T.C. Yalova Üniversitesi

Kalp Rahatsızlığı Tahmini

5 Mayıs 2020

Prepared by Oğuz Kaan PARLAK, 120101047. Eren SÖNMEZ, 160101030

1 Introduction

1.1 Background

Bu araştırma kalp rahatsızlığının daha hızlı tahminlenmesi için yapılmıştır. Projenin amacı insanlarda ölçülen çeşitli sağlık değerlerine göre kişinin kalp rahatsızlığına sahip olup olmadığını farklı yöntemler kullanılarak tahminlenmesidir.

Araştırmanın sonucunda hangi öğrenme yönteminin tahmin üzerinde ne kadar başarılı olduğu, hangi sağlık değerlerinin kalp rahatsızlığını daha çok etkilediğini ve raporlardaki parametreler ile hastalığın ilişkisi hakkında fikir sahibi olunması beklenilmektedir. Projede kullanılan veri seti 1025 kişinin raporlarından yararlanılarak dört farklı hastanenin veri tabanındaki raporlardan oluşturulmuştur. Raporları kullanılan hastalar bu dört hastanede kontrollerini yaptıran kişilerdir. Raporların toplandığı dört hastane şu şekildedir;

- Hungarian Institute of Cardiology, Budapeşte, Macaristan.
- University Hospital, Zürih, İsviçre.
- University Hospital, Basel, İsviçre.
- V. A. Medical Center, Long Beach and Cleveland Clinic Foundation [1].

Raporlar sağlık merkezlerinde gerçekleştirilen testlerin sonucunda düzenlenmiş olup kişiye dair toplam 76 farklı özellik içermektedir.

1.2 Objectives

- Kalp rahatsızlığı sınıflandırması yapabilme
- kNN algoritmasının uygulanması
- Naïve Bayes sınıflandırıcısının uygulanması
- Support Vector Machine algoritmasının uygulanması
- Random Forest algoritmasının uygulanması
- Rahatsızlığı en çok etkileyen özniteliklerin bulunması
- Özniteliklerin kalp rahatsızlığı üzerindeki dağılımlarının bulunması

2 Methodology

2.1 Research Questions

1. kNN Algoritması Kullanılırken k Değeri Kaç Olmalı?
2. Kullanılan Algoritmalardan Hangisi Rahatsızlığı Sınıflandırmada Daha İyi Sonuç Veriyor?
3. Hangi Özniteliklerin Sonuca Etkisi Daha Fazladır?
4. Özniteliklerin Veri Setindeki Dağıllımı Ne Şekildedir?

2.2 Sample

Veri seti 14 özniteliğe sahip 1025 veriden oluşmaktadır. Her bir veri kişilerin kontrolden geçtikten sonraki değerlerini içeren raporlardan alınan değerlerle oluşturulmuştur. Veri setin 29-77 yaş arası kişilerin raporlarından oluşmaktadır [2]. 1025 kişinin %31'I kadın, %69'u erkektir. Veri seti amacına uygun olarak kardiyoloji bölümünden alınan verileri kapsamaktadır. Veriler 4 farklı hastanenin kardiyoloji bölümünden alınan raporların birleştirilmesiyle oluşmaktadır. Veri seti amacına uygun olması için kalp rahatsızlığı şüphesiyle test edilen hastaların raporlarından oluşmuştur. Toplam 76 öznitelikten oluşan raporlar etkisiz özniteliklerin çıkarılması ile 14'e düşmüştür. Kişinin adı, soyadı, sosyal güvenlik numarası gibi öznitelikler kalp rahatsızlığı üzerine herhangi bir etkide bulunmadığı için çıkarılmıştır. Sonuç olarak veri setindeki öznitelikler şu şekildedir:

- age: Kişinin yaşı
- sex: Kişinin cinsiyeti(1=erkek, 0=kadın)
- cp(chest pain type): Kişinin göğüs ağrısı tipi(0'dan 3'e kadar sıralanmıştır.)
- trestbps: Kişinin dinlenmiş haldeki kan basıncı
- chol: Kişinin mg/dl cinsinden serum kolestrolü(mg/dl: desilitre başına milligram, belirli bir miktar kandaki belirli maddenin miktarını gösteren ölçüm)
- fbs: Kişinin aç karnına ölçülen kan şekeri(>120 mg/dl ise 1, değilse 0)
- restecg: Kişinin dinlendikten sonraki elektrokardiyografik sonuçları
- thalach: Kişinin elde edilen maksimum kalp atış hızı
- exang: Egzersize bağlı anjina(efor anjinası olarak da bilinir, anjina bir çeşit kalp rahatsızlığıdır.)
- oldpeak: ST depresyon(segment depresyonu, coroner arter hastalıkları tanı ve değerlendirilmesinde kullanılan fonksiyonel testlerden biri)
- slope: ST segmentinin eğimi
- ca: Kişinin büyük damar sayısı(0'dan 3'e kadar sıralanmıştır.)
- thal: talasemi tanısı(normal 3, fixed defect 6 ve reversable defect 7 olarak sıralanmıştır. Kalıtımsaldır.)
- target: Kalp rahatsızlığı tanısı(sağlıklı 0, hasta 1 şeklinde.)

2.3 Data Collection

Veriler kişilerin raporlarındaki değerlerden oluşmaktadır. Kişilerin raporları tedavi gördükleri hastane tarafından veritabanında kayıt altında tutulmuştur. Daha sonra veri setinde kullanılan raporlar hastanelerin veritabanlarından toplanmıştır. Her hastane için veritabanının oluşturulmasını ve raporların toplanmasını 4 farklı kişi gerçekleştirmiştir. Bu 4 kişi kalp rahatsızlığı üzerinde çalışmalardan bulunan kişilerdir [3]:

- Budapeşte, Prof. Andras Janosi
- Zürih, William Steinbrunn, M.D.
- Basel, Matthias Pfisterer, M.D.
- Long Beach ve Cleveland, Robert Detrano, M.D.

4 hastanenin veritabanından toplanan veriler daha sonra birleştirilerek veri setindeki güncel halini almıştır. Veri seti yüz yüze ya da online değil kişilerin raporlarından oluşturulmuştur.

2.4 Data Analysis

Veriler online olarak university of california Irvine(Kaliforniya üniversitesi) web sitesinden indirilmiştir. Bilgisayarda .csv formatında tutulmaktadır. CSV(comma separated values) uzantısı virgüllerle ayrılmış değerlerden oluşan bir dosya türüdür. Verilerin analizi için veri bilimi uygulamalarında sıkça kullanılan Python ve R geliştirme ortamı Anaconda üzerinde Jupyter Notebook ide kullanılmıştır. Veriler Python programlama dili kullanılarak analiz edilmiştir. Verilerin analizinde:

- kNN(En yakın komşu)
- Naïve Bayes
- Support Vector Machine
- Random Forest

olmak üzere 4 farklı algoritma kullanılmıştır. Analizden önce veri setine uygulanan önişleme yöntemi ile nitel verilerin de analizde kullanılması sağlanmıştır. Ön işleme yöntemi nitel verileri sonuca olan etkilerinin ve anlamlarının bozulmadan nicel veriye dönüşmesini gerçekleştirmiştir.

2.5 Limitations

Veri setini oluşturan raporlar kalp rahatsızlığı şüphesiyle kontrolden geçen insanların raporlarıdır. Veri seti 1025 rapordan oluşmaktadır. Veri seti %70 train ve %30 test şeklinde bölünmüştür. 1025 verinin 717'l train(eğitim) verisi, 308'si ise test verisi olarak ayrılmıştır.

Kalp rahatsızlığı gibi bir çok farklı parametreye bağlı olabilen bir hastalık için küçük bir veri setidir. Ancak veri seti hastanelerin veri tabanlarından toplandığı için raporların içindeki veriler yanlış ya da eksik değildir ve tam olarak kalp rahatsızlığı tahmini için kullanılabilecek parametreleri kapsamaktadır. Yine de sağlık alanında yapılan bir çalışmanın daha genel olarak kabul görmesi ve daha kesin yorumlar yapılabilmesi için daha büyük bir veri seti üzerinde çalışılmalıdır. Görece küçük veri seti, yaptığımız çalışmada elde edilen sonuçlar üzerinde daha kesin yorum yapmamızı kısıtlayabilir.

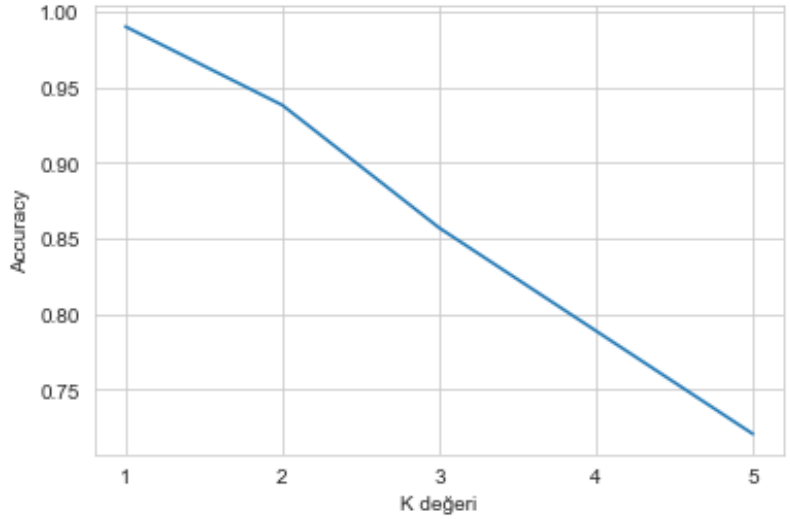
2.6 Research Design

Research Question	Method Used to Answer Question
1. kNN Algoritması Kullanılırken k Değeri Kaç Olmalı?	- k Nearest Neighbor
2. Kullanılan Algoritmalardan Hangisi Rahatsızlığı Sınıflandırmada Daha İyi Sonuç Veriyor?	- k Nearest Neighbor - Naïve Bayes - Support Vector Machine - Random Forest
3. Hangi Özniteliklerin Sonuca Etkisi Daha Fazladır?	- Random Forest Feature İmportance
4. Özniteliklerin Veri Setindeki Dağılımı Ne Şekildedir?	- Random Forest - Statistical Methods - Seaborn

3 Results

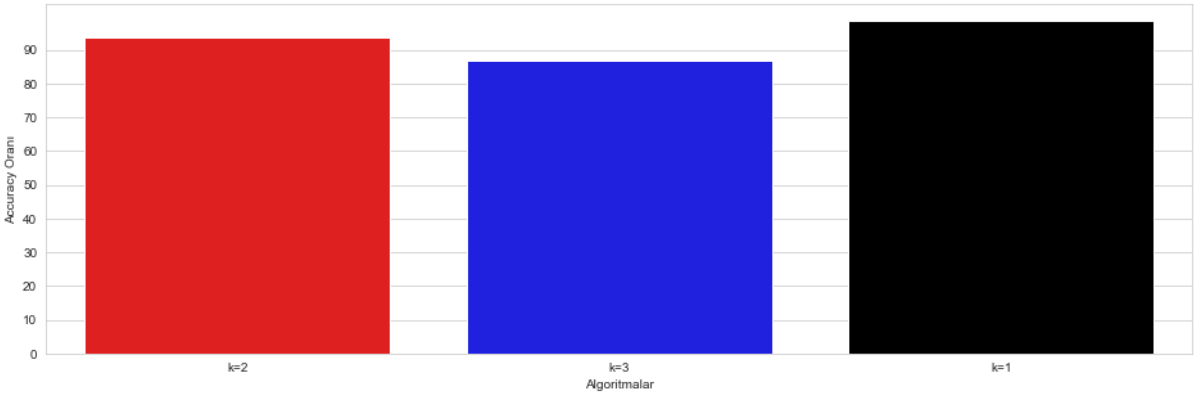
3.1 kNN Algoritması Kullanılırken k Değeri Kaç Olmalı?

Sınıflandırma için kNN algoritması kullanıldığında sistemin accuracy oranı seçilen k değerine göre değişiklik göstermektedir. k değeri, hedefe en yakın kaç komşuya bakılacağı anlamına gelmektedir ve bu komşularla olan mesafelere göre hedef değer bir sınıfa atanır. Bu araştırmada bir döngü yardımıyla k=1'den k=5'e kadar değerlerin sırasıyla hangi accuracy oranını verdiği gözlemlenmiştir. Tablo 1.deki grafik k değerini seçerken karar vermemizi sağlamıştır.



Tablo 1. Accuracy Ve k Değeri Arasındaki İlişki

kNN algoritması en iyi değeri k=1 için vermiştir ki k=1 değeri her zaman en yüksek değeri verecektir. Ancak k=1 değeri seçildiği zaman "overfitting" problem meydana gelmektedir. Bu yüzden projede k=2 veya k=3 ile karşılaştırılmıştır. 1,2 ve 3 değerleri arasındaki karşılaştırmaya tablo 2. Üzerinden bakarsak;

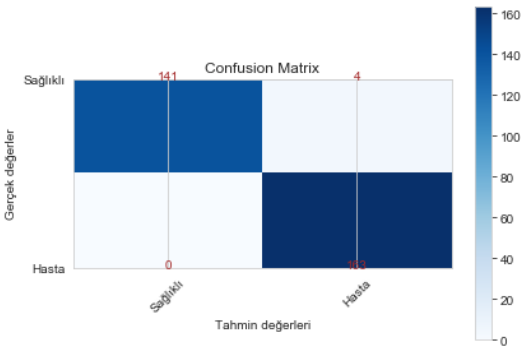


Tablo 2. kNN Algoritması İçin k Değerlerinin karşılaştırılması

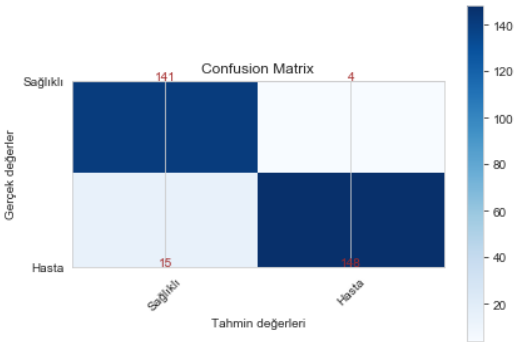
k değerleri için accuracy oranlarına bakarsak;

- k=1 için train accuracy oranı: %100, test accuracy oranı: %98.70
- k=2 için train accuracy oranı: %99.30, test accuracy oranı: %93.83
- k=3 için train accuracy oranı: %97.91, test accuracy oranı: %86.69

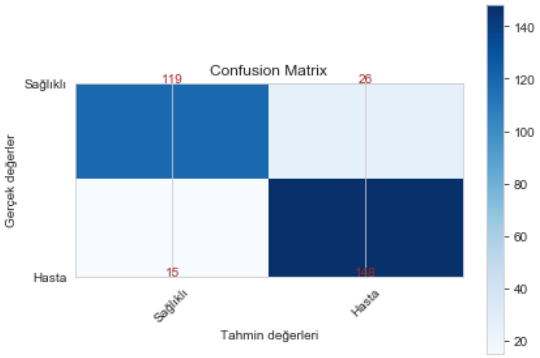
şeklindedir. Bu değerlerden yola çıkarak projede kNN algoritması için k=2 değeri seçilmiştir. Veri seti train ve test için %70'e %30 şeklinde ayrılmıştır. 1025 verinin 717'I train(eğitim) verisi, 308'si ise test verisi olarak ayrılmıştır. k değerlerini hata matrisi üzerinden incelersek;



- k=1 için kNN algoritması test verisi olarak kullanılan 308 kişiden; 145 sağlıklı bireyin 141'ini sağlıklı ve 163 hasta bireyin 163'ünü hasta olarak sınıflandırmıştır.



- k=2 için kNN algoritması test verisi olarak kullanılan 308 kişiden; 145 sağlıklı bireyin 141'ini sağlıklı ve 163 hasta bireyin 148'ini hasta olarak sınıflandırmıştır.



- k=3 için kNN algoritması test verisi olarak kullanılan 308 kişiden; 145 sağlıklı bireyin 119'unu sağlıklı ve 163 hasta bireyin 148'ini hasta olarak sınıflandırmıştır.

Hata matrisleri üzerinde fark edileceği üzere k için 1 değeri kullanıldığı zaman sistem hasta bireylerin(0) hepsini her zaman doğru sınıflandırmaktadır. Bu yüksek accuracy oranı için daha iyidir ancak sistem sadece bir çıkış değerini sürekli doğru tahmin edebiliyorsa aşırı öğrenmeye kayabilir.

Araştırmamızın sonucunda kNN algoritması için k=2 değeri seçilmiştir. k=1 değerine nazaran sistemin accuracy oranını az bir miktar düşürse de aşırı öğrenme durumu oluşmaması için seçilmesi daha uygundur.

3.2 Kullanılan Algoritmalardan Hangisi Rahatsızlığı Sınıflandırmada Daha İyi Sonuç Veriyor?

Kalp rahatsızlığı veri setinden sınıflandırma yaparken 4 adet algoritma kullanılmıştır. Bunlar sırasıyla kNN, Naïve Bayes, SVM ve Random Forest algoritmalarıdır. Bu algoritmaların temel tanımlarını yaparsak:

kNN algoritması, denetimli öğrenme algoritmalarından biridir. Belirlenen k değerine göre yeni sınıflandırılacak verinin etrafındaki komşularla olan uzaklıkları hesaplanarak bir sınıfa atanması şeklinde açıklanabilir. Projemiz için k değerini 2 olarak ele alacağız ve bu sebeple algoritma en yakın iki komşuya göre sınıflandırma yapacaktır [4].

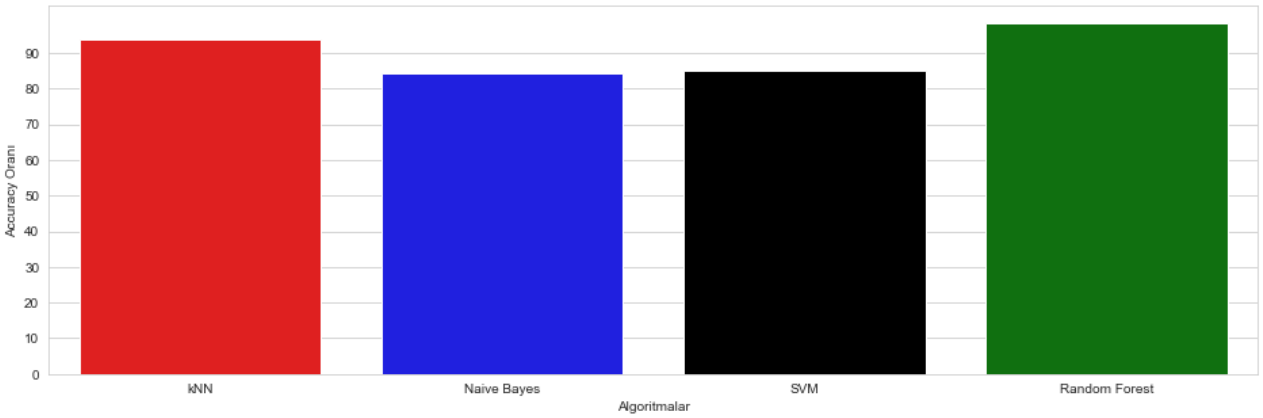
Naïve Bayes sınıflandırıcısının temeli Bayes teoremine dayanır. Tembel(Lazy) bir öğrenme algoritmasıdır aynı zamanda dengesiz veri setlerinde de çalışabilir. Algoritmanın çalışma şekli; bir eleman için her durumun olasılığını hesaplar ve olasılık değeri en yükse olana göre sınıflandırır. Az bir eğitim verisiyle bile başarılı işler çıkarabilir. Test setindeki bir değerin eğitim kümesinde gözlemlenemeyen bir değeri varsa olasılık değeri sıfır verir yani tahmin yapmaz [5].

Support Vector Machine algoritması, temel olarak iki sınıfa ait verileri birbirinden en uygun şekilde ayırmak için kullanılır. Bunun için karar sınırları ya da diğer bir ifadeyle hiper düzlemler belirlenir. Sınıflandırma için bir düzlemde bulunan iki grup arasında bir sınır çizilerek iki grubu ayırmak mümkündür. Bu sınırın çizileceği ye rise iki grubun da üyelerine en uzak olan yer olmalıdır. SVM bu sınırın nasıl çizileceğini belirler. Parametre almayan bir sınıflayıcıdır. Dağılım hakkında herhangi bir ön bilgi veya varsayım yoktur. Eğitim setlerinde girdi ve çıktılar eşlenir. Eşler aracılığıyla test setlerinde ve yeni veri setlerinde girdi değişkenini sınıflayacak karar fonksiyonları elde edilir [6].

Random Forest algoritması, birleştirilmiş karar ağaçları temelli sınıflandırma ve regresyon yöntemidir. Verilen bir örnek için orman içindeki her bir ağaçta sınıflandırma işlemi gerçekleştirilir. Sonrasında orman, oylama işlemi ile örneğin sınıfını belirler. Rastgele orman algoritmasında birden çok karar ağacının ortaya koyduğu sonuçlar bir araya getirilerek, orman adına tek bir karar verilerek daha güvenilir tahminler gerçekleştirilmektedir [7].

Projede kullanılan algoritmaların train ve test accuracy değerlerinin yanında ayrıca hata matrisleri de karşılaştırılmıştır.

Tablo 3.'te test sonucunda oluşan accuracy değerlerini karşılaştırılması görülmektedir.



Tablo 3. Algoritmaların Test Accuracy Oranlarının Karşılaştırılması

Tabloda görülen değerlerin sayısal olarak tam hali şu şekildedir:

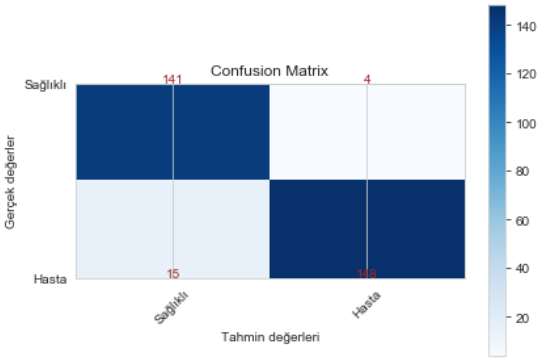
- kNN için test accuracy oranı: %93.83
- Naïve Bayes için test accuracy oranı: %84.42
- SVM için test accuracy oranı: %85.06
- Random Forest için test accuracy oranı: %98.38

Projede kullanılan dört algoritmanın train accuracy oranları ise şu şekildedir:

- kNN için train accuracy oranı: %99.30
- Naïve Bayes için train accuracy oranı: %84.52
- SVM için train accuracy oranı: %85.08
- Random Forest için train accuracy oranı: %99.86

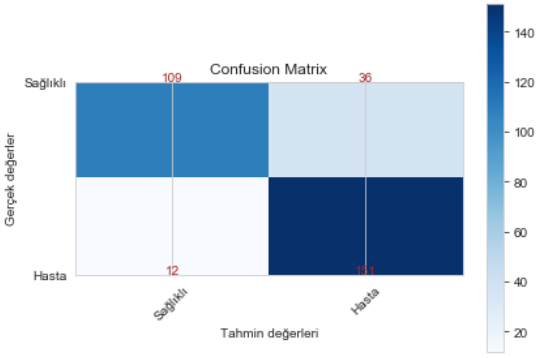
Train ve test değerlerini karşılaştırırsak kNN dışında diğer algoritmalar yakın değerler vermiştir. kNN algoritması train için %99 doğruluk verirken test için %93'de kalmıştır. En yüksek doğruluk oranlarını veren train için %99 ve test için %98 ile Random Forest olmuştur.

kNN algoritmasında k değerini seçerken accuracy dışında hata matrisi üzerinden de kıyaslama yapılmıştır. Bu sonuçlar üzerinden Random Forest algoritması en iyi sınıflandırmayı yapmıştır diyebiliriz ancak hata matrislerini ve precision, recall ve f-1 score değerlerini de kıyaslayarak daha doğru bir seçim yapmamız gerekmektedir.



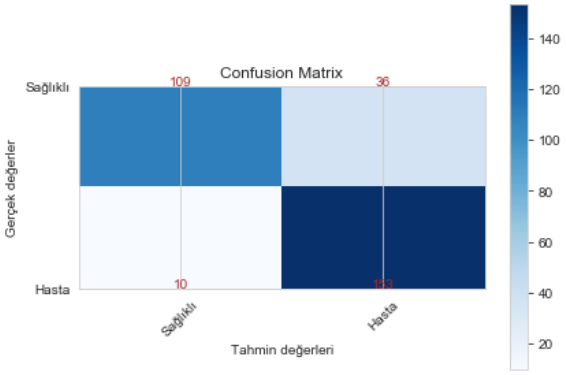
	precision	recall	f1-score	support
Sağlıklı	0.90	0.97	0.94	145
Hasta	0.97	0.91	0.94	163
accuracy			0.94	308
macro avg	0.94	0.94	0.94	308
weighted avg	0.94	0.94	0.94	308

- kNN algoritması test verisi olarak kullanılan 308 kişiden; 145 sağlıklı bireyin 141'ini sağlıklı ve 163 hasta bireyin 148'ini hasta olarak sınıflandırmıştır.



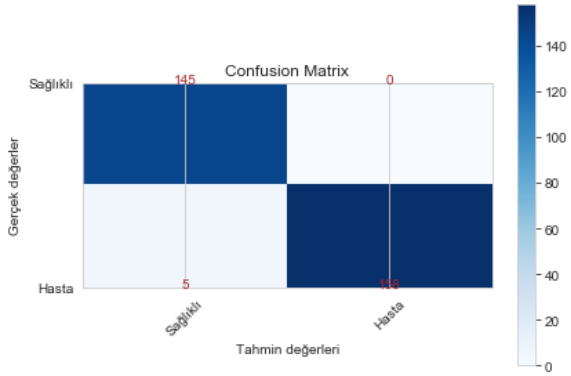
	precision	recall	f1-score	support
Sağlıklı	0.90	0.75	0.82	145
Hasta	0.81	0.93	0.86	163
accuracy			0.84	308
macro avg	0.85	0.84	0.84	308
weighted avg	0.85	0.84	0.84	308

- Naïve Bayes test verisi olarak kullanılan 308 kişiden; 145 sağlıklı bireyin 109'unu sağlıklı ve 163 hasta bireyin 1151'ini hasta olarak sınıflandırmıştır.



	precision	recall	f1-score	support
Sağlıklı	0.92	0.75	0.83	145
Hasta	0.81	0.94	0.87	163
accuracy			0.85	308
macro avg	0.86	0.85	0.85	308
weighted avg	0.86	0.85	0.85	308

- SVM test verisi olarak kullanılan 308 kişiden; 145 sağlıklı bireyin 109'unu sağlıklı ve 163 hasta bireyin 153'ünü hasta olarak sınıflandırmıştır.



	precision	recall	f1-score	support
Sağlıklı	0.97	1.00	0.98	145
Hasta	1.00	0.97	0.98	163
accuracy			0.98	308
macro avg	0.98	0.98	0.98	308
weighted avg	0.98	0.98	0.98	308

-

- Random Forest test verisi olarak kullanılan 308 kişiden; 145 sağlıklı bireyin 145'ini sağlıklı ve 163 hasta bireyin 158'ini hasta olarak sınıflandırmıştır.

Hata matrisleri ve sınıflandırma ölçütleri de incelendiğinde Random Forest algoritması test verisi üzerinde en iyi sınıflandırmayı yapmıştır. Bu yüzden train, test accuracy oranları ve hata matrisleri üzerinden incelendiğinde Random Forest bu proje için en iyi sonuçları veren algoritma olmuştur.

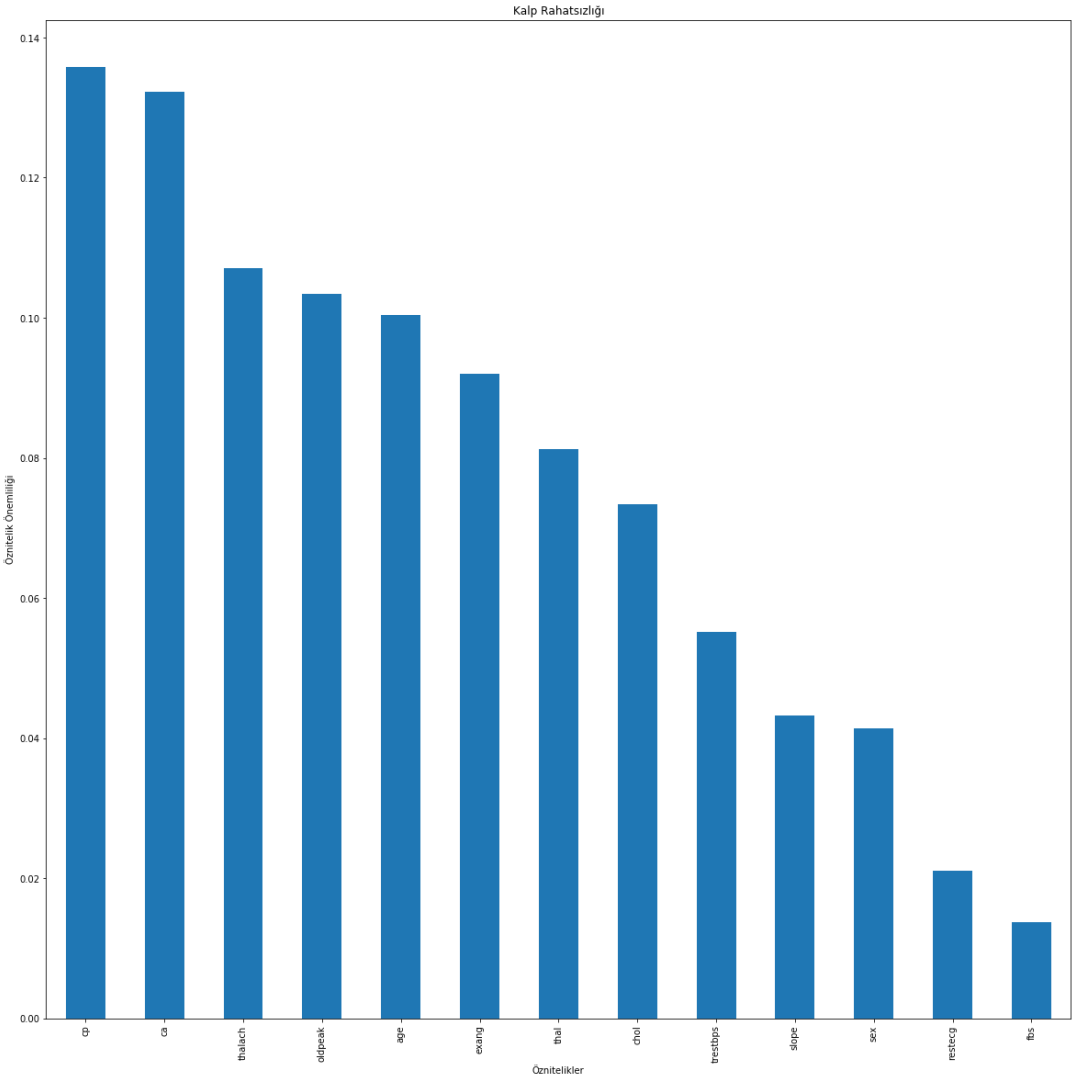
3.3 Hangi Özniteliklerin Sonuca Etkisi Daha Fazladır?

Veri setinde toplam 14 öznitelik bulunmaktadır. Bu öznitelikler şu şekildedir:

- age: Kişinin yaşı
- sex: Kişinin cinsiyeti(1=erkek, 0=kadın)
- cp(chest pain type): Kişinin göğüs ağrısı tipi(0'dan 3'e kadar sıralanmıştır.)
- trestbps: Kişinin dinlenmiş haldeki kan basıncı
- chol: Kişinin mg/dl cinsinden serum kolestrolü(mg/dl: desilitre başına milligram, belirli bir miktar kandaki belirli maddenin miktarını gösteren ölçüm)
- fbs: Kişinin aç karnına ölçülen kan şekeri(>120 mg/dl ise 1, değilse 0)
- restecg: Kişinin dinlendikten sonraki elektrokardiyografik sonuçları
- thalach: Kişinin elde edilen maksimum kalp atış hızı
- exang: Egzersize bağlı anjina(efor anjinası olarak da bilinir, anjina bir çeşit kalp rahatsızlığıdır.)
- oldpeak: ST depresyon(segment depresyonu, coroner arter hastalıkları tanı ve değerlendirilmesinde kullanılan fonksiyonel testlerden biri)
- slope: ST segmentinin eğimi
- ca: Kişinin büyük damar sayısı(0'dan 3'e kadar sıralanmıştır.)
- thal: talasemi tanısı(normal 3, fixed defect 6 ve reversable defect 7 olarak sıralanmıştır. Kalıtımsaldır.)
- target: Kalp rahatsızlığı tanısı(sağlıklı 0, hasta 1 şeklinde.)

Bu 14 özniteliğin 13'ü giriş 1'i çıkış değeridir. Veri setimizde 13 giriş değeri vardır. Ancak 13 öznitelik de sonucu aynı şekilde etkilememektedir.Bunlardan bazıları kalp rahatsızlığı tanısı için diğerlerine göre daha etkilidir.

Random Forest algoritması kullanarak öznitelikleri önem derecesine göre sıralamak mümkündür(Random Forest feature importance). Bu projede Random Forest algoritmasından yararlanarak yapmış olduğumuz araştırma sonucunda elde ettiğimiz tablo şu şekildedir:



Tablo 4. Özniteliklerin Sonuca Etkileri

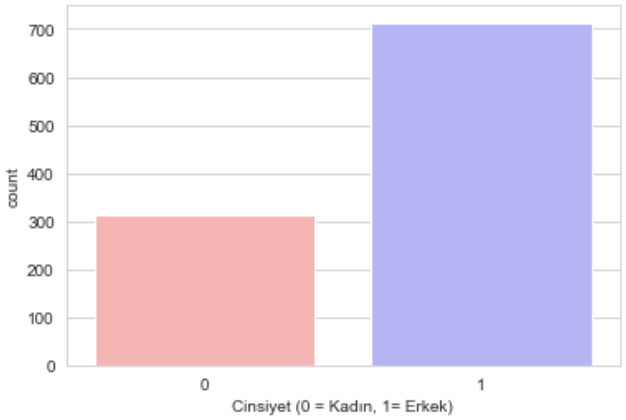
Grafikte görülen sonuca etkilerinin sıralı olarak gösterimi ise şu şekildedir:

1. cp: 0.1357
2. ca: 0.1322
3. thalach: 0.1070
4. oldpeak: 0.1034
5. age: 0.1003
6. exang: 0.0920
7. thal: 0.0812
8. chol: 0.0733
9. trestbps: 0.0551
10. slope: 0.0431
11. sex: 0.0414
12. restecg: 0.0210
13. fbs: 0.0136

Bu grafikten yola çıkarak kalp rahatsızlığı sınıflandırmasında sonucu diğerlerinden daha fazla etkileyen 5 öznitelik sırasıyla şu şekildedir: cp, ca, thalach, oldpeak, age. Diğerlerine nazaran daha az etkileyen öznitelikler ise sex, fbs ve restecg'dir.

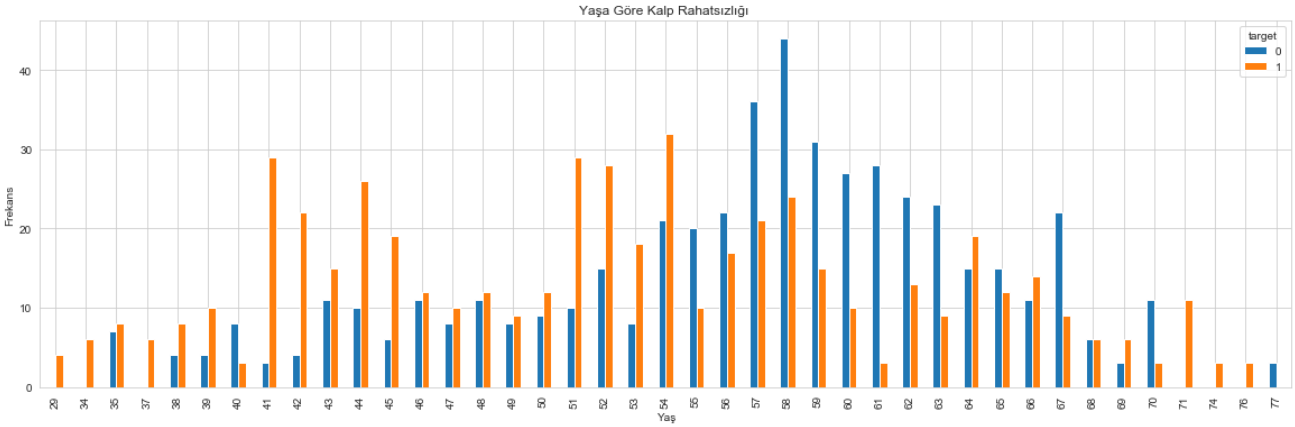
3.4 Özniteliklerin Veri Setindeki Dağılımı Ne Şekildedir?

Veri seti 1025 veriden oluşmaktadır. Bunların 526'sı hasta 499 tanesi ise sağlıklı bireylerden oluşmaktadır. Sağlıklı bireyler tüm veri setinin %48.68'ini oluştururken hasta bireyler %51.32'sini oluşturmaktadır. Bireylerin 713'ü(%69.56) erkek 312'si(%30.44) kadınlardan oluşmaktadır. Veri setindeki Kadın-Erkek dağılımı şu şekildedir:



Tablo 5. Kadın-Erkek Dağılımı

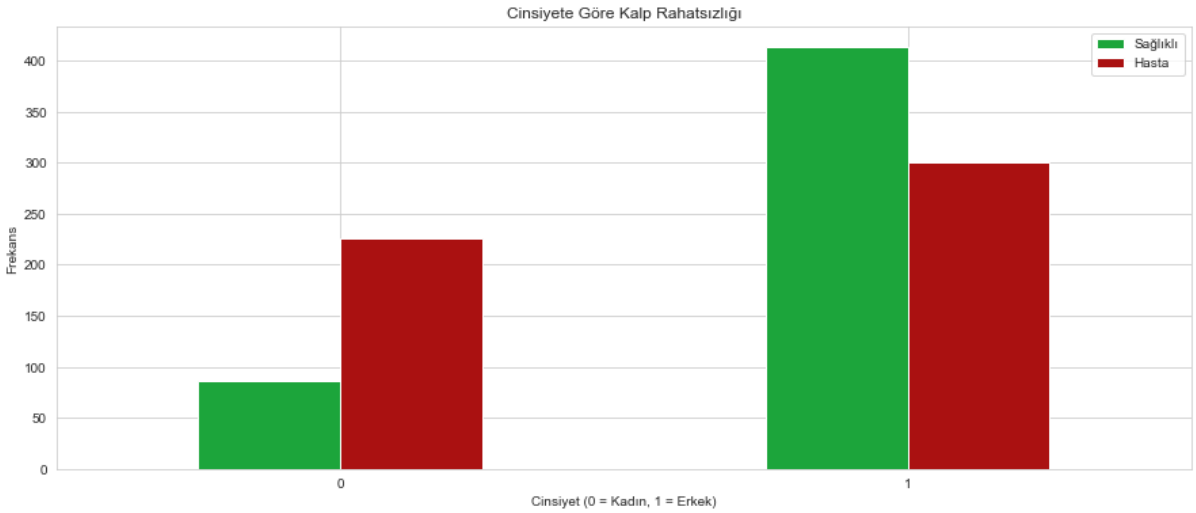
Yaşa göre Kalp rahatsızlığı dağılımı:



Tablo 6. Yaşa Göre Kalp Rahatsızlığı Dağılımı

Kalp rahatsızlığı tanısı konan hasta sayısı 55 yaşa kadar aradaki dalgalanmaları göz ardı edersek sağlıklı bireylerden yüksek çıkmıştır. Daha sonra belirli bir yaşa kadar sağlıklı birey sayısı hasta sayısından yüksek kalmıştır. Bu veriler sadece kullandığımız veri setindeki bireyler için geçerlilik göstermektedir.

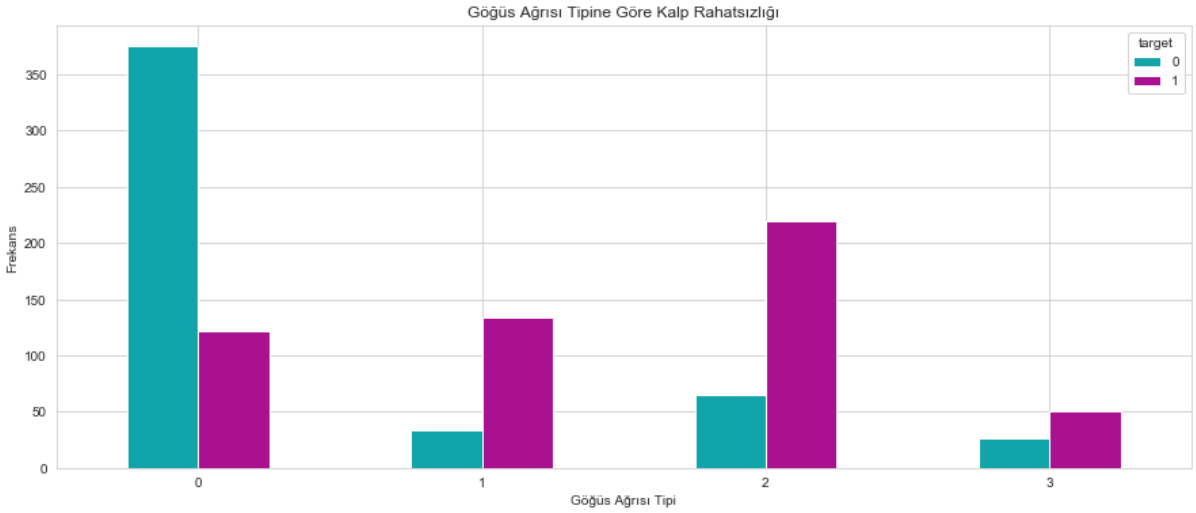
Cinsiyete göre Kalp rahatsızlığı dağılımı:



Tablo 7. Cinsiyete Göre Kalp Rahatsızlığı Dağılımı

Veri setindeki kadınlarda hasta kişi sayısı sağlıklı kişi sayısının 3 katı kadardır. Erkeklerde ise sağlıklı kişi sayısı hasta kişi sayısının 1.5 katı kadardır. Bu dağılımlar projede kullanılan veri setindeki dağılımlardır. Aradaki oranların bu şekilde olması erkek sayısının kadın sayısının 2 katı olmasıyla da açıklanabilir.

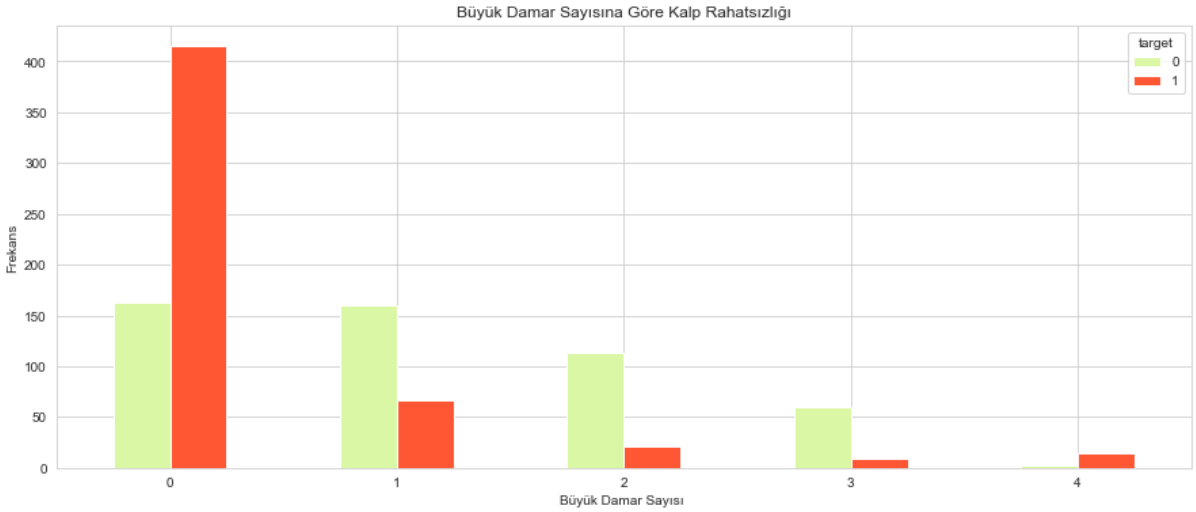
Göğüs ağrısı tipine göre Kalp rahatsızlığı dağılımı:



Tablo 8. Göğüs Ağrısı Tipine Göre Kalp Rahatsızlığı Dağılımı

Göğüs ağrısı tipi veri setinde 0-3 arasında sayısal değerler şeklinde ifade edilmektedir. Güçsüzden kuvvetliye doğru 0: en düşük, 3 ise en kuvvetli ağrıdır. Bunun dışında göğüs ağrısı 0 tipine sahip olanların büyük çoğunluğu sağlıklı iken 1,2 ve 3 ağrı tipine sahip olan insanların çoğunluğu hastalığa sahiptir.

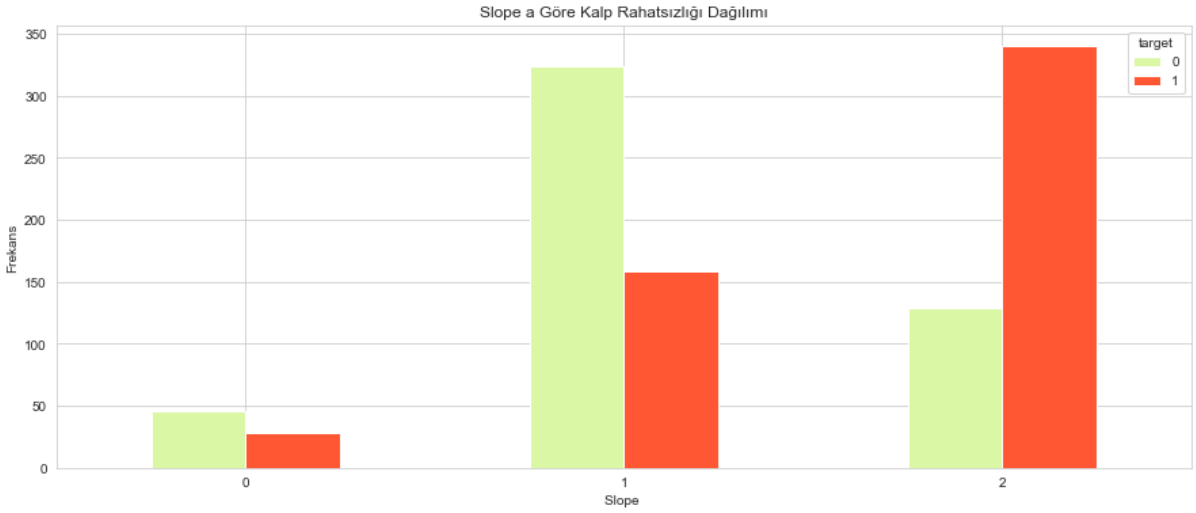
Büyük damar sayısına(ca) göre Kalp rahatsızlığı dağılımı:



Tablo 9. Büyük Damar Sayısına Göre Kalp Rahatsızlığı Dağılımı

Veri setindeki kişilerin büyük çoğunluğunun büyük damar sayısı 0'dır. Büyük damar sayısı 0 olan kişilerin %80'i hasta %20'si ise sağlıklıdır. 4'e kadar büyük damar sayısı artarken hasta kişi sayısı azalmaktadır. 4'te ise hasta sayısı %100'dür ancak 4 büyük damara sahip olan kişi sayısının 50'nin altında olduğunu unutmamak gerekmektedir.

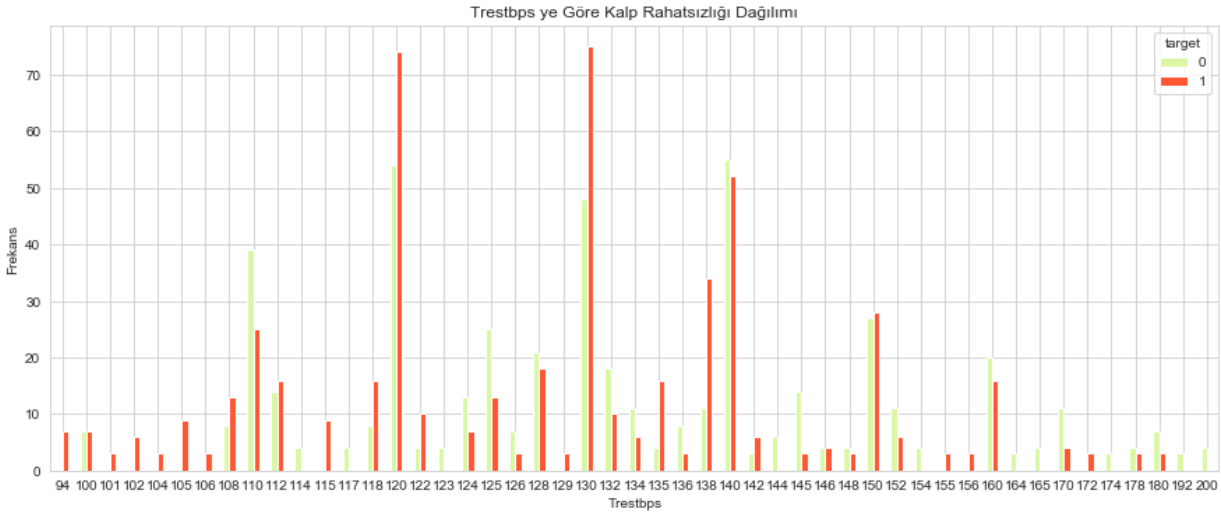
Slope'a göre kalp rahatsızlığı dağılımı:



Tablo 10. Slope'a Göre Kalp Rahatsızlığı Dağılımı

Slope değeri 0 ve 1 iken sağlıklı kişi sayısı hasta sayısının 2 katı kadardır. Ancak slope sayısı 2 olan kişiler de ise tam tersi bir durum söz konusudur, hasta sayısı sağlıklı kişi sayısının 2 katı kadardır.

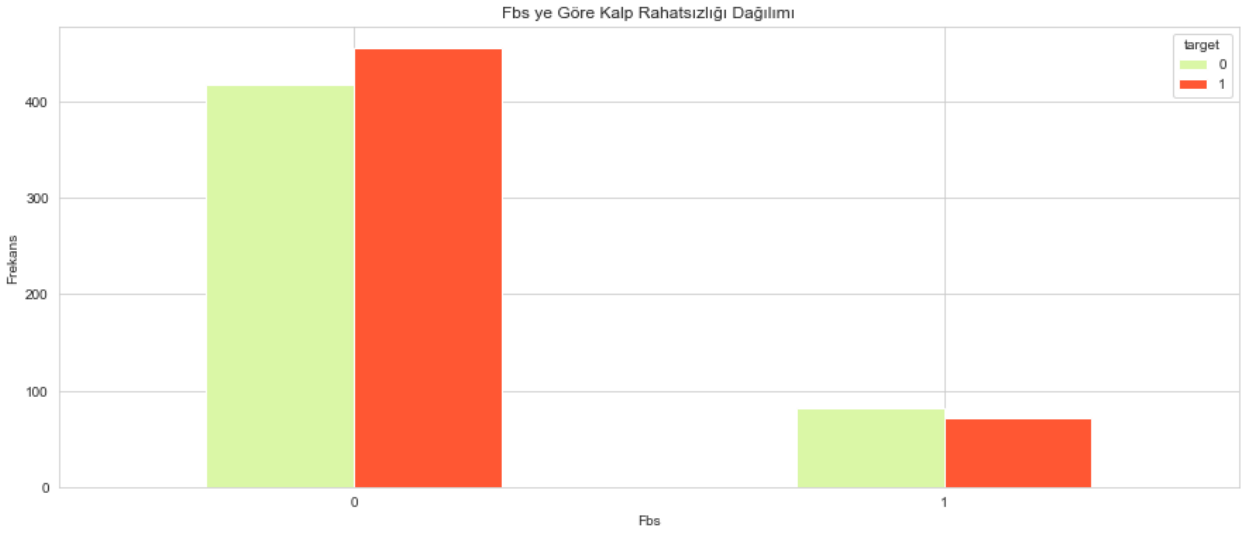
Trestbps'ye göre kalp rahatsızlığı dağılımı:



Tablo 11. Trestbps'ye Göre Kalp Rahatsızlığı Dağılımı

Dinlenmiş kan basıncı incelendiğinde sürekli bir dalgalanma halinde olduğu görülmektedir. Ancak 150'den sonraki değerlerde sağlıklı kişi sayısı hasta kişi sayısından fazla olmuştur.

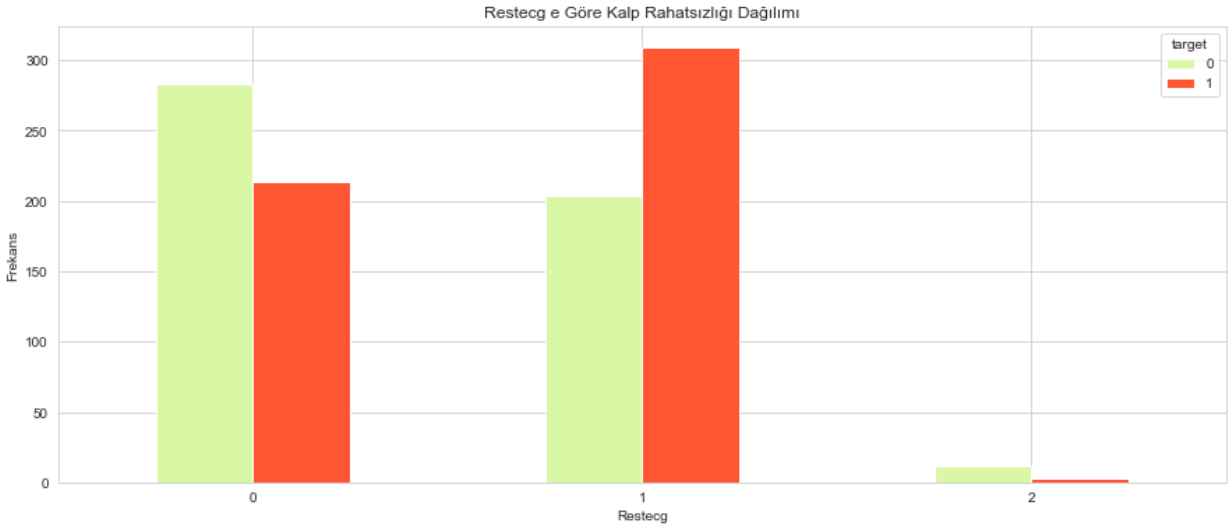
Fbs'ye göre kalp rahatsızlığı dağılımı:



Tablo 12. Fbs'ye Göre Kalp Rahatsızlığı Dağılımı

Fbs değerinin sonuca etkisinin az olduğu görülmektedir. Ancak veri setinin çoğunluğunda fbs değeri 0'dır. Fbs 0 iken hasta kişi sayısı sağlıklı kişi sayısından biraz fazladır, 1 olduğu durumda ise sağlıklı kişi sayısı hasta kişi sayısından çok az fazladır. Anlaşılacağı üzere bu öznitelik sonucu çok az etkilemektedir.

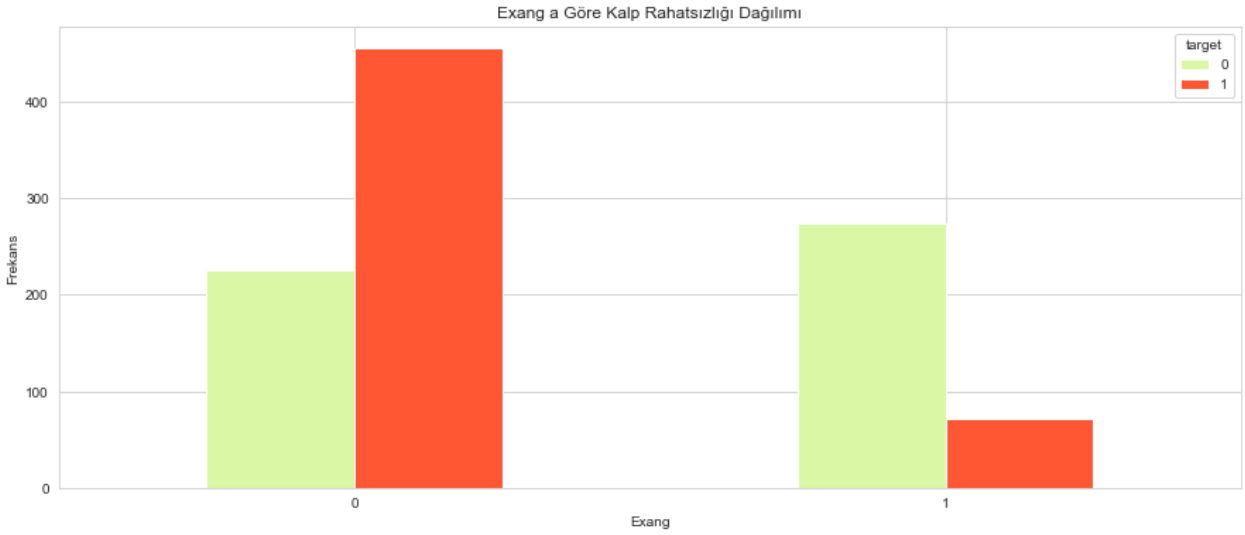
Restecg'e göre kalp rahatsızlığı dağılımı:



Tablo 13. Restecg'e Göre Kalp Rahatsızlığı Dağılımı

Veri setinin %95'inde restecg değeri 0 veya 1'dir. Buna göre restecg değeri 0 olduğunda sağlıklı kişi sayısı hasta kişi sayısından fazla, 1 olduğunda ise hasta kişi sayısı sağlıklı kişi sayısından fazladır.

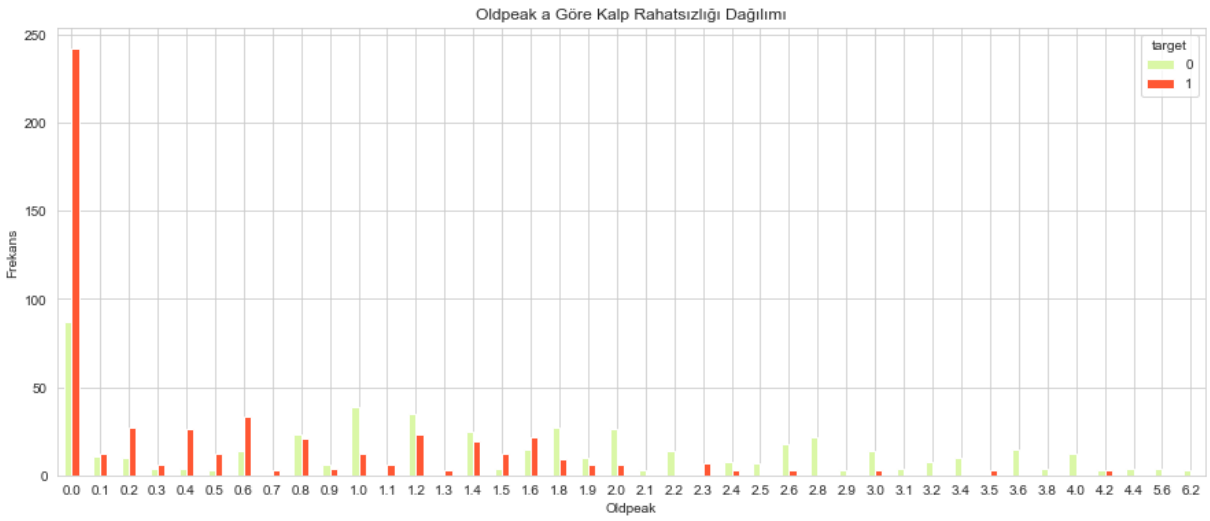
Exang'a göre kalp rahatsızlığı dağılımı:



Tablo 14. Exang'a Göre Kalp Rahatsızlığı Dağılımı

Exang değeri 0 iken hasta kişi sayısı sağlıklı kişi sayısının 2 katı kadardır. Değer 1 olduğunda ise sağlıklı kişi sayısı hasta kişi sayısının 3 katına kadar çıkmaktadır. Arada oluşan büyük farklar sınıflandırmada bu özniteliğin önemini arttırmaktadır.

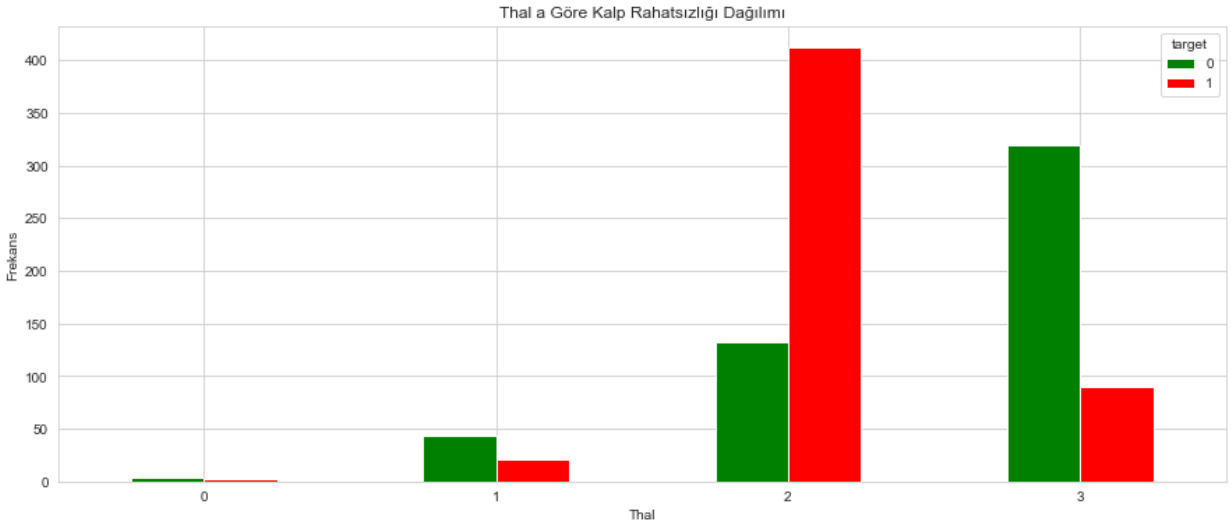
Oldpeak'e göre kalp rahatsızlığı dağılımı:



Tablo 15. Oldpeak'e Göre Kalp Rahatsızlığı Dağılımı

Oldpeak değeri 0.0 ile 6.2 arasında değişmektedir. Ancak büyük çoğunluk 0.0 değerindedir. Bu değere sahip kişilerde hasta kişi sayısı sağlıklı kişi sayısının 3 katı kadardır. Oldpeak değeri arttığında hasta kişi sayısı azalmaktadır.

Thal'a göre kalp rahatsızlığı dağılımı:



Tablo 16. Thal'a Göre Kalp Rahatsızlığı Dağılımı

Veri setinde thal değeri büyük çoğunlukla 2 ve 3 olarak görülmektedir. Thal değeri 2 olduğunda hasta kişi sayısı sağlıklı kişi sayısının 3 katına çıkmaktadır. Thal değeri 3 olduğunda ise sağlıklı kişi sayısı hasta kişi sayısının 3 katına çıkmaktadır. Bu farklılık sayesinde thal özniteliği sınıflandırmada büyük önem arz eden öznitelikler arasına girmektedir.

4 Discussion

Büyük damar sayısı, maksimum kalp atış hızı ve anjina tanısı sınıflandırma probleminde en büyük öneme sahip özniteliklerdir ancak beklenmeyecek şekilde öznitelikler arasında bulunan kan şekeri, cinsiyet ve dinlenme sonrası elektrokardiyografik sonuçların sınıflandırmada etkisi diğer özniteliklere nazaran çok daha düşük kalmıştır.

Sınıflandırmada kullanılan Random Forest algoritması diğer üç algoritmaya karşın en iyi doğruluk oranını vermiştir. Naïve Bayes ise en düşük doğruluk oranına sahip sınıflandırmayı yapmıştır. Ancak kullanılan dört algoritma da %80'in üzerinde accuracy oranı vererek test verisinin çoğunluğunu doğru sınıflandırmıştır.

kNN algoritmasında k değeri olarak 1 seçildiğinde elde edilen yüksek accuracy oranlarına rağmen her iterasyonda sadece bir sınıfı tamamen doğru sınıflandırması sebebiyle k değeri 2 olarak seçilmiştir. Bu değer herhangi bir sınıf için tamamen doğru değer vermemiştir ancak bu da accuracy oranını düşürmüştür.

Elde edilen sonuçlar 1025 kişilik veri seti üzerinde yapılan çalışmalarla ortaya çıkmıştır. Veri setindeki özniteliklerin kalp rahatsızlığı üzerine etkisinin daha iyi görülebilmesi için daha büyük bir veri setinde çalışma yapılabilir. Bazı özniteliklerin dengesiz dağılımı da aynı şekilde verimi azaltmış olabilir. Veri setindeki kadın sayısı erkek sayısına göre çok düşük kalmaktadır. Kalp rahatsızlığı üzerinde ciddi etkiye sahip olabilecek sigara kullanımı, spor yapma sıklığı vs. gibi öznitelikler de veri setinde bulunmamaktadır.

5 Recommendations and Conclusion

Her yıl sadece Amerika'da yaklaşık 647.000 kişi kalp rahatsızlıklarından hayatını kaybetmektedir [8]. Bir çok hastalıkta olduğu gibi kalp rahatsızlıklarında da erken teşhis önem arz etmektedir. Gerçekleştirdiğimiz projede elimizdeki verileri kullanarak kalp rahatsızlıklarını sınıflandırdık ve sistemi test verileri üzerinde sınıflandırma yapmasını sağlayarak tahminde bulunduk.

Elde edilen sonuçlardan yola çıkarak, sistem farklı öğrenme yöntemleri ve algoritmalar ile kişinin raporunda bulunan, ölçülen değerlerine göre kişinin kalp rahatsızlığına sahip olup olmadığını yüksek bir oranda doğru tahmin edebildi. Ancak bu çalışma görece küçük bir veri seti üzerinde gerçekleştirildi. Kalp rahatsızlığının tahmininde yapılacak diğer çalışmalarda daha verimli ve daha genel kabul görecek sonuçlar elde etmek için;

- Daha büyük veri setlerinde çalışma yapmak
- Daha farklı öğrenme algoritmaları kullanarak daha verimli sonuç almak veya algoritmaları birbirleriyle karşılaştırarak veri setindeki başarılarını ölçmek
- Veri setinde bulunan özniteliklerin üzerine kalp sağlığını iyi veya kötü anlamda doğrudan etkileyebilecek öznitelikler eklemek(sigara kullanımı vs. gibi)
- Dünyanın farklı bölgelerindeki hastanelerden veri toplayarak o bölgeye özgü beslenme alışkanlığından, iklim koşullarına kadar bir çok farklı parametrenin kalp rahatsızlığına etkisini araştırmak
- Kalp sağlığı üzerinde geliştirilecek bir projede bu konuda uzman kişilerden oluşan(sağlık doktoru, kalp uzmanları) bir ekip kurmak ve bu sonuçların uzman kişiler tarafından yorumlanmasını sağlamak

belirtilen maddeler uygulanabilir.

6 References

[1] Andras Janosi, William Steinbrunn, Matthias Pfisterer, Robert Detrano, 01/07/1988. Heart Disease Data Set. University of California Irvine. Available at: https://archive.ics.uci.edu/ml/datasets/Heart+Disease

[2] Andras Janosi, William Steinbrunn, Matthias Pfisterer, Robert Detrano, 01/07/1988. Heart Disease Data Set. University of California Irvine. Available at: https://archive.ics.uci.edu/ml/datasets/Heart+Disease

[3] Andras Janosi, William Steinbrunn, Matthias Pfisterer, Robert Detrano, 01/07/1988. Heart Disease Data Set. University of California Irvine. Available at: https://archive.ics.uci.edu/ml/datasets/Heart+Disease

[4] Anonymous, Last Update: 14/04/2020. k nearest neighbors algorithm. Web. Available at: https://en.wikipedia.org/wiki/K-nearest_neighbors_algorithm

[5] Anonymous, Last Update: 01/05/2020. Naïve Bayes classifier. Web. Available at: https://en.wikipedia.org/wiki/Naive_Bayes_classifier

[6] Anonymous, Last Update: 05/05/2020. Support-vector machine. Web. Available at: https://en.wikipedia.org/wiki/Support-vector_machine

[7] Anonymous, Last Update: 02/05/2020. Random forest. Web. Available at: https://en.wikipedia.org/wiki/Random_forest

[8] Kenneth D. Kochanek, M.A.; Jiaquan Xu, M.D.; Sherry L. Murphy, B.S.; Arialdi M. Minin˜o, M.P.H.; and Hsiang-Ching Kung, Ph.D. 29/11/2011. Deaths: Final Data for 2009. National Vital Statistics Reports Volume 60, Number 3. Available at: https://www.cdc.gov/nchs/data/nvsr/nvsr60/nvsr60_03.pdf

7 Appendices

7.1 Kullanılan Yazılım Geliştirme Ortamları

Anaconda (2019.10 version) platformu üzerinde Jupyter Notebook IDE(6.0.1 version)

7.2 Kullanılan Yazılım Dili

Python (3.7.4 version)

7.3 Veri Seti İçin CSV Düzenleme

Microsoft Office Excel 2016

7.4 Kullanılan Kütüphaneler

Numpy (1.16.5 version)

Pandas (0.25.1 version)

Matplotlib (3.1.1 version)

Sklearn (0.21.3 version)

Seaborn (0.9.0 version)

Itertools

7.5 Yardımcı Web Siteleri

https://www.wikipedia.org

www.kaggle.com

www.stackoverflow.com

www.towardsdatascience.com

www.medium.com

www.tinyurl.com/OruntuTanima2019