**MAKİNE ÖĞRENMESİ SINIFLANDIRMA ALGORİTMALARI İLE KALP YETMEZLİĞİ KLİNİK KAYITLARI VERİ SETİ ANALİZİ**

**Ömer Faruk YILDIZ**

**Biyomedikal Mühendisliği-Yüksek Lisans**

**2028142006**

**BMM723-Biyomedikal Mühendisliğinde Yapay Sinir Ağları ve Uygulamaları**

İçindekiler

[**1.** **GİRİŞ** 2](#_Toc58837977)

[**2.** **Veri Setine Genel Bir Bakış** 2](#_Toc58837978)

[**3.** **Veri Temizliği ve Aşırı Veri Kontrolü** 3](#_Toc58837979)

[**4.** **Veri Setinde Sınıflandırma İşlemleri** 3](#_Toc58837980)

[**5.** **Değerlendirme** 4](#_Toc58837981)

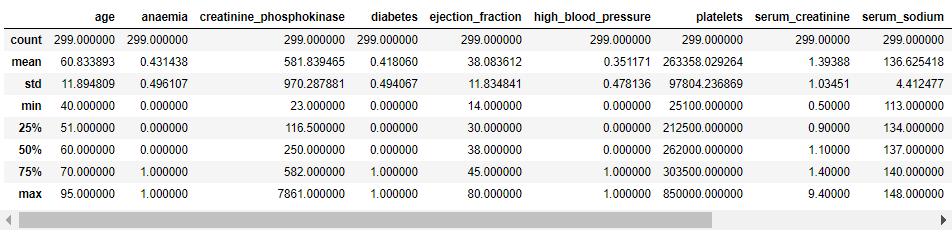
# **GİRİŞ**

Yapılan bu çalışmada klinik bulgular ile elde edilen kalp yetmezliği hastalığı ve bu hastalıkta etkisinin bulunduğu düşünülen bazı değerler aracılığıyla makine öğrenmesi sınıflandırma algoritmaları kullanılarak bireyin kalp yetmezliği riskinin olup olmadığı araştırılmıştır. Elde edilen bu veri seti  <https://www4.stat.ncsu.edu/> veri bankası tarafından elde edilmiştir. Çalışılan bu veride 299 satır ve 13 sütun bulunmaktadır. Veri setindeki sütunların nelere karşılık geldiği aşağıda maddeler halinde verilmiştir.

* age: Hastanın yaşı
* anaemia: Hasta anemi hastası mı? 1-evet,0-hayır
* creatinine\_phosphokinase: Kreatin fosfokinez değeri
* diabetes: Hasta şeker hastası mı? 1-evet,0-hayır
* ejection\_fraction :
* high\_blood\_pressure : Yüksek tansiyon var mı? 1-evet,0-hayır
* platelets :
* serum\_creatinine : Serum Kreatin değeri
* serum\_sodium : Serum Sodyum değeri
* sex : Cinsiyet
* smoking: Sigara kullanıyor mu ? 1-evet,0-hayır
* time: Zaman
* DEATH\_EVENT: Hasta yaşıyor mu?

# **Veri Setine Genel Bir Bakış**

Analiz yapılacak olan bu veri setine genel olarak bakıldığında 299 x 13 bir matristen oluşmaktadır. Veri seti info() komutu ile incelendiğinde alınan bütün değerlerin sayısal veriler olduğu obje türünde değer olmadığı görülmektedir. Bu verilerden 4’ 0 ve 1’den oluşan kategorik veriler geriye kalan değerler ise normal aralıklı değerler olduğu görülmektedir.

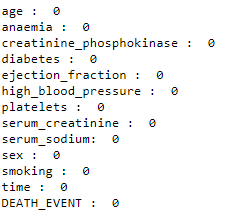


Şekil 1 Veri seti hakkında genel bilgiler

Veri seti describe komutu ile incelendiğinde Şekil-1’de gösterilen tablo elde edilmektedir.

# **Veri Temizliği ve Aşırı Veri Kontrolü**

Kalp yetmezliği veri seti incelendiğinde yapılması gereken ilk adımlardan birisi eksik verilerin olup olmadığıdır. Her sütun ayrı ayrı analiz edildiğinde kayıp verilerin olmadığı görülmektedir.



Şekil 2 Kayıp verilerin kontrolü

Veri setini analiz etmek için yapılması gereken bir diğer işlem ise aşırı verilerin olup olmadığıdır. Data setteki her sütun için aşırı veri kontrolü yapılmıştır. Bu kontrol bazı sütunlar için z skoru hesaplanarak ortaya çıkarılmış bazı sütunlar için ise ilk ve son çeyrekliklere bakılmıştır. Sonuç olarak veri seti içerisinde aşırı verilere rastlanmamış bu yüzden herhangi bir işlem gerçekleştirilmemiştir.

# **Veri Setinde Sınıflandırma İşlemleri**

Kalp yetmezliği veri ile seti makine öğrenmesi sınıflandırma yöntemlerinden bazıları aracılığıyla tahminlerde bulunulmuştur. Bu algoritmalar KNN, SVM, Decision Tree, Random Forest algoritmalarıdır. Bu algoritmaların her birisi için ayrı ayrı skorlar skorlar hesaplanmıştır.

Veri setinde öncelikle tahmin edilen sütun olan “DEATH\_EVENT” sütunu yani y değerleri alınmış ve alınan bu y değeri bütün data setten çıkarılarak x değeri yani bağımsız değişkenlerimiz elde edilmiştir. Daha sonra elde edilen bu veriler kullanılarak gerekli kütüphaneler eklenip işlemler devam ettirilmiştir.

Veri seti öncelikle KNN algoritması ile analiz edilmiştir. Bu işlemler yapılırken x ve y dataları test ve eğitim verileri olarak ayrılmıştır. Bu verilerin %20’si test için ve %80’lik kısmı da eğitim için ayrılmıştır. Bu algoritma ile x-test verileri tahmin edilmeye çalışılmış ve skor elde edilmiştir (Şekil-3).



Şekil 3 KNN ile elde edilen skor

İkinci olarak SVM algoritması ile x-test verileri kullanılarak tahmin yapılmış ve skor olarak 0.77 elde edilmiştir (Şekil-4). Daha sonra aynı veriler yani x-test verileri ile decision tree algoritması ile bir tahmin yapılmış ve bu tahmin sonucunda 0.8 skoru elde edilmiştir (Şekil-5).



Şekil 4 SVM skoru



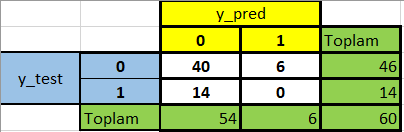
Şekil 5 Decision tree skoru

Tahmin yapılan bu üç algoritmadan sonra son olarak rondom forest algoritması ile bir tahmin gerçekleştirilmiştir. Bu tahmin sonucunda rondom forest algoritmasının skoru 0.93 olarak bulunmuştur.

# **Değerlendirme**

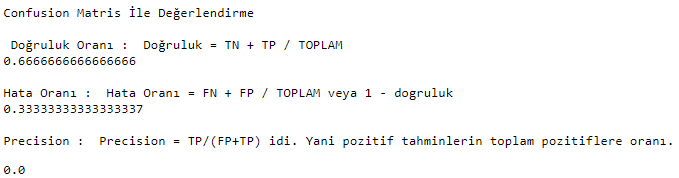
Makine öğrenmesinde sınıflandırma işlemlerinde kullanılan farklı değerlendirme metodları bulunmaktadır. Bunların en bilinenlerinden biri ise hata matrisidir (confusion matrix). Bu matris aracılığıyla yapılan sınıflandırma işleminin doğruluk oranı, hata oranı ve precision değerleri elde edilebilir.

KNN algoritması ile yapılan sınıflandırma işlemi sonucunda elde edilen hata matrisi aşağıda gösterilmektedir.



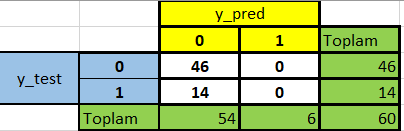
Şekil 6 KNN hata matrisi

KNN hata matrisinden elde edilen sonuçlar Şekil-7 de gösterilmektedir. Bu sonuçlara göre KNN algoritmasına ait doğruluk oranı %67 olarak elde edilmiştir. Bu matrise ait hata oranı ise %33 olarak bulunmuştur.



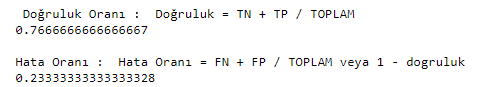
Şekil 7 KNN hata matrisi değerlendirme

SVM algoritması ile yapılan işlemde elde edilen hata matrisi Şekil-8 de gösterilmektedir.



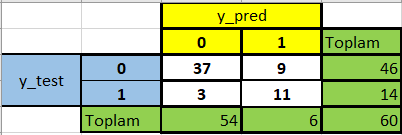
Şekil 8 SVM algoritması hata matrisi

Elde edilen bu hata matrisine göre SVM algoritması ile ulaşılan doğruluk oranı %77, hata oranı ise %22 olarak bulunmuştur.

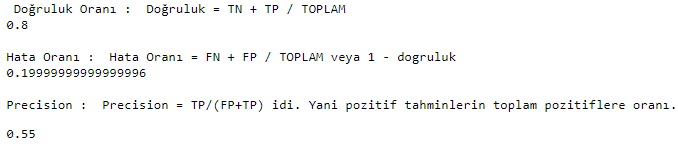


Şekil 9 SVM algoritması doğruluk ve hata oranları

Bu veri seti için yapılan başka bir çalışma ise decision tree algoritması için hata matrisini oluşturmaktır (Şkeil-10). Elde edilen bu matriste 0-0 eşleşmeler ve 1-1 eşleşmelerin sayısının artmasından da anlaşılacağı üzere bu algoritmada doğruluk yüzdesi diğer iki algoritmaya göre yükselmiştir. Decision tree algoritmasına ait doğruluk oranı %80 ve hata oranı %20 olarak bulunmuştur.

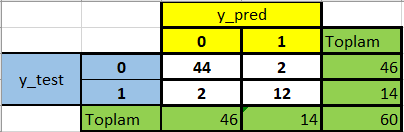


Şekil 10 Decision tree hata matrisi

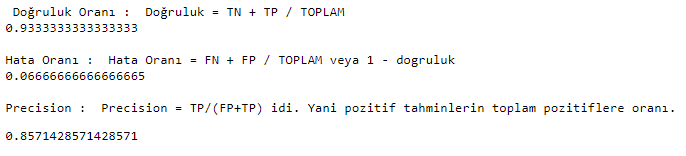


Şekil 11 Decision tree hata matrisi değerlendirmesi

Son olarak veri setinden rondom forest algoritması ile yapılan tahminler sonucunda bu algoritmaya ait hata matrisi oluşturulmuştur (Şekil-12). Oluşturulan bu matriste de görüleceği üzere doğruluk eşleşmeleri diğer üç algoritmaya göre artmıştır böylece doğruluk yüzdesi oldukça düşmüştür. Rondom forest algoritmasına ait doğruluk yüzdesi %93, hata oranı ise %7 olarak bulunmuştur (Şekil-13).

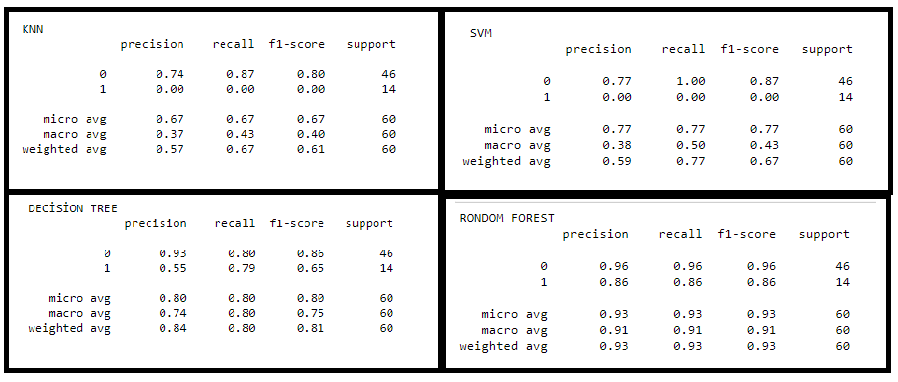


Şekil 12 Rondom forest hata matrisi



Şekil 13 Rondom forest hata ve doğruluk oranları

Uygulanan bu dört algoritma için ayrıca hasta olan bireyleri doğru tespit etme oranı diyebileceğimiz duyarlılık değeri, hasta olarak tahmin edilen bireylerin gerçekte kaçının hasta olduğunu belirten kesinlik değeri ve bu iki değer arasındaki ilişkiyi kullanarak yeni bir skor olarak elde edilen f-skoru hesaplanmıştır (Şekil-14).



Şekil 14 Duyarlılık, kesinlik ve F skoru

Hesaplanan bu değerlere göre en iyi sonuç veren algoritma rondom forest ve en düşük değeri veren algoritma ise KNN algoritmasıdır.

Değerlendirilen bu dört algoritma karşılaştırıldığında en düşük doğruluk oranına sahip algoritma KNN algoritması iken en yüksek doğruluk oranına sahip algoritma rondom forest algoritmasıdır. Bu bilgilere dayanarak veri seti biraz daha geliştirilerek doğruluk oranı biraz daha yukarıya çekilebilir ve böylece biyomedikal alanında kullanılabilir.