Avrupa Bilim ve Teknoloji Dergisi
Özel Sayı 24, S. 112-118, Nisan 2021

Araştırma Makalesi

www.ejosat.com ISSN:2148-2683

European Journal of Science and Technology
Special Issue 24, pp. 112-118, April 2021

Research Article

Kalp Yetmezliği Hastalarının Sağ Kalımlarının Sınıflandırma Algoritmaları ile Tahmin Edilmesi

Ezgi Aktaş Potur[1*], Nihal Erginel[2]

[1*] Gazi Üniversitesi, Mühendislik Fakültesi, Endüstri Mühendisliği Bölümü, Ankara, Türkiye, (ORCID: 0000-0003-0192-8655), ezgiaktas@gazi.edu.tr
[2] Eskişehir Teknik Üniversitesi, Mühendislik Fakültesi, Endüstri Mühendisliği Bölümü, Eskişehir, Türkiye (ORCID: 0000-0001-6231-9904), nerginel@eskisehir.edu.tr

(2nd International Conference on Access to Recent Advances in Engineering and Digitalization (ARACONF)-10–12 March 2021)

(**DOI:** 10.31590/ejosat.902357)

ATIF/REFERENCE: Aktaş Potur, E., Erginel, N. (2021). Kalp Yetmezliği Hastalarının Sağ Kalımlarının Sınıflandırma Algoritmaları ile Tahmin Edilmesi. *Avrupa Bilim ve Teknoloji Dergisi*, (24), 112-118.

Öz

Kalp yetmezliği, son yıllarda giderek yaygınlaşan kronik bir hastalıktır. Hastaların ölüm oranları çok yüksektir ve bu durum hastalığın en ciddi kalp hastalıklarından birisi olduğunu göstermektedir. Hastaların hayatta kalma oranı meme kanseri, prostat kanseri ve bağırsak kanseri gibi kanser türlerine göre daha düşüktür. Kalp yetmezliği ile yaşayan hastaların sağ kalımlarının tahmin edilmesinin kritik önemi vardır. Sağ kalım tahmini ile en önemli risk faktörlerinin belirlenmesi ve hastalığın erken aşamada teşhisi sağlanabilir. Veri madenciliği teknikleri son yıllarda klinik verilerin analiz edilmesi ve sınıflandırılması üzerinde büyük gelişim göstermiş, hekimlere ve hastalara faydalar sağlamıştır. Bu çalışmada kalp yetmezliği hastalarının sağ kalımlarının tahmin edilmesi amacıyla Naive Bayes, lojistik regresyon, çok katmanlı algılayıcı, destek vektör makineleri ve J48 karar ağacı sınıflandırma yöntemleri WEKA'da bulunan InfoGainAttributeEval, CfsSubsetEval ve ReliefFAttributeEval öznitelik seçim yöntemleri kullanılarak değerlendirme ölçütleri açısından karşılaştırılmıştır. Değerlendirme ölçütü olarak doğru sınıflandırma oranı, F-ölçütü ve Kappa istatistiği metrikleri kullanılmıştır. En yüksek sınıflandırma başarısına sahip sınıflandırıcı %90 doğru sınıflandırma oranı ile çok katmanlı algılayıcı olmuştur.

Anahtar Kelimeler: Veri Madenciliği, Kalp Yetmezliği, Sınıflandırma.

Predicting Survival of Heart Failure Patients via Classification Algorithms

Abstract

Heart failure is a chronic disease that has become increasingly common in recent years. Patients' mortality rates are very high, indicating that the disease is one of the most serious heart diseases. The survival rate of patients is lower than cancer types such as breast cancer, prostate cancer and bowel cancer. Predicting the survival of patients living with heart failure is critical. The most important risk factors can be determined and the disease can be diagnosed at an early stage via prediction of survival. Data mining techniques have made great progress in analyzing and classifying clinical data in recent years, providing benefits to physicians and patients. In this study, Naive Bayes, logistic regression, multilayer perceptron, support vector machines and J48 decision tree classification methods were compared in terms of evaluation metrics using InfoGainAttributeEval, CfsSubsetEval and ReliefFAttributeEval feature selection methods in WEKA. The accuracy rate, F-measure and Kappa statistics metrics were used as evaluation metrics. The classifier with the highest classification success was the multilayer perceptron with 90% correct classification rate.

Keywords: Data Mining, Heart Failure, Classification.

[*] Sorumlu Yazar: ezgiaktas@gazi.edu.tr

1. Giriş

Kalp yetmezliği, kalp kasının vücudun ihtiyaç duyduğu kan ve oksijeni sağlayacak düzeyde kanı vücuda gönderemediği durumda gelişen kronik bir rahatsızlıktır. Kalp yetmezliğinin başlıca sebepleri koroner arter hastalığı, yüksek kan basıncı ve kalp krizi geçmişinin bulunmasıdır [1, 2]. Amerika'da 2030 yılına kadar kalp yetmezliğinin %46 artış göstererek 8 milyonun üzerine çıkacağı öngörülmektedir. Türkiye'de kalp yetmezliği hastalığına sahip kişi sayısı 2 milyonun üzerindedir. Kalp yetmezliği ile yaşayan insanların sağ kalım oranları meme kanseri, bağırsak kanseri ve prostat kanserine kıyasla daha düşüktür. Kalp yetmezliği hastalığına sahip insanların en az bir kez hastaneye yatış oranı %83'tür. Bu hastaların %50'si yoğun bakım ünitelerinde izlenmektedir [3].

Sağlık alanında toplanan veri miktarı her geçen gün artmaktadır. Artan veri miktarı, tespit edimesi zor olan gizli bilgi ve ilişkilerin ortaya çıkarılması ihtiyacını doğurmuştur. Tıbbi verilerin analiz edilmesinde veri madenciliği teknikleri hayati öneme sahiptir. Kalp hastalığının artan seyri ve yüksek ölüm oranları araştırmacıları veri madenciliği teknikleri ile hastalıkların mümkün ölçüde önlenmesi, erken aşamada teşhis edilmesi ve hastane ölümlerinin önüne geçilebilmesi için çalışmalar yapmaya teşvik etmiştir [3, 4].

Saqlain vd. çalışmalarında kalp yetmezliği ile yaşayan hastaların 1 yıl ve daha fazla süre sağ kalımlarının tahmin edilmesi için bir model önermişlerdir. Naive Bayes algoritması ile %86,7 doğru sınıflandırma oranına ulaşmışlardır [5]. Jagad vd. koroner arter hastalığının erken dönemde teşhis edilebilmesi için Naive Bayes, karar ağacı ve sinir ağları sınıflandırıcılarının performansları değerlendirmişlerdir. Üç algoritma arasında en hızlı algoritma Naive Bayes olmuştur. Hatayı en küçükleyen sinir ağları ise görece daha yüksek hesaplama zamanına sahiptir. Naive Bayes ve sinir ağı için ulaşılan doğru sınıflandırma oranları sırasıyla %86 ve %85,7'dir [6]. Küçükakçalı vd. veri madenciliği yöntemlerinden birliktelik kurallarını temel alan ilişkisel sınıflandırmayı kullanarak kalp yetmezliğine bağlı ölüm olaylarının tahmin edilmesini amaçlamışlardır. 299 örnek ve 13 öznitelikten oluşan kalp yetmezliği veri seti kullanılarak gerçekleştirilen çalışmada doğru sınıflandırma oranı, dengeli doğruluk, duyarlılık, özgüllük, pozitif prediktif değer, negatif prediktif değer ve F ölçütü değerlendirme ölçütleri için sırasıyla 0,866, 0,819, 0,688, 0,951, 0,868, ve 0,865 ve 0,767 sonuçlarına ulaşılmıştır [7]. Chicco ve Jurman çalışmalarında 299 örnekten oluşan kalp yetmezliği veri setini kullanarak hastaların sağ kalım oranlarının tahmini ve en önemli risk faktörlerinin ortaya çıkarılması adına sınıflandırma yöntemlerini analiz etmişlerdir. Lojistik regresyon sınıflandırıcısı ile %83,8 doğru sınıflandırma oranına ulaşmışlardır. [8]. Gürfidan ve Ersoy kalp yetmezliği hastalarının klinik bilgilerini ve yaşamlarına ait bilgileri içeren UCI web sitesinden alınan kalp yetmezliği veri setini kullanarak kalp hastalığına bağlı ölüm oranlarının değerlendirilmesini, hastaların ve hekimlerin erken tanıya yönlendirilmesini amaçlamışlardır. Sınıflandırma başarıları %73 ile %83 arasında değişen 6 farklı sınıflandırma algoritması içinde en başarılı sınıflandırıcının destek vektör makineleri olduğu sonucuna varmışlardır [9].

Bu çalışmada 299 örnekten oluşan kalp yetmezliği veri seti kullanılarak hastaların 4-285 gün arasında değişen, ortalama 130 günlük takip süresi içindeki sağ kalımları tahmin edilmiştir. Sınıflandırma işlemi öncesinde WEKA yazılımında bulunan InfoGainAttributeEval, CfsSubsetEval ve ReliefFAttributeEval öznitelik seçim yöntemlerinden yararlanılarak daha önce bu veri setinin kullanıldığı çalışmalara göre sınıflandırma başarısının arttırılması amaçlanmıştır. Naive Bayes, lojistik regresyon, çok katmanlı algılayıcı, destek vektör makineleri ve J48 karar ağacı sınıflandırma algoritmalarının performansları doğru sınıflandırma oranı, F-ölçütü ve Kappa istatistiği değerlendirme ölçütleri açısından karşılaştırılmıştır.

2. Materyal ve Metot

Bu çalışmada UCI web sitesinden alınan kalp yetmezliği veri seti kullanılmıştır. Veri seti 13 öznitelik ve 105'i kadın, 194'ü erkek olan 299 hastaya ait kayıt içermektedir. Hastalardan 96'sı takip edildiği süre içerisinde hayatını kaybetmiştir. Veri setine ilişkin ayrıntılı bilgi Tablo 1'de yer almaktadır.

Tablo 1. Çalışmada Kullanılan Öznitelikler ve Açıklamaları

Öznitelik Adı	Öznitelik Tanımı	Öznitelik Türü	Veri Türü
Yaş	Hasta yaşı	Girdi	Nümerik
Anemi	Kırmızı kan hücrelerinde veya hemoglobinde gerçekleşen azalma	Girdi	Kategorik(0:Yanlış, 1:Doğru)
Yüksek Tansiyon	Hastada hiper tansiyon olması	Girdi	Kategorik(0:Yanlış, 1:Doğru)
Kreatinin Fosfokinaz (CPK)	Kandaki kreatinin fosfokinaz enzim seviyesi	Girdi	Nümerik
Diyabet	Hastanın diyabetinin olması	Girdi	Kategorik(0:Yanlış, 1:Doğru)
Ejeksiyon Fraksiyonu	Kalbin her kasılmasında kalpten çıkan kan miktarı	Girdi	Nümerik
Trombosit	Kandaki trombositler	Girdi	Nümerik
Cinsiyet	Hasta cinsiyeti	Girdi	Kategorik(0:Kadın, 1:Erkek)
Serum Kreatinin	Kandaki serum kreatinin seviyesi	Girdi	Nümerik
Serum Sodyum	Kandaki serum sodyum seviyesi	Girdi	Nümerik
Sigara İçme Durumu	Hastanın sigara içme durumu	Girdi	Kategorik(0:Yanlış, 1:Doğru)
Süre	Gün bazında takip süresi	Girdi	Nümerik
Ölüm Olayı	Hastanın takip süresi içinde ölümü	Çıktı	Kategorik(0:Yanlış, 1:Doğru)

2.1. Öznitelik Seçimi

Çalışmada WEKA yazılımında bulunan InfoGainAttributeEval, CfsSubsetEval ve ReliefFAttributeEval yöntemleri en etkili özniteliklerin ortaya çıkarılması amacıyla kullanılmıştır. Öznitelikleri derecelendirmek için kullanılan InfoGain yöntemi entropi kavramına dayanmaktadır. Bu yöntem ile hedef sınıfa göre hesaplanan bilgi kazancı ile özniteliğin önemi ölçülmektedir. Entropi "H" ile ifade edilirse bilgi kazancı şu şekilde hesaplanmaktadır [10, 11]:

Bilgi kazancı(Sınıf, Öznitelik)= H(Sınıf)- H(Sınıf | Öznitelik) (1)

Best fit arama algoritmasından yararlanan CfsSubsetEval yöntemi ise sınıf etiketi ile en yüksek ilişkiye sahip özniteliklerin seçilmesi için öznitelik alt kümesinin değerini özniteliklerin her birinin bireysel tahmin yeteneğini ve öznitelikler arasındaki fazlalık derecesini dikkate alarak değerlendirmektedir [12, 13]. ReliefFAttributeEval yöntemi bir örneği tekrar tekrar örnekleyerek özniteliğin önemini değerlendiren ağırlık tabanlı bir öznitelik seçim yöntemidir. Her bir öznitelik, sınıf ile ilişkisine göre ağırlıklandırılmaktadır. Bu yöntemde başlangıçta tüm ağırlıklar sıfır olarak ayarlanmıştır. Ağırlık hesaplaması için rastgele seçilen örnekler kullanılmaktadır. Her bir tekrarda rastgele bir i örneği seçilmekte ve bu örneğin her bir öznitelik değerinin örneğe en yakın örnekler arasında ne derece iyi bir ayrım yaptığı tahmin edilmektedir. Algoritma her bir özniteliğin ağırlığını tekrarlı olarak güncellemekte ve en yüksek ağırlığa sahip belirli sayıda öznitelik seçilmektedir [14, 15].

2.2. Sınıflandırma Algoritmlaları

Kalp yetmezliği hastalarının sağ kalımlarının tahmini için gerçekleştirilen bu çalışmada Naive Bayes, destek vektör makineleri, lojistik regresyon, çok katmanlı algılayıcı ve J48 karar ağacı sınıflandırma algoritmaları kullanılmıştır.

2.2.1. Naive Bayes Algoritması

Naive Bayes, Thomas Bayes'in Bayes teoremini temel alan istatistiksel bir sınıflandırma algoritmasıdır. Özniteliklerin belirli bir sınıf üzerindeki etkisinin diğer özniteliklerin aldığı değerlerden bağımsız olduğu (koşullu bağımsızlık) varsayımının yapıldığı bu sınıflandırıcı ile koşullu sınıf olasılıkları hesaplanarak sonuç değişkeni tahmin edilmektedir [16, 17].

Koşullu bağımsızlık varsayımı sayesinde, rassal değişkenlerin tüm kombinasyonları için koşullu sınıf olasılığının hesaplanması yerine yalnızca verilen bir sınıf etiketi için her bir rassal değişkenin koşullu olasılığı hesaplanmaktadır. Naive Bayes sınıflandırıcısı ile test verilerinin sınıflarının tahmin edilmesi için kullanılan ifade şu şekildedir [17]:

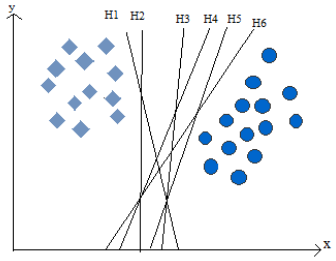
$$(Y|X) = \frac{P(Y)\prod_{i=1}^{n} P(X_i|Y)}{P(X)} \quad (2)$$

Burada sınıf etiketi Y ile, i. özniteliğin aldığı değer X_i (i=1,…, n) ile gösterilmiştir.

2.2.2. Destek Vektör Makineleri

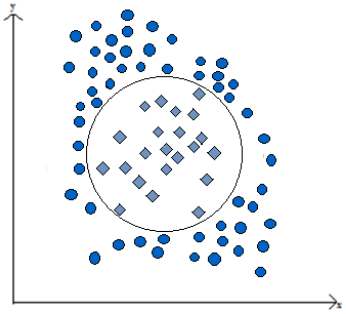
Destek vektör makineleri, sınıflandırma problemlerinin çözümü için verileri marjini en büyükleyecek en uygun düzlem veya hiper düzlem ile ayırmayı amaçlayan bir sınıflandırıcıdır. Marjin, iki farklı sınıfa ait birbirine en yakın veri noktaları arasındaki uzaklık olarak tanımlanmaktadır. Bu noktalar destek vektörleri olarak adlandırılmaktadır. Çok boyutlu uzayda sınıflandırma hatasını en küçükleyen en büyük marjinli hiper düzlemin belirlenmesi amaçlanmaktadır [16].

Doğrusal olarak ayrılabilen veriler bir düzlem ile ayrılırken doğrusal olarak ayrılamayan veriler doğrusal olmayan haritalama yöntemi ile yüksek boyutlu bir uzaya aktarılmakta, burada hiper düzlem ile sınıflara ayrıldıktan sonra veri noktalarının girdi uzayına iz düşümleri alınmaktadır [18]. Şekil 1'de doğrusal olarak ayrılabilen verilerin sınıflandırılmasına yönelik bir örnek görsele yer verilmiştir.



Şekil 1. Doğrusal Olarak Ayrılabilen Verilerin Sınıflandırılması

Şekil 2'de doğrusal olarak ayrılamayan verilerin doğrusal olmayan haritalama yöntemi ile sınıflandırıldığı ve veri noktalarının girdi uzayına iz düşümlerinin alındığı bir görsel yer almaktadır.



Şekil 2. Doğrusal Olarak Ayrılamayan Verilerin Sınıflandırılması

2.2.3. Lojistik Regresyon

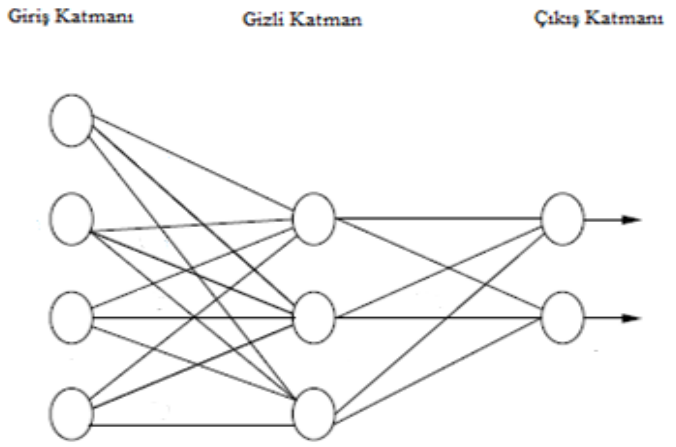
Lojistik regresyon, yanıt değişkeninin iki veya daha fazla kategoriden oluştuğu durumlarda kullanılan istatistiksel bir sınıflandırma yöntemidir. Problemin türüne göre ikili, çok kategorili ve sıralı olmak üzere çeşitli lojistik regresyon modelleri kullanılmaktadır. Lojistik regresyon modelinde de diğer regresyon modellerinde olduğu gibi yanıt değişkeni ile bir dizi bağımsız değişken arasındaki ilişkinin ortaya konmasını amaçlanmaktadır. Lojistik regresyon pek çok yönden doğrusal regresyon modeline benzese de yanıt değişkeninin kesikli yapıda olmasıyla diğer regresyon türlerinden ayrılmaktadır [19].

Yanıt değişkeni doğrusal regresyonda sürekli bir değere sahipken lojistik regresyonda kullanılan lojistik fonksyion sayesinde 0 ile 1 arasında değer almaktadır. Lojistik fonksiyondan elde edilen değere göre örneğin ait olduğu sınıf tahmin edilmektedir.

2.2.4. Çok Katmanlı Algılayıcı

Çok katmanlı algılayıcı, insan beynininin bilgiyi işleme sürecini taklit eden bir yapay sinir ağı türüdür. Tek katmanlı algılayıcıların sadece doğrusal problemlerin çözümünde kullanılmasından kaynaklanan yetersizliklerin giderilmesi için geliştirilmiştir. Çok katmanlı algılayıcılar giriş katmanı, çıkış katmanı ve gizli katman olmak üzere üç çeşit katmandan

oluşmaktadır. Dışarıdan gelen bilgilerin toplandığı giriş katmanı kendisine gelen bilgileri işlenmek üzere gizli katmana iletmektedir. Aradaki gizli katmanlardan çıkış katmanına ulaşan bilgiler ise tahmin sonuçları üretildikten sonra sistemden ayrılmaktadır. Şekil 3'te çok katmanlı algılayıcılarda bilgi aktarım sürecinin işleyişini gösteren bir örnek verilmiştir [20,21].



Şekil 3. Çok Katmanlı Algılayıcı Yapısı

2.2.5. J48 Karar Ağacı Algoritması

C4.5 karar ağacının WEKA yazılımındaki karşılığı olan J48 karar ağacı, bilginin keşfedilme sürecinde veri madenciliğinde kullanılan güçlü sınıflandırıcılardan birisidir. J48 algoritması ile karar kuralının oluşturulabilmesi için bilgi kazancının hesaplanması gerekmektedir. Bilgi kazancının hesaplanmasında entropi kavramından yararlanılmaktadır. Entropi kavramı veri setindeki düzensizliği ifade etmek için kullanılmaktadır. Kök düğümü ve karar düğümlerinin oluşturulması için her özniteliğin bilgi kazancı ölçülmekte ve elde edilen değerlere göre en iyi ayırıcı öznitelikler belirlenmektedir [22]. Bu algoritma ile eğitim veri seti kullanılarak karar ağacı oluşturulduktan sonra kök düğümünden yapraklara kadar karar kuralı doğrulrusunda ilerlenerek her bir yeni örneğin sınıf etiketi tahmin edilmektedir.

2.3. Değerlendirme Kriterleri

Eğitim veri seti ile sınıflandırma modelleri oluşturulduktan sonra test verisi kullanılarak sınıflandırıcıların performansları test edilmektedir. Sınıflandırıcıların performanslarının test edilebilmesi için çeşitli değerlendirme kriterleri hesaplanmaktadır. Bu çalışmada değerlendirme kriterlerinin hesaplanması için karmaşıklık matrisinden yararlanılmıştır. Karmaşıklık matrisi Tablo 2'de verilmiştir.

Tablo 2. Karmaşıklık Matrisi

		Tahmin Edilen Sınıf	
		Pozitif	Negatif
Gerçek Sınıf	Pozitif	GP	YN
	Negatif	YP	GN

Gerçek Pozitif (GP), gerçek sınıfı pozitif olan bir örneğin doğru sınıflandırıldığı durumları göstermektedir.

Yanlış Negatif (YN), gerçek sınıfı pozitif olan bir örneğin yanlış sınıflandırıldığı durumları göstermektedir.

Yanlış Pozitif (YP), gerçek sınıfı negatif olan bir örneğin yanlış sınıflanırıldığı durumları göstermektedir.

Gerçek negatif (GN), gerçek sınıfı negatif olan bir örneğin doğru sınıflandırıldığı durumları göstermektedir.

Karmaşıklık matrisinin hücrelerindeki değerler kullanılarak doğru sınıflandırma oranı, kesinlik, duyarlılık, F ölçütü ve Kappa istatistiği değerlendirme ölçütleri hesaplanabilmektedir.

Doğru sınıflandırma oranı, doğru sınıflandırılan örneklerin tüm örneklere oranı ile bulunmaktadır. Formülasyonu aşağıdaki eşitlikte verilmiştir [23].

$$\text{Doğru Sınıflandırma Oranı} = \frac{\text{GP+GN}}{\text{GP+YN+YP+GN}} \quad (3)$$

Kesinlik, pozitif olan ve doğru tahmin edilen örneklerin pozitif olarak tahmin edilen örneklerin toplamına oranıdır. Eşitlik (4)'teki gibi hesaplanmaktadır [23].

$$\text{Kesinlik} = \frac{\text{GP}}{\text{GP+YP}} \quad (4)$$

Duyarlılık, pozitif olan ve doğru tahmin edilen örneklerin pozitif örneklerin toplamına oranıdır [23].

$$\text{Duyarlılık} = \frac{GP}{GP+YN} \quad (5)$$

F-ölçütü eşitlik (4) ve (5)'teki değerlerin harmonik ortalaması alınarak hesaplanmaktadır. F-ölçütünün hesaplanmasına ilişkin formül eşitlik (6)'da verilmiştir [23].

$$\text{F ölçütü} = \frac{2\times Duyarlılık\times Kesinlik}{Kesinlik+Duyarlılık} \quad (6)$$

Tesadüfi faktörleri de hesaba katan Kappa istatistiği, güvenilirliğin bir ölçüsüdür. Kappa istatistiğine ait formül eşitlik (7)'de verilmiştir.

$$\text{Kappa istatistiği} = \frac{\text{Gözlenen Doğruluk}-\text{Beklenen Doğruluk}}{1-\text{Beklenen Doğruluk}} \quad (7)$$

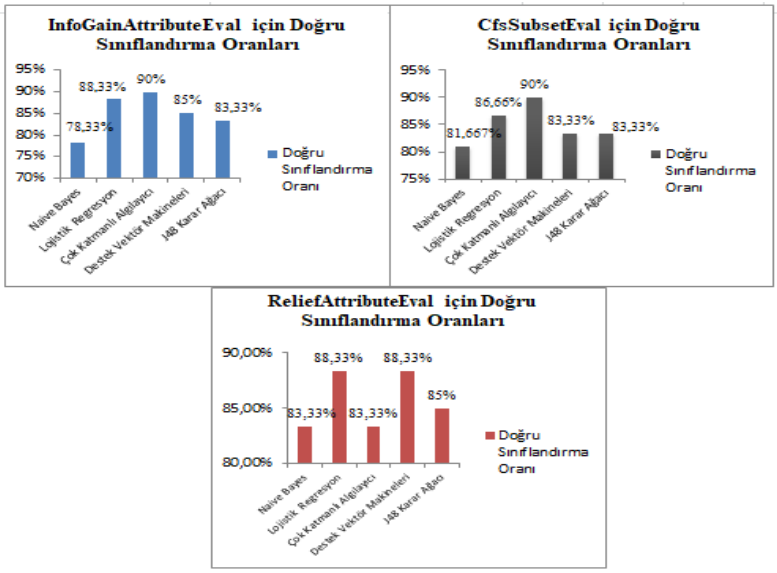
Kappa istatistiği hesaplamasında 0 ile 1 arasında bir değer elde edilmektedir. 0.00-0,20 arasında elde edilen Kappa istatistiği, önemli olmayacak düzeyde uyum olduğunu, 0,21-0,40 düşük düzeyde uyum olduğunu, 0,41-0,60 orta derecede uyum olduğunu, 0,61-0,80 iyi derecede uyum olduğunu ve 0,81-1,00 ileri derecede uyum olduğunu göstermektedir [24].

3. Araştırma Sonuçları ve Tartışma

Bu çalışmada kalp yetmezliği veri seti kullanılarak Naive Bayes, Lojistik Regresyon, Destek Vektör Makineleri, J48 Karar Ağacı ve Çok Katmanlı Algılayıcı sınıflandırıcılarının performansları değerlendirilmiştir. Veri setindeki örneklerin %80'i eğitim, %20'si test verisi olarak ayrılmıştır. Sınıflandırma öncesinde veri ön işleme aşamasında veri standardizasyonu yapılmıştır. InfoGainAttributeEval, CfsSubsetEval ve ReliefFAttributeEval öznitelik seçim yöntemlerinden yararlanılmıştır. InfoGainAttributeEval yöntemi ile yaş, ejeksiyon fraksiyonu, serum kreatinin, serum sodyum ve süre öznitelikleri; CfSubsetEval yöntemi ile yaş, ejeksiyon fraksiyonu, serum kreatinin ve süre öznitelikleri; ReliefFAttributeEval yöntemi ile ejeksiyon fraksiyonu, diyabet, anemi, cinsiyet ve süre öznitelikleri seçilmiştir.

Her bir öznitelik seçim yöntemi için sınıflandırma algoritmalarından elde edilen doğru sınıflandırma oranları Şekil 4'te verilmiştir. Öznitelik seçim yöntemleri ve değerlendirme kriterleri için sınıflandırma algoritmalarından elde edilen sonuçlar Tablo 3-7 arasında yer almaktadır. Değerlendirme kriterleri incelendiğinde en başarılı sınıflandırıcının 5 adet özniteliğin kullanıldığı InfoGainAttributeEval ve 4 adet özniteliğin kullanıldığı CfsSubsetEval öznitelik seçim yöntemi ile %90 doğru sınıflandırma oranının elde edildiği çok katmanlı algılayıcı sınıflandırıcısı olduğu görülmüştür. Her iki öznitelik seçim yöntemi ile de en yüksek Kappa istatistiği değeri olan 0,78'e ulaşılmıştır. Güvenilirliğin bir ölçüsü olan bu değer, beklenen ve gözlenen doğruluk arasında iyi derecede uyum olduğunu göstermektedir. F-ölçütü değeri InfoGainAttributeEval ve CfsSubsetEval öznitelik seçim yöntemleri için 0,86 olarak hesaplanmıştır. CfsSubsetEval yöntemi ile InfogainAttributeEval yöntemine göre daha az sayıda öznitelik kullanılmıştır. Çok katmanlı algılayıcı algoritması ile daha önce bu veri setininin kullanıldığı [7], [8] ve [9] referanslarına göre daha yüksek bir sınıflandırma başarısı elde edilmiştir.

ReliefFAttributeEval yöntemi ile en başarılı sınıflandırıcılar %88,33 doğru sınıflandırma oranı ile destek vektör makineleri ve lojistik regresyon olmuştur. Destek vektör makineleri için F-ölçütü ve Kappa istatistiği değerleri sırasıyla 0,81 ve 0,73 iken lojistik regresyonda 0,82 ve 0,74 olarak hesaplanmıştır.



Şekil 4. Öznitelik Seçim Yöntemleri ve Sınıflandırma Algoritmalarından Elde Edilen Sonuçlar

Tablo 3. Lojistik Regresyona Göre Elde Edilen Sınıflandırma Sonuçları

Öznitelik Seçim Yöntemi	Lojistik Regresyon		
	Doğru Sınıflandırma Oranı	F-Ölçütü	Kappa İstatistiği
InfoGainAttributeEval	%88,33	0,82	0,74
CfsSubsetEval	%86,67	0,79	0,70
ReliefFAttributeEval	%88,33	0,82	0,74

Tablo 4. Naive Bayes Algoritmasına Göre Elde Edilen Sınıflandırma Sonuçları

Öznitelik Seçim Yöntemi	Naive Bayes		
	Doğru Sınıflandırma Oranı	F-Ölçütü	Kappa İstatistiği
InfoGainAttributeEval	%78,33	0,61	0,48
CfsSubsetEval	%81,67	0,71	0,58
ReliefFAttributeEval	%83,33	0,76	0,63

Tablo 5. Çok katmanlı Algılayıcıya Göre Elde Edilen Sınıflandırma Sonuçları

Öznitelik Seçim Yöntemi	Çok Katmanlı Algılayıcı		
	Doğru Sınıflandırma Oranı	F-Ölçütü	Kappa İstatistiği
InfoGainAttributeEval	%90,00	0,86	0,78
CfsSubsetEval	%90,00	0,86	0,78
ReliefFAttributeEval	%83,33	0,76	0,63

Tablo 6. Destek Vektör Makinelerine Göre Elde Edilen Sınıflandırma Sonuçları

Öznitelik Seçim Yöntemi	Destek Vektör Makineleri		
	Doğru Sınıflandırma Oranı	F-Ölçütü	Kappa İstatistiği
InfoGainAttributeEval	%85	0,76	0,65
CfsSubsetEval	%83,33	0,72	0,61
ReliefFAttributeEval	%88,33	0,81	0,73

Tablo 7. J48 Karar Ağacına Göre Elde Edilen Sınıflandırma Sonuçları

Öznitelik Seçim Yöntemi	J48 Karar Ağacı		
	Doğru Sınıflandırma Oranı	F-Ölçütü	Kappa İstatistiği
InfoGainAttributeEval	%83,33	0,75	0,63
CfsSubsetEval	%83,33	0,75	0,63
ReliefFAttributeEval	%85	0,78	0,67

4. Sonuç

Kalp yetmezliği, hastalıktan kaynaklanan ölüm oranının yüksekliği ve son yıllarda giderek artan bir seyir göstermesi sebebiyle en ciddi kalp hastalıklarından biri olarak görülmektedir. Bu çalışmada kalp yetmezliği hastalarının sağ kalımlarının yüksek bir doğrulukla tahmin edilmesi için 299 kayıt ve 13 öznitelikten oluşan kalp yetmezliği veri seti kullanılarak 5 farklı sınıflandırıcı ve 3 farklı öznitelik seçim yönteminin performansı değerlendirilmiştir.

Çalışmanın sonucunda değerlendirme ölçütlerine göre en başarılı sınıflandırıcının %90 doğru sınıflandırma oranına sahip olan çok katmanlı algılayıcı olduğu görülmüştür. Öznitelik seçim yöntemleri karşılaştırıldığında InfoGainAttributeEval ve CfsSubsetEval yöntemleri kullanılarak seçilen öznitelikler ile en yüksek sınıflandırma başarısına ulaşılmıştır. CfsSubsetEval yöntemi ile daha az sayıda öznitelik kullanılmıştır. Aynı veri setinin kullanıldığı [7], [8] ve [9] kaynaklarına göre daha yüksek bir doğru sınıflandırma oranına ulaşılmıştır. Gelecek çalışmalarda veri sayısının arttırılması, farklı öznitelik seçim yöntemlerinden yararlanılması ve sınıflandırma algoritmalarının birlikte değerlendirilmesiyle oluşturulacak bütünleşik sınıflandırıcılar ile daha başarılı sonuçlar elde edilebilir.

Kaynakça

[1] Türk Kardiyoloji Derneği, Resmi web sitesi, https://tkd.org.tr/kalp-yetersizligi-calisma-grubu/sayfa/toplum_icin_bilgiler, Erişim Tarihi, 01.02.2021

[2] American Heart Association, Causes and Risks for Heart Failure, https://www.heart.org/en/health-topics/heart-failure/causes-and-risks-for-heart-failure, Erişim Tarihi, 02.02.2021.

[3] Tokgözoğlu, L., Yılmaz, M.B., Abacı, A., Altay, H., Atalar, E., Aydoğdu, S., Bozkurt, E., Çavuşoğlu, Y., Eren, M., Sarı, İ., Selçuk, T., Temizhan, A., Ural, D., Zoghi, M. (2015). Türkiye'de kalp yetersizliği yol haritasi kalp yetersizliğinin ve buna bağlı ölümlerin önlenmesi amacıyla geliştirilebilecek politikalara illişkin öneriler. TKD, 1-31.

[4] Patel, J., Upadhyay, T. and Patel, S. (2015). Heart disease prediction using machine learning and data mining technique. International Journal of Computer Science & Communication, 7(1), 129-137.

[5] Saqlain, M., Hussain, W., Saqib, N., Khan, M. (2016). Identification of heart failure by using unstructured data of cardiac patients. 45th International Conference on Parallel Processing Workshops, 426-431.

[6] Jagad, H., Kandawalla and Nair, S. (2015). Detection of Coronary Heart Diseases using Data Mining Techniques. International Journal on Recent and Innovation Trends in Computing and Communication, 3(1).

[7] Küçükakçalı, Z., Çiçek, I., Güldoğan, E., Çolak, C. (2020). Assessment of associative classification approach for predictıng mortality by heart failure. The Journal of Cognitive Systems, 5(2), 41-45.

[8] Chicco, D. and Jurman, G. (2020). Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone. BMC Medical Informatics and Decision Making, 20(1), 1-16.

[9] Gürfidan, R. and Ersoy, M. (2021). Classification of death related to heart failure by machine learning algorithms. Advances in Artificial Intelligence Research, 1(1), 13-18.
[10] Phyu, T., Oo, N. (2016). Performance Comparison of Feature Selection Methods. MATEC Web of Conferences.
[11] Gnanambal, S., Thangaraj, M., Meenatchi, V.T., Gayathri, V. (2018). Classification algorithms with attribute selection: an evaluation study using WEKA. Int. J. Advanced Networking and Applications, 9(6), 3640-3644.
[12] Çavuşoğlu, Ü. ve Kaçar, S. (2019). Anormal Trafik Tespiti için Veri Madenciliği Algoritmalarının Performans Analizi. Akademik Platform Mühendislik ve Fen Bilimleri Dergisi, 7 (2), 205-216.
[13] Zaffar, M., Hashmani, M.A., Savita, K.S. (2017). Performance analysis of feature selection algorithm for educational data mining. In: IEEE Conference on Big Data and Analytics (ICBDA), 7(12).
[14] Rosario, S.F. and Thangadurai, K. (2015). RELIEF: Feature selection approach, International Journal of Innovative Research & Development, 4(11), 218-224.
[15] Zaffar, M., Savita, K.S., Hashmani, M.A., Rizvi, S. (2018). A study of feature selection algorithms for predicting students academic performance. International Journal of Advanced Computer Science and Applications, 9(5), 541-549.
[16] Han, J., Kamber, M. and Pei, J. (2012). Data mining: Concepts and techniques. (3rd Edition). Waltham: Morgan Kaufmann.
[17] Tan, P. N., Steinbach, M. and Kumar, V. (2006). Introduction to data mining. USA: Addison-Wesley.
[18] Akşehirli, Ö., Ankaralı, H., Aydın, D., Saraçlı, Ö. (2013). Tıbbi tahminde alternatif bir yaklaşım: Destek vektör makineleri. Turkiye Klinikleri Journal of Biostatistics, 5(1), 19-28.
[19] Hosmer, D.W., Lemeshow, S., Sturdivant, R.X. (2013). Applied Logistic Regression. (3rd Edition). John Wiley & Sons.
[20] Giudici, P. (2003). Applied data mining: Statistical methods for business and industry. New York: J. Wiley.
[21] Öztemel, E. (2012). Yapay sinir ağları. (3.baskı). İstanbul: Papatya Yayıncılık.