Kalp Krizi Analizi İçin En Başarılı Sınıflandırma Algoritmasının Belirlenmesi
Determining The Most Successful Classification Algorithm For The Heart Attack Analysis

Aysu Yıldız Çaldıran , aysuyildizcaldiran@gmail.com

Öz

Pek çok alanda etkili bir şekilde kullanılan veri madenciliğinin, günümüzde eğitim alanındaki uygulamaları hızla artmaktadır. Veri madenciliği yöntemleri ile eldeki veriler sınıflandırılarak, gruplandırılarak ya da veriler arasında ilişkiler, bağıntılar, istatistiksel sonuçlar oluşturularak modeller oluşturulur. Oluşturulan model, oluşturulduğu veri kümesinde olmayan yeni bir kayıt geldiğinde, yeni gelen kayıt hakkında tahminleme yapar. Yapılan tahminlerin doğruluk derecesi oluşturulmuş modelin veri üzerindeki başarısını ortaya koyar. Bu çalışmada Kalp Krizi Analizi Ve Tahmini Veri Kümesi kullanılmıştır.Bu veri setine sınıflandırma algoritmaları uygulanarak en başarılı algoritma tespit edilmiştir. En başarılı algoritma tespit edilirken Doğruluk, Ortalama Mutlak Hata(MAE), Kök Hata Kareler Ortalaması(RMSE) göz önüne alınmıştır. Sonuç olarak kalp krizi veri setini sınıflandırmada en başarılı algoritmanın Gaussian Naive Bayes olduğu tespit edilmiştir.

Abstract

The applications of data mining, which is used effectively in many fields, in the field of education are increasing rapidly. With data mining methods, models are created by classifying and grouping the available data, or by creating relationships, correlations and statistical results between the data. The created model makes predictions about the new record when a new record arrives that is not in the dataset it was created. The accuracy of the predictions shows the success of the model on the data. In this study, Heart Attack Analysis and Prediction Dataset was used. By applying classification algorithms to this dataset, the most successful algorithm was determined. While determining the most successful algorithm, Accuracy, Mean Absolute Error (MAE), Root Error Mean of Squares (RMSE) were taken into consideration. As a result, it has been determined that Gaussian Naive Bayes is the most successful algorithm in classifying the data set obtained from student projects.

Anahtar Sözcükler: Veri Madenciliği, Eğitimde Veri Madenciliği, Veri Sınıflandırma, Sınıflandırma Algoritmaları, Birliktelik Kuralı, Apriori Algoritması.

1.GİRİŞ

Kalp krizi; kalp damarındaki plakların aniden yırtılması ve kalbi besleyen atardamarlarda gelişen herhangi bir ani tıkanma kalp kasının yeterince oksijen alamamasına neden olarak kalp dokusunda hasara yol açabilir. Kalbe kan akışından sorumlu olan damarların duvarlarında yağ yapıdaki kolesterol gibi maddeler birikir ve plak olarak adlandırılan yapıları oluşturur.
Plaklar zaman içinde çoğalarak damarı daraltır ve üzerlerinde çatlaklar oluşur. Bu çatlaklarda meydana gelen pıhtılar veya duvardan kopan plaklar damarları tıkayarak kalp krizine neden olabilir.

Erken zamanda ve doğru bir müdahale yapılarak damar açılmazsa kalp dokusu kaybı meydana gelir. Kayıp, kalbin pompalama gücünü azaltır ve kalp yetmezliği oluşur.Oksijensiz kalan kalp kası hücreleri bir süre sonra ölmeye başlar. Bu sürece kalp krizi (miyokart infarktüsü) adı verilir.

Kalp krizi belirtilerinde göğüs ağrısı veya göğüs de baskı hissi, göğüsdeki ağrının vücudun farklı bölgelerine yansıyan ağrısı,terleme,halsizlik,nefes darlığı,sersemlik,düzensiz kalp atışları gibi olaylar görülmektedir.[1]

Dünyada en fazla ölüm, kalp ve ona bağlı rahatsızlıklardan kaynaklı hastalıklar nedeniyle meydana geliyor. Dünya Sağlık Örgütü'nün (DSÖ) verilerine göre, her yıl ortalama 17 milyon kişi kalp ve damar hastalıkları sebebiyle hayatını kaybediyor.[2]

Tüm bunlar göz önüne alındığında kalp krizinin erken ön görülebilmesi insan sağlığı açısından çok önem arz ediyor.

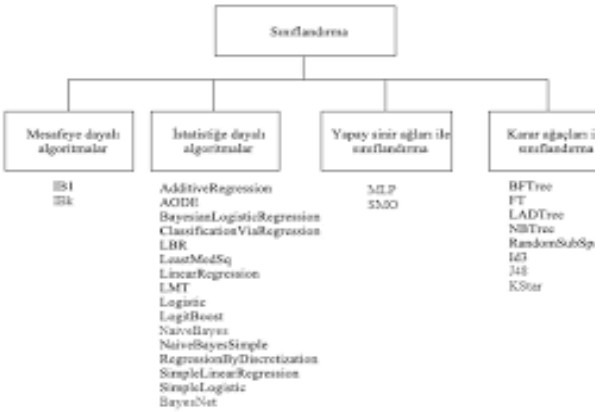
Bu çalışmada Heart Attack Analysis & Prediction Dataset kullanılarak sınıflandırma sıkça kullanılan sınıflandırma yöntemleri kullanılarak kişinin kalp krizi geçirme olasılığını sınıflandırmada(olumlu /olumsuz) en başarılı olan sınıflandırma algoritması tespit edilmiştir.

Ayrıca veriler arasındaki ilişkiler incelenmiştir. Sınıflandırmalarda algoritmaların başarıları değerlendirilirken Doğruluk değeri, Ortalama Mutlak Hata(MAE), Kök Hata Kareler Ortalaması(RMSE) vb. başarı değerleri göz önüne alınmıştır.

Elde edilen sonuçlar Bölüm 4’de yer almaktadır.

2.TEKNİKLER

Sınıflandırma sıkça kullanılan bir yöntemdir. Veri setinde bulunan her örneğin bir dizi niteliği vardır ve bu niteliklerden biri de sınıf bilgisidir. Sınıflandırma bir öğrenme algoritmasına dayanır, hangi sınıfa ait olduğu bilinen veri kümesi (eğitim kümesi) eğitim amacıyla kullanılır ve bir model oluşturulur. Oluşturulan model öğrenme kümesinde yer almayan veri kümesi (deneme kümesi) ile denenerek başarısı ölçülür. Oluşturulan bu model kullanılarak hangi sınıfa ait olduğu bilinmeyen bir kayıt için bir sınıf belirlenebilir. Aşağıda sınıflandırma algoritmaları çerçevesinde hiyerarşik yapı ve bu yapı üzerinden kullanılan algoritmalar Şekil 1'de[3] gösterilmiştir.



Şekil-1: Sınıflandırma algoritmasının hiyerarşik yapısı

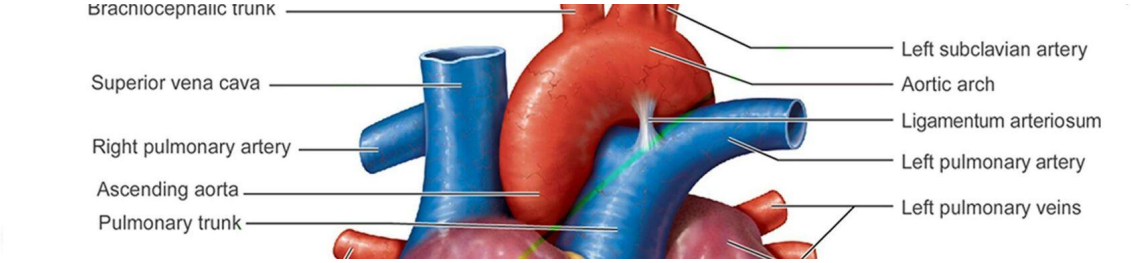
Çalışma kapsamında mesafeye dayalı algoritmalar ile sınıflandırma, istatistiğe dayalı algoritmalar ile sınıflandırma, yapay sinir ağı ile sınıflandırma ve karar ağaçları ile sınıflandırma da kullanılan bazı algoritmalar incelenmiştir (Tablo 1).

Tablo-1: Kullanılan veri madenciliği teknikleri

Sınıflandırma Teknikleri	Sınıflandırma Algoritması
Mesafeye Dayalı Algoritmalar	**IBk;** En yakın K-Komşu (K-Nearest Neighbors) algoritmasıdır. Bu algoritma sınıflandırma için kullanılır. K tabanlı komşuların uygun değerini çapraz doğrulama ile seçebilir. Ayrıca mesafe ağırlıklandırabilir.
	KStar; K*, örnek tabanlı bir sınıflandırıcıdır. Bazı benzerlik fonksiyonlarıyla belirlendiği gibi, eğitim örnekleriyle aynı olan sınıfa istinaden, test örneğinin sınıfıdır. Diğer örnek tabanlı öğrenenlerden entropi tabanlı mesafe fonksiyonu kullanması yönüyle farklıdır [4].
İstatistiğe Dayalı Algoritmalar	**Gaussian Naive Bayes (NB)**; Naive Bayes sınıflandırıcı Bayes teoremine dayanmaktadır. Bu teorem bir rassal değişken için olasılık dağılımı içinde koşullu olasılıklar ile marjinal olasılıklar arasındaki ilişkiyi gösterir. Veri kümesindeki her özelliğin sınıflama problemine eşit katkıda bulunduğu ve katkıların birbirinden bağımsız olduğu varsayıldığında basit bir sınıflama olan NB sınıflayıcısı kullanılabilir.

3. Materyal

Çalışma “Heart Attack Analysis & Prediction Dataset” kullanılarak ilerlenmiştir. [5]



Veri seti öznitelikleri :

- Age: Hastanın yaşı
- Sex: Hastanın cinsiyeti
- exang: Egzersize bağlı anjina (1 = evet; 0 = hayır)
- ca: Büyük gemi sayısı (0-3)
- cp : Göğüs Ağrısı tipi Göğüs ağrısı tipi
    - Değer 1: Tipik anjina
    - Değer 2: Atipik anjina
    - Değer 3: Anjin olmayan ağrı
    - Değer 4: Asemptomatik
- trtbps : Kan basıncı (mmHg cinsinden)
- chol : BMI sensörü aracılığıyla getirilen mg/dl cinsinden kolesterol
- fbs : (açlık kan şekeri > 120 mg/dl) (1 = doğru; 0 = yanlış)
- rest_ecg : Elektrokardiyografik sonuçları
    - Değer 0: Normal
    - Değer 1: ST-T dalgası anormalliğine sahip olmak (T dalgası inversiyonları ve/veya > 0.05 mV ST yükselmesi veya çökmesi)
    - Değer 2: Estes kriterlerine göre olası ve ya kesin sol ventrikül hipertrofisini gösteriyor
- thalach: Ulaşılan maksimum kalp atış hızı
- target: 0= daha az kalp krizi geçirme şansı 1= daha fazla kalp krizi geçirme şansı

Veri seti 10 sütun 303 satırdan oluşmaktadır.

4. Uygulama

Sınıflandırma için GaussianNB , KNeighborsClassifier , SVC(Support Vector Classifier) gibi sınıflandırma algoritmaları analiz edilerek en başarılı algoritma bulunmuştur.En başarılı algoritma bulunurken başarı metriklerinin sonuçları dikkate alınmıştır.Ancak sadece doğruluk metriklerine bakarak algoritma seçimi yapmamız sağlıklı değildir. Doğruluk değerinin yanı sıra MAE ve RMSE kriterleri de göz önüne alınmıştır. MAE ve RMSE negatif eğilimlidirler ve düşük değerler daha iyidir. Bu yüzden düşük MAE değerine sahip algoritmalar daha iyi algoritmalar olarak değerlendirilir.

Değerlendirmelerin yapılabilmesi için veri seti indirilip Colab platformunda görüntülenmiştir. Veri setinden bir kısmı şekil 2. de görebilirsiniz.

	age	sex	cp	trtbps	chol	fbs	restecg	thalachh	exng	oldpeak	slp	caa	thall	output
0	63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
1	37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
2	41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
3	56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
4	57	0	0	120	354	0	1	163	1	0.6	2	0	2	1
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...

Şekil-2: Veri seti

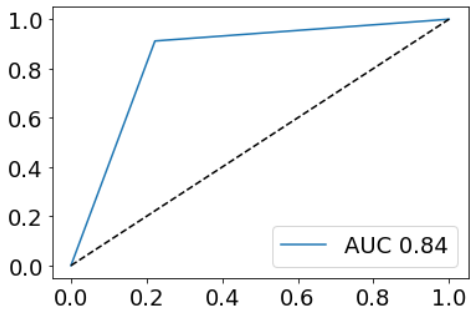
Veriler incelendiğinde 303 adet veri ve bunların 165 adedi olumlu , 138 adedi olumsuz çıktılar verdiği görülmüştür. Veri setinde eksik değerli veri veya kayıp veri görülmemiştir.

Verilerin birbiri arasındaki ilişkiler tablolar yardımıyla incelendiğinde kalp krizi geçirme olasılığının yüksek olduğu durumlar aşağıdaki gibi saptanmıştır.

- 40-60 yaş arasında
- Erkeklere oranla kadınlarda
- Göğüs ağrısı tipine göre değer
- 3-Anjin olmayan ağrı görülenlerde
- Kolestrol seviyesinin 200 ile 280 arasında olanlarda
- Elektrokardiyografik sonuçlarında ST-T dalgası anormalliğine sahip olanlarda

vb. durumlarda kalp krizi geçirme olasılığının yüksek olduğu görülmüştür.Tür dönüşüm gereksinimlerini tamamlayıp veri ön işleme adımını başarıyla bitirdikten sonra veri setini test ve eğitim olarak istediğimiz oranda 2 kümeye bölüyoruz.

Eğitim veri setini GaussianNB, KNeighborsClassifier, SVC algoritmalarıyla eğitimini tamamladıktan sonra her bir algoritmanın modelini üretiyoruz. Üretilen modelleri test veri kümesi ile testini tamamladıktan sonra doğruluk metriklerini kullanarak değeri hesaplanıyor. Ancak doğruluk ölçütü tek başına yorumlanırsa değerlendirme yanlış sonuçlara götürebilir. Bu ölçütü MAE , R-Kare,MSE, MedAE ve RMSE ölçütleriyle beraber ele almak gerekir. Bu ölçütler çerçevesinde değerlendirildiğinde, sınıflandırmada en başarılı algoritma Doğruluk, MAE, MSE,RMSE,R-Kare, MedAE değerleri sırasıyla %84, 0.15, 0.15 ,0.38, 0.40 ,0.00, olan GaussianNB algoritmasıdır. MAE ve RMSE sonuçları oldukça küçük olup oldukça gelecek vadediyor.GaussianNB algoritmasının AUC grafiğini şekil 3. de görebilirsiniz.



Şekil-3 : AUC grafiği

KAYNAKÇA

[1] Internet: Kalp Krizi Nedir ?
,https://www.medicalpark.com.tr/kalp-krizi-belirtileri-nelerdir/hg-1851

[2] Internet: INDEPENDENT
https://www.indyturk.com/node/545246/sa%C4%9Flik/gen%C3%A7-ya%C5%9Fta-kalp-krizi-kaynakl%C4%B1-%C3%B6l%C3%BCmler-artt%C4%B1%E2%80%A6-kalp-damar-uzman%C4%B1-art%C4%B1%C5%9F-oran%C4%B1#:~:text=Bilindi%C4%9Fi%20gibi%20d%C3%BCnyada%20en%20fazla,kaynakl%C4%B1%20hastal%C4%B1klar%20nedeniyle%20meydana%20geliyor.&text=D%C3%BCnya%20Sa%C4%9Fl%C4%B1k%20%C3%96rg%C3%BCt%C3%BC'n%C3%BCn%20(DS%C3%96,damar%20hastal%C4%B1klar%C4%B1%20sebebiyle%20hayat%C4%B1n%C4%B1%20kaybediyor.

[3] Internet : Google

[4] Bırtıl, F. S. Kız meslek lisesi öğrencilerinin akademik başarısızlık nedenlerinin veri madenciliği tekniği ile analizi, Yüksek Lisans Tezi, Afyon Kocatepe Üniversitesi, Fen Bilimleri Enstitüsü, 2011.

[5] Internet : Kaggle ,
https://www.kaggle.com/datasets/rashikrahmanpritom/heart-attack-analysis-prediction-dataset