Akciğer Kanseri Hastalarının Farklı Özelliklerine Göre Hayatta Kalma Olasılıkları ile İlgili Tahmin

Estimation for Survival of Lung Cancer Patients Depending on Different Attributes

Tuğba SARAÇ
Bilgisayar Mühendisliği Bölümü, Başkent Üniversitesi, Ankara, Türkiye
tsarac@tai.com.tr

***Özetçe*— Günümüzde, gelişen teknoloji, artan veri sayısı ve veri çeşitliliği ile birlikte, bu verilerin kaydedileceği elektronik kayıt ortamlarının maliyetindeki düşüş sayesinde pek çok alanda çok büyük veri yığınları oluşmuştur. Bu veri yığınlarından anlamlı bilgiye ulaşabilmek için çeşitli çalışmalar yapılmaktadır. Bu çalışmada, akciğer kanseri teşhisi koyulan hastaların farklı parametrelerine (özellik) ait veri seti, üç farklı algoritma ile anlamlı bilgiye dönüştürülmüş ve hastaların operasyondan sonra bir sene içerisinde hayatta kalıp kalamama durumlarına yönelik tahminler yapılmıştır. Buna göre en iyi sonuç, k=5 ve Cross Validation=10 parametreleri ile elde KNN Algoritması ile elde edilmiştir.**

Anahtar Kelimeler—Algoritma;Veri Madenciliği; Tahmin.

***Abstract*—Nowadays, huge volumes of data are available thank to developing technology, variety of data and reduction in electronic storage cost of data. Various studies are carried out in order to transform data to meaningful information. In this paper, data set which belongs to lung cancer patients were transformed meaningful information. post-operative life expectancy within one year r operation has been estimated. It is observed that best result has been achieved by using KNN Algorithm (k=5 and Cross Validation=10).**

Keywords—Algorithm; Data Mining; Estimation.

I. GİRİŞ

Teknolojideki hızlı gelişim, büyük miktarda verinin toplanması, depolanması ve analiz edilmesi süreçlerini kolaylaştırarak bilgiye ulaşma süresini oldukça azaltmıştır. Bu sayede verinin bilgiye dönüştürülerek, doğru zamanda karar alabilmesi için kullanıcılara sunulması hedeflenmiştir. Veri, bir durum hakkında birbiri ile bağlantısı henüz keşfedilmemiş varlıklardır. Bilgi ise, bu verinin bir anlam ifade edecek ve katma değer sağlayacak şekilde düzenlenerek anlamlı bir hale dönüştürülmüş halidir. Dünyada mühendislik ve tıp alanları başta olmak üzere pek çok alanda yararlı olabilecek verilerin keşfedilerek, bu verilen içindeki saklı örüntülerin ortaya çıkarılması ve bu örüntünün problemlerin çözümlerinde kullanılması gün geçtikçe yaygınlaşmaktadır.

Bu çalışmada, Wroclaw Thoracic Surgey Centre'da 2007-2013 yılları arasında toplanan akciğer kanseri teşhisi konularak akabinde operasyona alınan 470 hastaya ait veri seti kullanılmıştır. Bu veri seti, her bir hastanın operasyondan önce genel durumunu gösteren 16 özellik ile birlikte, 17. özellik olarak hastanın operasyondan sonra bir sene içinde hayatta kalıp kalmadığını gösteren veriyi içermektedir. Bu çalışmanın amacı, 470 gerçek veriden yola çıkarak bu veri setindeki saklı örüntüye göre, aynı teşhis koyulan yeni bir hastanın mevcut 16 özelliğinden yola çıkarak operasyondan sonra hayatta kalıp kalamama durumu ile ilgili bir tahmin yapmaktır.

Bu çalışmada mevcut veri seti, KNN Algoritması, Naive Bayes Algoritması ve Karar Ağacı yöntemi ile sınıflandırılmış ve sonuçları karşılaştırılmıştır. İlave olarak, bu verilerden sonucu en çok etkileyen özellikler **S**tatistical **P**ackage for the **S**ocial **S**ciences (SPSS) aracı kullanılarak Stepwise yöntemi ile seçilmiş ve seçilen bu veriler ile algoritmalar tekrar çalıştırılarak algoritmaların başarılarındaki değişiklikler gözlemlenmiştir.

Bu çalışmada kullanılan algoritmalar, Yeni Zelanda'daki Waikato Üniversitesi araştırmacıları tarafından makine öğrenmesi (machine learning) çalışmalarında kullanılmak üzere geliştirilen WEKA yazılımında çalıştırılmıştır.

Bu çalışmanın amacı, doktorlara kanser hastalarına uygulayacakları tedavide rehberlik etmektir. Önerilen yöntem ile, olası bir operasyondan sonra hastanın hayatta kalıp kalamayacağına dair yapılan tahmine göre, operasyon kararı alıp alma konusunda doktorlara ve hastalara kılavuzluk etmek amaçlamaktadır.

Ayrıca bu yöntem dünyada, yalnızca doktorlara ve hastalara kılavuzluk etmek amacı ile değil, hasta yakınlarına, sigorta şirketlerine, avukatlara ve sağlık politikası planlayıcılarına da karar aşamasında destek olmak üzere kullanılmaktadır.

Bu çalışma ASYU-2017 Konferansında sunulmuş ve sadece özet olarak
Bildiri Özetleri Kitabı (ISBN: 978-605-84722-8-0)'nda yayımlanmıştır.

II. VERİ SETİ ÖZELLİKLERİ

III. DATA SET SINIFLANDIRILMASI VE YORUMLANMASI

A. KNN Algoritması ile Sınıflandırma

KNN Algoritması, verilerin özelliklerden sınıflandırılmak istenen yeni bireyin daha önceki bireylerden k tanesine yakınlığına bakarak sınıflandırma yapan bir öğrenme algoritmasıdır. Çalışma prensibi basitçe, sınıflandırılma yapılmak istenen yeni verinin, kendisine en yakın k tane komşu hangi sınıfa dahil ise o sınıfa dahil edilmesi şeklinde özetlenebilir. Sınıflandırılmak istenen yeni verinin komşuları, elemanın kendisine olan uzaklıklarına göre tespit edilir.

KNN Algoritması, k=1, 3 ve 5 değerleri için Cross Validation=10 ve 20 olacak şekilde 6 kez çalıştırılmıştır. Burada bahsedilen k değeri, sınıflandırılmak istenen yeni veriye en yakın k komşu veri anlamına gelmektedir. Cross Validation değeri ise, veri setinin kaç alt kümeye bölüneceğini gösterir. KNN Algoritması alt kümelerden birisini eğitim kümesi olarak kabul ederek sistemi eğitir. Ardından bu eğitim sonucunu diğer bir alt küme üzerinde sınar. Bu işlemi belirtilen küme sayısı kadar tekrarlayarak sistemi iyileştirmeye çalışır.

Cross Validation=10, k=1 parametreleri ile çalıştırılan KNN Algoritmasının sonuçları Tablo-1'de verilmiştir.

Elde edilen sonuçlara göre, 470 veriden 363 adet veri doğru olarak sınıflandırılmış olup, doğruluk değeri %77 olarak hesaplanmıştır. Veri setine göre 70 hasta operasyon sonrasındaki bir yıl içinde hayatını kaybetmiştir. KNN Algoritması ise bu değeri 60 olarak tahmin etmiştir. Bir başka deyişle, 10 kişi algoritma tarafından yaşayacak şeklinde sınıflandırdığı halde bu 10 kişi de maalesef hayatını kaybetmiştir. Bu değerlere göre hayatta kalmama sınıflandırmasına ait Recall değeri 0.143 (10/70) olarak hesaplanmıştır. Benzer şekilde, veri setine göre 400 hasta operasyon sonrasında hayatta kalmış iken, bu değer KNN algoritması ile 353 olarak elde edilmiştir. Algoritma 47 kişiyi de hayatını kaybeder şeklinde sınıflandırdığı halde bu 47 kişi hayatta kalmıştır. Bu değerlere göre hayatta kalma sınıflandırmasına ait Recall değeri de 0.883 (363/400) olarak elde edilmiştir.

Cross Validation = 20, k=1 parametreleri ile KNN Algoritması çalıştırıldığında, "Correctly Classified Instances" değeri 360 değerine, doğruluk değerini de %76'ya düştüğü görülmüştür. Cross Validation = 10, k=3 parametreleri ile KNN Algoritması çalıştırıldığında, "Correctly Classified Instances" değeri 388 değerine, doğruluk değerini de %82'ye çıktığı görülmüştür. Elde edilen bu sonuçlara göre, 470 veriden 388 adet veri doğru olarak sınıflandırılmış olup, doğruluk değeri %82 olarak hesaplanmıştır. Veri setine göre 70 hasta operasyon sonrasındaki bir yıl içinde hayatını kaybetmiştir. KNN Algoritması ise bu değeri 63 olarak tahmin etmiştir. Bir başka deyişle, 7 kişi algoritma tarafından yaşayacak şeklinde sınıflandırdığı halde bu 7 kişi de maalesef hayatını kaybetmiştir. Bu değerlere göre hayatta kalmama sınıflandırmasına ait Recall değeri 0.10 (7/70) olarak hesaplanmıştır. Benzer şekilde, veri setine göre 400 hasta operasyon sonrasında hayatta kalmış iken, bu değer KNN algoritması ile 381 olarak elde edilmiştir. Algoritma 19 kişiyi de hayatını kaybeder şeklinde sınıflandırdığı halde bu 19 kişi hayatta kalmıştır. Bu değerlere göre hayatta kalma sınıflandırmasına ait Recall değeri de 0.95 (381/400) olarak elde edilmiştir.

Correctly Classified Instances				363	77.23 %	
Incorrectly Classified Instances				107	22.76 %	
TP Rate	FP Rate	Precision	Recall	ROC Area	PRC Area	Class
0,143	0,118	0,175	0,143	0,513	0,153	T
0,883	0,857	0,855	0,883	0,513	0,854	F
a b <-- classified as						
10 60 | a = T						
47 353 | b = F						

Tablo 1. Weka Sonuçları

Cross Validation = 20, k=3 parametreleri ile KNN Algoritması çalıştırıldığında, "Correctly Classified Instances" değeri 390 değerine, doğruluk değerini de %82.9'a çıktığı görülmüştür. Bu parametreler ile KNN Algoritması 63 kişinin hayatını kaybedeceğini, 383 kişinin ise hayatını kaybetmeyeceğini doğru olarak tahmin etmiştir.

Cross Validation = 10, k=5 parametreleri ile KNN Algoritması çalıştırıldığında, "Correctly Classified Instances" değerinin 400 değerine, doğruluk değerini de %85.1'a çıktığı görülmüştür**.** Bu parametreler ile KNN Algoritması 65 kişinin hayatını kaybedeceğini, 395 kişinin ise hayatını kaybetmeyeceğini doğru olarak tahmin etmiştir. Recall değerleri sıra ile 0,071 ve 0,988 olarak hesaplanmıştır.Cross Validation = 20, k=5 parametreleri ile KNN Algoritması çalıştırıldığında, "Correctly Classified Instances" değerinin 399 değerine, doğruluk değerini de %84.1'a düştüğü görülmüştür**.** Bu parametreler ile KNN Algoritması 65 kişinin hayatını kaybedeceğini, 394 kişinin ise hayatını kaybetmeyeceğini doğru olarak tahmin etmiştir. Recall değerleri sıra ile 0,071 ve 0,985 olarak hesaplanmıştır.

Bu çalışma ASYU-2017 Konferansında sunulmuş ve sadece özet olarak
Bildiri Özetleri Kitabı (ISBN: 978-605-84722-8-0)'nda yayımlanmıştır.

B. *Naive Bayes Yöntemi ile Sınıflandırma*

Bu sınıflandırma algoritması, olasılık ilkelerine göre tanımlanmış bir dizi hesaplama ile yeni bir elemanın sınıf üyelik olasılığını tahmin ederek istatistiksel bir kestirim yapar. Bu algoritmada sisteme yeni sunulan bir eleman, daha önce elde edilmiş olasılık değerlerine göre işlenir ve verilen test verisinin hangi kategoride olduğu tespit edilmeye çalışılır.

Bu çalışmada kullanılan veri seti, Naive Bayes sınıflandırma algoritması ile sınıflandırılmıştır. Bu sınıflandırma ile elde edilen sonuçların doğruluk değeri % 79 olarak elde edilmiştir. Bu algoritma, 70 kişinin 62 tanesini hayatını kaybedecek şeklinde doğru olarak sınıflandırırken, 8 tanesini yaşadığı halde hayatını kaybedecek şeklinde hatalı olarak sınıflandırmıştır. Benzer şekilde, veri setine göre 400 hasta operasyon sonrasında hayatta kalmış iken, bu değer Naive Bayes algoritması ile 364 olarak elde edilmiştir. Algoritma 36 kişiyi de hayatını kaybeder şeklinde sınıflandırdığı halde bu 36 kişi hayatta kalmıştır.

Veri seti, Cross Validation = 10 ve 20 olarak çalıştırılmıştır. Doğruluk ve Recall değerleri dikkate alındığında, Naive Bayes Algoritmasının KNN Algoritmasına göre daha kötü bir sonuç verdiği görülmüştür.

C. *Karar Ağacı Yöntemi ile Sınıflandırma*

Mevcut veri seti son olarak karar ağaçları yöntemi ile sınıflandırılmıştır. Bu yöntem sınıflandırma ve tahmin için sıkça kullanılan veri madenciliği yaklaşımlarından birisidir.

Bu yöntemde tüm veri kümesi bir kümeleme algoritması yardımı ile tekrar tekrar guruplara bölünür. Grubun tüm elemanları aynı sınıf etiketine sahip olarak kadar kümeleme işlemi derinlemesine devam eder.

Mevcut veri seti karar ağacı yöntemi ile çalıştırıldığında doğruluk değeri %84 olarak hesaplanmıştır. Bu algoritma, 70 kişinin 69 tanesini hayatını kaybedecek şeklinde doğru olarak sınıflandırırken, 1 tanesini yaşadığı halde hayatını kaybedecek şeklinde hatalı olarak sınıflandırmıştır. BENZER şekilde, veri setine göre 400 hasta operasyon sonrasında hayatta kalmış iken, bu değer Karar Ağacı yöntemi ile 394 olarak elde edilmiştir. Algoritma 6 kişiyi de hayatını kaybeder şeklinde sınıflandırdığı halde bu 6 kişi hayatta kalmıştır.

Buna göre her üç algotritma ile yapılan farklı denemelerde en iyi sonucun KNN Algoritması (**k=5 ve Cross Validation=10)** ile elde edildiği görülmüştür.

IV. STEPWISE YÖNTEMİ İLE SONUCU EN ÇOK ETKİLEYEN DEĞİŞKENİ TEPİT EDİLMESİ

Mevcut veri setinde yer alan veriler arasında sonucu en çok etkileyen verileri tespit etmek için Stepwise yöntemi (adım adım) kullanılmıştır. Bu yöntemde öncelikle en iyi tek değişkenli model ile başlanır ve her seferinde modele en yüksek katkıyı sağlayacak değişken ilave edilir. Yeni bir değişken ilavesi ile regresyonun hata kareleri toplamında meydana gelen değişmenin önemli olup olmadığına bakılır. Önceki adımda önemli bulunan bir değişken bir sonraki adımda önemsiz bulunabilir. Her adımda denkleme hangi yeni değişkenin katılmasının gerektiği, kısmi korelasyon katsayısı ile belirlenir. Bu yöntemde işlem, modele herhangi bir değişken ilave edilemez ya da atılamaz hale geldiğinde durdurulur.

Stepwise yöntemi ile sonucu en çok etkileyen değişkenler SPSS aracı ile tespit edilmiş ve aşağıdaki sonuçlar elde edilmiştir. Tespit edilen bu değişkenler yukarıda belirtilen üç algoritma ile tekrar işlenmiş ve elde ediln sonuçları karşılaştırılmıştır.

Buna göre sonucu en çok etkileyen özelliklerin X6, X11, OC, X14, X7, DSG olduğu tespit edilmiştir. Bu kez 16 özellik yerine, yalnızca bu özelliklere ait veriler ile KNN Algoritması, Naive Bayes Algoritması ve Karar Ağacı Yöntemi Algoritması tekrar çalıştırılmıştır. Buna göre, her üç algoritmadaki doğruluk değerleri ve tahmin edilen değerlerin karşılaştırılması Tablo 3'de verilmiştir. Stepwise yöntemine göre seçilen özellikler ile algoritmalar çalıştırıldığında, özellikle Naive Bayes algoritmasın da önemli bir iyileşme olduğu görülmektedir

Model Summary

Model	R	R Square	Adjusted R Square	Std.Error of the Estimate
1	0,140[a]	0,020	0,017	0,337
2	0,194[b]	0,038	0,032	0,334
3	0,227[c]	0,051	0,044	0,332
4	0,249[d]	0,062	0,052	0,331
5	0,271[e]	0,073	0,061	0,329
6	0,291[f]	0,085	0,070	0,328

a. Predictors: (Constant), X6
b. Predictors: (Constant), X6, X11
c. Predictors: (Constant), X6, X11,OC
d. Predictors: (Constant), X6, X11,OC,X14
e. Predictors: (Constant), X6, X11,OC,X14,X7
f. Predictors: (Constant), X6, X11,OC,X14,X7,DSG

Tablo 2.SPSS Sonuçları

Bu çalışma ASYU-2017 Konferansında sunulmuş ve sadece özet olarak
Bildiri Özetleri Kitabı (ISBN: 978-605-84722-8-0)'nda yayımlanmıştır.

	16 Özelliğe Göre Doğruluk Değeri True | False	Stepwise Yöntemi ile Seçilen Verilere göre Doğruluk Değeri True | False
KNN	% 85.1 5 | 65 5 | 395	% 84.68 0 | 70 2 | 398
Naive Bayes	% 79.14 8 | 68 36 | 364	% 84.25 4 | 66 8 | 392
Karar Ağacı	% 84.04 1 | 69 6 | 394	% 85.04 1 | 69 5 | 394

Tablo 3.Karşılaştırma

V. SONUÇ

Bu çalışmada, Wroclaw Thoracic Surgey Centre'da akciğer kanseri teşhisi konulan 470 hastaya ait 17 adet özellik kullanılarak üç farklı algoritma ile sınıflandırma yapılmış ve bu algoritmaların sınıflandırma başarıları karşılaştırılmıştır. Bu veri setindeki sınıflandırma, hastanın operasyondan sonraki bir yıl içerisinde hayatta kalıp kalamayacağı ile ilgili bir sınıflandırmadır. Çalışmada kullanılan algoritmalar, KNN Algoritması, Naive Bayes Algoritması ve Karar Ağacı Yöntemi Algoritması olup bu algoritmaların başarıları "Doğruluk" ve "Recall" değerlerine göre karşılaştırılmıştır.

Buna göre en iyi sonuç, k=5 ve Cross Validation=10 parametreleri ile elde KNN Algoritması olmuştur. Bu algoritma, 70 kişiden 65 kişinin hayatını kaybedeceğini, 400 kişiden de 395 kişinin hayatta kalacağını doğru bir şekilde tahmin etmiştir. Çalışmanın devamında, sonucu en çok etkileyen veriler Stepwise yöntemi ile tespiti edilerek yalnızca bu özellikler girdi olarak verilerek bu üç algoritma ile tekrar sınıflandırma yapılmıştır. Sonucu en çok etkileyen özellikler ile tekrar sınıflandırma yapıldığında, Navie Bayes algoritması ile elde edilen sonuçlarda doğruluk değerinin %79'dan %84'e çıktığı, Recall değerinde de önemli bir artış olduğu görülmüştür.

KAYNAKÇA

[1] T. H. Cormen, C. E. Leiserson, R. L. Rivest, C. Stein, "Introduction to Algorithms", Third Edition.

[2] M. Lubicz, K. Pawelczyk, A. Rzechonek, J. Kolodziej Wroclaw University of Technology, wybrzeze Wyspianskiego 27, 50-370, Wroclaw, Poland

[3] An Introduction to Statistical Learning with Applications in R, G. James, D. Witten, T. Hastie, R. Tibshirani

[4] Life Expectancy Post Thoracic Surgery, A. Abdulhamid, I. Bahtchevanov, P. Jia, Standford University

Bu çalışma ASYU-2017 Konferansında sunulmuş ve sadece özet olarak Bildiri Özetleri Kitabı (ISBN: 978-605-84722-8-0)'nda yayımlanmıştır.