Makine Öğrenmesi Algoritmaları ile Hastanın Hayatta Kalım Tahmini

Patient Survival Prediction with Machine Learning Algorithms

Mustafa Berkant Selek[1], Saadet Sena Egeli[2,3], Yalçın İşler[4]
[1]Ege Meslek Yüksekokulu, Ege Üniversitesi, İzmir, Türkiye
mustafa.berkant.seelek@ege.edu.tr
[2]Biyomedikal Teknolojiler Anabilim Dalı, İzmir Katip Çelebi Üniversitesi, İzmir, Türkiye
[3]İslerya Medikal ve Bilişim Teknolojileri, İzmir, Türkiye
egelisena@gmail.com
[4]Biyomedikal Mühendisliği Bölümü, İzmir Katip Çelebi Üniversitesi, İzmir, Türkiye
islerya@yahoo.com

Özetçe— Bu çalışmada, yoğun bakım ünitelerinde yatan hastaların, ilk 24 saatte yapılan tetkiklerine göre hayatta kalma durumları makine öğrenmesi algoritmalarıyla tahmin edilmiştir. Çalışmada, bir yıllık süre zarfında yaklaşık iki yüz hastaneden toplanan yoğun bakım hastalarının verileri kullanılmıştır. Algoritmalar Python ortamında koşturulmuştur. Çapraz Doğrulama yöntemi ile makine öğrenmesi modelleri karşılaştırılmış, en iyi sonuç veren Rastgele Orman algoritması kullanılmıştır. Kullanılan model %92,53 doğruluk oranı ile tahminlemeyi gerçekleştirmiştir.

Anahtar Kelimeler—makine öğrenmesi; yoğun bakım; hayatta kalım.

Abstract— In this study, the intensive care unit patient survival is predicted by machine learning algorithms according to the examinations performed in the first 24 hours. The data of intensive care patients collected from approximately two hundred hospitals over a period of one year were used. Algorithms are run in Python environment. Machine learning models were compared with the Cross-Validation method, and the random forest algorithm is used. The model made the prediction with 92,53% accuracy rate.

Keywords—machine learning; intensive care unit; patient survival.

I. GİRİŞ

Yoğun bakım üniteleri tüm tıbbi ve cerrahi uzmanlık alanlarındaki hastalarla ilgilenir. Bu nedenle yoğun bakım ünitelerinde yatan hastalar çok farklı popülasyonlardan oluşur [1]. Genellikle uzun bir hastalık öyküsü veya başka hayati hastalıkları olan hastalar, yoğun bakım ünitelerinde tedavi görürler. Bununla beraber, yoğun bakım ünitelerinde tedavi gören her hasta kurtarılamamaktadır. Hastanın kurtarılıp kurtarılamama durumu, birçok faktöre bağlıdır. Yoğun bakım ünitelerinde tedavi gören hastaların yaşı, başka bir hastalığının olup olmaması, cinsiyeti, kan değerleri gibi birçok faktöre bağlıdır [2].

Daha önce gerçekleştirilen çalışmalarda, yoğun bakım ünitelerinde tedavi gören hastaların hayatta kalma durumunu sistematik hastalıkların nasıl etkilediği [3], yoğun bakım ünitelerinde çalışan doktorların ve dahiliye uzmanlarının bu konuyla ilgili gerçekleştirdiği tahminler [4], yoğun bakım ünitelerinde tedavi gören kritik hastaların hayatta kalma durumları [5], bu hastaların cinsiyetlerinin hayatta kalma durumunu nasıl etkilediği [6] gibi konular çalışılmıştır.

Bu çalışmada ise, yüz otuz binden fazla hastanenin yoğun bakım ünitelerinden bir yıl boyunca toplanan veriler ile hayatta kalım tahmini yapılmıştır. Dokuz farklı makine öğrenmesi sınıflandırma modelleri çapraz doğrulama yöntemi ile karşılaştırılmış ve çapraz doğrulama doğruluk değerine göre aralarından en iyi sonuç veren model olan Rastgele Orman Sınıflandırma modeli tahminlemede kullanılmıştır. Tahminleme gerçekleştirildikten sonra model performansı, doğruluk oranı hesaplanarak ölçülmüştür.

II. MATERYAL VE METOT

A. Spyder

Spyder, Python dilinde bilimsel programlama için açık kaynaklı bir platformlar arası entegre geliştirme ortamıdır (IDE). Spyder, diğer açık kaynaklı yazılımların yanı sıra NumPy, SciPy, Matplotlib, Pandas, IPython, SymPy ve Cython dahil olmak üzere bilimsel Python yığınında önemli paketlerle bütünleşir. Python bilim adamları, mühendisler ve veri analistleri tarafından ve onlar için tasarlanmış güçlü bir bilimsel ortamdır. Bilimsel bir paketin veri keşfi, etkileşimli yürütme, derin denetim ve

başarılı görselleştirme özellikleri ile kapsamlı bir geliştirme aracının gelişmiş düzenleme, analiz, hata ayıklama ve profil oluşturma işlevlerin kombinasyonlarını gerçekleştirmek mümkündür. Spyder, Anaconda tümleşik platformunda bulunan bir programlama ortamıdır. Anaconda resmi sitesinden indirip kurulduktan sonra, Spyder bu platform içinde çalıştırılır ve Python dilinde kodlamaya imkan verir.

B. Veri Seti

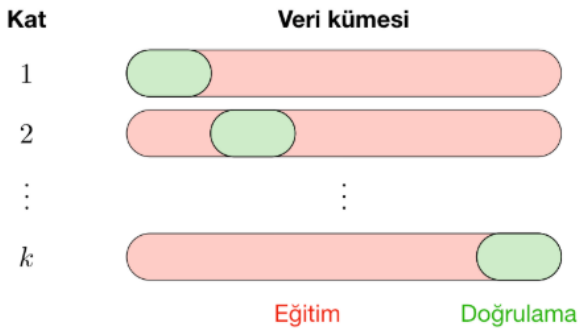
Çalışmada kullanılan veri seti, MIT'in GOSSIS topluluk girişimi tarafından Harvard Gizlilik Laboratuvarı'ndan gizlilik sertifikası ile yoğun bakım ünitelerinde tedavi gören hastalardan bir yıllık bir zaman dilimini kapsayan yüz otuz binden fazla hastaneden toplanan verilerle oluşturulmuştur. Bu veriler, Arjantin, Avustralya, Yeni Zelanda, Sri Lanka, Brezilya ve ABD'deki iki yüzden fazla hastaneyi kapsayan küresel bir çaba ve konsorsiyumun bir parçasıdır [7].

C. Veri Ön İşleme

Ham veri setleri makine öğrenmesi algoritmalarına girdi olarak verilmek için uygun formatta bulunmayabilir. Bunun için genellikle veri setleri makine öğrenmesi algoritmaları koşturulmadan önce bir ön işleme sürecine tabi tutulur [8]. Veri ön işleme sürecinde, veri setini daha uygun bir formata getirebilmek için veriyi biçimlendirme, temizleme, örnekleme, ölçekleme ayrıştırma gibi işlemler uygulanabilir [9].

D. Çapraz Doğrulama

Çapraz doğrulama veya k katlamalı çapraz doğrulama, bir makine öğrenmesi modelinde yapılan testin hatasını daha iyi tahmin edebilmek için model seçiminde kullanılan bir tekniktir. Çapraz doğrulamanın arkasındaki fikir, eğitim verileri setinden doğrulama kümeleri olarak bilinen örnek gözlem bölümlerini oluşturmaktır [10]. Bir modeli eğitim verilerine yerleştirdikten sonra, performansı, her yeni doğrulama kümesine karşı ölçülür ve daha sonra, yeni gözlemleri öngörmek istenildiğinde modelin nasıl performans göstereceğine ilişkin daha iyi bir değerlendirme elde edilir. Yapılacak bölüm sayısı, örnek veri kümesindeki gözlem sayısına ve önyargı varyansı dengelemesine ilişkin kararın, daha fazla bölünmenin daha küçük bir yanlılığa yol açmasına ve daha fazla varyansa bağlı olarak değişmesine bağlıdır [11].



Şekil 1.Çapraz Doğrulama Gösterimi

E. Rastgele Orman Sınıflandırıcı

Rastgele ormanlar, sınıflandırma ve regresyon sırasında çok sayıda karar ağacı oluşturarak sınıfların modunu (sınıflandırma) veya tek tek ağaçların ortalama tahmini (regresyon) sınıfını çıkarmak için kullanılan bir öğrenme yöntemidir [12].

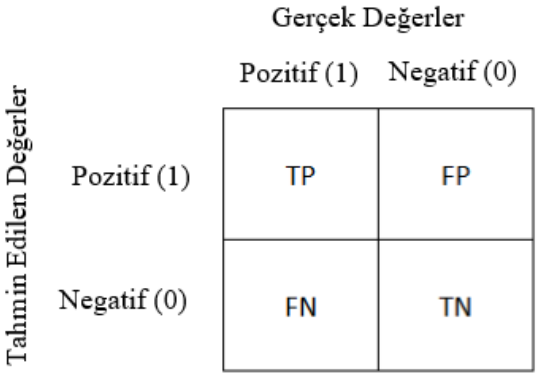
Rastgele ormanlar, ağaç tipi sınıflandırıcılar topluluğu olarak tanımlanabilir. Rastgelelik özelliği eklenerek Torbalama yönteminin geliştirilmiş bir versiyonudur. Rastgele ormanlar, tüm değişkenler arasından en iyi dalı kullanarak her bir düğümü dallara ayırmak yerine, her bir düğümde rastgele olarak seçilen değişkenler arasından en iyisini kullanarak her bir düğümü dallara ayırır. Her bir veri seti orijinal veri setinden yer değiştirmeli olarak üretilir. Sonra rastgele özellik seçimi kullanılarak ağaçlar geliştirilir. Geliştirilen ağaçlar budanmaz. Bu strateji rastgele ormanların doğruluğunu eşsiz yapar [13].

F. Performans Metrikleri

Sınıflandırma algoritmalarının performansını değerlendirmek için performans veya hata metrikleri kullanılır [14].

1. Karışıklık Matrisi

Bir karışıklık matrisi, bir sınıflandırma modelinin gerçek değerlerin bilindiği bir dizi test verisindeki performansını tanımlamak için kullanılan bir tablodur (Şekil 2). Karışıklık matrisinin kendisini anlamak nispeten basittir, ancak ilgili terminoloji kafa karıştırıcı olabilir [15]. Gerçek pozitif (TP), gerçek negatif (TN), yanlış pozitif (FP) ve yanlış negatif (FN) terimlerini içerir.



Şekil 2. Karışıklık Matrisi

2. Doğruluk Oranı

Doğruluk oranı, tüm veri noktalarından doğru tahmin edilen veri noktalarının sayısıdır. Daha resmi olarak, gerçek pozitiflerin ve gerçek negatiflerin sayısının gerçek pozitiflerin sayısına, gerçek negatiflere, yanlış pozitiflere ve yanlış negatiflere bölünmesiyle tanımlanır (Denklem 1). Gerçek pozitif veya gerçek negatif, algoritmanın sırasıyla doğru veya yanlış olarak doğru bir şekilde sınıflandırıldığını belirtir. Öte yandan yanlış pozitif veya yanlış negatif, algoritmanın yanlış sınıflandırıldığını belirtir [16]. Örneğin, algoritma yanlış bir veri noktasını doğru olarak sınıflandırırsa, yanlış pozitif olur.

$$Doğruluk\ Oranı = \frac{\mathrm{TP+TN}}{2TP+TN+FP+FN} \quad (1)$$

III. SONUÇLAR

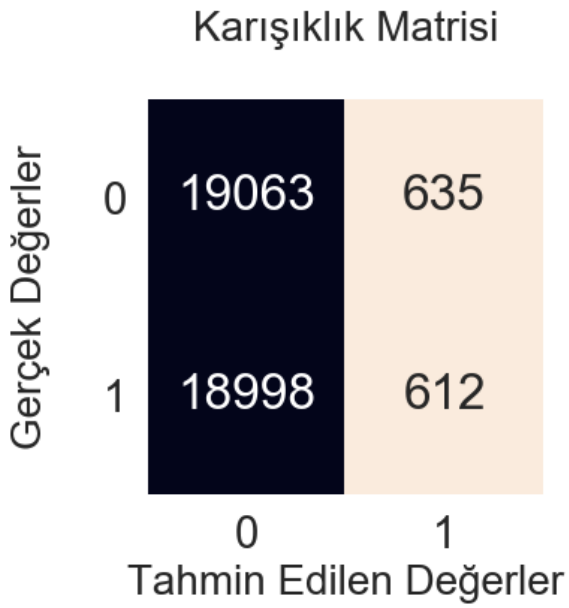
Düzenlemeye çalışmada kullanılan veri seti, test ve eğitim seti olmak üzere birbirinden ayrı iki farklı veri seti olarak elde edilmiştir. Modeller eğitim setiyle oluşturulup, model performansları test setiyle ölçülmüştür. Makine öğrenmesi algoritmaları koşturulmadan önce, veri setleri ön işleme sürecinden geçmiştir. Veri ön işleme ile çıktıyı etkilemeyen özellikler veri setinden çıkarılmış, veri setinde bulunan kategorik veriler tespit edilip etiketlenmiş ve veri setinde bulunan eksik veriler o sütunda bulunan verilerin medyanı ile telafi edilmiştir.

Çapraz doğrulama yöntemi ile rastgele orman, karar ağacı, k-en yakın komşu, Gauss naif bayes, doğrusal ayırma analizi, Bernoulli naif bayes, ekstra rastgele orman, pasif agresif ve kuadratik sınıflandırıcı modelleri karşılaştırılmış ve aralarında en iyi performans gösteren model olan rastgele orman sınıflandırıcı modeli çalışmada kullanılmıştır (Tablo 1). Çapraz doğrulama yöntemi kullanılırken k değeri 10 olarak alınmıştır.

Sınıflandırıcı Modeli	Çapraz Doğrulama Skoru
Rastgele Orman	50.3129
Karar Ağacı	50.3023
Gauss Naif Bayes	49.8499
Doğrusal Ayrıma Analizi	50.0738
Bernoulli Naif Bayes	50.0356
Ekstra Rastgele Orman	49.9516
Pasif Agresif	49.6159
Kuadratik	49.5981

Tablo I. Çapraz Doğrulama Sistemi

Rastgele orman sınıflandırıcı modeli ile yoğun bakım ünitelerinden tedavi gören hastaların hayatta kalımı tahmin edilmiş ve modelin karışıklık matrisi hesaplanmıştır (Şekil 3). Modelin doğruluk oranı hesaplandığında %92,53 doğruluk oranı ile tahminlemenin gerçekleştirildiği anlaşılmıştır.



Şekil 3. Karışıklık Matrisi

IV. TARTIŞMA

Bu çalışmada, yoğun bakım ünitelerinde tedavi gören hastaların hayatta kalımları, makine öğrenmesi algoritmalarından biri olan rastgele orman sınıflandırıcı modeliyle öngörülmüştür. Öngörü modeli, çapraz doğruluk yöntemiyle sekiz farklı makine öğrenmesi algoritma performansları karşılaştırılarak seçilmiştir. Rastgele orman

sınıflandırıcı modeli, %92,53 doğruluk oranı ile tahminlemeyi gerçekleştirmiştir.

KAYNAKÇA

[1] Nielsen, A. B., Thorsen-Meyer, H. C., Belling, K., Nielsen, A. P., Thomas, C. E., Chmura, P. J., ... & Spangsege, L. (2019). Survival prediction in intensive-care units based on aggregation of long-term disease history and acute physiology: a retrospective study of the Danish National Patient Registry and electronic patient records. The Lancet Digital Health, 1(2), e78-e89.

[2]Hollinger, A., Gayat, E., Féliot, E., Paugam-Burtz, C., Fournier, M. C., Duranteau, J., ... & Arrigo, M. (2019). Gender and survival of critically ill patients: results from the FROG-ICU study. Annals of intensive care, 9(1), 43.

[3] Raffin, T. A. (1989). Intensive Care Unit Survival of Patients with Systemic IIInessl-3. Am Rev Respir Dis, 140, S28-S35.

[4] Escher, M., Ricou, B., Nendaz, M., Scherer, F., Cullati, S., Hudelson, P., & Perneger, T. (2018). ICU physicians' and internists' survival predictions for patients evaluated for admission to the intensive care unit. Annals of intensive care, 8(1), 108.

[5]Simchen, E., Sprung, C. L., Galai, N., Zitser-Gurevich, Y., Bar-Lavi, Y., Levi, L., ... & Ekka-Zohar, A. (2007). Survival of critically ill patients hospitalized in and out of intensive care. Critical care medicine, 35(2), 449-457.

[6]Hollinger, A., Gayat, E., Féliot, E., Paugam-Burtz, C., Fournier, M. C., Duranteau, J., ... & Arrigo, M. (2019). Gender and survival of critically ill patients: results from the FROG-ICU study. Annals of intensive care, 9(1), 43.

[7]https://www.kaggle.com/c/widsdatathon2020/overview

[8] ŞEKER, Ş. (2013). İş zekası ve veri madenciliği.

[9] Wu, C. L., Chau, K. W., & Li, Y. S. (2009). Predicting monthly streamflow using data-driven models coupled with data-preprocessing techniques. Water Resources Research, 45(8).

[10] Browne, M. W. (2000). Cross-validation methods. Journal of mathematical psychology, 44(1), 108-132.

[11] Wong, T. T. (2015). Performance evaluation of classification algorithms by k-fold and leave-one-out cross validation. Pattern Recognition, 48(9), 2839-2846.

[12] Akar, Ö., & Güngör, O. (2012). Rastgele orman algoritması kullanılarak çok bantlı görüntülerin sınıflandırılması. Jeodezi ve Jeoinformasyon Dergisi, ss, 139-146.

[13] Akar, Ö., & Güngör, O. (2012). Rastgele orman algoritması kullanılarak çok bantlı görüntülerin sınıflandırılması. Jeodezi ve Jeoinformasyon Dergisi, ss, 139-146.

[14] Tağıl, Ş. (2006). Change of habitat fragmentation and quality in the balıkesir plain and its surroundings with landscape pattern metrics (1975-2000).

[15] Townsend, J. T. (1971). Theoretical analysis of an alphabetic confusion matrix. Perception & Psychophysics, 9(1), 40-50.