Avrupa Bilim ve Teknoloji Dergisi
Özel Sayı 28, S. 1112-1116, Kasım 2021

Araştırma Makalesi

www.ejosat.com ISSN:2148-2683

European Journal of Science and Technology
Special Issue 28, pp. 1112-1116, November 2021

Research Article

Makine Öğrenimi Algoritmaları Kullanarak Kalp Hastalıklarının Tespit Edilmesi

Mustafa Coşar[1*], Emre Deniz[2]

[1*] Hitit University, Faculty of Engineering, Departmant of Computer Engineering, Çorum, Turkey, (ORCID: 0000-0001-6482-4592), mustafacosar@hitit.edu.tr
[2] Hitit University, Faculty of Engineering, Departmant of Computer Engineering, Çorum, Turkey, (ORCID: 0000-0001-6482-0000), emredeniz@hitit.edu.tr

(1st International Conference on Applied Engineering and Natural Sciences ICAENS 2021, November 1-3, 2021)

(**DOI:** 10.31590/ejosat.1012986)

ATIF/REFERENCE: Coşar, M., Deniz, E. (2021). Makine Öğrenimi Algoritmaları Kullanarak Kalp Hastalıklarının Tespit Edilmesi. *European Journal of Science and Technology*, (28), 1112-1116.

Öz

Kardiyovasküler hastalıklar, her yıl tahminen 17.9 milyon can kaybına neden olmaktadır. Toplam ölüm miktarının büyük bir çoğunluğunu oluşturan kalp hastalıkları için erken tanı ve tedaviler önemli bir yer kapsamaktadır. Uzun zamandır tıp alanında gerçekleştirilen çalışmalar, son çeyrek yüzyılda bilgisayar bilimlerinin hızlı yükselişi sayesinde makine öğrenmesi ve yapay zekâ gibi yeni tekniklerle desteklenerek daha başarılı hale getirilmiştir. Bu çalışmada kalp rahatsızlığını tespit etmek için örnek veri seti üzerinde makine öğrenmesi teknikleri uygulanmış ve sonuçlar karşılaştırılmıştır. İlk olarak veri seti analiz edilmiştir. Hangi verilerin kalp rahatsızlığına dair işaretlerde bulunabileceği belirtilmiştir. Ardından üç farklı makine öğrenmesi yöntemleri kullanılarak örnek bir model oluşturulmuş ve kalp rahatsızlığı olan bireyler tespit edilmiştir. Elde edilen sonuçlar karşılaştırıldığında, Random Forest algoritması ile %88'lik bir doğruluk oranı ile daha başarılı olduğu gözlemlenmiştir. Bunu sırasıyla %85'lik bir doğrulukla ile Lojistik Regresyon ve %70'lik bir doğruluk ile kNN algoritması takip etmiştir. Bulgular, kalp rahatsızlığının temel birkaç veri ile kolayca tespit edilebileceğini göstermektedir.

Anahtar Kelimeler: Makine öğrenmesi, Veri madenciliği, Kalp hastalığı tespit etme, Medikal veri analizi.

Detection of Heart Diseases Using Machine Learning Algorithms

Abstract

Cardiovascular diseases cause an estimated 17.9 million deaths each year. Early diagnosis and treatments have an important place for heart diseases, which make up the majority of the total number of deaths. Studies carried out in the field of medicine for a long time have been made more successful by being supported by new techniques such as machine learning and artificial intelligence, thanks to the rapid rise of computer science in the last quarter century. In this study, machine learning techniques were applied on the sample data set to detect heart disease and the results were compared. First, the data set was analyzed. It has been noted which data may be indicative of heart disease. Then, a sample model was created using three different machine learning methods and individuals with heart disease were identified. When the obtained results are compared, it has been observed that the Random Forest algorithm is more successful with an accuracy rate of 88%. This was followed by Logistic Regression with 85% accuracy and kNN algorithm with 70% accuracy, respectively. The findings how that heart disease can be easily detected with a few basic data.

Keywords: Machine learning, Data mining, Detect heart disease, Medical data analysis.

[*] Corresponding Author: mustafacosar@hitit.edu.tr

1. Giriş

Son yıllarda bilişim teknolojilerinin her sektörde yaygın bir şekilde kullanıldığı bilinmektedir. Özellikle sağlık sektöründe hastalıkların belirlenmesinde makine öğrenmesi tekniklerinin kullanımı her geçen gün artmaktadır (Selvi, 2019; Taşdelen, 2019). Konu canlılar ve canlıların sağlığı olduğunda hastalıkların önceden belirlenerek erken teşhis edilmesi tedavide başarım oranını yükseltmekte ve hayat kurtarmaktadır. Erken teşhis aşamasında bilişim teknolojilerinin gözde alanlarından olan makine öğrenme teknikleri faydalı ve başarılı sonuçlar vermektedir. Makine öğrenme teknikleri ya da algoritmaları olarak bilinen Yapay Sinir Ağları (YSA), Derin Öğrenme (DL), Karar Ağaçları (DT), Sınıflandırma ve Regresyon modelleri (CART, Logistic Regression-LR, K-Nearest Neighbours- KNN), Random Forest (RF), Bulanık Mantık (FL), Genetik Algoritmalar (GA), Destek Vektör Makineleri (SVM) ve Uzman Sistemler (US) gibi pek çok teknik kullanılmaktadır.

Literatürde sağlık sektörünün her bir alanında bu tür çalışmalara rastlamak mümkündür. Fan ve diğ. (2011)' nin WBCD veri seti üzerinde SVM, kNN ve NaviBayes gibi yöntemleri kullanarak göğüs kanserinin tespiti çalışmasını yapmışlardır. Araştırmanın sonuçlarına göre SVM'de 0.776, kNN'de 0.737 ve NaviBayes algoritması ile 0.702 oranında başarım sağlanmıştır.

Bektaş ve Babur (2016) çalışmalarında Makine Öğrenmesi Teknikleri yardımıyla Kent Ridge 2 veri seti üzerinde Meme Kanserini teşhis etmeye çalışmışlardır. Bu araştırmanın bulgularına göre YSA algoritmaları ile başarım oranları 0.814 bulunurken, DVM algoritmasında 0.845 ve Rastgele Orman algoritması ile 0.907 olarak bulunmuştur.

Jain ve Singh (2018) çalışmalarında diyabet hastalığının tespiti için makine öğrenme algoritmalarından geleneksel sınıflandırma yöntemi ve SVM algoritmasını kullanmışlardır. Sonuç olarak 0.924 oranında bir başarım ile hastalığı tespit etmişlerdir.

Haq ve arkadaşlarının (2018) yapmış olduğu çalışmada çeşitli sınıflandırma algoritmalarını karşılaştırmalı olarak analiz etmişlerdir ver en başarılı doğruluk oranına Lojistik Regresyon ile ulaşıldığını belirtmişlerdir. Çalışmanın bulgularına göre, SVM algoritması %88, DT algoritması %76 ve kNN algoritması ise %69 oranında başarım sağlamıştır.

Saygın ve Baykara (2021) çalışmalarında Makine Öğrenmesi tekniklerinde özellik seçimi kullanarak karaciğer yetmezliğini teşhis etmeye çalışmışlardır. Indian Liver Patient Dataset (ILPD) üzerinde yaptıkları uygulamada Hafif Gradyan Güçlendirme Makinesi Sınıflandırıcısı-LGBM %82.12, Çok Katmanlı Algılayıcı-MLP %81.13, DT %81.13, SVM %77.87 ve Lojistik Regresyon-LR %77.80 şeklinde başarım oranlarıyla hastalığı tespit etmişlerdir

Benzer olarak, Potur ve Erginel (2021), 299 kalp hastalığı riski bulunan hasta veri seti üzerinde makine öğrenme algoritmaları ile bir tahmin uygulaması yapmıştır. Uygulamada, LR, Naive Bayes-NB, MLP, SVM ve Karar Ağacı algoritmaları kullanmışlardır. Araştırmanın sonuçlarına göre, LR %88, NB %78, MLP %90, SVM %85ve son olarak Karar Ağacı %83 oranında tahmin başarısı göstermiştir.

Yukarda anlatılan örnek çalışmaların birçoğunda makine öğrenme algoritmalarından DT, LR, SVM, kNN ve NaiveBayes algoritmalarının kullanıldığı görülmektedir. Bu algoritmaların başarım oranlarını veri setinin yapısı, set içerisndeki veri miktarı ve araştırmada kullanılan araçların etkilediği varsayılmaktadır. Bu nedenle veri seti seçimi temel parametre olarak değerlendirilebilir. Bu parametreyi kontrol altına almak için gerçek hasta grubu ve onların verileri üzerinde bir araştırma yapılarak sonuçların karşılaştırması yapılabilir.

Bu çalışmada, kalp rahatsızlığını tespit etmek için KYTVS örnek veri seti üzerinde makine öğrenmesi teknikleri uygulanmış ve sonuçlar karşılaştırılmıştır. İlk olarak veri seti analiz edilmiştir. Hangi verilerin kalp rahatsızlığına dair işaretlerde bulunabileceği belirtilmiştir. Ardından üç farklı makine öğrenmesi yöntemleri kullanılarak örnek bir model oluşturulmuş ve kalp rahatsızlığı olan bireyler tespit edilmeye çalışılmıştır.

Araştırmanın II. bölümünde, kullanılan veri seti, hasta demografik özellikleri, hasta olup olmamaları ve belirti gösterip göstermedikleri gibi parametreler ortaya konmuştur. III. bölümde veri seti üzerinde makine öğrenme algoritmaları uygulanarak hastalığın tespit başarım oranları belirlenmeye çalışılmıştır. IV. bölümde algoritmaların öne çıkan özellikleri bulgular ışığında sonuç bölümünde makine öğrenme algoritmalarının sağlık sektöründe ne oranda başarım elde ettikleri, veri setlerinin genel durumu ve gelecekte bizleri nelerin beklediği gibi konular hakkında genel bir değerlendirme yapılmıştır.

2. Materyal and Yöntem

2.1. Veri Seti

Bu çalışma, Kalp Yetmezliği Tahmin Veri Seti (KYTVS) (Kaggle, 2021) kullanılarak gerçekleştirilmiştir. Bu veri seti, daha önce kullanılan ama birbirinden bağımsız olarak kullanılan 5 farklı veri setini içermektedir. Bu farklı veri setleri, 11 farklı ortak özellik üzerinden birleştirilerek asıl veri setini oluşturmuştur.

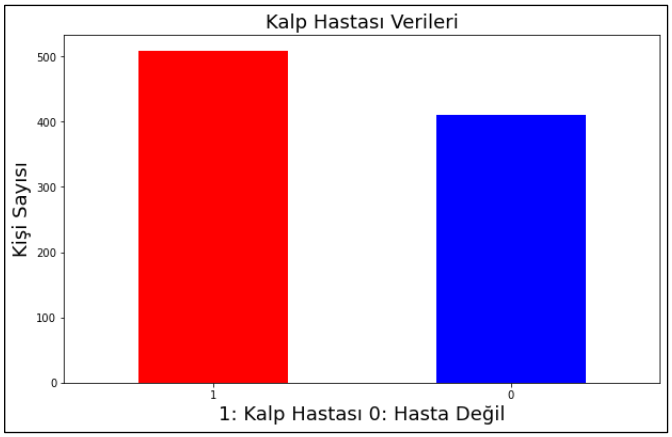
Tablo 1. KYTVS Özellikleri

Özellik	Açıklama	Örnek
Age	Hastanın Yaşı	30
Sex	Hastanın Cinsiyeti	M
ChestPainType	Göğüs Ağrısı Tipi	ATA
RestingBP	Dinlenme Kan Basıncı	140
Cholesterol	Kolesterol	214
FastingBS	Açlık Kan Şekeri	1
RestingECG	Dinlenme Elektrokardiyo Sonuçları	Normal
MaxHR	Maksimum Nabız	156
ExerciseAngina	Egzersize Bağlı Ağrı	Yes
Oldpeak	Depresyonda Ölçülen Sayısal Değer	1.5
ST_Slope	Tepe Egzersizinin Eğimi	Flat
HeartDisease	Çıktı Sınıfı	0, 1

Kullanılan veri setinde 11 farklı özellik ve bunlara ek olarak kalp rahatsızlığının olup olmadığını belirten özellik bulunmaktadır. Bu özellikler Tablo 1' de detaylı olarak gösterilmiştir.

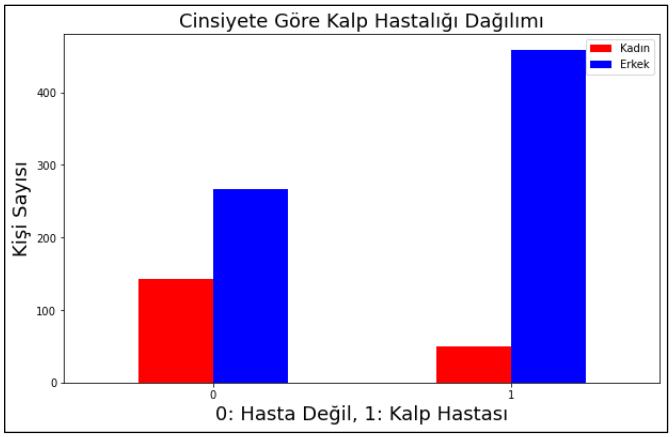
Veri seti içerisinde bulunan hasta verilerine ait toplam hasta ve sağlıklı birey sayısı Şekil 1'de gösterilmiştir. Şekil 1'de görüldüğü üzere toplam 918 birey içerisinden 508 hastalık

belirtisi gösteren birey, 410 kişi ise sağlıklı birey olarak sınıflandırılmaktadır.



Şekil 1. Sağlıklı ve Kalp Hastası Birey Dağılımı

Şekil 2'de cinsiyete göre kalp hastaların dağılımı gösterilmiştir. Görüldüğü üzere erkeklerin kalp hastalığına sahip olma oranı, kadınlara göre daha yüksektir.



Şekil 2. Cinsiyete Göre Hastalık Dağılımı

Tablo 2'de göğüs ağrısı tipine göre bireylerin hasta olup olmadığı karşılaştırılmıştır.

Tablo 2. Göğüs Ağrısı Çeşidi ve Hastalık Durumu

Ağrı Türü	Hastalıklı Birey Adedi	Sağlıklı Birey Adedi
ATA	24	149
NAP	72	131
TA	20	26
ASY	392	104
Toplam	**508**	**410**

Tablo 2 incelendiğinde, kalp hastalığına sahip olanların büyük çoğunluğunun 'ASY' tipi göğüs ağrısında şikayetçi olduğu görülmektedir.

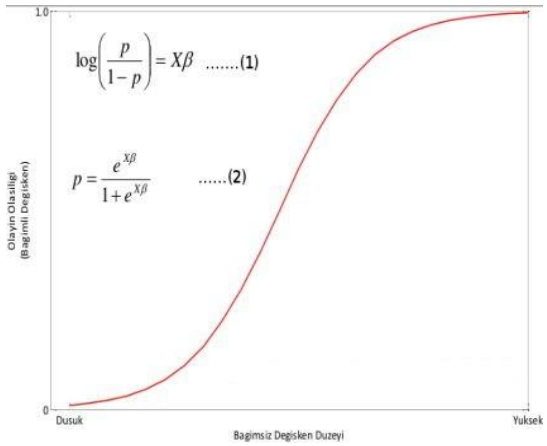
Veri seti üzerinde hastalık belirtileri olarak adlandırılan bu parametreler, python programlama dilinde yazılan bir modül ile sınıflandırma ve regresyon tahminleri yapılır hale getirilmiştir. Oluşturulan bu model sayesinde makine öğrenme algoritmalarından LR, kNN ve RF algoritmaları formül ve fonksiyonları kullanılarak tahmin doğruluk oranları tespit edilmeye çalışılmıştır.

2.2. Yöntem

Kalp Hastalığı olan bireyleri tespit edebilmek için üç farklı makine öğrenmesi modeli oluşturulmuştur. Bunlar sırasıyla Lojistik Regresyon, kNN ve Random Forest algoritmalarıdır. Bu algortmalar sınıflandırıcı ve değişkenler üzerindeki ilişkiyi ortaya koyan analiz yöntemleridir.

Sınıflandırma kategorik değerler üzerinde örüntü kurma, sınıflama ve tahmin etme işlerini yaparken regresyon analizi ise süreklilik veya kesikli durum gösteren değişkenlerin birbirleri arasındaki ilişkiyi belirlemeyi sağlar (Çalış ve diğ. 2014).

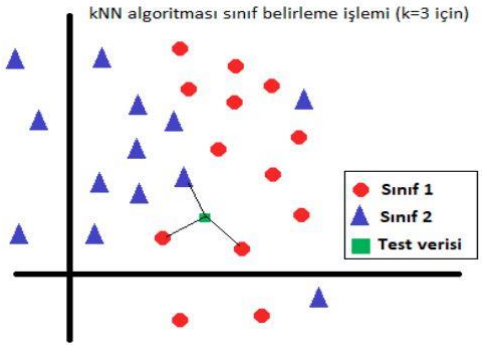
Lojistik Regresyon, değişkenlerin birbirlerini ne yönde etkilediklerini araştıran bir sınıflandırma algoritmasıdır. Bir olayda, bağımlı değişken iki veya daha fazla sınıfa sahip iken, bağımsız değişkenler sürekli veya kategorik bir yapıyı içeriyor ise, bağımsız değişkenler ile bağımlı değişken arasındaki ilişkiyi araştıran bir yöntemdir (Haq ve diğ. 2018; Akgül ve Çelik, 2020).



Şekil 3. Lojistik Regresyon (Budak ve dig. 2013)

Şekil 3'te örnek bir lojistik regresyon analiz fonksiyonu ve eğrisi görülmektedir. Bu algoritmada, bağımlı değişken olan p' nin β sabitli x bağımsız değişkenleri ile arasındaki ilişkinin belirlenmesi sağlanmış olur.

K-En Yakın Komşuluk (K-Nearest Neighbors-kNN) algoritması denetimli bir makine öğrenme algoritmasıdır (Wu ve diğ. 2007). Bu algoritma genellikle sınıflandırma problemlerinin çözümünde kullanılan bir yöntemdir (Nayak ve diğ. 2020). kNN, veri seti içerisindeki büyük miktarda eğitim verisini işleyerek test verisi üzerinde sınflama ve tahmin yapma olanağı sunan bir algoritmadır.

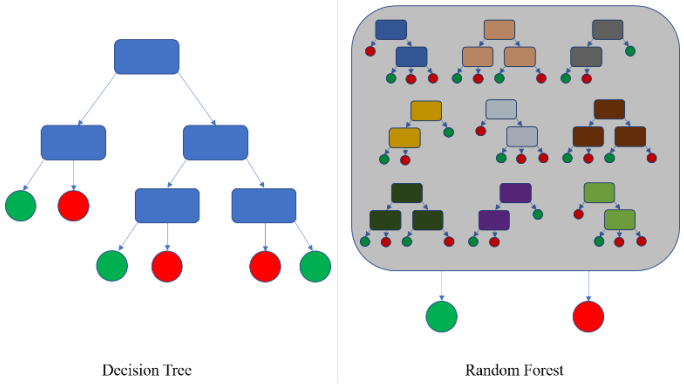


Şekil 4. kNN Algoritması ile Sınıf Belirleme Modeli (Çavuşoğlu ve Kaçar, 2019)

Şekil 4'te k=3 için sınıf belirleme işlemiş yapılmıştır. K değerine en yakın üç komşu belirlenerek bu komşuların hangi

sınıfa ait olduğu bulunmuş olur. Eğitim verisinde bu sınıflar belirlendikten sonra test verisinde bu modele uygun olarak sınıflama ve tahmin yapılmış olur.

Random Forest, sınıflama ve regresyon özelliklerini bir arada barındıran bir algoritmadır. Bu algoritma Şekil 5'te görüldüğü gibi karar ağaçlarının birleşiminden oluşan bir algoritmadır.



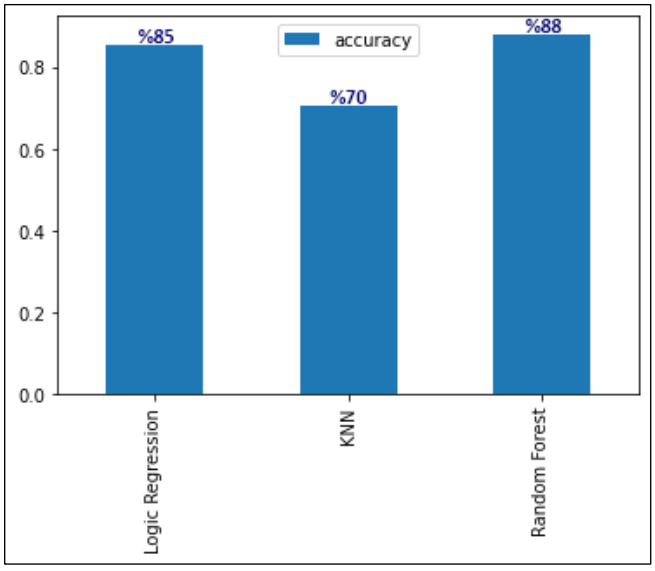
Şekil 5. Karar Ağacı ve Randon Forest Algoritması (WikiMedia,2020)

Bir sınıflama için kullanılan karar ağaçları arasında doğruluğu ve bağımsızlığı en yüksek ağaçların bir arada kullanılması tercih edilmektedir (Veranyurt ve diğ. 2020). Bu durumda ilk akla gelen ise RF algoritması olmaktadır.

3. Bulgular

Bu çalışmada, klinik verilerinden oluşturulmuş veri seti üzerinden hastaların kalp rahatsızlığı olup olmadığı makine öğrenme teknikleriyle tespit edilmiştir. Hastalık ile ilgili olarak veri seti içerisindeki hastalık belirtilerinin hastalık teşhisinden ne oranda etkili olduğu bazı algoritmalar yardımıyla sınıflandırılmıştır.

Şekil 6'da üç farklı sınıflandırma algoritması sonucu elde edilen sonuçlar gösterilmektedir. Görüldüğü üzere en başarılı sonuçlar, Random Forest algoritması ile elde edilmiştir. Lojistik Regresyon ile yapılan sınıflandırma Random Forest'a biraz daha yakın görünse de kNN algoritmasının diğer iki algoritmaya göre çok daha düşük başarı oranı sağladığı görülmektedir.



Şekil 6. Algoritmaların Hastalık Tespitindeki Doğruluk Oranları

Hastaların söyledikleri şikayetler, onların cinsiyet ve fiziksel durumları göz önüne alındığında kalp hastalığını doğrulayan en önemli etkenin göğüs ağrısı tipi olduğu gözlemlenmiştir. Hastalığın tespitinde kullanılan makine öğrenme algoritmaları Tablo 3'te özetlenmiştir.

Tablo 3. Makine Öğrenme Algoritmaları ile Kalp Rahatsızlığı Tespiti Doğruluk Oranları

Algoritma	Doğruluk Oranı (%)
Random Forest	88
kNN	70
Lojistik Regresyon	85

Ayrıca erkeklerin kadınlara göre daha sık kalp hastalığına yakalandığı veri setinin analizinde görülmüştür. Veri setini içeren tüm 11 özellik göz önüne alındığında, test verileri üzerinde bize en başarılı sınıflandırmayı yapan ise %88 doğruluk oranı ile Random Forest algoritması olmuştur. %85 ile Lojistik Regresyon Random Forest'a yakın doğrulukta başarı sağlamıştır. kNN ile yapılan sınıflandırma ise %70 ile doğruluk sağlamıştır.

4. Sonuç

Bu çalışmada kardiyovasküler hastalıklar için oluşturulmuş olan ve beş farklı veri setinin birleştirilmiş hali olan KYTVS veri seti kullanılmıştır. Bu veri set içerisinde hastaların demografik özellikleri ve fiziksel şikayetleri yer almaktadır. Bu veri seti üzerinde yaptığımız analizler doğrultusunda makine öğrenme algoritmalarının hastalığı tahmin etmede iyi sonuçlar verdiği gözlenmiştir.

Ayrıca, veri setine uygulanan makine öğrenme algoritmalarından Random Forest algoritmasının tahmin etme oranı olarak en yüksek sonucu verdiği belirlenmiştir.

Farklı veri setleri ve farklı makine öğrenme algoritmaları üzerinde yeni çalışmalar yapılarak hastalıkların teşhisinde daha dağru sonuçlar elde edilebilir. Ayrıca disiplinler arası çalışmalarda gerçek veriler kullanılarak algoritmaların doğruluklarının karşılaştırılması sağlanabilir. Özellikle sağlık alanında ciddi sonuçlar doğruabilecek hastalıkların önceden teşhis ve tadavisinde makine öğrenme algoritmalarının sağlık çalışanlarına faydalı olabileceği düşünülmektedir.

References

Akgül, G., Çelik, A., (2020). Hipotiroidi Hastalığı Teşhisinde Sınıflandırma Algoritmalarının Kullanımı, *Bilişim Teknolojileri Dergisi*, Cilt: 13, Sayı: 3, Temmuz 2020, doi: 10.17671/gazibtd.710728

Bektaş, B., Babur, S. (2016). Makine Öğrenmesi Teknikleri Kullanılarak Meme Kanseri Teşhisinin Performans Değerlendirmesi, *TıpTekno"16 Tıp Teknolojileri Kongresi*, 27-29 Ekim, Antalya.

Budak, İ., Şen, B. ve Yıldırım, M.Z. (2013), Lojistik Regresyon ile Bilgisayar Ağlarında Anomali Tespiti, *Akademik Bilişim Konferansı 2013*, Akdeniz Üniversitesi, 23-25 Ocak 2013, Antalya.

Çalış, A., Kayapınar, S., Çetinyokuş, T. (2014). Veri Madenciliğinde Karar Ağaci Algoritmalari ile Bilgisayar ve İnternet Güvenliği Üzerine Bir Uygulama, *Endüstri Mühendisliği Dergisi*, Cilt: 25, Sayı: 3-4, Sayfa: 2-19.

Çavuşoğlu, Ü., Kaçar, S. (2019). Anormal Trafik Tespiti için Veri Madenciliği Algoritmalarının Performans Analizi, *Academic Platform Journal of Engineering and Science* 7-2, 205-216, 2019.

Fan, C.Y., Chang, P.C., Linb, J.J., Hsiehb, J.C. (2011). A hybrid model combining casebased reasoning and fuzzy decision tree for medical data classification. *Applied Soft Computing*, 11(2011), pp.632-644

Haq, A., Li, J., Memon, M.H., Nazir, S. and Sun, R., (2018). A Hybrid İntelligent System Framework For The Prediction Of Heart Disease Using Machine Learning Algorithms, *Mobil Information Systems*, Volume 2018, Article ID 3860146, 21 pages, doi: 10.1155/2018/3860146

Jain, D., Singh, V. (2018). Feature selection and classification systems for chronic disease prediction: A review. *Egyptian Informatics Journal*, 19 (2018), pp.179–189, doi: 10.1016/j.eij.2018.03.002

Kaggle, (2021). Heart Failure Prediction Dataset | Kaggle. Available at: https://www.kaggle.com/fedesoriano/heart-failure-prediction (accessed Oct. 9, 2021).

Nayak, S., Panda, M., Palai, G. (2020). Realization of optical ADDER circuit using photonic structure and KNN algorithm, *Optik*, Volume 212, June 2020, 164675, doi: 10.1016/j.ijleo.2020.164675

Potur, E.A., Erginel, N., (2021). Kalp Yetmezliği Hastalarının Sağ Kalımlarının Sınıflandırma Algoritmaları ile Tahmin Edilmesi, *European Journal of Science and Technology* Special Issue 24, pp. 112-118, April 2021, doi: 10.31590/ejosat.902357

Saygın, E., Baykara, M., (2021). Karaciğer Yetmezliği Teşhisinde Özellik Seçimi Kullanarak Makine Öğrenmesi Yöntemlerinin Başarılarının Ölçülmesi, *Fırat Üniversitesi Müh. Bil. Dergisi*, 33(2), 367-377, 2021, doi:10.35234/fumbd.832264.

Selvi, O., (2019). Göğüs Kanseri Teşhisinde Farklı Makine Öğrenmesi Tekniklerinin Performans Karşılaştırması, *European Journal of Science and Technology (EJOSAT)*, Year 2019, Issue 16, 176-185, doi:10.31590/ejosat.553549.

Taşdelen, D., (2019). Veri Madenciliği Uygulamaları, Yayımlanmamış Yüksek Lisans Tezi, Ankara Üniversitesi Fen Bilimleri Enstitüsü.

Veranyurt, Ü. , Deveci, A. , Esen, M. F. ve Veranyurt, O. (2020). Makine Öğrenmesi Teknikleriyle Hastalık Sınıflandırması: Random Forest, K-Nearest Neighbour ve Adaboost Algoritmaları Uygulaması. *Uluslararası Sağlık Yönetimi ve Stratejileri Araştırma Dergisi*, 6 (2) , 275-286.

WikiMedia Commons. (2020). Decision Tree vs. Random Forest, [İnternet], Erişim Tarihi: 10 Ekim, 2021, Available at: https://commons.wikimedia.org/wiki/File:Decision_Tree_vs._Random_Forest.png

Wu, X., Kumar, V., Quinlan, R., Ghosh, J. *et al.* (2007), Top 10 Algorithms in Data Mining. *Knowledge and Information Systems.* 14**,** pp.1–37. doi: 10.1007/s10115-007-0114-2