INTERNATIONAL SCIENTIFIC AND VOCATIONAL JOURNAL (ISVOS JOURNAL)

Vol.: 2 Issue: 2 Date: December 2018 Received: 28.11.2018 Accepted: 12.12.2018 Final Version: 25.12.2018

ISVOS Journal 2018, 2(2):80-85

A Comparison of Five Methods for Missing Value Imputation in Data Sets

Pınar Cihan[a,1]

[a]Namık Kemal Üniversitesi, Çorlu Mühendislik Fakültesi, Bilgisayar Mühendisliği Bölümü, Tekirdağ

Abstract

The missing values in the data sets do not allow for accurate analysis. Therefore, the correct imputation of missing values has become the focus of attention of researchers in recent years. This paper focuses on a comparison of most reliable and up to date estimation methods to imputing the missing values. Imputation of missing values has a very high priority because of its impact on next pre-processing, data analysis, classification, clustering, etc. Root mean square error (RMSE) value, classification accuracy and execution time are used to evaluate the performances of most popular five methods (mean, k-nearest neighbors, singular value decomposition, bayesian principal component analysis and missForest). When RMSE and classification accuracy values of methods were compared, it has observed that missForest method outperformed other methods in all datasets.

Keywords: "Missing value imputation, k-nearest neighbor, singular value decomposition, bayesian principal component analysis, missForest"

1. Giriş

Bilimsel araştırmalarda üzerinde çalışılmak istenilen veriler her zaman istenildiği gibi eksiksiz bir şekilde toplanılamayabilir. Günümüzde eksiksiz veri seti pek mümkün olmayıp, kayıp değerler birçok gerçek dünyada görülen yaygın bir dezavantajdır. Veri setlerindeki eksik değerler; değerin veri kümesine kayıt edilmemesi, veri kümesindeki aykırı, gürültülü veya tutatsız değerlerin veri kümesinden silinmesi sonucunda oluşabilmektedir. Veri kümelerinde değeri olmayan bu eksiklikler, kayıp/eksik değer olarak adlandırılmaktadır. Örneğin anket çalışmalarında; katılımcıların bazı sorulara cevap vermeme veya verilen cevapların bir kısmının yanlış olduğu bilindiğinde veri eksikliği görülmektedir. Endüstriyel bir süreçte; izleme veya veri toplayıcı ekipmanlarının arızalanması, veri toplayıcıları ile merkezi yönetim sistemi arasındaki iletişimin kesintiye uğraması, arşivleme sistemi sırasında başarısızlık nedeniyle eksik veriler ortaya çıkabilir. Otomatik konuşma tanımada; çok yüksek düzeyde gürültüyle bozulmuş konuşma örnekleri eksik veri olarak değerlendirilir. DNA mikrodizileri ile biyoloji araştırmasında; gendeki çizik veya kontamine numuneler gibi çeşitli nedenlerden ötürü gen verileri eksik olabilir. Tıbbi teşhiste; örneğin hekim kesin sonucu verecek veya tanı ile alakalı olmayan test isteyebilir veya ölçmenin zor/zararlı olduğu durumlarda veri eksik olabilir[1]. Hayvancılıkta; hayvanların tartım, ölçüm ve diğer işlemler için bir araya getirilmesi veya kanlarının alınarak bunların analiz edilmesi oldukça zahmetli ve masraflı olmasından veri eksikliği sıklıkla görülmektedir[2].

Veri setlerindeki eksik değerler yapılacak analizler için sorun teşkil etmektedir. Çünkü klasik ve modern istatistiksel yöntemlerin hemen hemen hepsi veri setinin eksiksiz olduğu varsayımı altında geliştirilmiştir [3]. Bu nedenle son yıllarda eksik değerlerin tahmin edilmesi popüler bir konu haline gelmiştir. Bunun temel sebebi araştırmacıların kaliteli veriye ulaşma isteğidir. Çünkü eksik değerler düzgün bir şekilde ele alınamaz ise araştırma sonuçlarının geçerliliği azalabileceği gibi bazı durumlarda araştırma başarısızlığa bile uğrayabilmektedir[4-7].

Eksik değerleri tamamlamak için yöntem seçilmeden önce, veri setindeki kayıp veri mekanizması göz önüne alınarak uygun yöntem ile eksik değerler tamamlanmalıdır. Aksi halde, kayıp veri çalışması, mekanizma tarafından uygun olmayan yöntemle çözümlenebilir ve yanlı kestirimler elde edilebilir. Little ve Rubin kayıp veri mekanizmasını üç temel kategoriye ayırmıştır. Bunlar; Tamamıyla Rassal Olarak Kayıp (Missing Completely at Random, MCAR), Rassal Olarak Kayıp (Missing at Random, MAR) ve Rassal Olmayan Kayıp (Missing not at Random, MNAR)[8].

Bu çalışmada, Irvine Machine Learning Repository of the University of California (UCI) veri tabanından alınan dört farklı medikal veri seti kullanılmıştır. Veri setlerinde yaklaşık %5, %10, %15, %20 ve %25 oranlarında MCAR yapay veri

[1] Corresponding author. Tel.: +0-282-250-2456; fax: +0-282-250-9924 .
E-mail address: pkaya@nku.edu.tr

oluşturulmuştur. Daha sonra veri setlerindeki bu eksik değerler ortalama, en yakın k-komşu (k-nearest neighbor, kNN)[9], tekil değer ayrışımı (singular value decomposition, SVD)[9], bayes temel bileşen analizi (bayesian principal component analysis, bPCA)[10] ve MissForest[11] eksik değer hesaplama yöntemleri ile tamamlanarak eksik değerleri tamamlamada en başarılı yöntem belirlenmiştir. Yöntemlerin başarıları karşılaştırılırken karekök ortalama hata, sınıflandırma doğruluğu ve çalışma süresi kriterleri göz önüne alınmıştır. Çalışma R programlama kullanılarak gerçekleştirilmiştir.

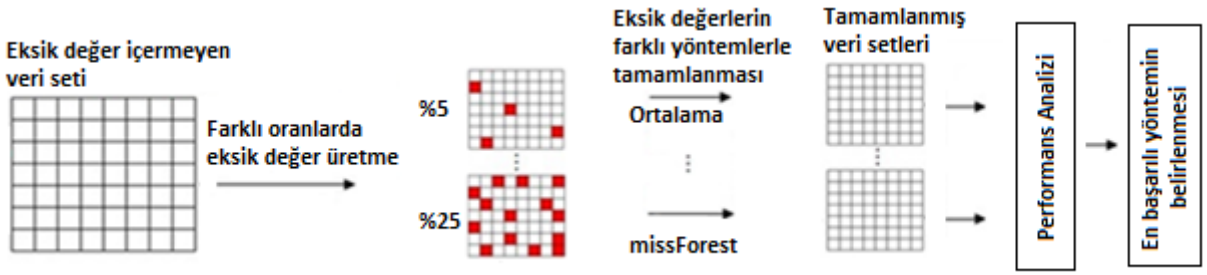
2. Materyal ve Yöntemler

2.1. Eksik Değer Tamamlama Yöntemleri

Günümüzde veri setindeki eksik değerleri tamamlamada hala belli bir algoritmanın üstünlüğü kanıtlanamamıştır. Bu nedenle veri setindeki eksik değerleri tamamlama günümüzün güncel problemlerinden biri olup, bu problemin üstesinden gelebilmek için birçok yöntem geliştirilmiş ve geliştirilmeye devam etmektedir[12].

Veri kümelerindeki eksik değerlerin nasıl tamamlandığı önemli bir konu olup veri madenciliği sonuçlarını etkilemektedir. Bu çalışmada, veri setindeki eksik değerleri tamamlamada en başarılı yöntemi belirlemek için 5 farklı yöntem (ortalama, kNN, SVD, bPCA, ve missForest) ile eksik değerler tamamlanmıştır. Bu beş yöntem kısaca şunlardır; Ortalama (mean imputation): Eksik değerler, eksik olmayan değerlerin ortalaması ile tamamlanır. kNN: En yakın komşu yöntemi aslında bir sınıflandırma yöntemi olup kayıp değerlerin çözümlenmesinde bu sınıflandırma mantığı kullanılmaktadır. En yakın k-komşu algoritması ile kayıp değer atama yöntemi, gözlemlerin birbirlerine olan yakınlıkları üzerine kuruludur. Eksik değer içermeyen değişkenler arasında mesafe ölçümü yaparak 'k' en yakın gözlemin ortalaması ile tamamlanır. Yani eksik olmayan özelliklere en çok benzeyen k tane özellik seçilir[13]. Bu yöntemi uygulamak için R'da 'VIM' paketindeki 'kNN' fonksiyonu kullanılmıştır. SVD: Yinelemeli yöntemlerden olan bu yöntemi uygulamak için R'da 'bcv' paketindeki 'impute.svd' fonksiyonu kullanılmıştır. BPCA: Temel bileşen analizi (Principal Component Analysis, PCA) için yinelemeli iyileştirme algoritması olan beklenti –ençoklama (Expectation – Maximization, EM) algoritması ile Bayesçi modelini birleştirir [10]. Bu yöntemi uygulamak için R'da 'pcaMethods' paketinde, 'pca' fonksiyonundaki 'bpca' yöntemi kullanılmıştır. MissForest: Eksik veriyi tamamlarken rastgele orman yöntemini kullanır. Veri kümesindeki değişkenlerin geri kalanını kullanarak her değişken için rastgele bir orman modeli oluşturur ve bu değişkenin eksik değerlerini tahmin etmek için onu kullanır[13]. Bu yöntemi uygulamak için R'da 'missForest' paketindeki 'missForest' fonksiyonu kullanılmıştır.

Çalışmada eksik değerleri tamamlamada izlenilen süreç Şekil 1'de sunulmuştur. Eksik değer içermeyen veri setleri üzerinde %5, %10, %15, %20 ve %25 oranlarında MCAR mekanizmasına sahip yapay eksik değerler üretilmiştir. Farklı oranlarda eksik değer içeren bu beş veri setindeki eksik değerler ortalama, kNN, SVD, bPca ve missForest yöntemleri ile tamamlanmıştır. Daha sonra bu yöntemlerin performanslarını değerlendirmek için RMSE, sınıflandırma doğruluğu ve yöntemlerin çalışma süreleri karşılaştırılmıştır.



Şekil 1. Akış şeması

2.2. Veri Seti

Algoritma başarısını doğrulamak için, çeşitli boyutlarda ve çeşitli örnek büyüklüklerine sahip dört farklı medikal veri kümesi (liver disorders, diabetes, breast cancer ve WDBC) UCI veri tabanından seçilmiştir[14]. Bu veri setleri literatürde sıkça kullanılmaktadır[15-18]. Seçilen veri setleri orta büyüklükte olup hepsi kategorik olmayan özelliklere sahiptir. Veri setlerinin temel özellikleri Tablo 1'de özetlenmiştir.

Tablo 1. Veri setlerinin özellikleri

Veri seti	Özellik	Örnek sayısı	Sınıf etiketi sayısı
Karaciğer bozuklukları(Liver disorders)	6	345	2
Diyabet (Diabetes)	8	768	2
Meme kanseri (Breast cancer)	9	683	2
Meme kanseri Wisconsin (WDBC)	30	569	2

2.3. Değerlendirme Kriterleri

Eksik değer hesaplama yöntemlerinin başarıları karşılaştırılırken RMSE, sınıflandırma doğruluğu ve yöntemlerin çalışma süreleri ölçütleri kullanılmıştır. Veri setlerindeki eksik değerler ortalama, kNN, SVD, bPca ve missForest yöntemleriyle tamamlandıktan sonra rasgele orman (Random Forest, RF) algoritması kullanılarak sınıflandırma yapılmıştır. Sınıflandırma yapılırken 10-kat çapraz geçerlilik kullanılmış olup 50 tekrar yapılmıştır. kNN yönteminde k = 5 seçilmiştir. RF yönteminde eğitim 500 ağaç ile gerçekleştirilmiştir. Tüm deneyler R programı kullanılarak gerçekleştirilmiştir.

3. Deneysel Sonuçlar

UCI veri tabanından alınan dört farklı veri seti üzerinde %5, %10, %15, %20 ve %25 oranlarında MCAR mekanizmasına sahip yapay eksik değerler üretilmiştir. Veri setlerindeki bu eksik değerleri tamamlamada hangi yöntemin daha az hata yaptığını gözlemleyebilmek için RMSE değerleri hesaplanmıştır. Elde edilen RMSE sonuçları Tablo 2'de özetlenmiştir.

RMSE, hata miktarının büyüklüğünü ifade ettiği için bu değerlerin düşük olması yöntemin başarısını arttırmaktadır. RMSE değeri hesaplanırken eksik değer içermeyen veri seti (UCI'dan alınmış orijinal eksiksiz veri kümesi) ile tamamlanmış veri seti karşılaştırılarak hata hesaplanır. Tablo 2 incelendiğinde veri setindeki eksiklik oranı arttıkça RMSE değerinde de genel olarak yükselme olduğu görülmektedir. Normal şartlar altında veri setindeki eksiklik oranı arttıkça hatanın artması beklenir ancak eksik değerler rastgele üretildiği için hangi özelliklerde eksiklik olduğu da sonucu etkilemektedir. Eğer üretilen eksiklik küçük değerlere sahip özelliklerde fazla ise RMSE daha küçük, büyük değerlere sahip özelliklerde daha fazla ise RMSE'e sonucu daha büyük çıkacaktır. Bu nedenle çalışma sonucunda ortalama değerler de ele alınmıştır. Yöntemlerin ortalama RMSE sonuçlarına bakıldığında tüm veri setleri için en düşük RMSE değerine sahip, yani en başarılı yöntemin missForest olduğu görülmektedir.

Tablo 2. RMSE sonuçları

	RMSE				
Veri seti ve eksiklik oranları	**Ortalama**	**kNN**	**SVD**	**bPca**	**missForest**
Veri seti 1: Liver disorders					
%5	22,771	21,029	25,745	29,656	17,504
%10	22,831	20,452	24,988	23,545	19,212
%15	21,823	19,891	24,412	24,015	17,619
%20	20,946	20,847	24,323	25,538	16,805
%25	27,397	23,465	28,338	25,197	21,731
Ortalama	23,154	21,137	25,561	25,590	**18,574**
Veri seti 2: Diabetes					
%5	55,527	41,173	55,694	46,920	33,736
%10	52,302	38,188	54,115	48,808	35,767
%15	47,724	42,595	53,987	45,265	39,109
%20	50,977	44,349	55,061	47,702	40,143
%25	48,451	41,586	55,736	45,822	38,246
Ortalama	50,996	41,578	54,919	46,903	**37,400**
Veri seti 3: Breast cancer					
%5	3,293	1,927	1,843	1,950	1,559
%10	3,342	1,945	1,789	1,929	1,586
%15	3,467	1,948	1,864	2,024	1,659
%20	3,394	1,975	1,857	2,023	1,644
%25	3,255	2,016	1,954	2,099	1,740
Ortalama	3,350	1,962	1,861	2,005	**1,638**
Veri seti 4: WDBC					
%5	447,333	440,550	1386,417	443,140	65,867
%10	327,997	312,710	1220,476	322,339	99,457
%15	340,939	323,779	1181,179	329,031	246,780

%20	336,117	324,592	1064,229	331,746	220,764
%25	305,562	292,324	917,993	307,611	203,328
Ortalama	351,590	338,791	1154,059	346,773	**167,239**

Dört farklı veri setindeki eksik değerler 5 yöntem ile tamamlandıktan sonra rasgele orman yöntemi ile sınıflandırılmış olup sınıflandırma doğruluğu (classification accuracy) sonuçları Tablo 3'de özetlenmiştir. Tüm yöntemlerin doğruluk değerleri, orijinal veri kümesine yakındır ve kayıp oran arttıkça doğruluk değerinde düşüş olmaktadır. Örneğin liver disorders veri kümesinde, ortalama, kNN, SVD, bPca ve missForest yakın doğruluk değerine sahip olup sırasıyla; 0.6985, 0.7100, 0.6989, 0.7082 ve 0.7138'dir. Veri setlerindeki eksik değer oranı %25'e ulaştığında, doğruluk değerleri sırasıyla 0.6580, 0.6443, 0.6457, 0.6352 ve 0.6995'e düşmektedir. Tüm yöntemlerin sonuçlarına genel olarak bakıldığında, bütün veri setlerinde missForest yöntemi diğer yöntemlerden daha başarılı sonuçlar üretmiştir.

Tablo 3. Sınıflandırma doğruluğu sonuçları

Veri seti ve eksiklik oranları	**Doğruluk (50 deney sonucu; ortalama ± standart sapma)**				
	Ortalama	**kNN**	**SVD**	**bPca**	**missForest**
Veri seti 1: Liver disorders					
0	0,6985 ± 0,0415	0,7100 ± 0,0408	0,6989 ± 0,0359	0,7082 ± 0,0338	0,7138 ± 0,0327
%5	0,6829 ± 0,0432	0,6933 ± 0,0422	0,6786 ± 0,0374	0,6685 ± 0,0470	**0,7101 ± 0,0388**
%10	0,6872 ± 0,0403	0,6878 ± 0,0282	0,6618 ± 0,0383	0,6685 ± 0,0411	**0,7086 ± 0,0376**
%15	0,6941 ± 0,0414	0,6652 ± 0,0340	0,6732 ± 0,041	0,6652 ± 0,0386	**0,7065 ± 0,0311**
%20	0,6819 ± 0,0397	0,6588 ± 0,0363	0,6460 ± 0,0394	0,6584 ± 0,0502	**0,7049 ± 0,0286**
%25	0,6580 ± 0,0447	0,6443 ± 0,0240	0,6457 ± 0,0347	0,6352 ± 0,0420	0,6995 ± 0,0392
Veri seti 2: Diabetes					
0	0,7558 ± 0,0239	0,7552 ± 0,024	0,7553 ± 0,0268	0,7581 ± 0,0233	0,7631 ± 0,02030
%5	0,7516 ± 0,0219	0,7538 ± 0,0297	0,7427 ± 0,0209	0,7471 ± 0,0198	**0,7588 ± 0,0266**
%10	0,7403 ± 0,0239	0,7512 ± 0,0271	0,7306 ± 0,0257	0,7334 ± 0,0279	**0,7514 ± 0,0272**
%15	0,7397 ± 0,0232	0,7514 ± 0,0227	0,7350 ± 0,0204	0,7354 ± 0,0218	0,7464 ± 0,02810
%20	0,7285 ± 0,0236	0,7420 ± 0,025	0,7113 ± 0,0227	0,7190 ± 0,0224	**0,7428 ± 0,0232**
%25	0,7231 ± 0,0300	0,7276 ± 0,0229	0,7083 ± 0,0271	0,7039 ± 0,0248	**0,7336 ± 0,0243**
Veri seti 3: Breast cancer					
0	0,9704 ± 0,0098	0,9693 ± 0,0097	0,9693 ± 0,0138	0,9683 ± 0,0108	0,9717 ± 0,00870
%5	0,9666 ± 0,0115	0,9671 ± 0,0104	0,9681 ± 0,0113	0,9678 ± 0,0096	**0,9709 ± 0,0104**
%10	0,9650 ± 0,0108	0,9640 ± 0,0091	0,9671 ± 0,0122	0,9684 ± 0,0107	**0,9688 ± 0,0112**
%15	0,9642 ± 0,0113	0,9634 ± 0,0098	0,9654 ± 0,0107	0,9637 ± 0,0107	**0,9662 ± 0,0103**
%20	0,9601 ± 0,0135	0,9620 ± 0,0114	0,9640 ± 0,0101	0,9624 ± 0,0128	0,9601 ± 0,01150
%25	0,9589 ± 0,0129	0,9571 ± 0,0101	0,9574 ± 0,0105	0,9586 ± 0,0125	**0,9592 ± 0,0109**
Veri seti 4: WDBC					
0	0,9547 ± 0,0137	0,9535 ± 0,0131	0,9557 ± 0,0133	0,9587 ± 0,0141	0,9564 ± 0,01570
%5	0,9531 ± 0,0179	0,9522 ± 0,0152	0,9520 ± 0,0155	0,9519 ± 0,0178	**0,9537 ± 0,0187**
%10	0,9528 ± 0,0139	0,9518 ± 0,0133	0,9515 ± 0,0163	0,9453 ± 0,0172	0,9518 ± 0,01810
%15	0,9467 ± 0,0187	0,9501 ± 0,0138	0,9479 ± 0,0181	0,9365 ± 0,0124	**0,9506 ± 0,0177**
%20	0,9431 ± 0,0166	0,9452 ± 0,0154	0,9351 ± 0,0187	0,9351 ± 0,0156	**0,9498 ± 0,0162**
%25	0,9395 ± 0,0161	0,9426 ± 0,0161	0,9298 ± 0,0199	0,9377 ± 0,0162	**0,9487 ± 0,0168**

Veri kümelerindeki eksik değerleri tamamlamada yöntemlerin çalışma süreleri Tablo 4'de verilmiştir. Veri setlerindeki eksiklik oranı arttıkça, bu değerlerin tamamlanması daha fazla zaman almaktadır. Dört veri setinde de eksik değerleri tamamlamada en yavaş yöntemin missForest olduğu görülmektedir. Çünkü bu yöntem karar ağaçlarından bir orman oluşturmaktadır. Ayrıca çalışmada, yöntemin eğitimi 500 ağaç gibi büyük bir sayı ile yapılmıştır. Buna rağmen missForest yönteminin eksik değerleri tamamlama süresi yaklaşık 1-33 saniye arasında değişmektedir. Bu nedenle MissForest yönteminin eksik değerleri tahmin

etmedeki başarısı göz önüne alındığında çalışma süresi göz ardı edilebilir seviyededir. Ancak devasa büyüklükteki veri kümelerindeki eksik değerleri tamamlamada bunun bir dezavantaj olabileceğini araştırmacılar göz önünde bulundurarak SVD gibi yöntemler tercih edilebilir.

Tablo 4. Yöntemlerin çalışma süresi

Veriseti ve eksiklik oranları	Çalışma süresi (saniye)				
	Ortalama	**kNN**	**SVD**	**bPca**	**missForest**
Veri seti 1: Liver disorders					
%5	0.047	0.164	0.016	0.426	1.253
%10	0.072	0.222	0.016	0.413	1.301
%15	0.066	0.275	0.031	0.440	1.366
%20	0.080	0.315	0.019	0.465	1.680
%25	0.097	0.388	0.031	0.479	1.862
Veri seti 2: Diabetes					
%5	0.082	0.411	0.031	1.370	5.895
%10	0.079	0.728	0.031	1.563	6.100
%15	0.082	0.864	0.031	1.844	6.472
%20	0.121	1.164	0.039	1.629	6.977
%25	0.243	1.521	0.047	1.995	7.139
Veri seti 3: Breast cancer					
%5	0.062	0.398	0.062	1.277	5.559
%10	0.075	0.691	0.031	0.935	6.159
%15	0.106	1.017	0.016	1.007	6.609
%20	0.093	1.407	0.026	1.200	5.804
%25	0.127	1.701	0.016	1.430	7.391
Veri seti 4: WDBC					
%5	0.092	2.201	0.700	1.111	23.663
%10	0.122	3.156	0.172	1.782	21.816
%15	0.136	4.964	0.167	1.726	26.914
%20	0.137	5.146	0.172	2.368	24.588
%25	0.164	5.95	0.222	2.303	33.438

4. Tartışma ve Sonuç

Bu çalışmada veri setlerindeki eksik değerleri tamamlamada ortalama, kNN, SVD, bPca ve missForest yöntemlerinin başarısı karşılaştırılmıştır. UCI veri tabanından alınan liver disorders, diabetes, breast cancer ve WDBC veri kümeleri üzerinde %5, %10, %15, %20 ve %25 oranlarında MCAR mekanizmasına sahip yapay eksik değer üretilmiştir. Daha sonra bu eksik değerler, beş yöntem ile tamamlanmıştır. Yöntemlerin eksik değerleri tamamlamadaki performansları RMSE, sınıflandırma doğruluğu ve çalışma süresi kriterlerine göre analiz edilmiştir. Çalışma sonucunda, tüm veri setleri için en düşük RMSE ve en yüksek sınıflandırma doğruluğu ile missForest yönteminin en başarılı yöntem olduğu belirlenmiştir. Çalışma süresine bakıldığında ise diğer yöntemlere göre daha yavaş olan missForest yöntemi 1 ile 33 saniye arasında süre harcamaktadır. MissForest yönteminin üstün başarısı göz önüne alındığında, küçük ve orta boyutlu veri setlerinde bu sürenin göz ardı edilebileceği seviyede olduğu düşünülmektedir. Son yıllarda evrimsel yöntemler kullanılarak eksik değerlerin tahmini yapılmaktadır[19]. İleriki çalışmalarda eksik değerler evrimsel yöntemlerle tamamlanarak başarı kıyaslanabilir.

Kaynaklar

[1] A. Saygili, S. Albayrak, "A new computer-based approach for fully automated segmentation of knee meniscus from magnetic resonance images", *Biocybernetics and Biomedical Engineering*, Cilt. 37, s. 432-442, 2017.

[2] P. Cihan, E. Gökce, O. Kalipsiz, "A Review of Machine Learning Applications in Veterinary Field". *Kafkas Univ Vet Fak Derg*, 23(4), s. 673-680, 2017. DOI: 10.9775/kvfd.2016.17281

[3] J.W. Osborne, "Best practices in data cleaning", California: Sage Publication, Inc., s. 596, 2013.

[4] A.G. Di Nuovo, "Missing data analysis with fuzzy C-Means: A study of its application in a psychological scenario", *Expert Syst Appl*, Cilt. 38, s. 6793-6797, DOI: 10.1016/j.eswa.2010.12.067, 2011.

[5] C. Bergmeir, J.M. Benitez, "On the use of cross-validation for time series predictor evaluation", *Inform Sciences*, Cilt. 191, s. 192-213, DOI: 10.1016/j.ins.2011.12.028, 2012.

[6] J. Van Hulse, and T.M. Khoshgoftaar, "Incomplete-case nearest neighbor imputation in software measurement data", *IRI 2007: Proceedings of the 2007 IEEE International Conference on Information Reuse and Integration*, s. 630-637 DOI: 10.1109/IRI.2007.4296691, 2007.

[7] S. Genc, F.E.Boran, D. Akay, and Z.S. Xu, "Interval multiplicative transitivity for consistency, missing values and priority weights of interval fuzzy preference relations", *Inform Sciences*, Cilt. 180, s. 4877-4891, DOI: 10.1016/j.ins.2010.08.019, 2010.

[8] R.J.A. Little, and D.B. Rubin, "Statistical Analysis with Missing Data", 333. John Wiley & Sons, 2014.

[9] O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, vd. "Missing value estimation methods for dna microarrays", Bioinformatics, Cilt. 17, s. 520-525, 2001.

[10] S. Oba, M.A. Sato, I. Takemasa, M. Monden, K.I. Matsubara, and S. Ishii, "A bayesian missing value estimation method for gene expression profile data", *Bioinformatics*, Cilt. 19, s. 2088-2096, DOI: 10.1093/bioinformatics/btg287, 2003.

[11] D.J. Stekhoven, and P. Buhlmann, "Miss Forest - nonparametric missing value imputation for mixed-type data", *Bioinformatics*. Cilt. 28, s. 112-118, DOI: 10.1093/bioinformatics/btr597, 2012.

[12] T. Marwala, "Computational intelligence for missing data imputation, estimation and management:knowledge optimization techniques", *Information Science Reference*, Hershey PA, 2009.

[13] P. Cihan, "Veri madenciliği yöntemleriyle hayvan hastalıklarında teşhis, prognoz ve risk faktörlerinin belirlenmesi". Yıldız Teknik Üniversitesi, Fen Bilimleri Enstitüsü, Doktora Tezi, 101s, İstanbul, 2018.

[14] Lichman, M. UCI Machine Learning Repository, http:// archive.ics.uci.edu/ml (Erişim Tarihi: 15.03.2018).

[15] L. Breiman, Random forests, in: Machine Learning, Kluwer Academic Publishers, 2001.

[16] R.S. Marko, "Improving random forests", In: European conference on machine learning, Springer, Berlin, Heidelberg, 2004.

[17] H. Kim, H. Kim, H. Moon, and H. Ahn, "A weight-adjusted voting algorithm for ensembles of classifiers", *J. Korean Stat. Soc.*, Cilt. 40, s. 437-449, DOI: 10.1016/j.jkss.2011.03.002, 2011.

[18] H.B. Li, W. Wang, H.W. Ding, and J. Dong, "Trees weighting random forest method for classifying high-dimensional noisy data", *e-Business Engineering (ICEBE) in: 2010 IEEE 7th International Conference on IEEE*, s. 160–163, 2010.

[19] R.D. Priya, R. Sivaraj, and N.S. Priyaa, "Heuristically repopulated Bayesian ant colony optimization for treating missing values in large databases". *Knowledge-Based Systems*, Cilt. 133, s. 107-121, 2017.