Avrupa Bilim ve Teknoloji Dergisi
Özel Sayı 29, S. 87-92, Aralık 2021

Araştırma Makalesi

www.ejosat.com ISSN:2148-2683

European Journal of Science and Technology
Special Issue 29, pp. 87-92, December 2021

Research Article

Makine Öğrenmesi Algoritmaları Kullanılarak Prostat Kanseri Tümör Oluşumunun İncelenmesi

Nesrin Aydın Atasoy[1*], Ahmet Demiröz[2]

[1*] Karabük Üniversitesi, Mühendislik Fakültesi, Bilgisayar Mühendisliği Bölümü, Karabük, Türkiye, (ORCID:0000-0002-7188-0020), nesrinaydin@karabuk.edu.tr
[2] Çankırı Karatekin Üniversitesi, Meslek Yüksekokulu, Bilgisayar Teknolojileri Bölümü, Çankırı, Türkiye, (ORCID:0000-0001-5739-896X), demiroz@karatekin.edu.tr

(International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT) 2021 – 21-23 October 2021)

(**DOI:** 10.31590/ejosat.1018897)

ATIF/REFERENCE: Atasoy, N. A. & Demiröz, A. (2021). Makine Öğrenmesi Algoritmaları Kullanılarak Prostat Kanseri Tümör Oluşumunun İncelenmesi. *Avrupa Bilim ve Teknoloji Dergisi*, (29), 87-92.

Öz

Makine öğrenmesi, bir algoritma veya yöntem kullanarak ham verilerden kalıpları çıkaran bir yapay zeka türüdür. Makine öğrenmesinin temel odak noktası, bilgisayar sistemlerinin açıkça programlanmadan veya insan müdahalesi olmadan deneyimlerden öğrenmesine olanak sağlamaktır. Trafik uyarıları, sosyal medya, ulaşım, ürün önerileri, sanal kişisel asistanlar, otonom arabalar, dinamik ücretlendirme, google çeviri, çevrimiçi video akışı, dolandırıcılık tespiti ve daha birçok kullanım alanı olmakla beraber tıp alanında teşhis ve tedavi süreçlerinde de sıklıkla kullanılmaktadır. Elde edilen tıbbi sonuçlar hastanın yaşam kalitesini arttırmak ve hastalığın durumunu takip etmek için alanında uzman kişilere yardımcı olabilmektedir. İnsanlar için çok çeşitli hastalıklar olmakla birlikte kanser yüksek riskli hastalıkların başında gelmektedir. Prostat kanseri, akciğer kanserinden sonra erkeklerde ikinci sırada yer almaktadır. Yapılan literatür araştırmalarında Prostat Spesifik Antijen, Gleason Skor, Androjen Hormonu ve T Aşaması prostat kanser tespitinde önemli girdiler olmakla beraber yeterli olmadıkları görülmüştür. Bu çalışmada çok boyutlu kanser genomik verilerini keşfetmek için açık bir platform olan cBioPortal veritabanından klinik veriler elde edilmiştir. Elde edilen verilerin daha anlaşılır ve işlenebilir hale getirilmesi için veri ön işleme işlemi gerçekleştirilmiştir. Prostat kanseri olan hasta takiplerinde tümörlü/tümörsüz durumu tahmin edilerek makine öğrenmesi algoritmalarından K-En yakın komşular, Rassal ağaçlar, Gradyan artırma, Destek vektör makinesi, Lojistik regresyon, Naive bayes ve Karar ağaçları sınıflandırma algoritmalarının performansı değerlendirilmiştir. Yapılan önceki çalışmalarda çoğunlukla Rassal ağaçlar algoritmasının daha iyi performans gösterdiği görülmüştür. Ancak klinik verilerle yaptığımız çalışmada sıklıkla kullanılan yedi sınıflandırıcı arasında Gradyan artırma algoritması ile %85.37 doğrulukla daha iyi sonuçlar elde edilmiştir. Özellik seçimi yapılmadan elde ettiğimiz klinik verilerde özellik seçimi ile en iyi alt kümenin seçilmesi işlemi yapılarak sonuçlar iyileştirilebilir.

Anahtar Kelimeler: Biyoinformatik, Gradyan artırma, Makine öğrenmesi, Prostat kanseri.

Examination of Prostate Cancer Tumor Formation Using Machine Learning Algorithms

Abstract

Machine learning is a type of artificial intelligence that extracts patterns from raw data using an algorithm or method. The focus of machine learning is to enable computer systems to learn from experience without being explicitly programmed or human intervention. Traffic alerts, social media, transportation, product recommendations, virtual personal assistants, autonomous cars, dynamic pricing, google translation, online video streaming, fraud detection and many other uses are also frequently used in diagnosis and treatment processes in the medical field. The medical results obtained can help experts in the field to improve the life quality of the patient and to follow the status of the disease. Prostate cancer ranks second in men after lung cancer. In the literature, it has been seen that Prostate Specific Antigen, Gleason score, androgen hormone and T stage prostate cancer are important inputs, but they are not sufficient. In this

[*] Sorumlu Yazar: Karabük Üniversitesi, Mühendislik Fakültesi, Bilgisayar Mühendisliği Bölümü, Karabük, Türkiye, ORCID:0000-0002-7188-0020, nesrinaydin@karabuk.edu.tr

study, clinical data were obtained from the cBioPortal database, which is an open platform to explore multidimensional cancer genomic data. Data preprocessing was realized for to make the obtained data more understandable and processable. The performance of K-Nearest neighbors, Random trees, Gradient boosting, Support vector machine, Logistic regression, Naive Bayes, and Decision trees classification algorithms from machine learning algorithms was evaluated by estimating the tumor/no-tumor status in the follow-ups of patients with prostate cancer. In previous studies, it has been seen that the Random trees algorithm mostly performs better. However, among the seven classifiers that are frequently used in our study with clinical data, better results were obtained with the Gradient boosting algorithm with an accuracy of 85.37%. Results can be improved by selecting the best subset with feature selection in the clinical data we obtained without feature selection.

Keywords: Bioinformatics, Gradient boost, Machine learning, Prostate cancer.

1. Giriş

Makine öğrenmesine dayalı karar destek sistemlerinin görüntü tanıma, metin okuma, konuşma tanıma, dolandırıcılık tespiti ve öneri sistemleri dahil olmak üzere oldukça geniş kapsamlı kullanım alanları vardır. Eltanashi ve Atasoy (2020) makine öğrenmesi yaklaşımını konuşma tanıma için kullanmışken, Tasdelen ve Sen (2021) miRNA'ları sınıflandırmak için kullanmıştır. Yalnızca karmaşık analizleri kolaylaştırmak için değil, aynı zamanda maliyet ve zamandan tasarruf etmek içinde kullanılmaktadır. Son zamanlarda insanlar ve elektronik cihazlar tarafından birçok veri oluşturulmaktadır. Veri miktarı büyüdükçe, veri analitiği o kadar karmaşık hale gelmekte ve makine öğrenmesinin önemi de artmaktadır (Smiti, 2020).

Makine öğrenmesi tıp alanında da kullanılmaktadır. Prostat kanseri üzerine yapılan çalışmalarda Regnier-Coudert vd. (2012) prostat kanserini patolojik olarak evrelemek için Prostat Spesifik Antijen (PSA), gleason skor ve klinik aşamadan oluşan partin tablosu ile karşılaştırmışlar ve Naive Bayes (NB) ile daha iyi sonuçlar elde etmişlerdir. Ge, Gao ve Chen (2015) prostat kanser tanılaması yapmak için oluşturulan modellerde yaş, toplam PSA, serbest PSA, serbest PSA oranı, boy, kilo, vücut kitle indeksi, prostat hacmi, PSA yoğunluğu verilerini kullanarak Lojistik Regresyon (LR) ile %86 ve Yapay Sinir Ağı (YSA) ile %87 başarı elde etmişlerdir.

Xiao vd. (2017) transrektal ultrason bulguları, yaş ve prostat spesifik antijenin serum seviyeleri hakkındaki veriler kullanılarak Rassal Ağaçlar (RA) algoritması ile prostat kanseri %83.10 doğruluk, %65.64 duyarlılık, %93.83 özgüllük ve %86.72 hassasiyetle tahmin edilmiştir. Çalışmalarında değişkenlerin tek başlarına yeterince güçlü olmadıklarını göstermişlerdir. Wen, H., Li, S., Li, W., Li, J. Ve Yin, C. (2018) yaptıkları çalışmada Gözetim, Epidemiyoloji ve Nihai Sonuçlar (SEER) prostat kanseri veri tabanında hastaların hayatta kalma sürelerini tahmin etmişlerdir. İki model üzerinde YSA %85.6, NB %71, Karar Ağacı (KA) %85 ve Destek Vektör Makinesi (DVM) %85.5 doğrulukla algoritmaların başarımını sağlamışlardır.

Nitta vd. (2019) klinik ve patolojik veriler kullanmış, YSA, DVM ve RA ile prostat kanser tespiti yaparak PSA yoğunluğu ve PSA hızından daha iyi tahmin yapan makine öğrenmesi modelleri sunmuşlardır. Srivenkatesh (2020) prostat kanseri tahmininde RA ve LR ile %90 doğruluk elde etmiştir. Murtojärvi (2020) Cox regresyon modellerinden en az mutlak büzülme ve seçim operatörü (Lasso) ve Greedy (açgözlü) iki değişken seçim yöntemi uygulayarak, hayatta kalma tahmin maliyetini önemli ölçüde azaltmışlardır. Çalışmalarında Lasso, en yüksek doğruluğu daha iyi verirken, Greedy yöntem ise düşük maliyet için daha iyi sonuç vermiştir. Lee vd. (2020) prostat kanserinin tekrarlamasını tahmin eden web tabanlı bir klinik karar destek sistemi önermişlerdir. KA, RA, YSA, LR ve Gradient Arıtma (GA) sınıflandırıcı algoritmaları modellenmiş ve çalışmada GA sınıflandırıcısı daha iyi performans göstermiştir.

Karunamuni vd. (2020) Cox regresyon modeliyle Afrika ırkına özgü prostat kanser tehlike skorunda iyileştirme yapmışlardır ve prostat kanser olasılığının ırka göre değiştiği vurgulanmıştır. Syed vd. (2020) prostat kanser hastalarını risk gruplarına ayırmak için T evresi, gleason skoru ve PSA gibi faktörleri kullanarak bilgi ve tedavi seçimi için çalışma yapmışlar ve %80'lerde başarı sağlamışlardır. Kaur, Doja ve Ahmad (2020) tedavi gören metastatik prostat kanser hastalarının her tedavi seti arasındaki zaman aralıklarıyla birlikte hastalara verilen tedavi sırasını analiz etmişlerdir. Boosting trees algoritmasının %84.5 doğrulukla daha iyi sonuçlar verdiği gösterilmiştir. Lasheras vd. (2020) prostat kanseri teşhisi için bir dizi özellikten en iyilerini seçebilen DVM ve genetik algoritmanın kullanıldığı hibrit bir algoritma önermişlerdir. Önerilen algoritma da 0.91 duyarlılık ve 0.87 özgüllük sağlanmış, kullanılan niteliklerin etki düzeyi sunulmuştur.

Deng, Li ve Guan (2020) klinik veriler ve patolojik sonuçlardan tedavinin devam edilip edilmeme durumu değerlendirmiş ve RA temel öğrenici olarak seçilmiştir. Auffenberg vd. (2019) prostat kanser hastaları için klinik verileri ve makine öğrenmesi yöntemlerini kullanarak tedavi kararlarını görüntülemek için web tabanlı bir sistem geliştirmişlerdir. Kişiselleştirilmiş bir tedavi tahmini sağlamak için klinikopatolojik ve demografik özellikleri RA kullanılarak bir öngörücü model geliştirilmiştir.

Yapılan literatür çalışmalarında prostat kanserinin teşhisinde prostat spesifik antijeni, yaş, gleason skor önemli belirteçler olmakla birlikte tek başlarına yeterince verimli olmadıkları görülmüştür. Özellikle PSA ve testosteron hormonunun da ne kadar etkili olduğu tam olarak kesinlik kazanmamıştır. Androjen baskılama tedavisi veya diğer tedavilerde PSA ve testosteronla beraber diğer klinik, patolojik sonuçlar ve kişisel özelliklerin değerlendirilmesi ihtiyacı devam etmektedir.

Bu çalışmada prostat kanseri olan hastalara ait yaş, ilk patolojik tanı gün sayısı, lenf düğüm aşaması (N stage), tümör aşaması (T stage), Tümör durumu (bağımlı değişken), kanser geçmişi, ırk, radyasyon terapi, hayatta kalma durumu ve süresi (ay), hastalığa özgü hayatta kalma durumu ve süresi (ay), hastalıksız durum ve süresi (ay), ilerlemesiz durum ve süresi (ay) klinik verileri kullanılarak tümör sınıflandırması gerçekleştirilmiştir.

2. Materyal ve Metot

Biyoinformatik; biyoloji, bilgisayar bilimi ve bilgi teknolojilerinin birleşiminden oluşan bir disiplin olarak

tanımlanmıştır. Yani biyolojik bilginin bilgisayar yardımı ile incelenmesi ve işlenmesidir. Disiplinler arası bir bilim olan biyoinformatik, biyolojik veriyi depolama teknikleri ve depodan bulma teknikleri geliştirir, düzenler ve analiz eder. Makine öğrenmesi yaklaşımları kullanılarak da hasta verileri üzerinde verilerin elde edilmesi, ön işlenmesi, normalizasyon veya standardizasyon yapılması, özellik seçimi veya boyut azaltma, sınıflandırma, değerlendirme ve hayatta kalma analizi yapılarak biyoinformatik analiz yapılabilmektedir.

Biyoinformatik analizlerde hastane ve laboratuvarlar kendi oluşturdukları klinik veya genomik verileri kullanabildikleri gibi, elde edilen verileri ilgili merkez ve enstitüler aracılığıyla çevrimiçi olarak paylaşarak kullanılabilmesine olanak sağlamaktadır. Ulusal Kanser Enstitüsü (Grossman vd., 2016), cBioPortal (Gao vd., 2013), Kanser Genom Atlası (Weinstein vd., 2013), Ulusal Biyoteknoloji Bilgi Merkezi (National Library, 2021), Gözetim, Epidemiyoloji ve Nihai Sonuçlar (SEER, 2021) çevrimiçi olarak veri sağlanabilecek ve sıklıkla kullanılan veri tabanlarıdır.

2.1. Prostat Kanserinde Klinik Veri Seti ve Değişken Seçimi

2.1.1. Veri Seti ve Ön İşleme Adımları

Çalışmada farklı veri tabanlarından hem klinik hem de genomik veri sağlayan cBioPortal platformundan Kanser Genom Atlas'ın sunmuş olduğu prostat kanser klinik verileri kullanılmıştır. Yaş, ilk patolojik tanı gün sayısı, Lenf düğüm aşaması (N stage), Tümör aşaması (T stage), Tümör durumu (bağımlı değişken), Kanser geçmişi, Irk, Radyasyon terapi, Hayatta kalma durumu ve süresi (ay), Hastalığa özgü hayatta kalma durumu ve süresi (ay), Hastalıksız durum ve süresi (ay), İlerlemesiz durum ve süresi (ay) gibi 16 özellik Tablo 1'de görüldüğü gibi belirlenmiştir. Tümör durumu bağımlı değişken olarak değerlendirilmiştir.

Klinik veriler üzerinde ön işleme olarak bağımlı değişkende eksik veriler setten çıkarılmıştır. Python programlama dili kullanılarak "LabelEncoder()" metodu ile her bir veriye alfabetik sıralamaya göre benzersiz bir tam sayı atanarak kategorik verilerin sayısal dönüşümü yapılmıştır. Bağımlı değişkende boş olan değerler setten çıkarılmıştır. Bağımsız değişkenlerde ise veri kaybı yaşamamak için o sütunun ortalama değerini alacak şekilde boş değerler düzenlenmiştir. String olarak alınan diğer sütunların sayısal dönüşümü "astype(float)" metoduyla yapılmıştır. Böylece, 16 özellik için 408 adet veri elde edilmiştir. 15 bağımsız ve 1 bağımlı değişken belirlenerek oluşturulan veri setinin bir kısmı Tablo 1'de görülmektedir. Bağımlı değişken olarak tümör durumu referans alınarak veri seti ilk önce %30 test ve %70 eğitim verisi olarak ayrılmış, ancak daha sonra %20 test ve %80 eğitim verisi ile daha iyi doğruluk sağladığı görüldüğü için bu değerler kullanılmıştır.

Tablo 1. Prostat kanseri örnek veri seti

Kişi_No	Tanı yaşı	İlk tanıdan geçen süre	N_aşaması	T_aşaması	Tümör Durumu	Ön Teşhis	Irk	Radyasyon tedavisi	Sağkalım durumu	Sağkalım süresi (Ay)	Hastalığa özgü sağkalım durumu	Hastalığa özgü sağkalım süresi	Hastalıksız durum	Hastalıksız geçen süre	İlerlemesiz durum	İlerlemesiz geçen süre
TCGA-2A-A8VL	51	621	0	1	0	0	1.927	0	0	20.416	0	20.416	0.0856	36.0844	0	20.416
TCGA-2A-A8VO	57	1701	0.165	3	0	0	1.927	0	0	55.923	0	55.923	0.0856	36.0844	0	55.923
TCGA-2A-A8VT	47	1373	1	5	0	0	1.927	1	0	45.139	0	45.139	0.0856	36.0844	0	45.139
TCGA-2A-A8VV	52	671	0	1	0	0	1.927	0	0	22.06	0	22.06	0.0856	36.0844	0	22.06
TCGA-2A-A8VX	70	1378	0	4	0	0	1.927	0	0	45.304	0	45.304	0.0856	36.0844	0	45.304
TCGA-2A-A8W3	69	863	0	3	0	0	1.927	0	0	28.372	0	28.372	0.0856	36,0844	1	6.51

2.1.2. Özellik Ölçeklendirme

Farklı ölçeklerde bulunan verileri, ortak bir sisteme taşımak ve karşılaştırılabilir hale getirmek için standardizasyon veya normalizasyon kullanılır. Veri standardizasyonu, öznitelikleri, ortalamaları 0 ve varyansı 1 olacak şekilde yeniden ölçeklendirme işlemidir. Standardizasyonu gerçekleştirmenin amacı, değerlerin aralığındaki farklılıkları bozmadan tüm özellikleri ortak bir ölçeğe indirgemektir.

Normalizasyon, özellikleri benzer bir ölçekte olacak şekilde dönüştürmek için kullanılır. Bu ölçek aralığı [0, 1] veya bazen [-1, 1] olarak ölçeklendirilir. Geometrik olarak dönüşüm n-boyutlu verileri n-boyutlu birim hiperküp haline getirir. Aykırı değerler olmadığında daha iyi performans göstermektedir. Bu çalışmada normalizasyon yapılarak daha iyi sonuçlar alınmıştır.

2.2. Makine Öğrenmesi Algoritmaları

Makine öğrenmesi, verileri iyileştirmek, açıklamak ve sonuçları tahmin etmek için verilerden yinelemeli olarak öğrenen çeşitli algoritmalar kullanır (Hurwitz ve Kirsch, 2018). Çalışmada cBioPortal'dan Kanser Genom Atlas prostat kanser klinik verileri üzerinde algoritmalara göre en iyi sonuçları veren parametreler GridSearchCV() metoduyla test edilerek uygulanmıştır. Oluşturulan modellerde belirtilen hiperparametre ve değerleri için tüm kombinasyonlar hesaplanır ve en uygun hiperparametre seti bulunur.

2.2.1. K-En Yakın Komşu

K-En yakın komşu (K-EYK) algoritması, sınıflandırma ve regresyon problemlerini çözmek için kullanılabilecek istatistiksel denetimli öğrenme tekniğidir. Eğitim setini ezberlemek ve ardından eğitim setindeki en yakın komşularının etiketlerine dayanarak herhangi bir yeni örneğin etiketini tahmin etmektir. Alan noktalarını tanımlamak için kullanılan özelliklerin, yakın noktaların aynı etikete sahip olma olasılığını artıracak şekilde etiketlemeleri ile ilgili olduğu varsayımına dayanmaktadır (Shalev ve David, 2014).

K-EYK'nun ilk adımı sırasında eğitim ve test verileri sisteme tanıtılır. K değeri belirlenerek en yakın veri noktaları seçilir. Öklid, Manhattan, Minkowski veya Hamming uzaklığı gibi yöntemlerle test verileri ile her eğitim verisi arasındaki mesafe hesaplanır. Mesafe değerine göre artan düzende sıralanır. Sıralanan diziden en üstteki K satırı seçilerek sınıflandırma yapılacaktır (Machine learning, 2021). Satırların en sık görülen sınıfını temel alarak test noktasına bir sınıf, sütunların en başarılı sınıfına bağlı olarak test noktasına bir sınıf atanır. Hamming uzaklık metriğinde GridSearchCV() metoduyla en iyi K değeri bulunmuş ve 10 kat çapraz doğrulama ile hesaplanmıştır. Burada Hamming mesafesi için $D_H = \sum_{i=1}^{k}(|x_i - y_i|)\,;\; x = y \rightarrow D = 0\; ve\; x \neq y \rightarrow D = 1$ formülü kullanılır.

2.2.2. Rassal Ağaçlar

RA, hem sınıflandırma hem de regresyon için kullanılan denetimli bir öğrenme algoritmasıdır. Daha çok sınıflandırma problemleri için kullanılmaktadır. Veri kümesinden rastgele örnekler seçilir. Her örnek için bir karar ağacı oluşturularak tahmin sonucu alınır. Tahmin edilen her sonuç için oylama yapılır. Nihai tahmin sonucu olarak en çok oylanan tahmin sonucu seçilerek en iyi çözüm elde edilir (Machine learning, 2021). Ormandaki ağaç sayısı n_estimators 15 ve bölünmenin kalitesini ölçmek için entropy parametre olarak belirlenmiştir. Entropi veri setinin düzensizliğinin ölçüsüdür ve dağılımın entropisi $H(x) = -\sum_{i=1}^{k} p_i \log_2 p_i$ formülü ile bulunur.

2.2.3. Gradyan Artırma

Gradyan artırma (GA), birçok zayıf öğreniciyi bir araya getirerek güçlü öğrenici elde etmek için kullanılmaktadır. Topluluk öğrenme ve yükseltme algoritmalarından biri olan gradyan artırma, tipik olarak karar ağaçları gibi zayıf tahmin modelleri topluluğu şeklinde bir tahmin modeli üreten regresyon ve sınıflandırma problemleri için denetimli bir makine öğrenme tekniğidir (Introduction to Machine, 2015). Her ağaçtan sonra bir iyileştirme yapmak için düğüm oluşturulmadan yaprak ile başlar. İlk tahmin ortalama değerdir. Her ağaç bir öncekinden ardışık düzende öğrenerek sığ ağaçlardan oluşan yeni modellerin eklendiği yinelemeli bir topluluk oluşturulur. Maliyet fonksiyonunu en aza indirmek için parametreler tekrarlanır. Güçlendirme aşamalarının sayısı n_estimators 100, criterion friedman_mse ve öğrenme oranı (α) 0.1 parametre değerleri GridSearchCV() metoduyla bulunmuştur. En iyi tahmin sonucunu bulmak için $Loss = \sum_{i=1}^{n} \frac{1}{2}(Y - \gamma)^2$ kayıp fonksiyonu kullanılmaktadır. Başlangıç değeri için $F_0(x) = \underset{\gamma}{argmin} \sum_{i=1}^{n} L(x_i, \gamma)$ ile γ göre türev alınır. γ tüm değerlerin aritmetik ortalamasıdır. $r_{im} = -[\frac{\partial L(Y_i, F(x_i))}{\partial F(x_i)}]_{F(x)=F_{m-1}(x)}$ ile her değerin tahmin değeri arasındaki fark bulunup R_{jm} hesaplanarak ilk eğitim yapılır. $\gamma_{jm} = \underset{\gamma}{argmin} \sum_{x_i} L(x_i, F_{m-1}(x_i) + \gamma)$ ile her yaprak için çıktı değerleri bulunur. $F_m(x) = F_{m-1}(x) + \alpha \sum_{j=1}^{J_m} \gamma_{jm} I(x \epsilon R_{jm})$ formülü ile değerler güncellenerek önceki modelin hatalarını azaltmak için eğitim gerçekleştirilir (Wikipedia contributors, 2021).

2.2.4. Destek Vektör Makinesi

DVM, sınıflandırma ve regresyon problemlerini çözmek için kullanılan makine öğrenme algoritmalarıdır. Genellikle sınıflandırma problemlerinde kullanılırlar. Örneklerin uzaydaki noktalar olarak temsil edilmesidir ve ayrı kategorilerin örnekleri olabildiğince geniş açık bir boşlukla bölünecek şekilde haritalanır. Yeni örnekler daha sonra aynı alana eşleştirilir ve boşluğun hangi tarafına denk geldiklerine bağlı olarak bir kategoriye ait oldukları tahmin edilir (Introduction to Machine, 2015). Kullanılan veriler iki boyutta doğrusal olarak ayrılabilir yapıda değildir. Boyut $z = x^2 + y^2$ formülü ile üç boyuta yükseltilerek veriler sınıflandırılmıştır. Bu nedenle kullanılacak çekirdek türü kernel rbf olarak belirlenmiştir.

2.2.5. Lojistik Regresyon

LR, bir veya daha fazla açıklayıcı değişkene dayalı ikili yanıtı tahmin etmek için kullanılır. Bir hedef değişkenin olasılığını tahmin etmek için kullanılan denetimli öğrenme sınıflandırma algoritmasıdır. Sonuçları tanımlayan olasılıklar, bir lojistik fonksiyon kullanılarak açıklayıcı değişkenlerin bir fonksiyonu olarak modellenir. Genelleştirilmiş doğrusal modelin özel bir durumu ve dolayısıyla doğrusal regresyona benzer bir durum olarak görülebilir (Introduction to Machine, 2015). Matematiksel olarak lojistik regresyon modeli, P (Y = 1)' i X' in bir fonksiyonu olarak tahmin eder. $h(x) = P(Y = 1|X)$ hipotezi iki olasılık sonucu toplamının $P(Y = 1|X) + P(Y = 0|X) = 1$ olmasıdır. GridSearchCV() metoduyla düzenleştirme parametresi 1.0 ve optimizasyon algoritması olarak saga belirlenmiştir.

2.2.6. Naive Bayes

NB algoritması, Bayes teoremini uygulamaya dayanan bir sınıflandırma tekniğidir. Bayes teoremi ise olasılığa dayalı sınıflandırma yöntemidir. Bayes teoremi bir sonucun sebebini ararken koşullu olasılıktan yararlanarak, sonucun hangi olasılıklarla, hangi sebeplerden kaynaklanmış olabileceğini bulmaya yardımcı olur. Test verisinden hareketle sistem öğrenmeyi gerçekleştirir ve en yüksek orana sahip olan örnek ilgili sınıfa dahil edilir. Temel fikir bir sınıftaki bir özelliğin varlığının, aynı sınıftaki diğer herhangi bir özelliğin varlığından bağımsız olduğudur (Kızılkaya ve Oğuzlar, 2018). Hesaplama kararlılığı için 'var_smoothing': np.logspace(0,-9, num=100) parametre değeri kullanılmıştır.

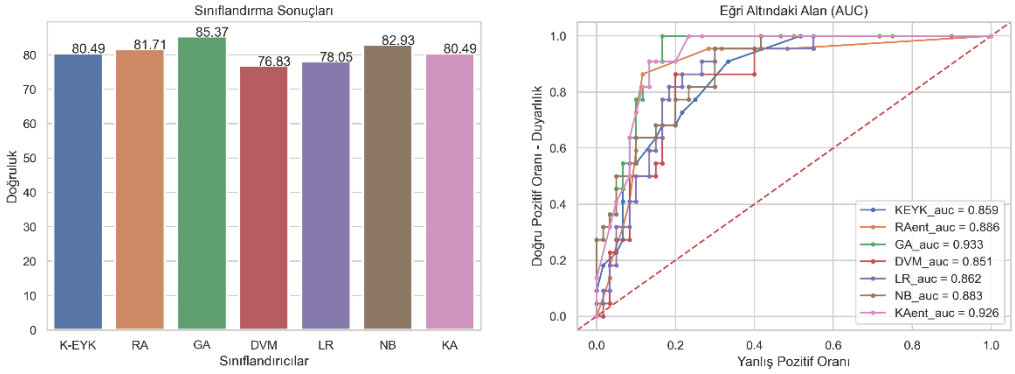
Bayes teoreminin matematiksel ifadesi; $P(y|x) = \frac{P(y)[P(x_1|y)P(x_1|y)\dots]}{P(y_1)P(x|y_1) + P(y_2)P(x|y_2)\dots} = \frac{P(y)\prod_{i=1}^{n} P(x_i|y)}{P(x)}$ şeklindedir. Burada *x*, veri setindeki her bir örneği; *y* ,kategori sayısını göstermektedir.

2.2.7. Karar Ağaçları

KA analizi birçok alanda uygulanabilen tahmine dayalı bir modelleme aracıdır. Bir kararın sonuçlarını göstermek için dallanma yapısı kullanır. Karar ağaçları, veri setini farklı koşullara göre farklı şekillerde bölebilen algoritmik bir yaklaşımla oluşturulabilir. Kararlar, denetimli algoritmalar kategorisine giren en güçlü algoritmalardır. Bir kararın olası sonuçlarını haritalamak için kullanılabilir. Her düğüm olası bir sonucu temsil eder (Hurwitz ve Kirsch, 2018). Hem sınıflandırma hem de regresyon görevleri için kullanılabilirler. Veri setinde bölünmenin kalitesini ölçmek için entropy parametresi kullanılmıştır. En ayırt edici özellik entropi ($H(x) = -\sum_{i=1}^{k} p_i \log_2 p_i$) ile bulunur ve kök olarak belirlenir. Daha sonra çocuk ve alt veri kümesi bulunur.

3. Deneysel Sonuçlar ve TartışmaAraştırma Sonuçları ve Tartışma

Çalışmada kullanılan her makine öğrenmesi algoritması için doğruluk, hata oranı, hassasiyet, duyarlılık, özgüllük, F1 skoru, Eğri Altında Kalan Alan (AUC) sınıflandırma metrikleri hesaplanmıştır. Şekil 2'de görülen değerlendirme sonuçlarına göre GA %85.37, NB %82.93, RA %81.71 olarak en iyi doğruluk değerlerine ulaşılırken, K-EYK %80.49, KA %80.49, LR %78.05, DVM %76.83 ile daha kötü doğruluk değerleri elde edilmiştir. AUC değeri sırasıyla GA, NB, RA, K-EYK, KA, LR, DVM için 0.933, 0.883, 0.886, 0.859, 0.926, 0.862 ve 0.851 olarak elde edilmiştir.



Şekil 1. Sınıflandırma doğruluk ve AUC sonuçları

4. Sonuç

Makine öğrenmesi, çeşitli alanlarda çok sayıda problemin üstesinden gelmeye yardımcı olmak için etkili yöntemler, teknikler ve araçlar sunmaktadır. Çoğu alanda olduğu gibi tıp alanında da teşhis ve tedavi süreçlerinde klinik verilerin değerlendirilmesinde makine öğrenmesi kullanılabilmektedir. Bu alanlardan birisi de prostat kanseridir. Literatürde yapılmış önceki çalışmalarda çoğunlukla RA algoritmasının daha iyi performans gösterdiği görülmüştür.

Bu çalışmada prostat kanseri olan hastalara ait klinik veriler üzerinde makine öğrenmesi algoritmaları ile tümörlü tümörsüz durumu sınıflandırılmış ve Gradyan artırma algoritması daha iyi performans göstermiştir. Lee vd. (2020) tarafından yapılan 5 yıllık hayatta kalma analizinde de %74 doğruluk ile Gradyan artırma algoritması daha iyi doğruluk sağlamıştır. Klinik verilerle yaptığımız çalışmada sıklıkla kullanılan yedi sınıflandırıcı arasında Gradyan artırma algoritması ile %85.37 doğrulukla daha iyi sonuçlar elde edilmiştir.

Kaynakça

Auffenberg, G. B., Ghani, K. R., Ramani, S., Usoro, E., Denton, B., Rogers, C., ... & Collaborative, M. U. S. I. (2019). *askMUSIC: leveraging a clinical registry to develop a new machine learning model to inform patients of prostate cancer treatments chosen by similar men.* European urology, 75(6), 901-907.

Deng, K., Li, H., & Guan, Y. (2020). *Treatment stratification of patients with metastatic castration-resistant prostate cancer by machine learning.* Iscience, 23(2), 100804.

Eltanashi, S., & Atasoy, F. (2020). *Proposed speaker recognition model using optimized feed forward neural network and hybrid time-mel speech feature.* ICATCES 2020 Proceeding Book, 130-140.

Gao, J., Aksoy, B. A., Dogrusoz, U., Dresdner, G., Gross, B., Sumer, S. O., Sun, Y., Jacobsen, A., Sinha, R., Larsson, E., Cerami, E., Sander, C., & Schultz, N. (2013).

Ge, P., Gao, F., & Chen, G. (2015, August). *Predictive models for prostate cancer based on logistic regression and artificial neural network.* In 2015 IEEE International Conference on Mechatronics and Automation (ICMA) (pp. 1472-1477). IEEE.

Goldenberg, S. L., Nir, G., & Salcudean, S. E. (2019). *A new era: artificial intelligence and machine learning in prostate cancer.* Nature Reviews Urology, 16(7), 391-403.

Grossman, Robert L., Heath, Allison P., Ferretti, Vincent, Varmus, Harold E., Lowy, Douglas R., Kibbe, Warren A., Staudt, Louis M. (2016). *Toward a Shared Vision for Cancer Genomic Data.* New England Journal of Medicine 375:12, 1109-1112

Hurwitz, J., & Kirsch, D. (2018). *Machine learning for dummies*. IBM Limited Edition, 75.

Integrative Analysis of Complex Cancer Genomics and clinical profiles using the cBioPortal. Science Signaling, 6(269). https://doi.org/10.1126/scisignal.2004088

Introduction to Machine Learning The Wikipedia Guide (p. 427). (2015). https://www.datascienceassn.org/sites/default/files/Introduction to Machine Learning.pdf

Karunamuni, R. A., Huynh-Le, M. P., Fan, C. C., Thompson, W., Eeles, R. A., Kote-Jarai, Z., ... & Practical Consortium. (2021). *African-specific improvement of a polygenic hazard score for age at diagnosis of prostate cancer.* International Journal of Cancer, 148(1), 99-105.

Kaur, I., Doja, M. N., & Ahmad, T. (2020). *Time-range based sequential mining for survival prediction in prostate cancer.* Journal of Biomedical Informatics, 110, 103550.

Kızılkaya, Y. M., & Oğuzlar, A. (2018). *Bazı denetimli öğrenme algoritmalarının R programlama dili ile kıyaslanması.* Dergi Karadeniz, 37(37), 90–98. https://doi.org/10.17498/kdeniz.405746

Lasheras, J. E. S., Lasheras, F. S., Donquiles, C. G., Tardón, A., Castaño-Vinyals, G., Palazuelos, C., ... & de Cos Juez, F. J. (2021). *Hybrid algorithm for the classification of prostate cancer patients of the MCC-Spain study based on support vector machines and genetic algorithms.* Neurocomputing, 452, 386-394.

Lee, S. J., Yu, S. H., Kim, Y., Kim, J. K., Hong, J. H., Kim, C. S., ... & Choi, I. Y. (2020). *Prediction system for prostate cancer recurrence using machine learning.* Applied Sciences, 10(4), 1333.

Machine learning with python tutorial in PDF. (n.d.). Retrieved September 28, 2021, from https://www.tutorialspoint.com/machine_learning_with_python/machine_learning_with_python_pdf_version.htm.

Murtojärvi, M., Halkola, A. S., Airola, A., Laajala, T. D., Mirtti, T., Aittokallio, T., & Pahikkala, T. (2020). *Cost-effective survival prediction for patients with advanced prostate cancer using clinical trial and real-world hospital registry datasets.* International journal of medical informatics, 133, 104014.

Nitta, S., Tsutsumi, M., Sakka, S., Endo, T., Hashimoto, K., Hasegawa, M., Hayashi, T., Kawai K., & Nishiyama, H. (2019). *Machine learning methods can more efficiently predict prostate cancer compared with prostate-specific antigen density and prostate-specific antigen velocity.* Prostate international, 7(3), 114-118.

Regnier-Coudert, O., McCall, J., Lothian, R., Lam, T., McClinton, S., & N'Dow, J. (2012). *Machine learning for improved pathological staging of prostate cancer: a performance comparison on a range of classifiers.* Artificial intelligence in medicine, 55(1), 25-35.

SEER (n.d.). Surveillance, epidemiology, and end results program. Retrieved October 1, 2021, from https://seer.cancer.gov/.

Shalev, S. & David, B. (2014). *Understanding machine learning: From theory to algorithms.* Cambridge: Cambridge University Press, pp. 258-267.

Smiti, A. (2020). *When machine learning meets medical world: Current status and future challenges.* Computer Science Review, 37, 100280.

Srivenkatesh, M. (2020). *Prediction of prostate cancer using machine learning algorithms.* Int. J. Recent Technol. Eng., vol. 8, no. 5, pp. 5353–5362.

Syed, K., Sleeman, W., Soni, P., Hagan, M., Palta, J., Kapoor, R., & Ghosh, P. (2021). *Machine-learning models for multicenter prostate cancer treatment plans.* Journal of Computational Biology, 28(2), 166-184.

Tasdelen, A., & Sen, B. (2021). *A hybrid CNN-LSTM model for pre-miRNA classification.* Scientific Reports, 11(1), 1-9.

U.S. National Library of Medicine. (n.d.). *National Center for Biotechnology Information.* Retrieved October 1, 2021, from https://www.ncbi.nlm.nih.gov/.

Weinstein, J. N., Collisson, E. A., Mills, G. B., Shaw, K. R., Ozenberger, B. A., Ellrott, K., Shmulevich, I., Sander, C., & Stuart, J. M. (2013). *The cancer genome Atlas Pan-Cancer Analysis Project.* Nature Genetics, 45(10), 1113–1120. https://doi.org/10.1038/ng.2764

Wen, H., Li, S., Li, W., Li, J., & Yin, C. (2018, December). *Comparision of four machine learning techniques for the prediction of prostate cancer survivability.* In 2018 15th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP) (pp. 112-116). IEEE.

Wikipedia contributors. (2021, October 17). *Gradient boosting*. Wikipedia. https://en.wikipedia.org/wiki/Gradient_boosting.

Xiao, L. H., Chen, P. R., Gou, Z. P., Li, Y. Z., Li, M., Xiang, L. C., & Feng, P. (2017). *Prostate cancer prediction using the random forest algorithm that takes into account transrectal ultrasound findings, age, and serum levels of prostate-specific antigen.* Asian journal of andrology, 19(5), 586.