SETSCI Conference Proceedings,
4 (1): 541-544, 2019

ISAS 2019
3rd International Symposium on Innovative Approaches in Scientific Studies
Section: Engineering and Natural Sciences
April 19 – 21, 2019, Ankara, Turkey

Kronik Böbrek Hastalığının Makine Öğrenmesi Teknikleri ile Sınıflandırılması

Mustafa İlker ERDURSUN[1*], Hasan ERBAY [2], Ömer Faruk AKMEŞE[3], İbrahim DOĞAN[4]

[1] Bilgisayar Teknolojileri / Osmancık Ömer Derindere MYO, Hitit Üniversitesi, Çorum, Türkiye
[2]Bilgisayar Mühendisliği / Fen Bilimleri Enstitüsü, Kırıkkale Üniversitesi, Kırıkkale, Türkiye
[3]Bilgisayar Teknolojileri / Osmancık Ömer Derindere MYO, Hitit Üniversitesi, Çorum, Türkiye
[4]Dahili Tıp Bilimleri / Tıp Fakültesi, Hitit Üniversitesi, Çorum, Türkiye
[*]Sorumlu Yazar: milkererdursun@hitit.edu.tr
[*]Konuşmacı: milkererdursun@hitit.edu.tr
Sunum / Bildiri Türü: Sözlü / Tam Metin

Özet – Teknolojinin ilerlemesi ile birlikte birçok veri dijital ortamlarda kayıt altına alınarak büyük veri yığınları ortaya çıkmıştır. Veri madenciliği sayesinde bu büyük veri yığınlarının içinden anlamlı ve yararlı bilgilerin ortaya çıkarılması için çalışmalar yapılmaktadır. Özellikle büyük veri yığınlarını analiz etmede klasik analiz yöntemlerinin yetersiz kalması veri madenciliği yöntemlerinin önemini arttırmıştır.

Her dönemde olduğu gibi günümüzün en önemli araştırma alanı olan tıp alanında da sürekli olarak hastalara ait veriler artarak kayıt altına alınmaktadır. Kayıt altına alınan veriler bazen tek başına anlamsız gibi görünürken diğer verilerle birlikte bütünsel olarak analiz edildiğinde gizli kalmış önemli bilgiler elde edilebilmektedir. Bu değerli bilgiler, sağlık sektörünün gelişmesine ve doktorların daha doğru bir şekilde teşhis verebilmesine yardımcı olmaktadır.

Bu çalışmada, Kronik Böbrek Hastalığı (KBH) veri seti üzerinde analiz yapılmıştır. Farklı modeller oluşturularak, bu modellerin veri üzerindeki tahmin sonuçları karşılaştırılmış ve bu sonuçlara bağlı olarak veriler üzerinde hangi modelin daha iyi sonuç verdiği belirtilmiştir.

Anahtar Kelimeler — Makine Öğrenmesi; Tıbbi Veri Madenciliği; Hastalık Tahmini; Sınıflandırma.

I. GİRİŞ

Veri madenciliği, daha önce bilinmeyen, önemsiz, pratik açıdan yararlı, insan faaliyetinin çeşitli alanlarında karar almak için gerekli olan bilginin yorumlanması için kullanılan ham verilerin keşfi sürecidir. Karar verme sürecine yardımcı olmak için bazı ön sistemler kullanılmaktadır. Bu sistemler, temizlik, veri entegrasyonu gibi yinelemeli bir ön işleme sırasını temsil eder ve veri seçimi, veri madenciliği ve bilgi temsilinin model tanımlamasını doğrular [1].

Tıbbi veri madenciliği önemli bir Veri Madenciliği alanıdır ve sağlık alanındaki uygulamalarından dolayı önemli araştırma alanlarından biri olarak kabul edilir. Genel sağlık kalitesini artırmak için makine öğrenimi ve veri madenciliği alanındaki en son başarılar biyomedikal araştırmalarda kullanılmaktadır. Birçok ülkede kalıcı tıbbi kayıt tutmak standart bir uygulama haline gelmiştir. Bu son tanı tekniklerine ek olarak, heterojen ve çok miktarda veri üretir. Tıbbi verilerin kötü yapılandırılmış doğası nedeniyle, Tıbbi Makine Madenciliği adı verilen depolanan verilerdeki mantıksal ilişkiyi tanımlamak için akıllı makine öğrenmesi ve veri madenciliği algoritmalarına ihtiyaç vardır. Tıbbi Veri Madenciliği, tıbbi veri kümelerindeki gizli kalıpları belirleme konusunda büyük bir yeteneğe sahiptir [2].

Biyomedikal veri setleri genellikle yüksek boyutlu özelliklerle ilişkilidir. Veri setlerini sağlayan klinik veri tabanları sistemik ve insan hataları içerebilir. Makine öğrenim şemalarının sınıflandırma doğruluğu, gürültülü doğadan, seyreklikten ve veri setindeki eksik değerlerden etkilenir. Bu nedenle, mevcut tıbbi teşhis araçlarının doğruluğunu arttırmaya ihtiyaç vardır. Buna ek olarak, tıbbi verilerin özellikleri ve değişkenlerin sayısı da yeni bir teknik geliştirmek için dikkate alınmalıdır [2].

Sağlık sektörünün karşılaştığı en büyük zorluk hizmet kalitesidir. Hastalığı doğru şekilde teşhis etmek ve etkili tedavileri uygulamak gerekir. Etkili karar verme için gizli bilgileri keşfetmeye maalesef "madenciliği olmayan" sağlık sektörü tarafından büyük miktarda sağlık verileri toplanmaktadır. Makine öğrenme teknikleri ilaç alanında kardiyovasküler hastalıklar, akciğer, meme karsinomaları vb. hastalıkların öngörülmesinde yoğun olarak kullanılmaktadır [3].

Daha önceden belirttiğimiz gibi veri madenciliği veya bilgi keşfi, verilerdeki geçerli, yeni, önceden bilinmeyen, potansiyel olarak yararlı ve nihayetinde anlaşılabilir kalıpları tanımlama işlemidir. Anlaşılabilir modeller, mevcut verileri açıklamak, taze verileri tahmin etmek veya sınıflandırmak, verileri özetlemek, böylece insanlarda verilerdeki daha derin modelleri keşfetmeye yardımcı olmak için kullanılır. Tıpta veri madenciliği, prognozun sağlanması ve hastalığın sınıflandırılmasının daha derinden anlaşılması için yüksek öneme sahip bir alandır [4].

Tıbbi karar destek sistemleri, klinisyenleri tanılarında desteklemek için tasarlanmıştır. Tipik olarak tıbbi verilerin analizi ve klinik uzmanlık bilgisi temeli üzerinde çalışırlar ve veri madenciliği yöntemleri vasıtasıyla, verilerde gömülü bilgileri almanın bir yolunu sağlar [5].

Modern tıp, hemen hemen her gün muazzam miktarda heterojen veri üretir. Bugün, en büyük zorluk bu çok büyük

miktardaki veriyi faydalı bilgiye dönüştürmektir. Tıbbi veriler, doktorun yorumunun yanı sıra görüntüler, sinyaller, sıcaklık, kolesterol seviyesi vb. gibi klinik bilgileri içerir [6]. İyi karar vermek için, makine öğrenmesi, hastanelerde bulunan veri tabanlarından alakalı verilerin çıkarılmasına yardımcı olur [7].

Tıbbi veri madenciliği, tıbbi alanın veri kümelerindeki gizli kalıpları keşfetme potansiyeline sahiptir. Bu verilerin düzenli bir şekilde toplanması gerekir. Bu toplanan veriler daha sonra bir hastane bilgi sistemi oluşturmak için entegre edilebilir. Veri madenciliği teknolojisi, verilerdeki yeni ve gizli kalıplara faydalı bir yaklaşım sunar [8].

Bu çalışma, Çorum Hitit Üniversitesi Erol Olçok Eğitim ve Araştırma Hastanesinde 216 hastadan elde edilen veriler üzerinden yapılmıştır. Veriler üzerinde Rapid Miner programı kullanılarak makine öğrenmesi yöntemleri uygulanmış ve hastalık tahmini yapılmıştır.

Uzman görüşü alınarak, hastalara tanı koymak için en önemli değişkenler olan Blood Urea Nitrogen (BUN), Kreatinin (Cr) ve Glomerular Filtration Rate (GFR) değişkenleri veri kümesinden çıkarılarak geriye kalan 28 değişkenle analiz yapılmıştır. Bu değişkenler arasında Ferritin, PTH (Paratiroid Hormonu), Sistolik Kan Basıncı (Büyük Tansiyon), Diyastolik Kan Basıncı (Küçük Tansiyon), Potasyum (K,) Kalsiyum (Ca), Hemoglobin (Hb), Hematokrit (Htc), Bikarbonat (HCO_3), Ürik Asit, Fosfor, Sodyum (Na), Albümin, C-reaktif protein (CRP), Glukoz, Vücut Kitle İndeksi (VKI), Diabetes Mellitus (DM) , Koroner Arter Hastalığı (KAH), Sigara kullanımı, Yaş bulunmaktadır. Farklı modeller oluşturularak, bu modellerin veri üzerindeki tahmin sonuçları karşılaştırılmış ve bu sonuçlara bağlı olarak bu veriler üzerinde hangi modelin daha iyi sonuç verdiği belirlenmiştir.

II. YÖNTEM VE TEKNİKLER

Yapılan çalışma ile KBH'nı tahmin etmek için RapidMiner programı kullanılarak makine öğrenmesine dayalı farklı sınıflandırma teknikleri uygulanmıştır. RapidMiner yazılımı, veri ön işleme, sınıflandırma, kümeleme, görselleştirme için modüller içerir. Ayrıca araştırma, eğitim ve uygulamalar için de kullanılmaktadır [10].

A. Sınıflandırma Teknikleri

Sınıflandırma, bir koleksiyondaki öğeleri hedef kategoriye veya sınıflara ayırmayı hedefleyen bir veri madenciliği işlevidir. Sınıflandırmanın amacı, verilerdeki her durum için hedef sınıfı doğru bir şekilde tahmin etmektir. Sınıflandırma modellerinde, bir dizi test verisinde öngörülen değerleri bilinen hedef değerlerle karşılaştırarak test edilir [1]. Sınıflandırma öğrenme problemlerinde, eğitilen sisteme bir dizi eğitim örneği ve buna karşılık gelen sınıf etiketleri verilir ve bir sınıflayıcı üretilir. Sınıflandırıcı, etiketlenmemiş bir örneği alır ve bunu bir sınıfa atar [9]. Sınıflandırıcıların performansı doğruluk, duyarlılık vb. iyi bilinen istatistiksel ölçütler kullanılarak değerlendirilebilir. Bu ölçümler doğru pozitif (TP), gerçek negatif (TN), yanlış pozitif (FP) ve yanlış negatif (FN) ile tanımlanır, sınıflandırma doğruluğu, pozitif ve negatif girdileri dikkate alarak doğru tahminlerin oranını ölçer [10]. Doğruluk, test verilerindeki gerçek sınıflandırmalarla karşılaştırıldığında model tarafından yapılan doğru tahminlerin yüzdesini ifade eder [5]. Aşağıdaki gibi hesaplanır:

$$Doğruluk(Accuracy) = (TP+TN)/N \quad (1)$$

N: Toplam Örnek Sayısı

B. Hata Matrisi(Confusion Matrix)

Bir hata matrisi, test verilerindeki gerçek sınıflandırmalarla karşılaştırıldığında model tarafından yapılan doğru ve yanlış tahminlerin sayısını gösterir [5].

C. Makine Öğrenmesi Algoritmaları ve Sonuçları

Bu çalışmada veri setindeki verilerin %70'i modeli eğitmek ve %30'u modeli test etmek için kullanılmıştır. Veri setine 31 tane değişken ve 216 hasta bulunmaktadır. Bu değişkenlerden grup değişkeni hedef değişken olup Hasta (1) ve Kontrol Grubu (0) değerlerini almaktadır. Geriye kalan 30 değişkenden 3 tanesi Uzman görüşü alındığında hastalığı belirleyen en önemli özellikler olduğu belirlenmiş ve bu çalışmada veri setinden çıkarılarak geriye kalan 27 değişken kullanılarak hastalık tahmini yapılmıştır. Kullanılan sınıflandırma tekniği ve programa göre bulunan hata matrisi sonuçları aşağıda gösterilmiştir.

1) Rasgele Orman(Random Forest)

Rastgele Orman bir topluluk öğrenme algoritmasıdır. Bir topluluk öğrenen yöntemi birçok bireysel öğrenici üretir ve sonuçları toplar [15]. Bu sınıflandırıcı, rastgele bir dizi özellik seçmeyi ve egzersiz verilerinin önyükleme örneği ile bir sınıflandırıcı oluşturmayı içerir. Bu şekilde çok sayıda ağaç (sınıflandırıcı) üretilir ve nihayet bir sınıfa bilinmeyen bir değer atamak için ağırlıklı oylama kullanılır [16]. Hata Matrisi sonucu Tablo I'de gösterilmiştir.

2) Naive Bayes

Öngörüler arasında bağımsızlık varsayımıyla Bayes Teorem'ine dayanan bir sınıflandırma tekniğidir. Basit bir ifadeyle, bir Naive Bayes sınıflandırıcısı, bir sınıftaki belirli bir özelliğin varlığının diğer herhangi bir özelliğin varlığı ile ilgisiz olduğunu varsayar [1]. Özellikler arasında bir bağımlılık olmadığını varsayan istatistiksel bir sınıflandırıcıdır. Sınıfın belirlenmesinde arka olasılığı maksimize etmeye çalışır. Naive Bayes kullanmanın avantajı, herhangi bir Bayesian yöntemi kullanmadan Naive Bayes modeliyle çalışabilmesidir. Naive Bayes sınıflandırıcıları birçok karmaşık gerçek dünya durumunda iyi çalışır [8]. Hata Matrisi sonucu Tablo II'de gösterilmiştir.

3) Destek Vektör Makineleri (Support Vector Machine)

Destek vektör makinelerinin (SVM) güçlü teorik temelleri ve mükemmel deneysel başarıları vardır [12]. SVM öğrenme, bazı güzel basit fikirlere dayanır ve örneklerden öğrenilen şeylerin ne anlama geldiğinin açık bir şekilde anlaşılmasını sağlar, pratik uygulamalarda yüksek performanslara yol açabilir [13]. Birçok uygulamada, SVM'nin geleneksel öğrenme makinelerinden daha yüksek performans sağladığı ve sınıflandırma problemlerini çözmek için güçlü araçlar olarak tanıtıldığı gösterilmiştir [14]. Hata Matrisi sonucu Tablo III'de gösterilmiştir.

4) Sinir Ağı (Neural Network)

Sinir Ağı modeli, biyoloji, işletme, denetim vb. alanlarda kullanılmaktadır. Çeşitli hastalıkların öngörülmesi ve teşhisi için veri vadenciliği teknikleri yaygın olarak kullanılmaktadır. En başarılı veri madenciliği araçlarından biri olan Sinir Ağları hastalığın öngörülmesinde kullanılır [2]. Yapay Sinir ağı biyolojik sinir ağımızın matematiksel bir modelidir. 3 katmandan oluşur; giriş katmanı, çıkış katmanı ve bazı ara bağlantı ağırlıklarına sahip gizli bir katman [7]. Hata Matrisi sonucu Tablo IV'de gösterilmiştir.

5) k-NN

Özellik alanındaki en yakın eğitim verilerine dayanarak nesneleri sınıflandırmak için bir yöntemdir. k-NN, bir örnek tabanlı öğrenimdir. En yakın komşu algoritması, tüm makine öğrenme algoritmalarının en basitleri arasındadır. Ancak k-NN algoritmasının doğruluğu, gürültülü veya alakasız özelliklerin varlığı veya özellik ölçeklerinin önemi ile tutarlı olmaması durumunda ciddi şekilde bozulabilir [8]. Hata Matrisi sonucu Tablo V'de gösterilmiştir.

6) *Lojistik* Regresyon (*Logistic Regression)*

İstatistik ve biyomedikal alanında Lojistik Regresyon güçlü ve köklü bir yöntemdir. Lojistik Regresyon, kategorik sonuçları ve bir açıklayıcı değişkenleri karşılaştırır [2]. Lojistik model olarak da adlandırılan lojistik regresyon, hedef değişken iki kategorili kategorik bir değişken olduğunda, örneğin aktif veya inaktif, sağlıklı veya sağlıksız, kazanma veya kaybetme gibi, kullanılabilecek bir öngörü modelidir [11]. Hata Matrisi sonucu Tablo VI'da gösterilmiştir.

TABLO I. RASGELE ORMAN (RANDOM FOREST) HATA MATRİSİ

	Doğru 0	**Doğru 1**	**Sınıf Hassasiyeti**
Tahmin 0	28	1	*96.55%*
Tahmin 1	0	36	*100.00%*
Sınıf Hatırlama	*100.00%*	*97.30%*	

TABLO II. NAİVE BAYES HATA MATRİSİ

	Doğru 0	**Doğru 1**	**Sınıf Hassasiyeti**
Tahmin 0	28	2	*93.33%*
Tahmin 1	0	35	*100.00%*
Sınıf Hatırlama	*100.00%*	*94.59%*	

TABLO III. SVM HATA MATRİSİ

	Doğru 0	**Doğru 1**	**Sınıf Hassasiyeti**
Tahmin 0	28	4	*87.50%*
Tahmin 1	0	33	*100.00%*
Sınıf Hatırlama	*100.00%*	*89.19%*	

TABLO IV. SİNİR AĞI (NEURAL NETWORK) HATA MATRİSİ

	Doğru 0	**Doğru 1**	**Sınıf Hassasiyeti**
Tahmin 0	34	1	*97.14%*
Tahmin 1	0	30	*100.00%*
Sınıf Hatırlama	*100.00%*	*96.77%*	

TABLO V. K-NN HATA MATRİSİ

	Doğru 0	**Doğru 1**	**Sınıf Hassasiyeti**
Tahmin 0	34	2	*94.44%*
Tahmin 1	0	29	*100.00%*
Sınıf Hatırlama	*100.00%*	*93.55%*	

TABLO VI. LOGİSTİC REGRESSİON HATA MATRİSİ

	Doğru 0	**Doğru 1**	**Sınıf Hassasiyeti**
Tahmin 0	21	2	*91.30%*
Tahmin 1	2	40	*95.24%*
Sınıf Hatırlama	*91.30%*	*95.24%*	

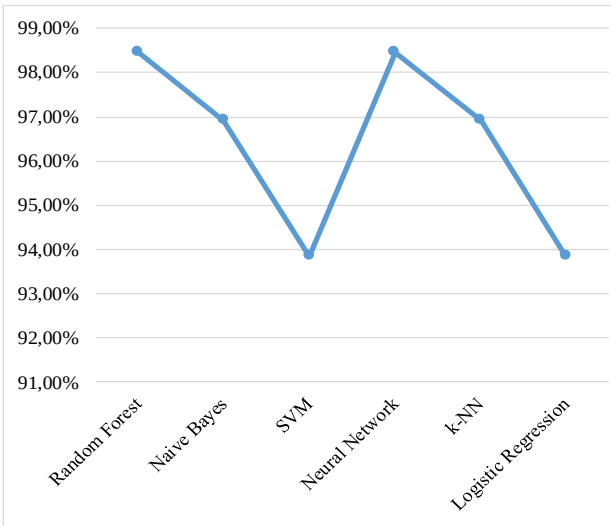
III. DOĞRULUK SONUÇLARININ KARŞILAŞTIRILMASI

Tablo VII'de, yukarıda anlatılan makine öğrenmesinde kullanılan sınıflandırma algoritmalarının sonucu ve Şekil 1'de sonuçların karşılaştırılması grafikle gösterilmiştir.

TABLO VII. ALGORİTMALAR VE DOĞRULUK ORANLARI

Sınıflandırma Algoritmaları	**Doğruluk Oranları**
Random Forest	*98.46%*
Naive Bayes	*96.92%*
SVM	*93.85%*
Neural Network	*98.46%*
k-NN	*96.92%*
Logistic Regression	*93.85%*

Şekil. 1. Doğruluk Oranı Karşılaştırma Grafiği



IV. SONUÇ VE ÖNERİLER

Bu çalışmada, kan numune verileri, sigara kullanımı, yaş, cinsiyet, tansiyon, diyabet durumu gibi klinik kayıtlar incelenerek kronik böbrek hastalığının olup olmadığı makine öğrenmesi yöntemleriyle yüksek doğruluk oranı sağlanarak bulunmuştur. Çalışmada tahmini veri madenciliği uygulaması yapılmış ve söz konusu yöntemlerin doğrulukları karşılaştırılmıştır. Farklı sınıflandırma algoritmalarının KBH veri seti ile olan doğruluğu RapidMiner yazılım desteği ile test edilmiştir. Çalışmaya 23 ile 84 yaş aralığında, 115'i erkek (%53), 101'i de kadın (%47) olmak üzere toplam 216 klinik kayıt alınmıştır. Veri seti 124 hasta (%57) ve 92 de kontrol grubu (%43) olan kayıtlardan oluşmaktadır. Çalışmada tahmin doğruluğunun yüksek olması en önemli faktör olarak ele alınmıştır. Random Forest ve Neural Network algoritmasının %98.46 ile diğer sınıflayıcılardan daha iyi olduğu tespit edilmiştir. Tanı için kullanılan algoritmaların hekime karar vermede yardımcı olabileceği düşünülmektedir. Gelecekte, farklı veri setlerinde uygulanabilir, kayıt sayısı artırılırsa daha iyi performans elde edilebilir.

KAYNAKLAR

[1] Meganathan, D., & Marudachalam, N. International Journal OF Engineering Sciences &Management Research.

[2] Raghavendra, S., & Indiramma, M. (2015). Classification and Prediction Model using Hybrid Technique for Medical Datasets. *analysis*, *127*(5).

[3] Singla, M., & Singh, K. Heart Disease Prediction System using Data Mining Clustering Techniques.

[4] Joshi, S., Shenoy, D., Rrashmi, P. L., Venugopal, K. R., & Patnaik, L. M. (2010, February). Classification of Alzheimer's disease and Parkinson's disease by using machine learning and neural network methods. In *Machine Learning and Computing (ICMLC), 2010 Second International Conference on* (pp. 218-222). IEEE.

[5] Peter, T. J., & Somasundaram, K. (2012, March). An empirical study on prediction of heart disease using classification data mining techniques. In *Advances in Engineering, Science and Management (ICAESM), 2012 International Conference on* (pp. 514-518). IEEE.

[6] Robu, R., & Hora, C. (2012, June). Medical data mining with extended WEKA. In *Intelligent Engineering Systems (INES), 2012 IEEE 16th International Conference on* (pp. 347-350). IEEE.

[7] Dewan, A., & Sharma, M. (2015, March). Prediction of heart disease using a hybrid technique in data mining classification. In *Computing for Sustainable Global Development (INDIACom), 2015 2nd International Conference on* (pp. 704-706). IEEE.

[8] Soni, J., Ansari, U., Sharma, D., & Soni, S. (2011). Predictive data mining for medical diagnosis: An overview of heart disease prediction. *International Journal of Computer Applications*, *17*(8), 43-48.

[9] Domingos, P., & Pazzani, M. (1997). On the optimality of the simple Bayesian classifier under zero-one loss. *Machine learning*, *29*(2-3), 103-130.

[10] Shrivas, A. K., & Yadu, R. K. (2017). An Effective Prediction Factors for Coronary Heart Disease using Data Mining based Classification Technique. *International Journal on Recent and Innovation Trends in Computing and Communication*, *5*(5), 813-816.

[11] Le Cessie, S., & Van Houwelingen, J. C. (1992). Ridge estimators in logistic regression. *Applied statistics*, 191-201.

[12] Tong, S., & Chang, E. (2001, October). Support vector machine active learning for image retrieval. In *Proceedings of the ninth ACM international conference on Multimedia* (pp. 107-118). ACM.

[13] Hearst, M. A., Dumais, S. T., Osuna, E., Platt, J., & Scholkopf, B. (1998). Support vector machines. *IEEE Intelligent Systems and their applications*, *13*(4), 18-28.

[14] Lin, C. F., & Wang, S. D. (2002). Fuzzy support vector machines. *IEEE transactions on neural networks*, *13*(2), 464-471.

[15] Alam, M. S., & Vuong, S. T. (2013, August). Random forest classification for detecting android malware. In *Green Computing and Communications (GreenCom), 2013 IEEE and Internet of Things (iThings/CPSCom), IEEE International Conference on and IEEE Cyber, Physical and Social Computing* (pp. 663-669). IEEE.

[16] Pal, M. (2005). Random forest classifier for remote sensing classification. *International Journal of Remote Sensing*, *26*(1), 217-222.