaminski, N., ... & Tseng, G. C. (2014). Missing value imputation in high-dimensional phenomic data: imputable or not, and how?. BMC bioinformatics, 15(1), 1-12.

[44] Luo, Y., Szolovits, P., Dighe, A. S., & Baron, J. M. (2018). 3D-MICE: integration of cross-sectional and longitudinal imputation for multi-analyte longitudinal clinical data. Journal of the American Medical Informatics Association, 25(6), 645-653.