KAYSERİ ÜNİVERSİTESİ
Sosyal Bilimler Dergisi
KAYSERİ UNIVERSITY JOURNAL OF SOCIAL SCIENCES

Makale Türü	Araştırma Makalesi	Yıl	2023	ss.	51-71
Gönderi Tarihi	30.05.2023	Cilt	5	DOI	10.51177/kayusosder.1307226
Kabul Tarihi	22.06.2023	Sayı	1		
Erken Görünüm Tarihi	23.06.2023	Ay	Haziran		
Yayın Tarihi	30.06.2023				

Makine öğrenmesi ile eksik veri tamamlama yöntemlerinin sınıflandırma performansına etkileri*Δ

The effects of missing data imputation methods with machine learning on classification performance

Şemsettin ERKEN[1]
Levent ŞENYAY[2]

Öz

Araştırma yapmak üzere toplanmış veri setlerindeki değerlerde eksiklerin olması sıklıkla karşılaşılan bir problemdir. Bu problemi çözmek adına literatürde, eksik değerlerin tamamlamasına ilişkin yöntemler bulunmaktadır. Bilgi teknolojileri ve veri yönetimindeki gelişmelerle birlikte ilgili probleme ilişkin yöntemler artmış ve makine öğrenmesi yöntemleri de eksik değerleri tamamlamada kullanılmaya başlanmıştır. Çalışma kapsamında, literatürde sıklıkla yararlanılan "Hitters" veri seti kullanılmıştır. Bu veri setindeki değerler, manipüle edilerek eksiltilmiş ve eksiltilen değerler Liste Boyunca Silme, Son Gözlemi İleri Taşıma, Ortalama Atama gibi temel eksik değer tamamlama yöntemlerinin yanı sıra Stokastik Regresyon, En Yakın k- Komşu algoritması, Random Forest algoritması ve Amelia algoritması gibi makine öğrenmesi yöntemleriyle tamamlanmıştır. Veri setinin eksiltilmemiş hali ve eksik değerleri, bahsedilen yöntemlerle tamamlanarak elde edilen veri setleri, WEKA paket programı kullanılarak Naive Bayes algoritmasıyla sınıflandırılmıştır. Sınıflandırma sonuçları, sınıflandırma süresi, doğruluk, kesinlik, duyarlılık, F-ölçütü ve ROC alanı performans değerlendirme kriterleriyle kıyaslanmıştır. Çalışmanın sonucunda, makine öğrenmesi yöntemlerinin, eksik veri tamamlamada ve sınıflandırma operasyonlarının performanslarını yükseltmede başarılı sonuçlar ortaya koyduğu görülmüştür.

Anahtar Kelimeler: Makine Öğrenmesi, Eksik Veri Tamamlama, Veri Madenciliği, Sınıflandırma.

Abstract

A common issue frequently encountered in research datasets is the presence of missing values. In the literature, a multitude of techniques for imputing missing values have been proposed. With advancements in information technology and data management, machine learning methods have emerged as viable approaches for addressing this problem. In this study, "Hitters" dataset, commonly utilized in literature, was employed. Manipulated values were introduced to create incomplete observations. In addition to fundamental techniques like Listwise Deletion, Last Observation Carried Forward, and Mean Imputation, machine learning methods, including Stochastic Regression, k-Nearest Neighbors algorithm, Random Forest algorithm, and Amelia algorithm, were employed to complete the missing values. The original dataset and the imputed datasets derived from these

Δ Yazarlar bu çalışmanın tüm süreçlerinin araştırma ve yayın etiğine uygun olduğunu, etik kurallara ve bilimsel atıf gösterme ilkelerine uyduğunu beyan etmiştir. Aksi bir durumda Kayseri Üniversitesi KAYÜSOSDER Dergisi sorumlu değildir.
* Bu çalışma, V.ASC 2023/Bahar Kongresinde bildiri olarak sunulmuştur.

[1] Doktora Öğrencisi, Dokuz Eylül Üniversitesi SBE Ekonometri ABD, semsettinerken@gmail.com
[2] Prof. Dr., Dokuz Eylül Üniversitesi İİBF Ekonometri Bölümü, levent.senyay@deu.edu.tr

methods were classified using the Naive Bayes algorithm within the WEKA software package. The classification outcomes were compared using performance evaluation criteria such as classification time, accuracy, precision, recall, F-measure, and ROC area. In conclusion, this study demonstrates that machine learning methods exhibit promising results in imputing missing values and enhancing classification performance.

Keywords: Machine Learning, Missing Data Imputation, Data Mining, Classification.

1. Giriş

Bilgi elde etme insanlık tarihi için en önemli konu olmuştur. İnsanlar, hayatta kalabilmek için yaşadıkları coğrafya içerisinde tecrübe ettikleri her türlü konuyu yaşamlarına entegre ederek yaşamışlardır. Bu durum, deneyim yoluyla elde ettikleri tecrübeleri veri kabul ederek bilgiye dönüştürmenin en temel örneklerindendir.

Günümüz teknolojisinde ve yaşamında bilgi, geçmişten olduğundan çok daha büyük öneme sahiptir. Bunun sebebi, teknolojik gelişmeler sayesinde, mevcut bilgiden çok daha fazla çıktı üretilmesi imkânıdır. Üretilen çıktı miktarı ve hızına nispeten veri üretim hızı çok daha yüksek seviyelerdedir. Bu çerçeveden bakıldığında, bilgiye erişimin anahtarı olan verinin önemi bir kat daha artmaktadır.

Verinin işlenmesi ve analizi, doğrudan yapılabilecek bir aksiyon değildir. Sağlıklı analiz ve sağlıklı çıktı sonuç elde etmek için verinin analize hazır hale getirilmesi gerekmektedir. Analize hazır hale getirilmesi gerekliliği, genel olarak verilerin analiz öncesinde birtakım problemler barındırmasındadır. Bahsi geçen problemler arasında sık karşılaşılan durumlardan biri veri setlerinde, eksik değerlerin olmasıdır. Bu eksiklik, farklı sebeplerden ortaya çıkmaktadır. Veri kaynağından dolayı eksiklikler, verinin saklanılması ve korunması konusundaki problemler, veriler arasındaki bazı ilişkisel durumlar, veri setinde eksik değerler olmasının sebepleri arasında sayılabilir.

Veri setindeki eksik değer probleminin, analiz öncesinde giderilmesi analiz etkinliği ve sağlıklı sonuçların elde edilmesi açısından hayatidir. Analizlerin sonuçlarının elde edilmesinin öncesinde, eksik verilerin var olduğu bir veri topluluğunun analizi sürecinde de çeşitli sorunlar ile karşılaşılmaktadır. Bu sorunların çözümlenmemesi, analizlerin başarısı anlamında, negatif olarak kümülatif bir etki doğurmaktadır. Bu noktadaki asıl üzerinde durulan konu, sadece eksik verilerin tamamlanması değil eksik verilerin, veri topluluğunun taşıdığı özellikleri yansıtan ve karakteristik unsurları üzerinde barındıracak bir şekilde tamamlanmasıdır. Bu sebeple literatürde, eksik değerleri tamamlamak için çeşitli yöntemler yer almaktadır. Bu noktada önemli olan düşünce, daha önce belirtildiği gibi eksik değerler tamamlanırken, tamamlanan değerin veri setinin karakteristiğini yansıtmasıdır. Yanlış bir yaklaşımla eksik değerleri tamamlanan veri setleri orijinalinden çok farklı özellikleri ve karakteristiği yansıtacaktır. Bu durum, analiz sonrasında elde edilecek sonuçların, gerçekçi olmayan bir yapıda olmasına sebep olacaktır.

Bir veri topluluğunda, eksik verilerin tamamlanmasından önce veri topluluğunun sahip olduğu birtakım özellik ve örüntülerinin ortaya çıkarılması oldukça önemlidir. Bu özelliklerin ışığında eksik verilerin tamamlanması, dolaylı olarak analizin sağlığı ve başarısını etkileyecek en önemli unsurdur. Ortaya çıkan örüntü ve özelliklere dayanarak yapılacak eksik veri tamamlama operasyonları ile sağlıklı analiz sonuçlarına ulaşılması mümkündür. Bu durumda, eksik değerlerin tamamlanmasının ve tamamlama yaklaşımının çok önemli olduğu ortadır. Gelişen bilgi ve veri teknolojileri, bu aşamada anahtar pozisyondadır. Bu amaçla, gelişen teknolojik araçlardan faydalanmak elzemdir.

Veri setlerinde eksik değerlerin tamamlanmasında, eksik olan değerlerin ilgili değişken değerlerinin ortalaması ya da medyanıyla tamamlanması, eksik değerden önceki ilk değerle tamamlanması veya eksik değerlerin bir anlamda yok sayılarak veri setinden silinmesi gibi çeşitli istatistiksel yöntemler bulunmaktadır. Bu yöntemlerle, eksik değerler tamamlansa da bahsedildiği gibi veri örüntülerinin ortaya çıkarılması konusundaki problemlerin çözümünde etkili olması tartışılabilir durumdadır.

Makine öğrenmesi yöntemleri oldukça geniş bir kullanım alanına sahiptir. Özellikle, gelişen bilgi ve bilgisayar teknolojileri ışığında kullanım alanı her geçen gün artmaktadır. Bu anlamda, makine öğrenmesi, bahsedilen veri setindeki eksik değerlerin tamamlanmasında da kullanılmaktadır. Makine öğrenmesi yöntemleriyle, veri topluluklarındaki karakteristik yapıyı ortaya çıkarıp bu yapıya uygun olarak eksik değerlerin tamamlanması, mevcut yöntemler alternatif olarak etkin sonuçlar üreten ve daha iyi performans sergileyen bir olanak sağlamaktadır. Bu sebeple, makine öğrenmesi yöntemleri, veriden bilgi üretmede, ilgili analizlerde ve literatürde çok büyük bir öneme sahiptir.

2. Literatür taraması

Makine öğrenmesi yöntemleri oldukça geniş bir kullanım alanına sahiptir. Bahsedildiği gibi bu alanlardan birisi de eksik veri tamamlamadır. Literatürde de görüleceği gibi, veri setlerinde sıklıkla karşılaşılan bir problem olan eksik değerlerin bulunması sorununun giderilmesi adına makine öğrenmesi yöntemlerine başvurulmuştur. Bu çalışmalarda, makine öğrenmesi yöntemlerinin ve makine öğrenmesi tabanlı yöntemlerin, eksik veri tamamlama operasyonlarında başarılı sonuçlar ortaya koyduğu gösterilmiştir.

Alamoodi vd. (2021), Ulusal Çocuk Gelişim Merkezi (Malezya)'nden elde ettiği 14 farklı zaman noktasına ait ve %80'den fazla eksik veriyi barındıran vücut kitle endeksi verilerini kullanarak önerdiği metodun, makine öğrenmesi yöntemleri ile birlikte kullanılmasıyla eksik değer oranının yüksek olduğu veri setlerinde de başarılı sonuçlar elde edildiğini göstermiştir.

Jerez vd. (2010) tarafından, İspanyol Göğüs Kanseri Araştırma Grubu'na ait 32 farklı hastaneden konulmuş olan 3679 kadına ait göğüs kanseri teşhisi verisini kullanarak makine öğrenmesine dayalı yöntemlerin eksik veri tamamlamaya çok daha uygun olduğu ve istatistiksel yöntemlere göre teşhis doğruluğunu anlamlı şekilde artırdığı tespit edilmiştir.

Emmanuel vd. (2021), İris ve Power Plan Fan veri setlerindeki %5'ten başlayarak %20'ye kadar ulaşan eksik değer oranlarında, makine öğrenmesi yöntemlerinden Random Forest tabanlı MissForest ve En Yakın k-Komşu algoritmasının, eksik veri tamamlamada oldukça başarılı sonuçlar ortaya koyduğunu göstermiştir.

Kaggle veri deposunda yer alan ve oldukça sık kullanılan DS2OS veri seti kullanılarak nesnelerin interneti alanında anomali ve saldırı tespiti içerikli araştırmada, Vangipuram vd. (2020) tarafından, eksik veri tamamlamada K-ortalamalar ve Bulanık K-ortalamalar gibi makine öğrenmesi algoritmaların daha iyi sonuç verdiği gösterilmiştir.

Kenyhercz ve Passalacqua (2016), Howells'in Kraniyometrik Veri Seti'nden faydalanılarak biyolojik bir süreç içeren çalışmasında, kullanılan diğer yöntemler arasından, makine öğrenmesi yöntemlerinden olan En Yakın k-Komşu algoritmasının en iyi sonucu verdiğini tespit etmiştir. İlgili çalışmada, gerçek değerler ile tamamlanmış değerler arasındaki mesafenin en düşük olan yöntemin En Yakın k-Komşu algoritması olduğu sonucu elde edilmiştir.

UCI veri deposunda erişime açık olan ve sık kullanıma sahip olan Dermatology, Pima, Wisconsin ve Yeast veri setleri kullanılarak veri setlerindeki gerçek değerler ile manipüle edilip makine öğrenmesi yöntemlerinden K-ortalamalar merkez tabanlı yöntemin en iyi sonuçları sunduğu, performans değerlendirme kriterlerince Raja ve Thangavel (2020) tarafından ortaya koyulmuştur.

Palanivinayagam ve Damaševicius (2023), PIMA Indian veri seti üzerinde yaptığı çalışmada makine öğrenmesi algoritmalarından destek vektör makinelerinin diyabet rahatsızlığının erken teşhisi konusunda yapılan sınıflandırma operasyonu performansını anlamlı bir şekilde artırdığını göstermiştir.

Thomas ve Rajavi (2021), makine öğrenmesi tabanlı eksik değer tamamlama teknikleriyle ilgili ortaya koydukları 2010 ve 2020 yıllarını kapsayan sistematik araştırmalarında, makine öğrenmesi tekniklerinin, veri setinin karakteristiğine uygun bir şekilde eksik verileri tamamlamada iyi olduğu ve kümeleme algoritmaları ile en yakın k-komşu algoritmalarının uygulama kolaylığı sağladığı sonucuna varmıştır.

Jadhav vd. (2019), UCI veri deposundaki açık erişimli Wine, Glass Identification, Concrete Comprehensive Strength, Indian Liver Patient ve Seeds veri setleri olmak üzere toplam 5 adet veri setinden yararlanmışlardır. Çalışmalarında, kullanılan ortalama atama, medyan atama, doğrusal regresyon, bayezyen doğrusal regresyon gibi yöntemler arasından makine öğrenmesi yöntemlerinden K-nn algoritmasının en iyi sonucu verdiğini göstermiştir.

3. Amaç ve yöntem

Bu çalışmada, veri analizi çalışmalarında, oldukça geniş bir kullanıma ve veri hacmine sahip olan "Hitters" veri seti kullanılmıştır. Bu veri setinde, 1986/1987 yıllarında Major Ligi beyzbol oyuncularına ait istatistikler yer almaktadır. Bu istatistiklerden birisi de sporcuların aldığı yıllık ücret (bin dolar)lerdir. Sporcuların yıllık ücretleri, A, B, C, D ve E şeklinde kodlanarak alınan en yüksek ücretlerden düşük ücretlere doğru kategorilere ayrılıp sınıflandırılmış bir veri şeklinde ifade edilmiştir.

Bu veri setindeki veriler, %5 oranında R programı kullanılarak manipüle edilmiş ve eksiltilmiştir. Eksiltilmiş değerler, Liste Boyunca Silme, Son Gözlemi İleri Taşıma, Ortalama Atama gibi temel eksik değer tamamlama yöntemleriyle ayrıca makine öğrenmesi yöntemlerinden, En Yakın k-Komşu Algoritması, Random Forest Algoritması, Stokastik Regresyon ve Amelia Algoritması yöntemleri kullanılarak tamamlanmıştır. Ardından bahsedilen yöntemlerle tamamlanmasıyla oluşmuş her bir veri setine ve "Hitters" veri setinin eksiltilmemiş haline, Naive Bayes Algoritması ile WEKA paket programı kullanılarak sınıflandırma operasyonu yapılmıştır. Böylece bu uygulamalarla, bahsedilen yöntemlerin, eksik değer tamamlama temelli sınıflandırma operasyonuna katkılarının karşılaştırılması amaçlanmaktadır.

Uygulamaların ardından, sınıflandırma için kullanılan süre, doğru sınıflandırma oranı, kesinlik, duyarlılık, F-ölçütü ve ROC (Receiver Operating Characteristic) alanı gibi performans değerlendirme kriterleri ile ilgili yöntemlerin eksik veri tamamlama performansları temelinde sınıflandırma operasyonuna katkıları kıyaslanmıştır.

4. Eksik veri

Veri setlerinde, eksik değerlerin bulunması çok sık rastlanan bir durumdur. Veri setinin analizden önce, sürecin ve sonucun etkinliği için eksik değer probleminin giderilmesi gerekmektedir. Eksik değerler, farklı sebeplerle ortaya çıkmaktadır. Bu noktada, eksik değerlerin türleri üç adet başlık altında incelenmektedir. Bu çerçevede eksik veri türleri; Missing Completely at Random (MCAR): tamamen rassal eksik veri, Missing at Random (MAR): rassal eksik veri ve Missing Not at Random (MNAR): rassal olmayan eksik veri şeklinde ifade edilmektedir.

MCAR eksik veri türü, değişkenin gözlenmiş değerleriyle ya da farklı bir değişkendeki değerlerle ilişkisi olmadan verinin eksik olması sonucu oluşan eksik veri türüdür. Diğer bir deyişle, bağımsızlığın olduğu bir durum söz konusudur ve bu durum oldukça önemlidir (Allison, 2009, ss. 72-74). Bu sebeple, eğer eksik veri türü MCAR ise eksik bir değerin olmadığı nitelikler veya değişkenler, kitlenin rastgele bir örneği şeklinde ifade edilebilir (Donders vd., 2006, s. 1088).

MAR eksik veri türünde, MCAR veri türündeki duruma kıyasla zayıf bir bağımsızlık söz konusudur. Eksik verinin eksik olması, ilgili nitelik veya değişkenden bağımsız iken farklı bir nitelik veya değişkenden kaynaklanmaktadır (Schaffer,1997, ss. 10-15).

MNAR eksik veri türünde, göz ardı edilemeyecek bir durum söz konusudur (Köse & Öztemur, 2014, s.402). Eğer eksiklik, bahsedilen MCAR ve MAR değil ise eksikliği, ilgili nitelik veya değişkenin kendisine bağlı olduğu MNAR türünde eksik veri türü olarak ifade etmek mümkündür (Enders, 2022, ss. 11-12).

Eksik veri türünün tespiti, eksik verilerin tamamlanmasında oldukça önemlidir. Verideki eksik verinin türünün tespit edilmesi, eksikliği tamamlama sürecinde yol haritası niteliğindedir. Eksikliğin sahip olduğu özellikler ışığında, eksik verilerin tamamlanarak verinin analize hazırlanması ve sağlıklı sonuçların elde edilmesi açısından büyük etkiye sahiptir (Abidin vd., 2018, ss. 442-443).

4.1. Eksik veri tamamlamada kullanılan temel yöntemler

Eksik veri tamamlamada, farklı yaklaşımları esas alan birçok yöntem mevcuttur. Bu kısımda, çalışmada kullanılmış olan temel eksik veri tamamlama yöntemleri ele alınmıştır.

- ***Liste boyunca silme yöntemi:*** veri setinde eksik değerlerin bulunduğu tüm veri kayıtlarının veri setinden çıkarılması şeklindeki uygulamadır. Bu yöntem, veri seti yoluyla yapılacak analizde, eksik değerlerin etkisinin ortadan kaldırılması ve pratikte veri setlerinde eksik değerlerin olmasının sıkça gözlenmesi nedeniyle tercih edilen yöntemler arasında yer almaktadır. Genel bir kullanıma sahip olsa da liste boyunca silme yöntemi, özellikle veri setinin eksik değerlerin yüksek sayılacak bir orana sahip olması durumunda oldukça büyük problemler yaratabilmektedir. Yüksek oranda eksik değerlerin olması çok sayıda veri kaydının silinmesi anlamına geleceğinden analiz sonuçlarında ciddi sapmalara yol açabilmektedir.
- ***Son Gözlemi İleri Taşıma:*** bu yöntem, eksik veri tamamlamada süreç anlamında uzun olan analizlerde tercih edilen yöntemlerdendir. Son gözlemi ileri taşıma, eksik değeri tamamlamasında eksik değer yerine eksik değerden önce gözlenmiş olan değerin tamamlanması şeklindeki operasyondur. Yöntemin kolay uygulanabilir olması, pratikte yaygın kullanıma sahip olmasının sebeplerindendir. Buna rağmen, literatürde, son gözlemi ileri taşıma yönteminin sonuçları etkinlik açısından tartışılmaktadır.
- ***Ortalama atama yöntemi:*** eksik verileri, veri setindeki ilgili değişkenin var olan değerlerine ait ortalamayı eksik olan değer yerine koyarak tamamlamaktadır. Literatürde, eksik veri tamamlama uygulamalarında oldukça sık bir kullanıma sahiptir.

5. Makine öğrenmesi

Gelişen ve gelişmeye devam eden teknoloji ile birlikte birçok alanda olduğu gibi bilgi ve veri teknolojileri alanında da ilerlemeler kaydedilmiştir. Buna paralel olarak bilgisayar teknolojisi ile birlikte yapay zekâ ve yapay öğrenme kavramları da literatürde ve pratik dünyada yerini almıştır. Bu kavramların bir karşılığı da makine öğrenmesi kavramıdır. Matematik ve istatistik tabanlı yaklaşımlarla verideki birtakım örüntüleri analiz eden makine öğrenmesi, buradan hareketle geleceğe yönelik açıklayıcı araçlar ve modeller oluşturmaktadır (Brynjolfsson & Mitchell, 2017, ss. 1530-1532). Makine öğrenmesinin en önemi özelliklerinden biri de bahsedilen aksiyonları insan unsurundan ve müdahalesinden bağımsız bir şekilde yapan bir yapay zekâ olmasıdır (Bi vd., 2018, s. 2222). Makine öğrenmesinde, öğrenme yöntemleri 4 başlık altında ifade edilebilir.

- ***Gözetimli öğrenme (Supervised Learning):*** Bu yöntem, gerçek ve doğru sonuçlardan oluşan bilgilerin girdi olarak sunulduğu, analizin bu şekilde etiketli veriler kullanılarak yapıldığı öğrenme türüdür. Etiketlenmiş verilerle bilgisayara öğrenme yaptırılarak tahmin için model ortaya koyulur.
- ***Gözetimsiz öğrenme (Unsupervised Learning):*** Bu yöntemde ise gözetimli öğrenmeden farklı olarak etiketli olmayan veriler sunulur. Bilgisayar, bu etiketsiz veriler kullanarak örüntü ve yapıları ortaya çıkarır. Bundan kaynaklı, bu öğrenme yaklaşımında, karmaşık ya da daha önceden kestirilemeyen sonuçlar ile karşılaşılabilir.
- ***Yarı gözetimli öğrenme (Semi-supervised Learning***): Gözetimli ve gözetimsiz öğrenmenin yaklaşımları kısmi ortak halde olarak yaklaşımları kullanılır. Özellikle gözetimli öğrenmede

karşılaşılan maliyet ve aynı şekilde zamansal sorunların giderilmesine olanak sağlamak için kullanılmaktadır.

- ***Takviyeli öğrenme (Reinforcement Learning):*** Bu yöntem, maksimum ödül odaklı bir oyuncunun, mevcut bir ortamda öğrenmesi şeklindedir. Oyuncu, ilgili ortamda belli bir aksiyonu gerçekleştirerek ortamın mevcut durumunu değiştirir ve yeni durumdaki alacağı ödülle göre yeni aksiyonlara yönelir. Bu yaklaşımla, alacağı ödülü en yüksekleyen stratejiyi ortaya koyar.

Makine öğrenmesi yöntemleri oldukça geniş bir kullanım alanına sahiptir. Gelişen bilgi ve bilgisayar teknolojilerine paralel olarak her geçen gün daha fazla alanda kullanılması ile birlikte algoritmik yapılarının da gelişmesi söz konusudur. Bu çalışma kapsamında kullanılmış olan makine öğrenmesi yöntemleri, bu kısımda açıklanmıştır.

En yakın k-Komşu Algoritması(K-nn): bu yöntem, mesafe tabanlı bir makine öğrenmesi yöntemidir. Sınıflandırma, regresyon gibi uygulamaların yanında eksik veri tamamlama alanında da oldukça geniş bir kullanıma sahiptir. K-nn, bahsi geçen mesafe dâhilinde kalan k adet komşu kaydın, hedef niteliğini dikkate alan ve benzerlik atayan yöntemdir (Mahesh, 2020, s. 385). K-nn yöntemi, verilerin tamamlanmasının ardından en yakın mesafede bulunan k adet gözlemi belirlemektedir. İlgili gözlemlerin belirlenmesi için Öklid, Manhattan ve Minkowski gibi uzaklık ölçülerini kullanılmaktadır (Zhang, 2016, s. 219). Bahsedilen örnekler arası uzaklıkların hesaplamasında kullanılan ölçülere ait fonksiyonlar verilmiştir.

2 boyutlu bir uzayda Pisagor teoreminin sonucu olarak Öklid uzaklığı elde edilmektedir. (1) nolu denklemle Öklid uzaklığı ifade edilmiştir.

$$d(i, j)=\sqrt{\sum_{k=1}^{p}(x_{ik}-x_{jk})^2} \tag{1}$$

Örnekler arasındaki mesafelerin mutlak değeri cinsinden ifadesi ise Manhattan uzaklığı olarak tanımlanmaktadır. Bu uzaklık, (2) nolu denklem ile gösterilmiştir.

$$d(i,j)=\sum_{k=1}^{p}(|x_{ik}-x_{jk}|) \qquad i, j=1, 2,\ldots, n; k=1, 2,\ldots, p \tag{2}$$

Minkowski uzaklığı, p tane nitelik veya değişken için denklem (3)'teki gibi tanımlanmaktadır. Denklem ele alındığında, m=1 için Manhattan ve m=2 için ise Öklid uzaklıkları ele edilebilmektedir.

$$d(i, j)=\left[\sum_{k=1}^{p}(|x_{ik}-x_{jk}|^{m})\right]^{\frac{1}{m}} \qquad i, j=1, 2,\ldots, n; k=1, 2,\ldots, p \tag{3}$$

Random Forest Yöntemi (RF): Random Forest algoritması, bir karar ormanı yapısıdır (Durmuş ve Güneri, 2019 s. 206). Eksik veri tamamlama, regresyon, sınıflama ve karar alma yöntemleri alanlarında sık kullanılan bu yöntem, birbiri ile ilişkisi bakımından önemsiz bir korelasyon düzeyinde veya bağımsız olan karar ağaçlarının bir araya getirilmesiyle oluşmaktadır. Bu karar ağaçları, bootstrap ile elde edilmiş örnekler yoluyla meydana gelmektedir. İlgili karar ağaçları, yine ilgili veri setinden bootstrap yöntemi ile elde edilen örneklerden oluşturulur. Veri setindeki bir veri kaydının sınıf, karar ağaçları tarafından oylanarak belirlenir.

Random Forest yönteminde, oluşturulmuş olan karar ağaçlarının, farklı veri alt kümelerinde eğitimi yapılır. Her ağaç, farklı alt kümelerde eğitilirken farklı bir özelliğe ilişkin alt kümeyi kullanıp optimum karar ağacı yapısını oluşturmayı amaçlar. Böylece, her ağaç farklı bir özelliğe odaklanacaktır. Eğitim sonrasında, yeni bir veri kaydı için her ağacın sınıflandırmayı yapmasının ardından elde edilen

sonuçlar oylama ile birleştirilir. Böylece, yüksek doğruluğun yanında, farklı ağaçlardaki verilere bağlı kusurların etkisi ve muhtemel sapmalar azaltılmaya çalışılır (Tang & Ishwaran, 2017, ss. 363-364).

Amelia Algoritması: eksik veri tamamlamada sık kullanılan bir yöntem olan bu yapı, bootstrap ve beklenti maksimizasyonu ile çoklu bir tamamlama stratejisiyle hareket etmekte olan istatistiksel bir modelleme aracıdır. Amelia algoritması da farklı örneklemlerle, eksik verileri olan alt setleri elde etmektedir. Birden çok modeli kullanarak, kullanılan her model için, bootstrap ve beklenti maksimizasyonu ile tamamlanmış olan eksik değerlerin ortalamasını alır ve elde edilen sonuçlar birleştirilmektedir (Honaker vd., 2011, ss. 1-2). Bootstrap, verilerden yeniden örnekleme yoluyla, örneklemin sahip olduğu istatistiksel dağılım hakkında bilgi sahibi olma imkânı sağlayan bir yöntemdir (Doğan, 2017, ss. 2-4). Beklenti maksimizasyonu ise tamamlanan eksik değerlerin, iterasyonlar boyunca güncellenip en yüksek olabilirlik tahminlerini bulmak için kullanılır (Doğru vd., 2016, s. 340). Böylece, Amelia algoritması, bahsedilen bootstrap ve beklenti maksimizasyonu kombinasyonları ile güvenilir sonuçlar üretmektedir.

Stokastik Regresyon: genel olarak Gauss Markov varsayımlarının sağlandığı ve gürültülü verilerin varlığında istatistiksel bir model olan stokastik regresyon, sürekli bir değişkenin tahminlerini, farklı bağımsız değişken ve normal dağılama sahip rassal hata terimi ile birlikte açıklanmak üzere kullanılmaktadır. Hata terimleri de birbirinden bağımsızdır. Böylece, en yüksek olabilirlik yöntemi ile model tahminlenir ve ilgili değişkenin gelecek değerleri için aralık tahminleri ortaya koyulabilmektedir. Eksik veri tamamlamada stokastik regresyonun kullanılması, eksik değerlerin, doğrusal regresyondaki doğru üzerindeki değerlerle tamamlanması nedeniyle hem verinin istatistiksel dağılımın etkilenmesi hem de kovaryanslarının gerekenden daha düşük tahminlemesi gibi sorunların giderilmesi için kullanılmaktadır. Literatürde yer alan çalışmalarda, bu yöntemle, tamamen rassal ve rassal eksik veri türünden eksiklikleri bulunan veri gruplarında, yansız parametre tahminlerinin elde edildiği ve doğrusal regresyon ile eksik veri tamamlama sonuçlarına göre çok daha iyi sonuçlara ulaşıldığı ifade edilmiştir (Baraldi & Enders, 2010, ss. 13-15).

Naive Bayes Algoritması: özellikle veri madenciliği, sınıflandırma ve metin madenciliği sık kullanılan bu algoritma, eksik veri tamamlamada da kullanılmaktadır. Naive Bayes, bayesyen olasılık tabanlı bir yöntemdir. Bu yöntemde, daha önceden sınıflandırılmış olan verilerden hareketle, sınıf değeri araştırılan nitelik veya değişkenin, tanımlanmış olan her değeri alma olasılığını hesaplanmaktadır. Ardından, hesaplanan en yüksek olasılık değerine sahip değerin, nitelik veya değişken değeri olarak ataması yapılmaktadır. Bu yöntem dâhilinde, veri setinde yer alan değişkenlerin birbirinden bağımsız veya değişkeler arası korelasyonun düşük olduğu varsayımı söz konusudur (Durmuş & Güneri, 2021, s. 98). Bu varsayım, analiz kolaylığı sağlamaktadır (Vembandasamy vd., 2015, s. 442). (4) nolu denklemde bahsedilen olasılık hesaplama fonksiyonu verilmiştir. Burada, payda kısmı eşit olacağından, (5) nolu ifadenin açıklamasında olduğu gibi pay değeri yüksek olan nitelik veya değişken değeri, ilgili veri kaydının sınıf değeri olarak atanır.

$$P(C_i|X)=\frac{P(X|C_i).P(C_i)}{P(X)} \tag{4}$$

$$\arg\max_{C_i}=\{P(X|C_i)P(C_i)\} \tag{5}$$

6. Performans değerlendirme kriterleri

Kullanılan yöntemlerin sınıflandırma performansları, "Confusion Matrix" isimli matristen hareketle elde edilen kriter değerlerine göre kıyaslanmaktadır. Tablo 1'de "Confusion Matrix" yapısı ifade edilmiştir.

Tablo 1

Confusion matrix yapısı

		Tahminlenen Sınıf	
		Sınıf 1	Sınıf 0
Gerçek Sınıf	Sınıf 1	X	Y
	Sınıf 0	Z	K

Veri kayıtlarına ilişkin gerçek ve tahminlenen sınıf sonuçlarına göre değer alan X, Y, Z ve K değerleri ile performans değerlendirme kriterleri hesaplanmaktadır. X: True Positive (TP), Y: False Negative (FN), Z: False Positive (FP) ve K: True Negative (TN) değerlerini göstermektedir (Oprea, 2014, ss. 250-251).

Sınıflandırma süresi: bu kriter, ilgili yöntemin mevcut veri setinin sınıflandırması için kullandığı süreyi saniye cinsinden ifadesidir.

Doğruluk oranı: sınıflandırma işleminde, doğru sınıflandırılmış kayıt sayısının tüm kayıt sayısına oranını ifade eden kriterdir. (6) nolu denklemde gibi hesaplanmaktadır.

$$\text{Doğruluk}=\frac{TP+TN}{TP+TN+FP+FN} \tag{6}$$

Kesinlik: analiz ile elde edilen sonuçların, aynı yöntem sonucunda elde edilmiş olanların tespitinde kullanılır. (7) nolu denklemle kesinlik değeri hesaplanmaktadır.

$$\text{Kesinlik}=\frac{TP}{TP+FP} \tag{7}$$

Duyarlılık: sınıflandırma operasyonu sonuçlarının, birbirine yakınlığını gösteren performans değeridir. (8) nolu denklem ile duyarlılık değeri elde edilmektedir.

$$\text{Duyarlılık}=\frac{TP}{TP+FN} \tag{8}$$

F-ölçütü, farklı iki kriterin bir anlamda kombinasyonundan oluşmaktadır. Duyarlılık ve kesinlik kriteri değerlerinin harmonik ortalaması bu kriteri oluşturmaktadır. Duyarlılık ve kesinlik geçerli ve değerlendirme anlamında önemli kriterler olsa da tek başlarına yeterli kıyaslama kalitesinde olmayabilirler. F-ölçütü, dengeli bir duyarlılık ve kesinlik durumunda yüksek değerler almaktadır. (9) nolu denklem F-ölçütü değerini vermektedir.

$$\text{F-ölçütü}=\frac{2 \text{ X Duyarlılık X Kesinlik}}{\text{Duyarlılık+Kesinlik}} \tag{9}$$

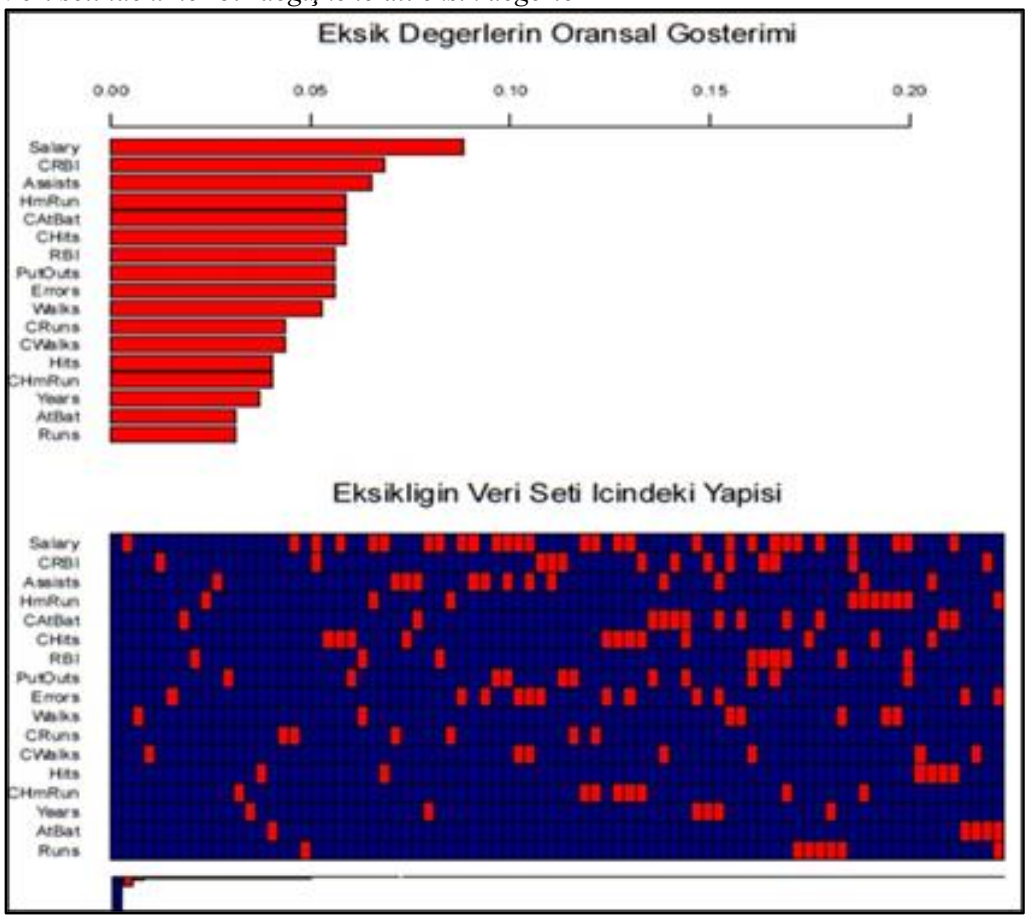
ROC alanı: veri madenciliği ve makine öğrenmesi operasyonların sık başvurulan kriterin performansı, belirttiği alan büyüklüğü ile doğru orantılıdır. 2 boyutlu uzayda TP değerinin ordinat ve FP oranının apsis olduğu bir grafiğin alanı, ROC alanı değer olarak ifade edilmektedir.

7. Bulgular

Çalışma kapsamında, "Hitters" veri seti %5 oranında manipüle edilerek rassal biçimde eksiltilmiştir. Şekil 1'de değişkenler bazında eksiltilen değerler gösterilmiştir.

Şekil 1

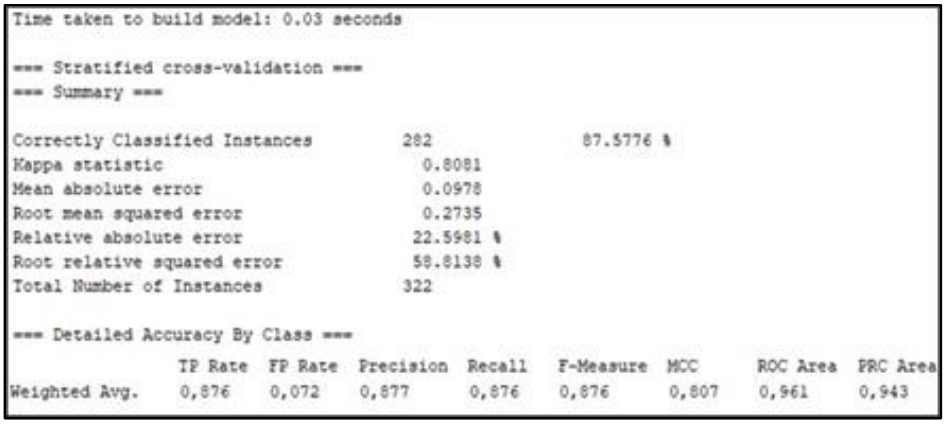
Veri setindeki her bir değişkene ait eksik değerler



İlk olarak, veri setinin orijinal hali, Naive Bayes algoritmasıyla sporcuların yıllık ücretlerinin, sınıflandırılmış veri halindeki değişkeni, sınıf niteliği olmak üzere sınıflandırılmış ve sonuçlar Şekil 2'de gösterilmiştir.

Şekil 2

Orijinal veri setinin sınıflandırma sonuçları



```
Time taken to build model: 0.03 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         282               87.5776 %
Kappa statistic                          0.8081
Mean absolute error                      0.0978
Root mean squared error                  0.2735
Relative absolute error                 22.5981 %
Root relative squared error             58.8138 %
Total Number of Instances              322

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area
Weighted Avg.  0,876    0,072    0,877      0,876    0,876      0,807    0,961     0,943
```

Tablo 2

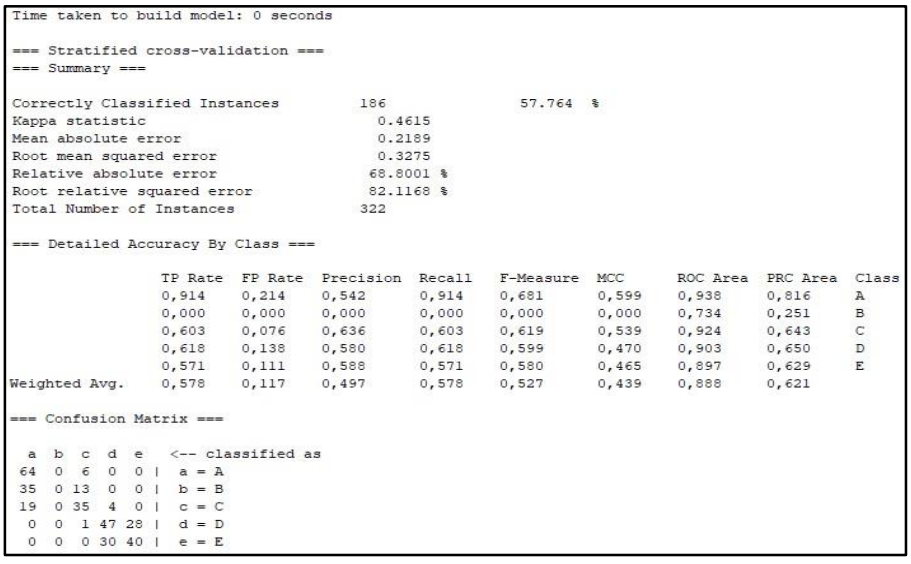
Orijinal veri seti sınıflandırması performans kriteri değerleri özeti

Sınıflandırma Sonuçları					
Sınıflandırması Süresi(Saniye)	**Doğruluk(%)**	**Kesinlik**	**Duyarlılık**	**F-Ölçütü**	**ROC Alanı**
0.03	87.5776	0.877	0.876	0.876	0.961

Şekil-3'te liste boyunca silme yöntemi ile oluşmuş olan veri setinin Naive Bayes algoritması ile sınıflandırılması sonucu performans kriterlerinin aldığı değerlere ait WEKA çıktısı verilmiştir.

Şekil 3

Liste boyunca silme ile tamamlanan veri setinin sınıflandırma sonuçları



```
Time taken to build model: 0 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         186               57.764  %
Kappa statistic                          0.4615
Mean absolute error                      0.2189
Root mean squared error                  0.3275
Relative absolute error                 68.8001 %
Root relative squared error             82.1168 %
Total Number of Instances              322

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0,914    0,214    0,542      0,914    0,681      0,599    0,938     0,816     A
               0,000    0,000    0,000      0,000    0,000      0,000    0,734     0,251     B
               0,603    0,076    0,636      0,603    0,619      0,539    0,924     0,643     C
               0,618    0,138    0,580      0,618    0,599      0,470    0,903     0,650     D
               0,571    0,111    0,588      0,571    0,580      0,465    0,897     0,629     E
Weighted Avg.  0,578    0,117    0,497      0,578    0,527      0,439    0,888     0,621

=== Confusion Matrix ===

  a  b  c  d  e   <-- classified as
 64  0  6  0  0 |  a = A
 35  0 13  0  0 |  b = B
 19  0 35  4  0 |  c = C
  0  0  1 47 28 |  d = D
  0  0  0 30 40 |  e = E
```

Şekil 4

Son gözlemi ileri taşıma ile oluşan veri setine ait sınıflandırma sonuçları

```
Time taken to build model: 0 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         220               68.323  %
Kappa statistic                          0.6024
Mean absolute error                      0.1282
Root mean squared error                  0.3227
Relative absolute error                 40.3034 %
Root relative squared error             80.9144 %
Total Number of Instances              322

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,800    0,083    0,727      0,800    0,762      0,693    0,963     0,882     A
                 0,813    0,058    0,709      0,813    0,757      0,714    0,957     0,797     B
                 0,655    0,011    0,927      0,655    0,768      0,742    0,937     0,854     C
                 0,368    0,045    0,718      0,368    0,487      0,421    0,915     0,732     D
                 0,843    0,202    0,536      0,843    0,656      0,557    0,918     0,876     E
Weighted Avg.    0,683    0,083    0,717      0,683    0,674      0,611    0,937     0,828

=== Confusion Matrix ===

  a  b  c  d  e   <-- classified as
 56 14  0  0  0 |  a = A
  7 39  0  0  2 |  b = B
  6  2 38  3  9 |  c = C
  5  0  3 28 40 |  d = D
  3  0  0  8 59 |  e = E
```

Şekil 4 ve Şekil 5'te sırasıyla, eksik değerlerin son gözlemi ileri taşıma ve ortalama atama yöntemleriyle tamamlanmasıyla oluşan veri setlerinin, Naive Bayes algoritması kullanılarak sınıflandırılması sonucu elde edilen performans kriterlerinin aldığı değerlere ilişkin WEKA çıktısı sunulmuştur.

Şekil 5

Ortalama atama ile oluşan veri setine ait sınıflandırma sonuçları

```
Time taken to build model: 0.04 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         259               80.4348 %
Kappa statistic                          0.7468
Mean absolute error                      0.0835
Root mean squared error                  0.2644
Relative absolute error                 27.4489 %
Root relative squared error             67.815  %
Total Number of Instances              322

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,865    0,081    0,672      0,865    0,756      0,711    0,962     0,874     A
                 0,756    0,046    0,705      0,756    0,729      0,689    0,950     0,775     B
                 0,792    0,040    0,922      0,792    0,852      0,780    0,969     0,951     C
                 0,730    0,035    0,836      0,730    0,780      0,733    0,944     0,863     D
                 0,913    0,040    0,792      0,913    0,848      0,824    0,986     0,946     E
Weighted Avg.    0,804    0,046    0,819      0,804    0,806      0,754    0,963     0,898

=== Confusion Matrix ===

  a  b  c  d  e   <-- classified as
 45  7  0  0  0 |  a = A
  6 31  4  0  0 |  b = B
 13  6 95  6  0 |  c = C
  2  0  4 46 11 |  d = D
  1  0  0  3 42 |  e = E
```

Stokastik regresyon ile eksik değerleri tamamlanmış olan veri setinin Naive Bayes algoritması ile sınıflandırılması sonucu elde edilen performans değerleri Şekil 6'da aynı şekilde sırasıyla veri setindeki eksik değerlerin K-nn, Random Forest ve Amelia algoritmalarını kullanılarak tamamlanmasıyla oluşan veri setlerinin sınıflandırılmasıyla ortaya çıkan performans değerlendirme kriterlerine ait değerler Şekil 7, Şekil 8 ve Şekil 9'da verilmiştir.

Şekil 6

Stokastik regresyon ile oluşan veri setine ait sınıflandırma sonuçları

```
Time taken to build model: 0.01 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         271               84.1615 %
Kappa statistic                          0.8013
Mean absolute error                      0.0727
Root mean squared error                  0.2259
Relative absolute error                 22.8044 %
Root relative squared error             56.5834 %
Total Number of Instances              322

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,955    0,055    0,821      0,955    0,883      0,853    0,983     0,974     A
                 0,887    0,015    0,922      0,887    0,904      0,886    0,946     0,929     B
                 0,719    0,016    0,920      0,719    0,807      0,775    0,920     0,882     C
                 0,795    0,049    0,838      0,795    0,816      0,759    0,967     0,914     D
                 0,867    0,065    0,754      0,867    0,806      0,761    0,966     0,915     E
Weighted Avg.    0,842    0,041    0,849      0,842    0,841      0,803    0,957     0,923

=== Confusion Matrix ===
  a  b  c  d  e   <-- classified as
 64  0  2  0  1 |  a = A
  0 47  1  2  3 |  b = B
  8  0 46  5  5 |  c = C
  4  4  0 62  8 |  d = D
  2  0  1  5 52 |  e = E
```

Şekil 7

K-nn ile oluşan veri setine ait sınıflandırma sonuçları

```
Time taken to build model: 0.02 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         285               88.5093 %
Kappa statistic                          0.8554
Mean absolute error                      0.054
Root mean squared error                  0.195
Relative absolute error                 16.9924 %
Root relative squared error             48.9155 %
Total Number of Instances              322

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,914    0,038    0,841      0,914    0,876      0,849    0,976     0,957     A
                 0,918    0,015    0,918      0,918    0,918      0,904    0,990     0,961     B
                 0,813    0,035    0,852      0,813    0,832      0,792    0,970     0,907     C
                 0,889    0,041    0,878      0,889    0,883      0,844    0,972     0,941     D
                 0,900    0,016    0,940      0,900    0,920      0,898    0,975     0,957     E
Weighted Avg.    0,885    0,030    0,886      0,885    0,885      0,855    0,976     0,944

=== Confusion Matrix ===
  a  b  c  d  e   <-- classified as
 53  0  4  0  1 |  a = A
  1 45  1  2  0 |  b = B
  7  1 52  4  0 |  c = C
  2  1  3 72  3 |  d = D
  0  2  1  4 63 |  e = E
```

Şekil 8

Random forest ile oluşan veri setine ait sınıflandırma sonuçları

```
Time taken to build model: 0.02 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         278               86.3354 %
Kappa statistic                          0.8285
Mean absolute error                      0.0562
Root mean squared error                  0.2011
Relative absolute error                 17.6587 %
Root relative squared error             50.4365 %
Total Number of Instances              322

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,897    0,024    0,910      0,897    0,904      0,878    0,984     0,969     A
                 0,940    0,022    0,887      0,940    0,913      0,897    0,991     0,957     B
                 0,862    0,078    0,737      0,862    0,794      0,741    0,973     0,930     C
                 0,765    0,033    0,886      0,765    0,821      0,770    0,980     0,948     D
                 0,897    0,015    0,929      0,897    0,912      0,894    0,981     0,956     E
Weighted Avg.    0,863    0,035    0,869      0,863    0,864      0,829    0,981     0,952

=== Confusion Matrix ===
   a  b  c  d  e   <-- classified as
  61  2  4  0  1 |  a = A
   0 47  2  1  0 |  b = B
   4  1 56  4  0 |  c = C
   2  1 13 62  3 |  d = D
   0  2  1  3 52 |  e = E
```

Şekil 9

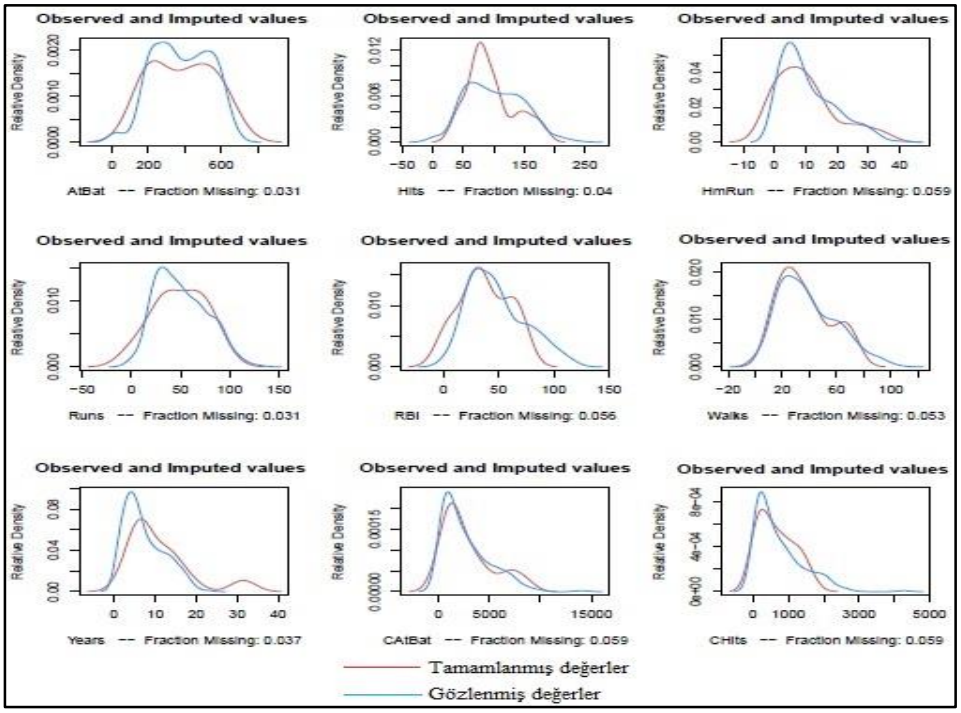
Amelia ile oluşan veri setine ait sınıflandırma sonuçları

```
Time taken to build model: 0.03 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         267               82.9193 %
Kappa statistic                          0.7853
Mean absolute error                      0.0781
Root mean squared error                  0.2366
Relative absolute error                 24.5441 %
Root relative squared error             59.3312 %
Total Number of Instances              322

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,929    0,044    0,855      0,929    0,890      0,860    0,976     0,956     A
                 0,917    0,026    0,863      0,917    0,889      0,869    0,964     0,933     B
                 0,724    0,042    0,792      0,724    0,757      0,707    0,918     0,845     C
                 0,789    0,037    0,870      0,789    0,828      0,779    0,966     0,916     D
                 0,800    0,067    0,767      0,800    0,783      0,722    0,960     0,900     E
Weighted Avg.    0,829    0,044    0,829      0,829    0,828      0,785    0,958     0,911

=== Confusion Matrix ===
 a  b  c  d  e   <-- classified as
65  1  2  1  1 |  a = A
 0 44  2  1  1 |  b = B
 8  1 42  2  5 |  c = C
 3  1  2 60 10 |  d = D
 0  4  5  5 56 |  e = E
```

Bunun dışında, iteratif bir yöntem olan Amelia algoritmasına ilişkin, iterasyonlar boyunca, her değişkene ait gözlemlere ve tamamlanmış değerlere ait sonuçlar Şekil 10'da sunulmuştur.

Şekil 10

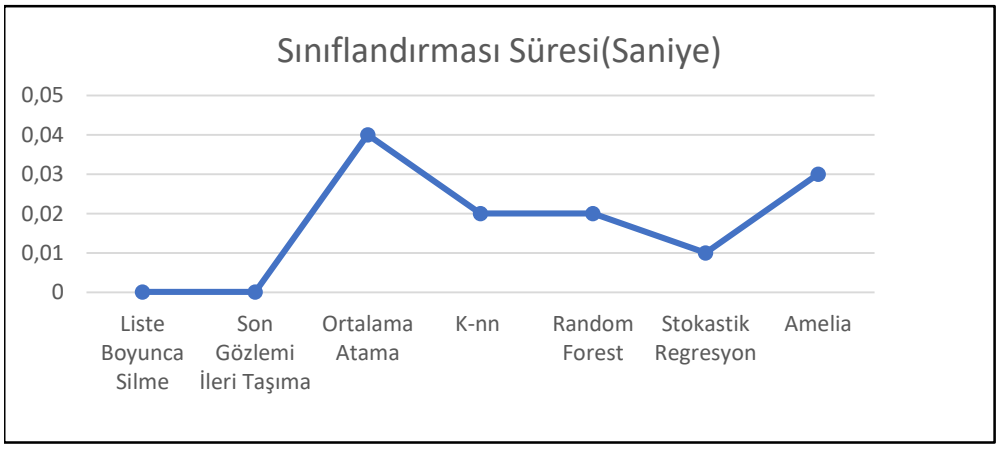
Amelia iterasyonları boyunca tamamlanan değerler



Eksik değerlerin tamamlanmasının ardından, manipüle edilen veri seti, daha önce bahsedilen yöntemlerle tamamlanmış ve yine Naive Bayes algoritmasıyla sınıflandırılmıştır. Manipüle edilerek eksiltilen veri setinin, bahsedilen yöntemlerce tamamlanıp sınıflandırma işlemi uygulanmasının ardından, performans değerlendirme kriterlerinin aldığı değerler grafiklerle gösterilmiştir.

Grafik 1

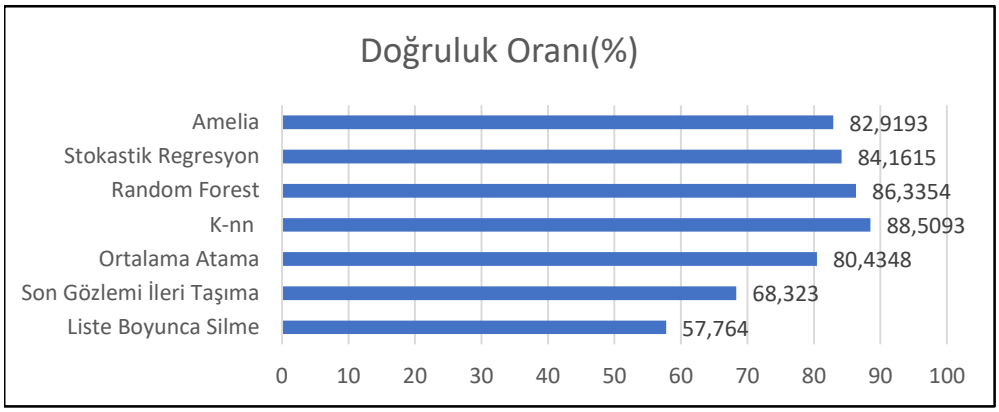
Sınıflandırma işlemlerinin sınıflandırma süresi değerleri



Grafik 1'de sınıflandırma sürelerine ait süreler incelendiğinde, sınıflandırma işlemini 0.04 saniye ile en uzun sürede tamamlayan yöntemin ortalama atama yöntemi ve en kısa sürede tamamlayan yöntemlerin ise liste boyunca silme ile son gözlemi ileri taşıma yöntemleri olduğu gözlenmektedir.

Grafik 2

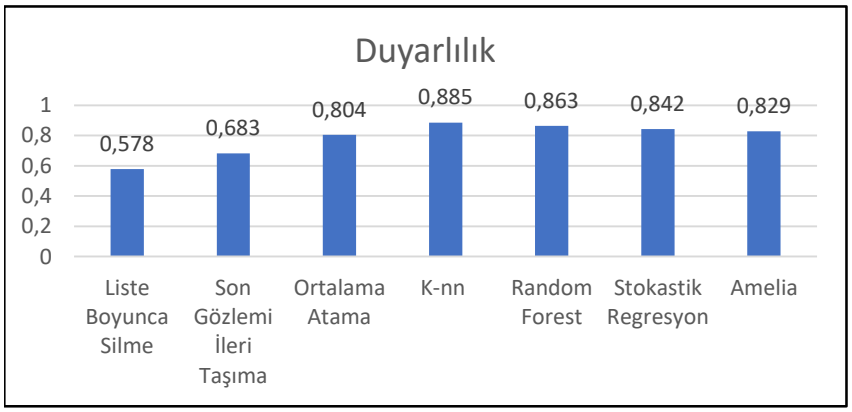
Sınıflandırma işlemlerine ait doğruluk oranı değerleri



Doğruluk oranı değerlerine bakıldığında, K-nn algoritmasının %88.5093 değeri ile tüm yöntemlerden daha yukarıda bir doğruluk oranı seviyesine sahip olurken liste boyunca silme yönteminin %57.764 ile en düşük doğruluk oranını seviyesinde kaldığı görülmektedir.

Grafik 3

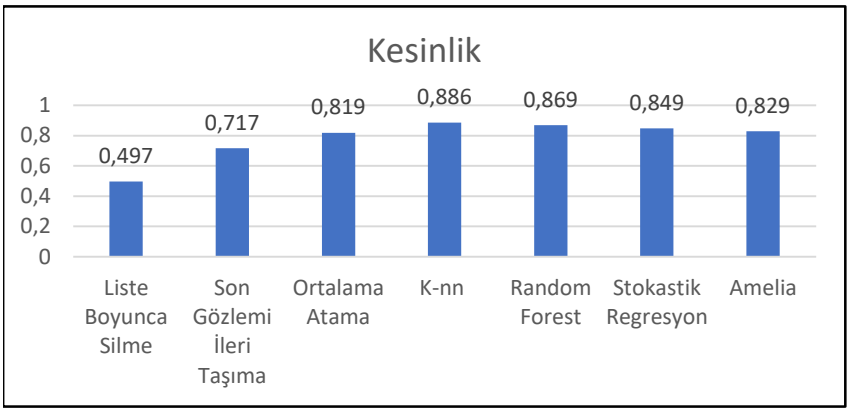
Sınıflandırma işlemlerine ait duyarlılık değerleri



Performans değerlendirme kriterlerinden duyarlılık kriterinin, değerleri bakımından en iyi performansın K-nn yöntemi ne ait olduğu, diğer tarafından nispeten en düşük performansın liste boyunca silme yöntemine ait olduğu gözlenmektedir.

Grafik 4

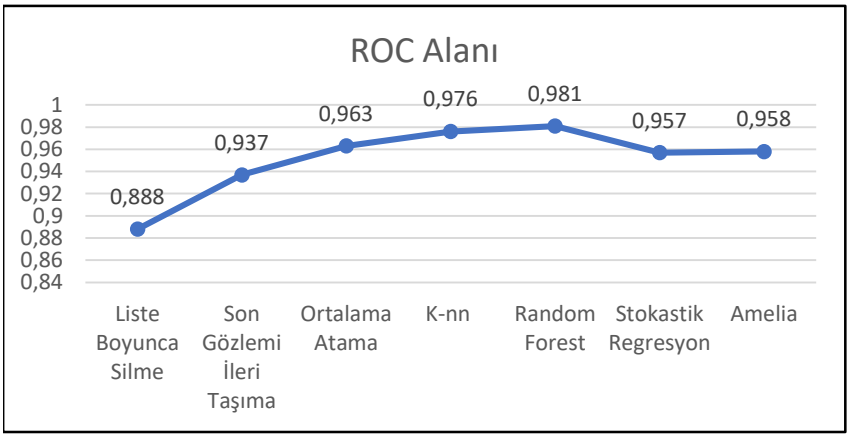
Sınıflandırma işlemlerine ait kesinlik değerleri



Grafik 4'te ki sonuçlar, en iyi kesinlik değerinin K-nn algoritmasına, yöntemler arasından en düşük kesinlik değerinin ise liste boyunca silme yöntemine ait olduğunu göstermektedir.

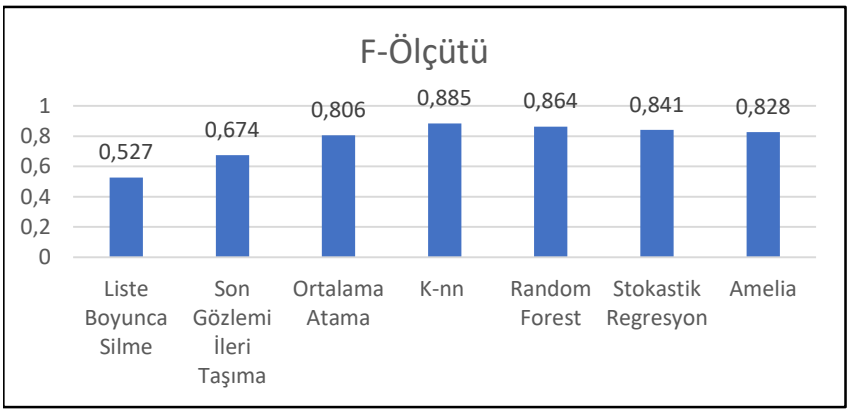
Grafik 5

Sınıflandırma işlemlerine ait ROC alanı değerleri



Grafik 6

Sınıflandırma işlemlerine ait F-ölçütü değerleri



Grafik 5'te en yüksek ROC alanı değerlerine ilişkin en iyi yöntemin Random Forest yöntemi olduğu sonucu çıkmıştır. F- ölçütü kriterlerine bakıldığında ise Grafik 6'da gözlendiği gibi en iyi yöntemin K-nn olduğu gözlenmiştir. Elde edilen sonuçların ardından, her bir yönteme ait performans değerleri Tablo 3'te özetlenmiştir.

Tablo 3

Sınıflandırma işlemlerine ilişkin performans değerleri

Algoritmalar	**Sınıflandırması Süresi(Saniye)**	**Doğruluk(%)**	**Kesinlik**	**Duyarlılık**	**F-Ölçütü**	**ROC Alanı**
Orijinal Veri Seti	0.03	87.5776	0.877	0.876	0.876	0.961
Liste Boyunca Silme	0.00	57.7640	0.497	0.578	0.527	0.888
Son Gözlemi İleri Taşıma	0.00	68.3230	0.717	0.683	0.674	0.937
Ortalama Atama	0.04	80.4348	0.819	0.804	0.806	0.963
K-nn	0.02	88.5093	0.886	0.885	0.885	0.976
Random Forest	0.02	86.3354	0.869	0.863	0.864	0.981
Stokastik Regresyon	0.01	84.1615	0.849	0.842	0.841	0.957
Amelia	0.03	82.9193	0.829	0.829	0.828	0.958

8. Sonuç

Bu çalışmada, "Hitters" veri seti kullanılmış ve R programlama dili ile veri seti manipüle edilerek %5 oranında rastgele şekilde eksiltilmiştir. Eksiltilen veri seti, yine R programlama diliyle Liste Boyunca Silme, Son Gözlemi İleri Taşıma, Ortalama Atama, K-nn, Random Forest, Stokastik Regresyon ve Amelia algoritmalarıyla tamamlanmıştır. Eksik veri tamamlama operasyonlarının ardından, orijinal veri seti ve veri setinin eksik veri tamamlama algoritmalarıyla tamamlanmış halleri WEKA paket programı kullanılarak Naive Bayes algoritması ile sınıflandırılmıştır. Sınıflandırma işlemi, sadece Naive Bayes algoritması ile yapıldığı için sınıflandırma sonuçları üzerinde farkı oluşturan unsur, sadece eksik verileri tamamlama algoritmalarıdır. Böylece, algoritmaların eksik veri tamamlama performansları ışığında sınıflandırma operasyonlarına etkilerinin kıyaslanması mümkün hale gelmiştir. Bu kıyaslama, sınıflandırması süresi(saniye), doğruluk(%), kesinlik, duyarlılık, F-Ölçütü ve ROC alanı şeklindeki performans değerlendirme kriterlerinin aldığı değerler ölçeğinde yapılmıştır.

0.04 saniye ile sınıflandırma işleminin en uzun süre aldığı veri seti, eksik değerlerin ortalama atama yöntemiyle tamamlandığı veri setidir. Diğer yöntemlerle tamamlanan veri setlerinde sınıflandırma işlemleri, en fazla orijinal veri setine ait sınıflandırma süresi kadar sürede gerçekleşmiştir. Liste boyunca silme ve son gözlemi ileri taşıma yöntemleriyle tamamlanarak oluşmuş veri setlerine ait sınıflandırma süresi diğer yöntemlere göre en kısa süren sınıflandırma operasyonları olmuştur. Bu durum, çok büyük hacimli veri setleriyle yapılan araştırmalarda, orijinal veri setinin sınıflandırma süresiyle kıyaslandığında, makine öğrenmesi yöntemlerinin ve temel yöntemlerden liste boyunca silme ile son gözlemi ileri taşıma yöntemlerinin daha az zaman maliyeti yaratacağını göstermektedir.

Doğruluk oranı kriteri değerlerine bakıldığında, genel olarak, eksik değerlerin makine öğrenmesi algoritmalarıyla tamamlanması ile oluşturulan veri setlerine ilişkin sınıflandırmalardaki doğruluk oranı performansının, diğer yöntemlere ait doğruluk oranlarına göre anlamlı ölçüde daha iyi

olduğu gözlenmektedir. Doğruluk oranına ilişkin diğer bir sonuç da makine öğrenmesi algoritmalarına ilişkin doğruluk oranlarının, orijinal veri setine ait sınıflandırmadaki doğruluk oranına çok yakın olması ve hatta K-nn algoritmasına ait doğruluk oranı gibi daha yüksek bir doğruluk oranı ortaya koymasıdır. Böylece, makine öğrenmesi yöntemleri ile eksik verilerin tamamlanmasının daha fazla veri kaydının doğru etiketlenmesi sağlayacağını ortaya çıkarmıştır.

Kesinlik, duyarlılık ve F-ölçütü kriterlerinin aldığı değerler incelendiğinde, doğruluk kriterindeki sonuçlara paralel olarak makine öğrenmesi yöntemlerinin ilgili performans değerlerine katkısı temel yöntemlerin katkısından oldukça yüksektir. Bu kriterler bazında, makine öğrenmesi yöntemlerinden K-nn en iyi performans katkısını sunan yöntem olmuştur. Böylece aynı yöntemle elde edilmiş olan analiz sonuçlarının yanı sıra birbirine yakın sınıflandırma sonuçlarının elde edilmesi anlamında makine öğrenmesi yöntemleri oldukça başarılı olduğu görülmüştür. Benzer şekilde, sınıflandırma işlemlerine ait ROC alanı değerlerinde, makine öğrenmesi yöntemleri üstün bir performans sergilemiş ve Random Forest algoritmasının ön plana çıktığı görülmüştür.

Temel eksik veri tamamlama yöntemleri arasından ise sınıflandırma operasyonuna yaptığı katkı anlamında, ortalama atama yönteminin öne çıkması söz konusu olsa da sınıflandırma işlemlerine ait performans değerlendirme kriterlerinin aldığı değerlere genel olarak bakıldığında, makine öğrenmesi yöntemleriyle eksik verilerin tamamlanmasının, temel yöntemlerle eksik veri tamamlanmasına göre sınıflandırma performansına anlamlı ve oldukça üstün bir katkı sağladığı gözlenmektedir. Yapılan uygulamada, performans katkısı anlamında, sırasıyla K-nn, Random Forest, Stokastik Regresyon ve Amelia algoritmaları şeklinde bir üstünlük gözlenmiştir. Böylece, K-nn ve Random Forest algoritmalarının ön plana çıktığı tespit edilmiştir.

Başka bir açıdan sonuçlara bakıldığında, makine öğrenmesi yöntemleriyle eksik verilerin tamamlandığı veri setlerine ait sınıflandırma performanslarının, veri setinin orijinal haline ilişkin sınıflandırma performansına çok yakın olması göze çarpmaktadır. Bu durum, makine öğrenmesi yöntemleriyle tamamlanan eksik değerlerin veri yapısına oldukça iyi uyum sağladığı ve veri setinin karakteristiğini daha iyi yansıttığı sonucunu göz önüne sermektedir. Ayrıca eksik değerlerin, K-nn algoritması ile eksik verilerin tamamlanmasından sonra yapılan sınıflandırma performansının, orijinal veriye yapılan sınıflandırma performansından anlamlı şekilde üstün olduğu görülmüştür. Böylece, ilgili algoritmanın sınıflandırma performansını güçlendirdiği sonucu elde edilmiştir.

9. Araştırmanın etik yönü

Yapılan bu çalışmada "Yükseköğretim Kurumları Bilimsel Araştırma ve Yayın Etiği Yönergesi" kapsamında uyulması belirtilen tüm kurallara uyulmuştur. Yönergenin ikinci bölümü olan "Bilimsel Araştırma ve Yayın Etiğine Aykırı Eylemler" başlığı altında belirtilen eylemlerden hiçbiri gerçekleştirilmemiştir.

Bu araştırmanın etik kurul izni gerektirmeyen araştırmalardan olduğunu beyan ederiz.

10. Çıkar çatışması beyanı

Bu çalışmada, sonuçları veya yorumları etkileyebilecek herhangi bir maddi veya diğer asli çıkar çatışması olmadığını beyan ederiz.

11. Katkı oranı

Yazarların makaleye eşit oranda katkı sağlamış olduğunu beyan ederiz.

KAYNAKÇA

Abidin, N. Z., Ismail, A. R., & Emran, N. A. (2018). Performance analysis of machine learning algorithms for missing value imputation. *International Journal of Advanced Computer Science and Applications*, 9(6), 442-447. https://dx.doi.org/10.14569/IJACSA.2018.090660

Alamoodi, A. H., Zaidan, B. B., Zaidan, A. A., Albahri, O. S., Chen, J., Chyad, M. A., Garfan, S., & Aleesa, A. M. (2021). Machine learning-based ımputation soft computing approach for large missing scale and non-reference data ımputation. *Chaos, Solitons & Fractals*, 151, 111236. https://doi.org/10.1016/j.chaos.2021.111236

Allison, P. D. (2009). *Missing data, handbook of quantitative methods in psychology* (Editor: Roger E. Millsap ve Alberto Maydeu-Olivares), Sage Publications.

Baraldi, A. N., & Enders, C. K. (2010). An introduction to modern missing data analyses. *Journal of School Psychology*, 48(1), 5-37. https://doi.org/10.1016/j.jsp.2009.10.001

Bi, Q., Goodman, K. E., Kaminsky, J., & Lessler, J. (2019). What is machine learning? A primer for the epidemiologist. *American Journal of Epidemiology*, 188(12), 2222-2239. https://doi.org/10.1093/aje/kwz189

Brynjolfsson, E., & Mitchell, T. (2017). What can machine learning do? Workforce implications. *Science*, 358(6370), 1530-1534. https://doi.org/10.1126/science.aap8062

Dogan, C. D. (2017). Applying bootstrap resampling to compute confidence intervals for various statistics with R. *Eurasian Journal of Educational Research*, 17(68), 1-18. https://dergipark.org.tr/en/download/article-file/623638

Doğru, F. Z., Bulut, Y. M., & Arslan, O. (2016). Finite mixtures of matrix variate t-distributions. *Gazi University Journal of Science*, 29(2), 335-341. https://dergipark.org.tr/tr/download/article-file/225490

Donders, A. R. T., Van Der Heijden, G. J., Stijnen, T., & Moons, K. G. (2006). A gentle introduction to imputation of missing values. *Journal of Clinical Epidemiology*, 59(10), 1087-1091. https://doi.org/10.1016/j.jclinepi.2006.01.014

Durmuş, B., & Güneri, Ö. İ. (2019). Data mining with R: An applied study. *International Journal of Computing Sciences Research*, 3(3), 201-216. https://doi.org/10.25147/ijcsr.2017.001.1.34

Durmuş, B., & Güneri, Ö. İ. (2021). A classification study for re-determination of the geographical regions: The case of Turkey. *International Journal of Applied Mathematics Electronics and Computers*, 9(4), 97-102. https://doi.org/10.18100/ijamec.988273

Emmanuel, T., Maupong, T., Mpoeleng, D., Semong, T., Mphago, B., & Tabona, O. (2021). A survey on missing data in machine learning. *Journal of Big Data*, 8(1), 1-37. https://doi.org/10.1186/s40537-021-00516-9

Enders, C. K. (2022). *Applied missing data analysis*. Guilford Publications.

Honaker, J., King, G., & Blackwell, M. (2011). Amelia II: A program for missing data. *Journal of Statistical Software*, 45(7), 1–47. https://doi.org/10.18637/jss.v045.i07

Jadhav, A., Pramod, D., & Ramanathan, K. (2019). Comparison of performance of data imputation methods for numeric dataset. *Applied Artificial Intelligence*, 33(10), 913-933. https://doi.org/10.1080/08839514.2019.1637138

Jerez, J. M., Molina, I., García-Laencina, P. J., Alba, E., Ribelles, N., Martín, M., & Franco, L. (2010). Missing data imputation using statistical and machine learning methods in a real breast cancer problem. *Artificial Intelligence In Medicine*, 50(2), 105-115. https://doi.org/10.1016/j.artmed.2010.05.002

Kenyhercz, M. W., & Passalacqua, N. V. (2016). Missing data imputation methods and their performance with biodistance analyses. *Biological Distance Analysis* (pp. 181-194). Academic Press. https://doi.org/10.1016/B978-0-12-801966-5.00009-3

Köse, I. A., & Öztemur, B. (2014). Kayıp veri ele alma yöntemlerinin t-testi ve ANOVA parametreleri üzerine etkisinin incelenmesi. *Abant İzzet Baysal Üniversitesi Eğitim Fakültesi Dergisi,* 14(1), 400-412. https://dergipark.org.tr/tr/download/article-file/16769

Mahesh, B.(2019). Machine learning algorithms-A review. *International Journal of Science and Research*, 9(1), 381-386.

Oprea, C. (2014). Performance evaluation of the data mining classification methods. *Information Society and Sustainable Development*, 1(Special Issue), 249-253. https://www.utgjiu.ro/revista/ec/pdf/2014-04.Special/45_Oprea%20Cristina.pdf

Palanivinayagam, A., & Damaševičius, R. (2023). Effective handling of missing values in datasets for classification using machine learning methods. *Information*, 14(2), 92. https://doi.org/10.3390/info14020092

Raja, P. S., & Thangavel, K. J. S. C. (2020). Missing value imputation using unsupervised machine learning techniques. *Soft Computing*, 24(6), 4361-4392. https://doi.org/10.1007/s00500-019-04199-6

Schaffer, J. L. (1997). *Analysis of incomplete multivariate data*. Chapman&Hall**.**

Tang, F., & Ishwaran, H. (2017). Random forest missing data algorithms. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 10(6), 363-377. https://doi.org/10.1002/sam.11348

Thomas, T., & Rajabi, E. (2021). A systematic review of machine learning-based missing value ımputation techniques. *Data Technologies and Applications*, 55(4), 558-585. https://doi.org/10.1108/DTA-12-2020-0298

Vangipuram, R., Gunupudi, R. K., Puligadda, V. K., & Vinjamuri, J. (2020). A machine learning approach for imputation and anomaly detection in iot environment. *Expert Systems*, 37(5), e12556. https://doi.org/10.1111/exsy.12556

Vembandasamy, K., Sasipriya, R., & Deepa, E. (2015). Heart diseases detection using naive bayes algorithm. *International Journal of Innovative Science, Engineering & Technology*, 2(9), 441-444. https://ijiset.com/vol2/v2s9/IJISET_V2_I9_54.pdf

Zhang, Z. (2016). Introduction to machine learning: K-nearest neighbors. *Annals of Translational Medicine*, 4(11), 218-224. https://doi.org/10.21037/atm.2016.03.37