**Models**

**Supervised learning :** cevabı verilmiş bir veri setinde çalışılacaktır.

**Regression analysis :**

İki tane değişken vardır. Bağımlı ve bağımsız.

Bağımlı : Tahmin edilen y değişkeni target’tır.

Bağımsız : x değerleri vardır örnek veri setidir.

Regresyon probleminde sürekli aralıkta olan sonuçlar tahmin edilmektedir.

**Clasification analysis :**

Sınıfları tahmin etmek için oluşturduğumuz algoritmalardır

Binary : iki çeşit sınıftan oluşuyorsa y değerlerimiz bunları tahmin etmek için kullanılır.

Multi class : ikiden fazla sınıftan oluşuyorsa y değerlerimiz bunları tahmin etmek için kullanılır.

**Linear Regresion :**

Doğrusal bir çizgi üzerinde değişen değerlerin tahmini diyebiliriz.

Örnekte tek değişkenli bir linear regression modeli bulunmaktadır (xi).

X bağımsız değişkeni ile y bağımsız değişkeni arasındaki bağımlı bir değişken var mı bunu doğrusal çizgi üzerinde modelde kullanıyoruz.

**Multiple linear regression :**

Veri setinde birden fazla verilerin olduğu zaman yine doğru üzerinde prediction yapılmasıdır.

**Polinominal regression :**

Yüksek dereceden eşitsizlikler demektir.

X değeri artarken artık y değeri artmıyor. Sürekli x artarken y değeri bazı azalmalar gözlemleniyor.

İ değeri instance değerleridir.

**Logistic regression :**

Verdiğimiz instance’larde olasılığın hesaplanması için kullanılır.

Linear regression gibi ağırların toplamı üzerinden input feature ların üzerinden ilerliyor. Fakat burada kullanılan logistic mantığı da sigmoid fonksiyonuna denktir. 0 ve 1 arası değerler alır.

**KNN :**

Non parametric öğrenme algoritması.

Bir parametre vermeden algoritmanın bi classification modeli ile veri setinde kesişmesine bakılır.

K değeri verilen noktanın çevresindeki en yakın komşularının sayısıdır.

1. Uzaklık hesaplanır. Soru işaretinin en yakın uzaklıkları hesaplanır.
2. Yakın komşular bulunur.
3. Sınıf için oy verilir.

Şekilde İnitial data : p1 noktası seçiliyor (yeni bir örnek). Önce uzaklıkları hesaplanır. Komşular buluunur ve verilen oylara göre sınıf belirleniyor.

Öklit ile hesaplama : noktanın x ve y eksenindeki değerleri alınyıor.

**SVM :**

Biraz daha acc değeri yüksek olan bir algoritmadır.

Kernal trick sayesinde, acc değeri daha da çok yükseliyor.

Buradaki asıl amaç, noktalar üzerinde hyper plane yani bir space oluşturuyor.

Svm hem classification hem de regression için kullanılmaktadır.

Çoklu değişken ve kategorik verilerde kolaylıkla kullanılabilir.

Ana fikir : linear çizgideki margin i en fazla alarak iki kümeyi sınıflara ayırmaya çalışıyoruz.

İki nokta arasındaki yere margin deniliyor ve margin değerini yükseltmeye çalışıyoruz.

Classification yaparken linear’lerin sol ve sağına bakıyoruz.

Kernels : Bazı türlerde ise, her zaman linear olmayabilir, polinominal bir değer alabiliyor. Burada kullanılacak şey, kernel trip’dir. Z axis’i oluşturuluyor.

1. Linear : dot production yapılıyor.
2. Polinominal : ikinci veya üçüncü dereceden bir denklem kullanılabilir.
3. RBF (Radial Basis Function - Gaussian) : sigmoid fonksiyonuna benzer bir fonksiyon kullanarak sınıflandırma yapmış oluyor. Daha fazla kullanılan bir kerneldir. Gama parametresi iki nokta arasındaki benzerliğin oranını hesap eder. 0 ve 1 arasında değişmektedir. Daha yüksek gama değeri verilen verisetleri model daha verimli çalışmaktadır.

Overfit : veri setinin aşırı uyumlu çıkmasıdır.

Underfit : veri setinin aşırı uyumsuzluğudur. Hesaplamaların yetersiz kaldığı bir durumdur.

Regression : epsilon değeri kullanılmaktadır. Epsilon değeri ile margin değeri içerisine yakınlaştırmaya çalışıyoruz. Burada yakınlığa bakarak ACC değerini yüksek tutmaya çalışıyoruz. Epsilon uzaklık değeridir.

1. Linear
2. Polinominal

Classification için SVC kütüphanesi, regression için SVR kullanılmaktadır.

Standart scaler kullanılıyor linear bir düzen oluşması için

**Decision Trees – Classification :**

Ağaç modelleri en üst kısımda köktür (root). Derinlik 0 en yüksek seviyedir yani root katmanıdır. Sorgulanan değer küçükse sol tarafa, büyükse sağ tarafa aktarılıyor.

Sample : Bir node’un sample değeri instance’ların sayısını belirtiyor.

Value değeri : verilen instance’da kaç tanesi bu class’a etki etmektedir.

Gini : impurity, o etiketin yanlış olma şansıdır.

Entropy : infortmation gain. O değişimin en küçük kalma değeri, saflık değeri.

**CART :** classificaton ve regression tree: her ikisinide barındırır.

**Decision Trees - Regression :**

Sürekli olan veriler üzerinde çalışılmaktadır.

Cost function : eğitime başlarken karşımıza çıkan hataları olabildiğince minimize etmeye çalışırız.

Regression’da cost function olarak MSE kullanılır. Classification’da gini kullanılır.

**Ensemble Model :**

Birleştirilmiş model demektir. Birden fazla modeli birleştirmektir. Örneğin Logistic regression, svm, random forest ve diğerleri gibi. Hepsi birlikte prediction yapıyor ve sonuçları voting yani oylama yöntemiyle karşımıza çıkan en iyi sonucu görmüş olabiliyoruz. Hepsinin ortak sonucunu almaktadır.

**Random Forest :**

Birden fazla ağaç kullanılıyor ve farklı regression kullanılarak sonuçlara gidilebilir.

Bagging ve pasting : bootstrap agregration – veri setini sampling yöntemiyle sınıflandırıcalara ayırıyor, distüribe ediyor. Pasting ise tek bir sınıflandırıcı için çalışmaktadır.