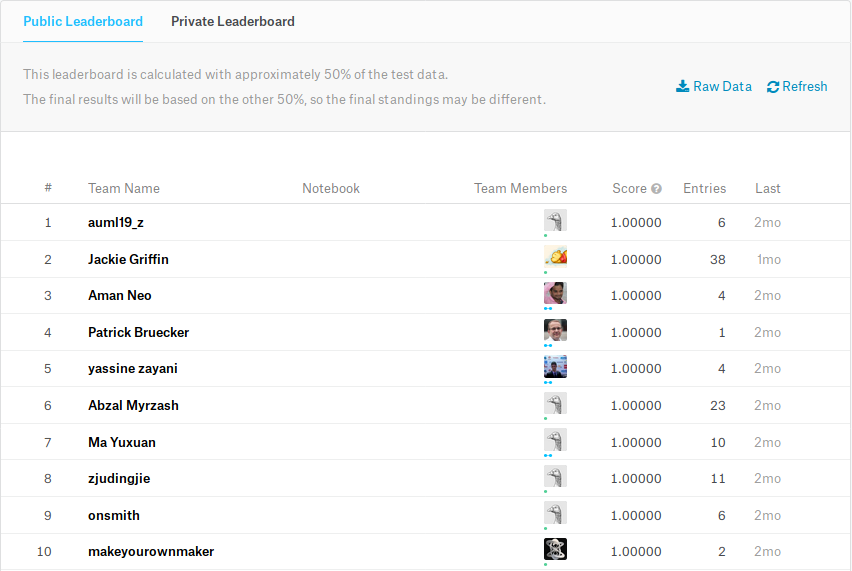
**RMS *Titanic***, [White Star Line](https://www.wikizeroo.org/index.php?q=aHR0cHM6Ly90ci53aWtpcGVkaWEub3JnL3dpa2kvV2hpdGVfU3Rhcl9MaW5l) şirketine ait Olympic sınıfı bir [transatlantik](https://www.wikizeroo.org/index.php?q=aHR0cHM6Ly90ci53aWtpcGVkaWEub3JnL3dpa2kvVHJhbnNhdGxhbnRpaw) yolcu gemisiydi. [Harland and Wolff](https://www.wikizeroo.org/index.php?q=aHR0cHM6Ly90ci53aWtpcGVkaWEub3JnL3dpa2kvSGFybGFuZF9hbmRfV29sZmY) ([Belfast](https://www.wikizeroo.org/index.php?q=aHR0cHM6Ly90ci53aWtpcGVkaWEub3JnL3dpa2kvQmVsZmFzdA), [İrlanda](https://www.wikizeroo.org/index.php?q=aHR0cHM6Ly90ci53aWtpcGVkaWEub3JnL3dpa2kvJUM0JUIwcmxhbmRh)) tersanelerinde üretilmiştir. [15 Nisan](https://www.wikizeroo.org/index.php?q=aHR0cHM6Ly90ci53aWtpcGVkaWEub3JnL3dpa2kvMTVfTmlzYW4) [1912](https://www.wikizeroo.org/index.php?q=aHR0cHM6Ly90ci53aWtpcGVkaWEub3JnL3dpa2kvMTkxMg) gecesi daha ilk seferinde bir buz dağına çarpmış ve yaklaşık iki saat kırk dakika içinde Kuzey Atlantik'in buzlu sularına gömülmüştür.[1912](https://www.wikizeroo.org/index.php?q=aHR0cHM6Ly90ci53aWtpcGVkaWEub3JnL3dpa2kvMTkxMg)'de yapımı tamamlandığında dünyanın en büyük buharlı yolcu gemisiydi. Batışı 1.514 kişinin ölümüyle sonuçlanmış ve en büyük deniz felaketlerinden biri olarak tarihe geçmiştir.

*Titanic*`in batışının yol açtığı büyük can kaybı oranı birçok nedene bağlanmaktaydı ama zamanla öne çıkan gerçek, geminin herkese yetecek kadar filika taşımıyor olmasıydı. *Titanic*`in tam kapasitesi 3,547 kişi olmasına rağmen gemideki filikaların toplam kapasitesi 1,178 kişiydi. Ayrıca kaza sırasında kadınlara ve çocuklara öncelik tanındığı için toplamda ölen erkek sayısı da çok orantısızdı.

*Titanic*`te zamanında mevcut olan en ileri teknolojiler kullanılmıştı. Birçok insan tarafından "batmaz gemi" olduğuna inanılıyordu ve bu inanış batmadan önce bu şekilde tanımlanmış ve lanse edilmişti. Bu derece ileri teknoloji ve eğitimli mürettebata rağmen batması birçok insanı şoke etti. Medya *Titanic'*in ünlü kurbanları ve batışı ile ilgili efsaneleri sürekli gündeme getirmeye devam etti. Bu tartışmaların sonucu denizcilik kanunun değişmesi oldu.

Amerikan donanmasından emekli [Robert Ballard](https://www.wikizeroo.org/index.php?q=aHR0cHM6Ly90ci53aWtpcGVkaWEub3JnL3dpa2kvUm9iZXJ0X0JhbGxhcmQ)'ın geminin enkazını 1985'te bulması *Titanic'*e olan ilgiyi artırdı ve bu ilgi günümüze kadar devam etti.

Konu en son olarak Kaggle sitesinde bir Makine Öğrenmesi yarışması olarak ortaya çıktı. Kaggle sitesindeki bu yarışmada bir çok farklı Makine Öğrenmesi algoritması ile çözümler yapıldı, yarışmacılar farklı sonuçlar ile sıralamada yer aldılar. Bazı yarışmacılar Wikipedia sitesindeki Titanic yolcu listesinden yararlanarak sistemde bir açık buldular. Gerçek yolcu listesini kendi tahmin algoritmalarına öğreterek kusursuz (%100) tahmin başarısına ulaştılar.



Veri Kümesi ve Problem Tanımı

Veri kümesi Kaggle tarafından yarışma için yayınlanmıştır. Bizden istenen çözüm verilen özniteliklere göre kişilerin hayatta kalma yada ölme durumunun tahmin edilebilmesidir. Verilen öznitelikler aşağıda tarif edilmiştir.

Verisetini incelemeye başlarsak içinde yer alan öznitelikler:

* Survived: Hayatta kaldı (1) or öldü (0)
* Pclass: Yolcunun sınıfı
* Name: Yolcunun ismi
* Sex: Yolcunun cinsiyeti
* Age: Yolcunun yaşı
* SibSp: Gemideki eş/kardeşlerinin sayısı
* Parch: Gemideki çocuk/ebeveynlerinin sayısı
* Ticket: Bilet numarası
* Fare: Bilet ücreti
* Cabin: Yolcu kabini
* Embarked: Biniş limanı

Survived: Hayatta kalmak yada ölmek 1,0 değerlerinden oluşur.

Pclass: 3 tane yolcu sınıfı vardır. Bunlar 1,2,3 diye gruplara ayrılır.

Name: Yolcu isimleri her yolcuya özeldir aynı isime sahip yolcu yoktur.

Sex: Cinsiyetler male ve female olmak üzere 2 tanedir.

Age: Yolcuların yaşları farklı değerlere sahiptir.

SipSb: Bir yolcunun gemideki toplam (kardeş+eş) sayısına bakılarak hazırlanmış değerdir.

Parch: Bir yolcunun gemideki toplam (ebeveyn+çocuk) sayısına bakılarak hazırlanmış değerdir.

Ticket: Her yolcu için farklı olan özel bilet numarasıdır.

Fare: Bilet ücretleri farklı liman ve yolcu tiplerine göre farklılık göstermektedir.

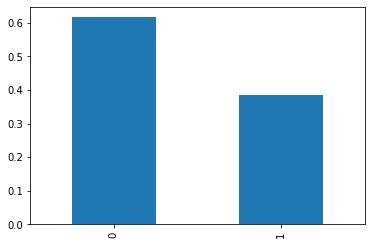
Cabin: Yolcuların sahip olduğu kabin numaralarıdır.

Embarked: 3 farklı liman vardır. (C = Cherbourg; Q = Queenstown; S = Southampton)

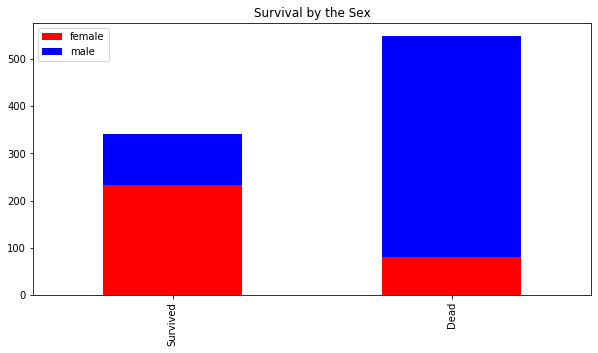
Title: Lakaplar. Lakaplar isimlerin başında yer alan farklı ünvanlardır.

Verisetinin grafiklerle tanıtılması.

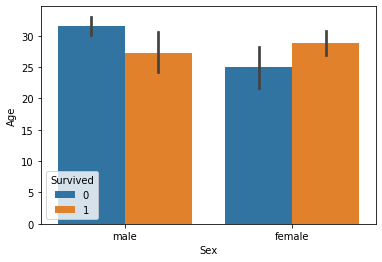
İlk grafiğimizde gemideki toplam hayatta kalma / ölüm oranı gözükmektedir. 0 ölümler 1 hayatta kalmaları temsil etmektedir. Grafiğe bakıldığında yaklaşık olarak %40 hayatta kalma oranı olduğunu görebiliriz.



İkinci grafikte ise cinsiyetlere göre toplam hayatta kalma yada ölüm sayılarını inceleyebiliriz. Gemide yolcular kurtarılırken kadın ve çocuklara öncelik verildiği için erkeklerin ölüm sayısının kadınlara göre çok daha yüksek olduğunu görebiliriz.



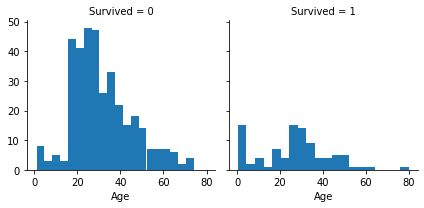
Yaşa ve cinsiyete göre ölüm durumları incelendiğinde ise erkeklerde yaşı daha genç olanların kadınlarda ise daha yaşlı olanların daha fazla hayatta kaldığını görmekteyiz.



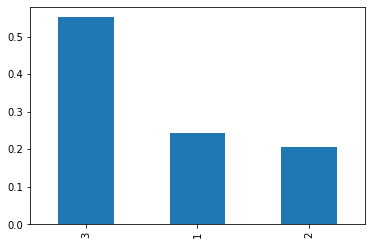
Yaşa göre hayatta kalma grafiğinin yoğunluğa göre dağılımı incelendiğinde ise en çok can kaybının 20 – 40 yaşları arasında olduğu görülmekte.



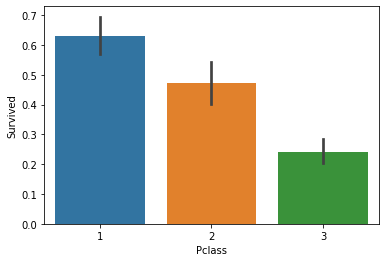
Yaşa göre hayatta kalma ile ilgili detaylı grafik.

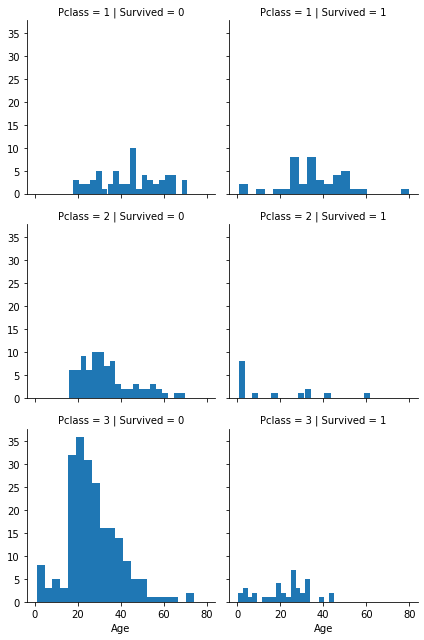


Yolcu sınıflarının dağılımı grafiğinde ise en çok 3. sınıf en az ile 2. sınıf yolcu olduğu görülmekte. Yolcuların %50’den fazlası 3. sınıf yolcudur.



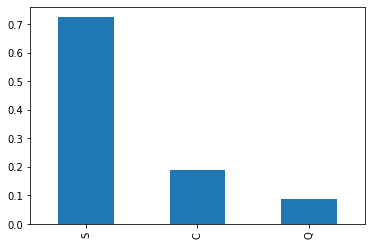
Farklı yolcu sınıflarının hayatta kalma oranı. 1. sınıf yolcuların hayatta kalma oranı diğer sınıflara göre daha yüksektir.



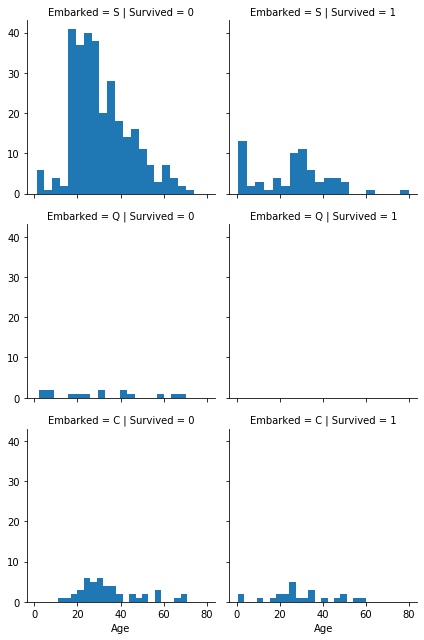


Farklı yolcu sınıfları ve yaşlara göre hayatta kalma ile ilgilim detaylı grafik.

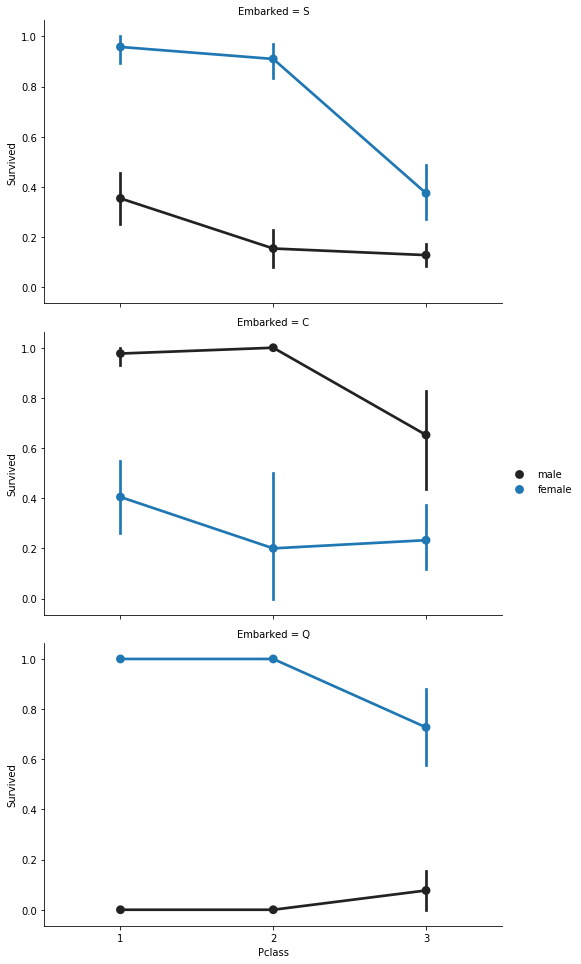
Farklı limanlardan binen yolcu sayıları. Southampton limanı yolcuların %70’den fazlasının bindiği liman olarak öne çıkmaktadır.



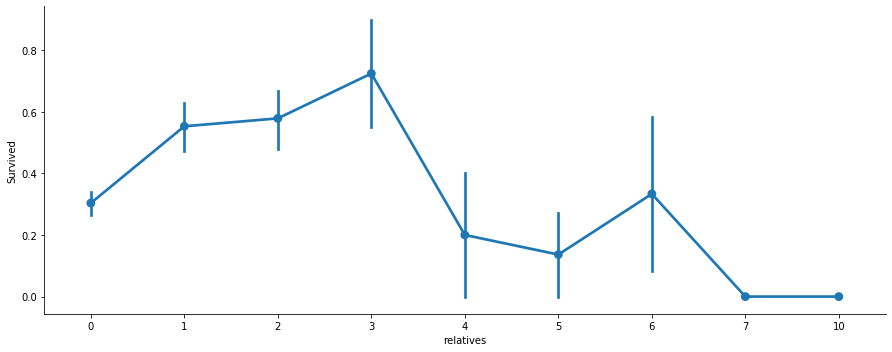
Farklı limanlardan binen yolcuların yaşlara göre hayatta kalma grafiği.



Aşağıdaki grafik ise limanlar,cinsiyet ve hayatta kalma durumları birleştirip yorumlayan bir grafiktir.

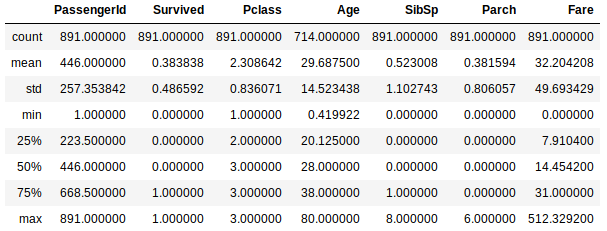


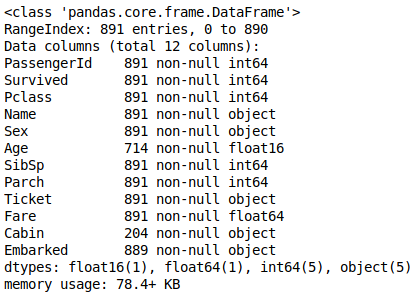
Gemideki akraba sayısına göre ölüm kalım durumlarını içeren grafik.



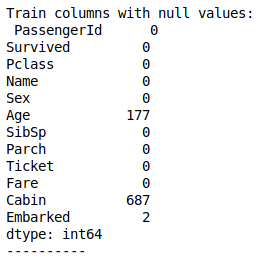
Veri önişleme

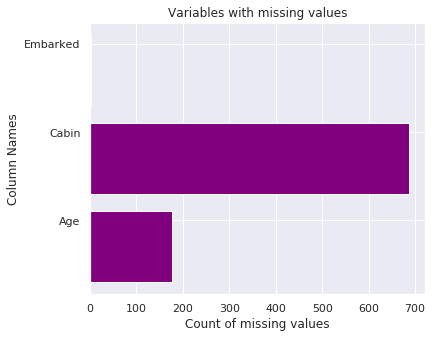
Öncelikli olarak verisetinin tanımı ve bilgisi incelenir.





Verisetindeki eksik veri sayıları öğrenilir.

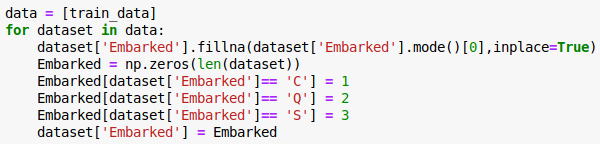




Verisetinden bazı öznitelikler atılır.



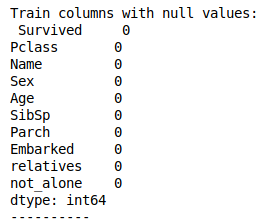
Biniş limanları 1,2,3 olarak yeniden adlandırılır.



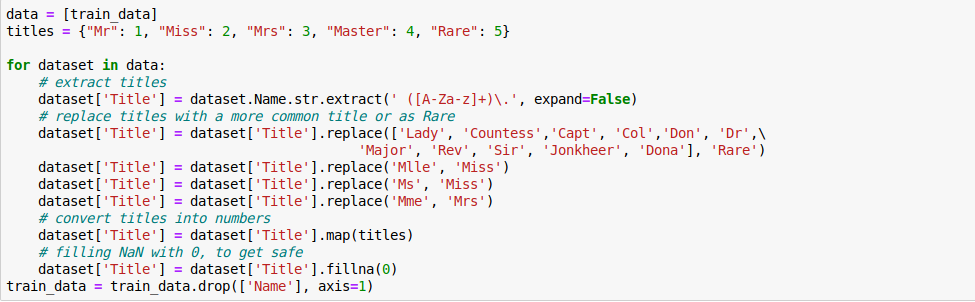
Eksik olan yaş verileri medyan ile doldurulur.



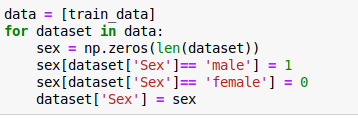
Eksik veri kalmadığını kontrol edilir.



Lakaplar önce kümelendirilir ondan sonra da sayısal değerlere dönüştürülür. Ondan sonra da isimler atılır.



Cinsiyetler 0,1 olarak değiştirilir.



**Kullanılan Modeller**

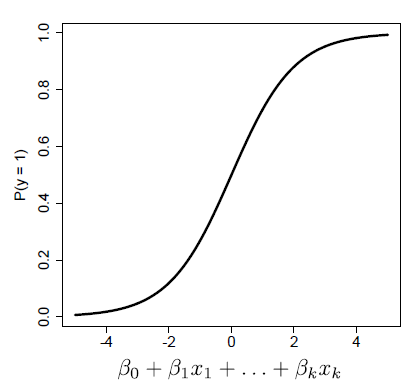
**Logistic Regression**

Logistic Regression ( Lojistik Regresyon ) sınıflandırma işlemi yapmaya yarayan bir regresyon yöntemidir. Kategorik veya sayısal verilerin sınıflandırılmasında kullanılır. Bağımlı değişkenin yani sonucun sadece 2 farklı değer alabilmesi durumda çalışır. ( Evet / Hayır, Erkek / Kadın, Şişman / Zayıf vs. ) Doğrusal sınıflandırma problemlerinde yaygın bir biçimde kullanılır.

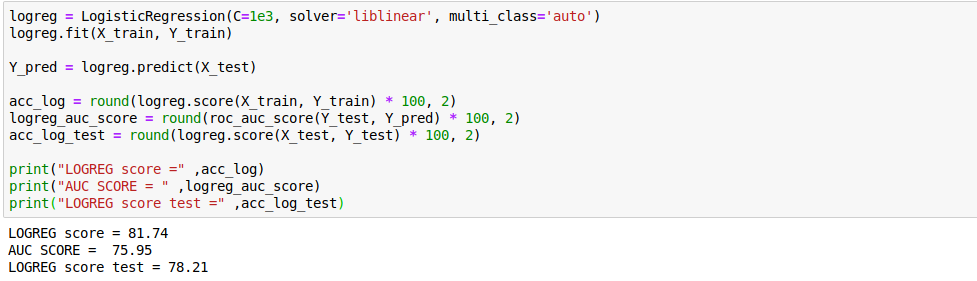
Logistic Regresyon için örnekler:

* *Email Spam Filtresi: Spam / Spam Değil.*
* *Dolandırıcılık kontrolü : Ödeme dolandırıcılık içerir , Evet / Hayır.*
* *Futbolcunun şut çekmesi : Gol oldu / Gol olmadı.*

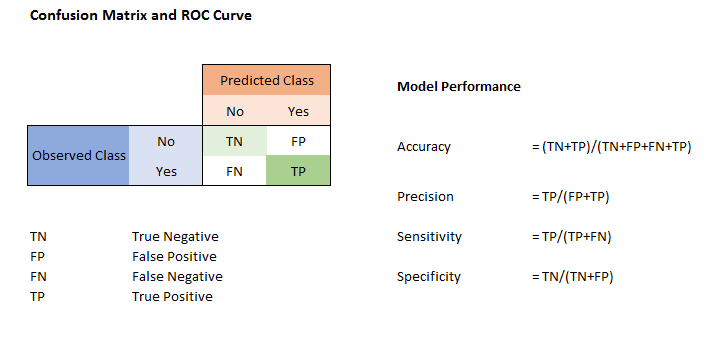
Lineer regresyon bize (-∞, +∞) arasında sonuçlar verir. Bizim problemimizde elde etmek istediğimiz bir olasılık değeri olduğu için değerleri S şeklindeki lojistik fonksiyonundan (sigmoid fonksiyonu) geçirip [0,1] arasına alıyoruz. Örneğin sonuç 0.8 çıktıysa şunu söylebiliriz: elimizdeki X değerleriyle (cinsiyet, yaş,akraba sayısı vs.) o kişi %80 olasılıkla hayatta kalmıştır(survived = 1). Günün sonunda da elde ettiğimiz olasılık değeri 0.5'ten büyükse sonucu 1 olarak küçükse 0 olarak etiketliyoruz.

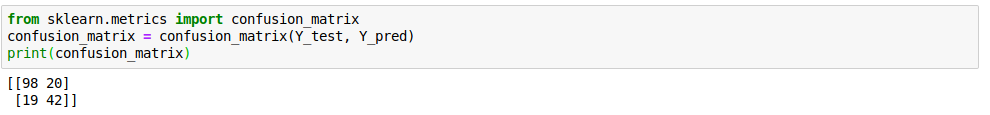


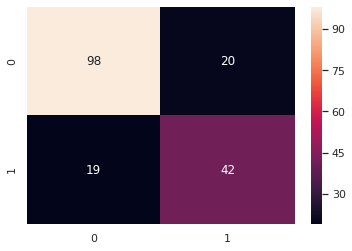
Model çözümü yaptırılır.



Karmaşıklık Matrisi ve ROC Eğrisi bize modelin performansı ile ilgili değerleri bulmamızda yardımcı olur. TN ve TP değerleri modelin doğru tahmin ettiği değerlerdir. Modelin accuracy (isabet) oranı (TN+TP)/(TN+TP+FN+FP) formülü ile bulunur.

Karmaşıklık matrisi

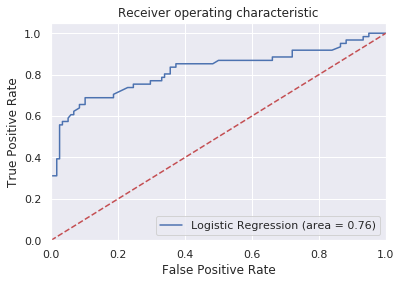


Karmaşıklık Matrisi’nin ısı haritası yöntemi ile gösterimi.

Üstteki accuracy oranının sağlaması bu matris yardımı ile yapabiliriz.

(98+42)/(98+42+20+19)= 0.7821

ROC Eğrisi: Ortadan geçen kesikli çizgi tamamen rassal olan bir modeli temsil eder. Bizim modelimiz o çizgiden ne kadar uzakta durursa (sol üst köşe) o kadar başarılı sayılır.



k-En Yakın Komşuluk — k Nearest Neighbours

k-en yakın komşuluk (KNN) algoritması hem sınıflandırma hem de regresyon problemlerinin çözümünde kullanılıyor olmakla birlikte, çoğunlukla sınıflandırma problemlerinin çözümünde kullanılmaktadır.

KNN algoritmaları, 1967 yılında T. M. Cover ve P. E. Hart tarafından önerilmiştir. Algoritma, sınıfları belli olan bir örnek kümesindeki verilerden yararlanılarak kullanılmaktadır. Örnek veri setine katılacak olan yeni verinin, mevcut verilere göre uzaklığı hesaplanıp, k sayıda yakın komşuluğuna bakılır. Uzaklık hesapları için genelde 3 tip uzaklık fonksiyonu kullanılmaktadır:

* “Euclidean” Uzaklık
* “Manhattan” Uzaklık
* “Minkowski” Uzaklığı’dır.

Minkowski Uzaklığı Denklemi:

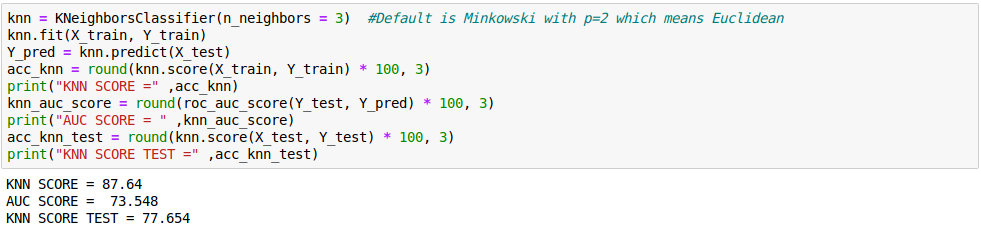


p = 2, olduğu durumlarda Minkowski uzaklığı Euclidean uzaklığına eşittir.

p = 1, olduğu durumlarda Minkowski uzaklığı Manhattan uzaklığına eşittir.

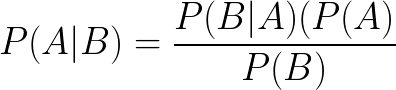
**KNN algoritmasının adımları:**

* İlk olarak k parametresi belirlenir. Bu parametre verilen bir noktaya en yakın komşuların sayısıdır. Örneğin: k=3 olsun. Bu durumda en yakın 3 komşuya göre sınıflandırma yapılacaktır.
* Örnek veri setine katılacak olan yeni verinin, mevcut verilere göre uzaklığı tek tek hesaplanır. İlgili uzaklık fonksiyonları yardımıyla.
* İlgili uzaklılardan en yakın k komşu ele alınır. Öznitelik değerlerine göre k komşu veya komşuların sınıfına atanır.
* Seçilen sınıf, tahmin edilmesi beklenen gözlem değerinin sınıfı olarak kabul edilir. Yani yeni veri etiketlenmiş (label) olur.



Modelde komşu sayısı 3 olarak alınmıştır. Çözüm yöntemi varsayılan olarak bırakılmış yani p=2 ‘ ye göre Minkowski denklemi kullanılmıştır. Bu da Euclidean denklemine eşittir.

**GAUSSIAN NAIVE BAYES**

****

P(A|B) = B olayı gerçekleştiği durumda A olayının meydana gelme olasılığıdır.

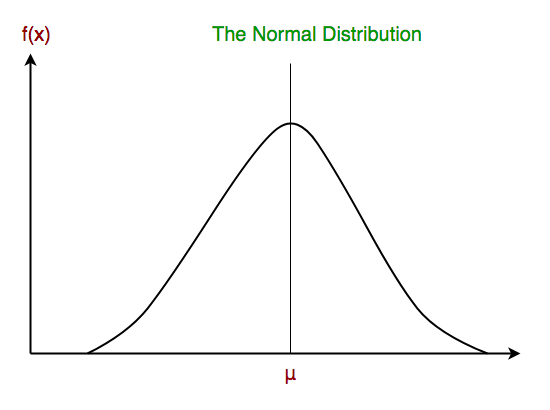
P(B|A) = A olayı gerçekleştiği durumda B olayının meydana gelme olasılığıdır.

P(A) = A olayının olma olasılığı.

P(B) = B olayının olma olasılığı.

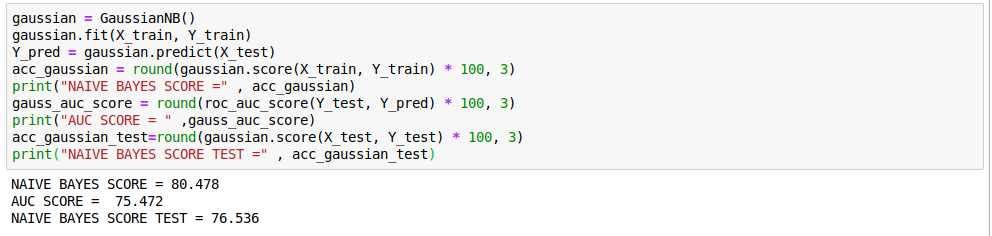
Naive Bayes sınıflandırıcıları Bayes’in teoremini temel alarak çalışmaktadır. Bir olayın gerçekleşme olasılığını önceden öğrenilmiş bağımlı şartlara göre hesaplamakta kullanılır.

Gaussian Naive Bayes sınıflandırıcısı öznitelikleri Gaussian dağılımına (Normal dağılım) göre sınıflandırarak tahmin yapar.



Gaussian Naive Bayes’e göre olasılık formülü:

Model çözümü:

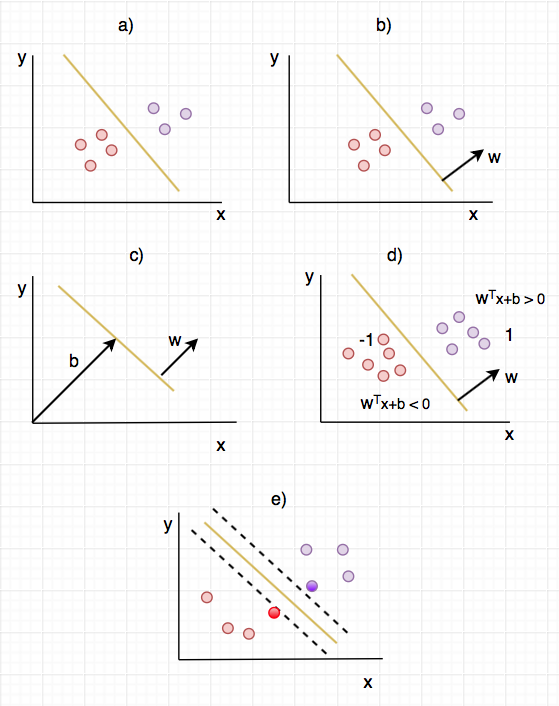


**Support Vector Machine (Destek Vektör Makineleri)**

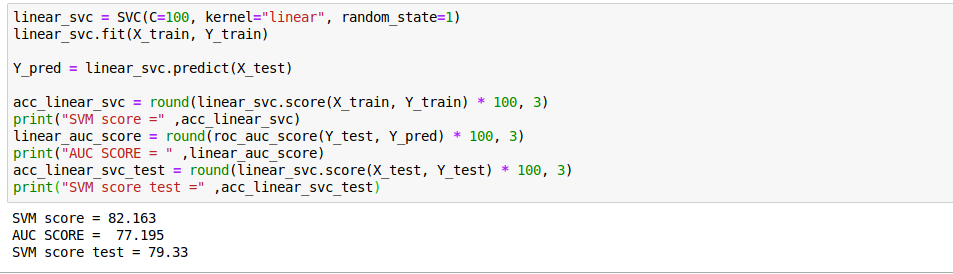
Sınıflandırma için bir düzlemde bulunan iki grup arasında bir sınır çizilerek iki grubu ayırmak mümkündür. Bu sınırın çizileceği yer ise iki grubun da üyelerine en uzak olan yer olmalıdır. SVM bu sınırın nasıl çizileceğini belirler.

Bu işlemin yapılması için iki gruba da yakın ve birbirine paralel iki sınır çizgisi çizilir ve bu sınır çizgileri birbirine yaklaştırılarak ortak sınır çizgisi üretilir.

Destek vektörlerinin (Şekilde kısım “e”) üzerinde bulunduğu ve kesikli çizgilerle gösterilmiş düzlemlere sınır düzlemleri denir. Sınır düzlemlerinin tam ortasından geçen ve her iki düzleme de eşit uzaklıkta bulunan düzlem ise hiper düzlem olarak ifade edilir. Şekilde (-1, +1) sınıf etiketlerini, w ağırlık vektörünü (hiper-düzlemin normali) ve b ise eğilim değerini göstermektedir.



Modelin uygulanması:



Karar Ağacı

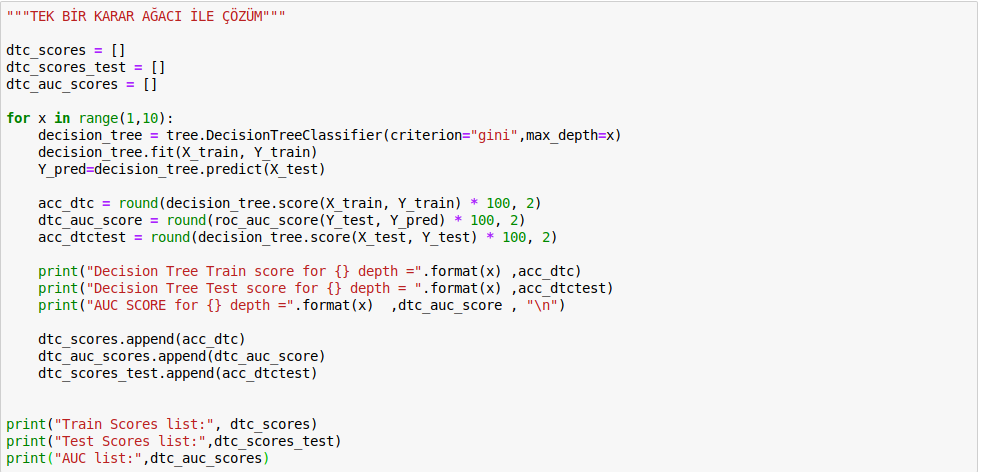
Karar ağacı algoritması, veri madenciliği sınıflandırma algoritmalarından biridir. Karar ağacı algoritması CART (Classification and Regression Tree) sınıfına girer.

Önceden tanımlanmış bir hedef değişkene sahiplerdir. Yapıları itibariyle en tepeden en aşağı inen bir strateji sunmaktadırlar.

Bir karar ağacı, çok sayıda kayıt içeren bir veri kümesini, bir dizi karar kuralları uygulayarak daha küçük kümelere bölmek için kullanılan bir yapıdır. Yani basit karar verme adımları uygulanarak, büyük miktarlardaki kayıtları, çok küçük kayıt gruplarına bölerek kullanılan bir yapıdır.

* Sınıflandırma problemleri : Sınıflandırma ağaçları (Classification Tree), bir kişinin harcamalarından eğitim düzeyinin tahmini gibi, hedef kümeyi çeşitli sınıflardan birisine yerleştirmeyi amaçlayan ve sınıf tanımı yapan problemlerdir.
* Regresyom problemleri : Regresyon ağaçları, sonuçta bir sınıf yerine sayısal bir değer döndüren veri madenciliği problemleridir. Örneğin babası kalp hastası olan ve sigara kullanan bir bireyin kalp krizi geçirme ihtimali.

Modelin Uygulanması



Rassal Orman (Random Forest)

Rassal orman, birden fazla karar ağacı ile oluşturur ve daha doğru ve istikrarlı bir tahmin elde etmek için onları birleştirir. Yeteri kadar ağaç sayısına ulaştığında ise aşırı öğrenme (Overfitting) sorunu çıkarmazlar.

Modelin Uygulanması

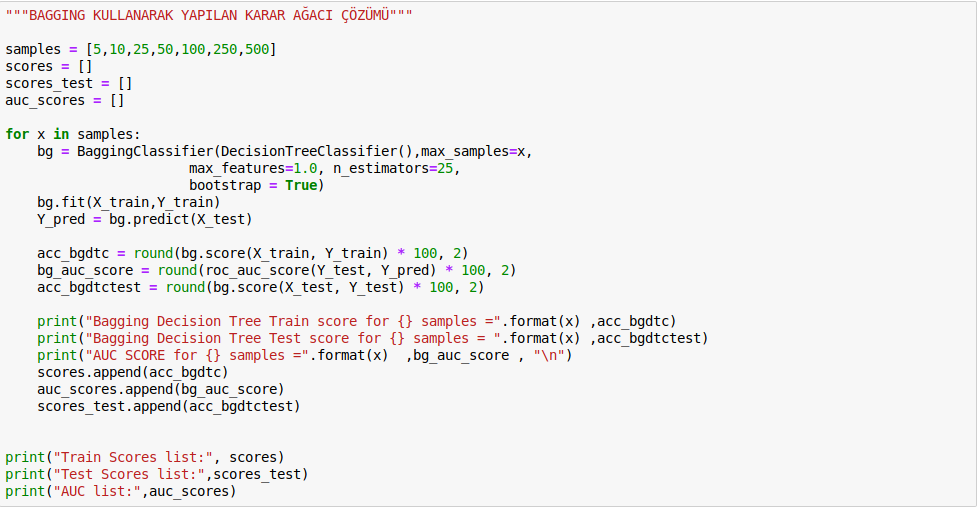


Torbalama(Bagging)

Birden çok makine öğrenmesi algoritmasının tahminlerini birleştirerek herhangi bir tekli modelden daha isabetli tahminler yapılmasını sağlayan bir topluluk yöntemidir.

Bootstrap Aggregation prosedürü varyansı yüksek olan algoritmalarının varyansının azaltılmasında kullanılabilir. Bu yüksek varyanslı algoritmalara örnek olarak CART (Sınıflandırma ve Regresyon Ağaçları) verilebilir.

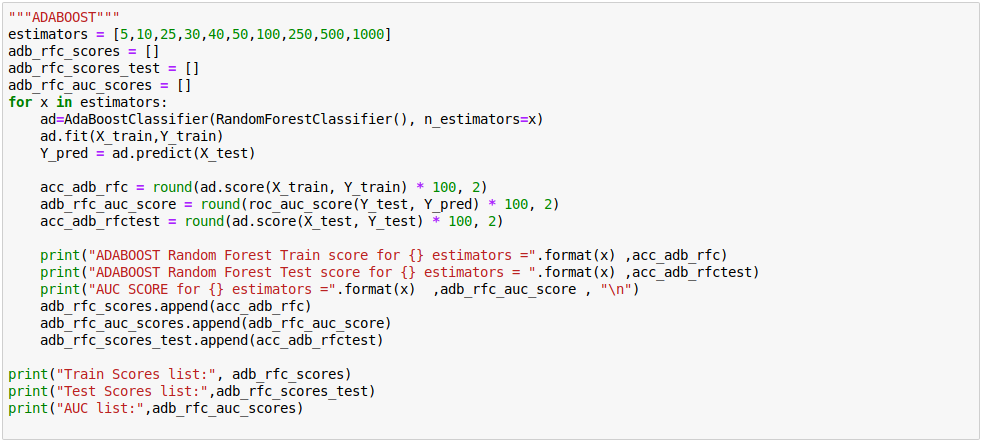
Modele uygulanması

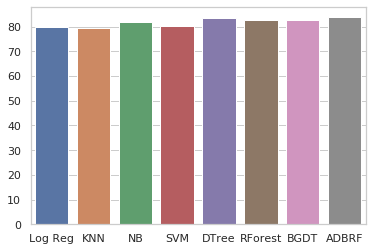


Adaboost

Adaboost sınıflandırıcısı zayıf olan sınıflandırıcı algoritmalarını birleştirerek güçlü bir sınıflandırıcı oluşturur. Tek bir algoritma sınıflandırmada hatalar yapabilir ama birden çok sınıflandırıcı bir araya getirilerek sınıflandırma yapılırsa bu hatalar azalır ve sınıflandırıcının isabeti artar.

Modelin uygulanması



Modellerin karşılaştırılması

