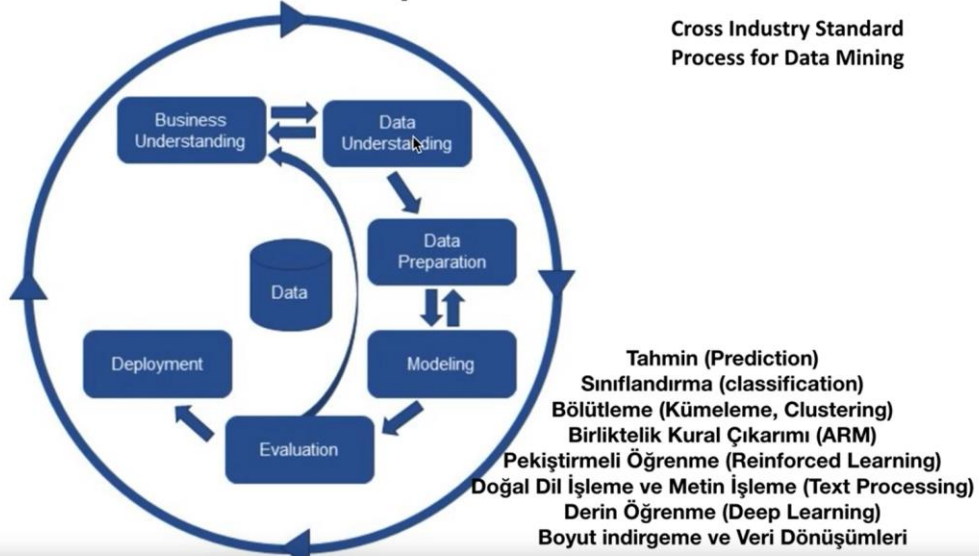


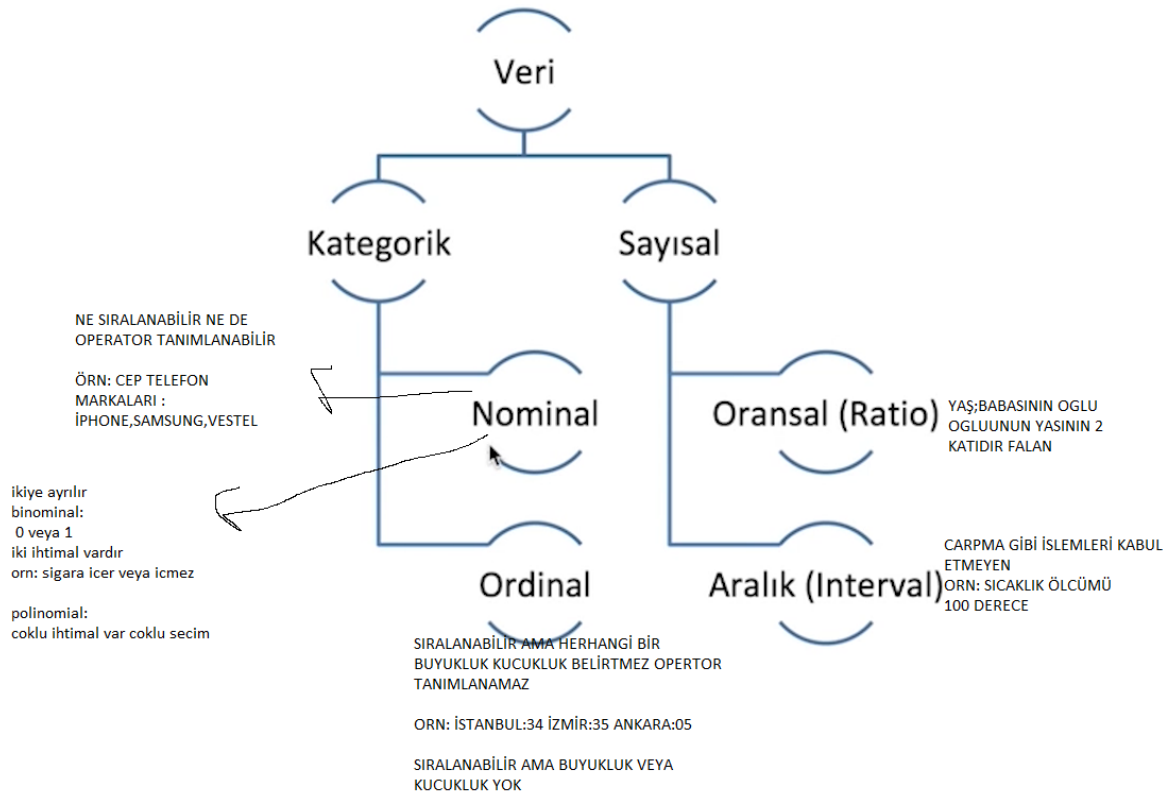
VERİ ÖNİŞLEME[DATA PREPROCESSING]

Metodoloji : CRISP-DM

Cross Industry Standard
Process for Data Mining



öğrenilen kısım : data preparation



A	B	C	D	E
ulke	boy	kilo	yas	cinsiyet
tr	130	30	10	e
tr	125	36	11	e
tr	135	34	10	k
tr	133	30	9	k
tr	129	38	12	e
tr	180	90	30	e
tr	190	80	25	e
tr	175	90	35	e
tr	177	60	22	k
us	185	105	33	e
us	165	55	27	k
us	155	50	44	k
us	160	58		k
us	162	59	41	k
us	167	62	55	k
fr	174	70	47	e
fr	193	90		e
fr	187	80	27	e
fr	183	88	28	e
fr	159	40	29	k
fr	164	66	32	k
fr	166	56	42	k
			28.45	

Buradaki ülke bilgilerini makinenin anlayacağı sayısal değerlere çevirmemiz gerekiyor; encode işlemi;

	TR	FR	US
TR	1	0	0
FR	0	1	0
US	0	0	1
TR	1	0	0
TR	1	0	0
FR	0	1	0
FR	0	1	0
US	0	0	1
US	0	0	1
TR	1	0	0
US	0	0	1
TR	1	0	0

Yukarıda yapılan işlem one-hot encoding (encode-> **kategorik verilerin sayısal**a dönüştürmek)

"Encode" işlemi, kategorik verilerin sayısal değerlere dönüştürülmesini sağlayan bir dönüşüm işlemidir. Örneğin, cinsiyet gibi bir kategorik değişkeni kodlamak için "One-Hot Encoding" veya "Label Encoding" gibi teknikler kullanılabilir.

- **One-Hot Encoding:** Kategorik değişkenlerin her bir farklı kategorisi için yeni bir sütun oluşturulur ve ilgili kategoriye ait olan gözlemler için bu sütunlar 1 ile kodlanırken, diğerleri 0 ile kodlanır.
- **Label Encoding:** Kategorik değişkenler, her bir kategoriye benzersiz bir sayısal değerle eşlenir. Her kategoriye farklı bir sayı atanır ve bu sayılarla değişken kodlanır.

```
ulke = veriler.iloc[:,0:1].values
print(ulke)

from sklearn import preprocessing # -> ön işleme

labelEnCode = preprocessing.LabelEncoder()

ulke[:,0]= labelEnCode.fit_transform(veriler.iloc[:,0]) #->fit_transform yöntemi, veriyi hem eğitir
# hem de dönüştürür bu nedenle genellikle
#eğitim verisi üzerinde kullanılır

print(ulke)

oneHotEnCode = preprocessing.OneHotEncoder()
ulke= oneHotEnCode.fit_transform(ulke).toarray()
print(ulke)
```

The figure consists of two side-by-side screenshots of the RStudio interface, showing data frames and their visualization.

Left Window: sonuc - Data frame

index	fr	tr	us	boy	kilo	yas	cinsiyet
0	0	1	0	130	30	10	e
1	0	1	0	125	36	11	e
2	0	1	0	135	34	10	k
3	0	1	0	133	30	9	k
4	0	1	0	129	38	12	e
5	0	1	0	180	90	30	e
6	0	1	0	190	80	25	e
7	0	1	0	175	90	35	e
8	0	1	0	177	60	22	k
9	0	0	1	185	105	33	e
10	0	0	1	165	55	27	k
11	0	0	1	155	50	44	k
12	0	0	1	160	58	28.45	k
13	0	0	1	162	59	41	k

Right Window: veriler - Data frame

index	ulke	boy	kilo	yas	cinsiyet
0	tr	130	30	10	e
1	tr	125	36	11	e
2	tr	135	34	10	k
3	tr	133	30	9	k
4	tr	129	38	12	e
5	tr	180	90	30	e
6	tr	190	80	25	e
7	tr	175	90	35	e
8	tr	177	60	22	k
9	us	185	105	33	e
10	us	165	55	27	k
11	us	155	50	44	k
12	us	160	58	nan	k

Verilerin test ve train olarak bolunmesi ;

```
from sklearn.model_selection import train_test_split

x_train, x_test, y_train, y_test = train_test_split(s,cinsiyetDF,test_size=0.33,random_state=0)
# x bağımsız y bağımlı degisken
# test_size ->%67 train %33 test
# hedef = bağımlı degisken
# x train ve x testleri kullanarak y[bağımlı degiskeni] tahmin ettircez
```

x_test - DataFrame

Index	fr	tr	us	boy	kilo	yas
20	1	0	0	164	66	32
10	0	0	1	165	55	27
14	0	0	1	167	62	55
13	0	0	1	162	59	41
1	0	1	0	125	36	11
21	1	0	0	166	56	42
11	0	0	1	155	50	44
19	1	0	0	159	40	29

y_test - DataFrame

Index	cinsiyet
20	k
10	k
14	k
13	k
1	e
21	k
11	k
19	k

x_train - DataFrame

Index	fr	tr	us	boy	kilo	yas
8	0	1	0	177	60	22
6	0	1	0	190	80	25
16	1	0	0	193	90	28.45
4	0	1	0	129	38	12
2	0	1	0	135	34	10
5	0	1	0	180	90	30
17	1	0	0	187	80	27
9	0	0	1	185	105	33
7	0	1	0	175	90	35
18	1	0	0	183	88	28
3	0	1	0	133	30	9
0	0	1	0	130	30	10
15	1	0	0	174	70	47
12	0	0	1	160	58	28.45

y_train - DataFrame

Index	cinsiyet
8	k
6	e
16	e
4	e
2	k
5	e
17	e
9	e
7	e
18	e
3	k
0	e
15	e
12	k

Evet, `train_test_split` fonksiyonunun genel kullanımı bu şekildedir. Bu fonksiyon, veri setini eğitim ve test setlerine bölmek için oldukça yaygın bir araçtır ve makine öğrenimi modelinin eğitilmesi ve değerlendirilmesi için veri bölme işlemini gerçekleştirmek için sıklıkla kullanılır.

Genel kullanım şu şekildedir:

python

Copy code

```
from sklearn.model_selection import train_test_split

x_train, x_test, y_train, y_test = train_test_split(s, cinsiyetDF, test_size=0.
```

Burada:

- `s`: Bağımsız değişkenleri (özellikleri) içeren veri setidir.
- `cinsiyetDF`: Bağımlı değişkeni (hedef değişkeni) içeren veri setidir.
- `test_size`: Veri setinin ne kadarının test seti olarak ayrılacağını belirten orandır. Genellikle 0 ile 1 arasında bir değerdir ve bu oran, test setinin toplam veri setine oranını ifade eder.
- `random_state`: Veri setinin rastgele bölünmesinde kullanılacak rastgele durumun belirlenmesini sağlayan bir değerdir. Bu, işlemi tekrarlandığında aynı bölünmenin elde edilmesini sağlar.

Bu satır, veri setini eğitim ve test setlerine böler ve sonuç olarak dört değişkene atanır: `x_train`, `x_test`, `y_train`, ve `y_test`. Bu değişkenler, sırasıyla eğitim setinin özellikleri, test setinin özellikleri, eğitim setinin hedefleri ve test setinin hedeflerini içerir. Bu şekilde, model eğitimi ve performans değerlendirmesi için gerekli veriler elde edilmiş olur.

Öznitelik ölçekleme; [verileri aynı dünyaya çevirmek birbirine yakın sayılar elde etmek][birbirine göre ölçeklenmiş oldu standart scaler kullanarak]

X_test - NumPy object array						
	0	1	2	3	4	5
0	1.29099	-0.377964	-1	0.4724	1.32854	-0.249913
1	-0.774597	-0.377964	1	0.549527	0.20439	-0.649773
2	-0.774597	-0.377964	1	0.70378	0.919757	1.58944
3	-0.774597	-0.377964	1	0.318147	0.613171	0.469836
4	-0.774597	2.64575	-1	-2.53554	-1.73732	-1.92932
5	1.29099	-0.377964	-1	0.626653	0.306586	0.549808
6	-0.774597	-0.377964	1	-0.221739	-0.306586	0.709752
7	1.29099	-0.377964	-1	0.0867674	-1.32854	-0.489829

X_train - NumPy object array						
	0	1	2	3	4	5
0	-0.632456	0.866025	-0.408248	0.450494	-0.296579	-0.247171
1	-0.632456	0.866025	-0.408248	1.00825	0.509655	0.0341619
2	1.58114	-1.1547	-0.408248	1.13696	0.912772	0.357695
3	-0.632456	0.866025	-0.408248	-1.60891	-1.18344	-1.18495
4	-0.632456	0.866025	-0.408248	-1.35148	-1.34468	-1.3725
5	-0.632456	0.866025	-0.408248	0.579207	0.912772	0.503051
6	1.58114	-1.1547	-0.408248	0.879537	0.509655	0.221717
7	-0.632456	-1.1547	2.44949	0.793728	1.51745	0.784384
8	-0.632456	0.866025	-0.408248	0.364686	0.912772	0.971939
9	1.58114	-1.1547	-0.408248	0.70792	0.832148	0.315495
10	-0.632456	0.866025	-0.408248	-1.43729	-1.50593	-1.46628
11	-0.632456	0.866025	-0.408248	-1.566	-1.50593	-1.3725
12	1.58114	-1.1547	-0.408248	0.321782	0.100538	2.09727
13	-0.632456	-1.1547	2.44949	-0.278876	-0.377202	0.357695

```
from sklearn.preprocessing import StandardScaler # -> verileri standartlaştırmaya yarar

sc=StandardScaler()

X_train = sc.fit_transform(x_train)
X_test = sc.fit_transform(x_test)
#Öznitelik ölçekleme; [verileri aynı dünyaya çevirmek birbirine yakın sayılar elde etmek]
#[birbirine göre ölçeklenmiş oldu]
# VERİLER STANDARTLAŞTIRILDI [StandardScaler]
```

Veri Önleme - ADIMLAR:

- 1- Veri yükleme: [pd.read_csv('eksikveriler.csv')]
- 2- Eksik veri tamamlama: [imputer= SimpleImputer(missing_values=np.nan, strategy='mean')]
- 3- Kategorik[nominal] veri donuşumu: [preprocessing.OneHotEncoder()]
- 4- Veri kümelerini birleştirme: [pd.concat([s,cinsiyetDF],axis=1)]
- 5- Verileri bölme: [train and test]
[x_train, x_test,y_train ,y_test =train_test_split(s,cinsiyetDF,test_size=0.33,random_state=0)]
- 6- Öznitelik ölçekleme: [StandardScaler().fit_transform(x_train)]



ChatGPT

Bu iki kod parçası arasındaki fark, "age" değişkeninin nasıl atanmış olduğudur.

İlk kod parçasında:

python

Copy code

```
age = data.iloc[:,3:4].values
```

Burada `iloc` metodu kullanılarak `data` veri çerçevesinin tüm satırlarını ve 3. sütununu (indeks numarasıyla 3) içeren bir alt veri çerçevesi seçilir. `.values` metodu bu alt veri çerçevesini bir NumPy dizisine dönüştürür. Bu şekilde `age` değişkeni 2 boyutlu bir NumPy dizisi olacaktır.

İkinci kod parçasında ise:

python

Copy code

```
age = data.iloc[:,3].values
```

Burada yine `iloc` metodu kullanılarak `data` veri çerçevesinin tüm satırlarını ve 3. sütununu seçeriz. Ancak bu sefer sadece bir sütun seçtiğimiz için çıktı 1 boyutlu bir NumPy dizisi olur.

Yani temel fark, ilk kod parçasında bir sütun seçilirken, ikinci kod parçasında bir sütun dilimleme işlemi yapılarak seçilmiştir.



One-hot encoding : ülke gibi herhangi bir sıralaması olmayan - cinsiyet

Label encoding : sıralaması olacak

"Regresyon" terimi, istatistik ve matematikte, bir deęişkenin dięer bir veya daha fazla deęişkenle ilişkisini modellemek için kullanılan bir tekniktir. Bu ilişkiyi açıklamak için kullanılan modeller, genellikle gözlemlenen verileri temsil eden matematiksel denklemlerdir. Regresyon analizi, deęişkenler arasındaki ilişkinin doğasını anlamak, tahmin yapmak ve sonuçları yorumlamak için kullanılır.

"Regresyon" terimi, İngiliz istatistikçi Francis Galton'un çalışmalarına dayanır. Galton, ebeveynlerin boy uzunluğunun çocukların boy uzunluğunu etkilediğini göstermek için çalışmıştır. Galton, çocukların boy uzunluğunun ebeveynlerin boy uzunluğunun ortalamasına "geri dönme eğilimi" gösterdiğini gözlemledi. Bu "geri dönme" veya "regresyon" terimi, Galton'un çalışmalarıyla popülerlik kazandı ve regresyon analizinin temelini oluşturdu.

Doğrusal regresyon, istatistiksel bir modelleme tekniğidir ve bir bağımlı deęişken ile bir veya daha fazla bağımsız deęişken arasındaki ilişkiyi açıklamak için kullanılır. Bu ilişkiyi bir doğru ile ifade eder. Genellikle, bir bağımsız deęişkenin bağımlı deęişken üzerindeki etkisini anlamak veya bir bağımlı deęişkenin tahminini yapmak için kullanılır.

Doğrusal regresyon modeli, veri setindeki gözlemleri en iyi şekilde temsil eden bir doğruyu bulmaya çalışır. Bu doğru, bağımsız deęişkenlerin katsayılarını (eğimleri) ve sabit bir terimi (kesme noktası) içerir. Model, bu katsayıları veri setine uygun olarak hesaplar.

Doğrusal regresyon, temelde iki tür veri arasındaki ilişkiyi anlamak için kullanılır. Örneğin, reklam harcamaları ile satışlar arasındaki ilişkiyi belirlemek veya bir kişinin yaşını ve kilosunu kullanarak sağlık durumunu tahmin etmek gibi durumlarda kullanılabilir.

Çoklu Doğrusal Regresyon (Multiple Linear Regression)

Basit Doğrusal Regresyon

$$y = \alpha + \beta x,$$

$$y_i = \alpha + \beta x_i + \varepsilon_i.$$

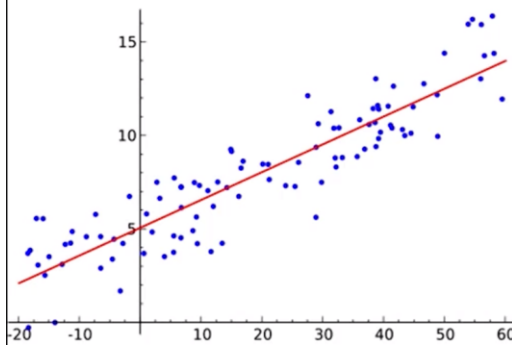
$$\text{Satış} = a + b (\text{Ay}) + e$$

Çoklu Doğrusal Regresyon

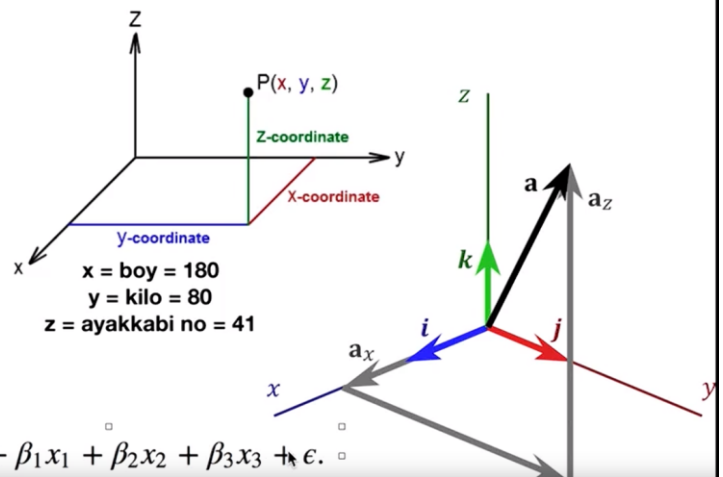
$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon.$$

$$\text{Boy} = a + b (\text{kilo}) + c (\text{yaş}) + d (\text{ayakkabı no}) + e$$

Çoklu Doğrusal Regresyon (Multiple Linear Regression)



$$y_i = \alpha + \beta x_i + \varepsilon_i.$$



$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon.$$

Kukla Değişken (Dummy Variable)

Kukla Değişken Tuzağı
Dummy Variable Trap

Boy	Kilo	Yaş	Cinsiyet
130	30	10	e
129	38	12	e
135	34	10	k
133	30	9	k
175	90	35	E

E	K
1	0
1	0
0	1
0	1
1	0

dummy değişken tuzağına dikkat edilmeli; burada cinsiyet kolonu ile E-K [encoding işlemi uygulanmış cinsiyet kolonu] aynı anda tabloda bulundurulursa öğrenme algoritmaları için **verimli olmayabilir**.

Aynı zamanda dummy değişkenin kendi içinde de bazı elemeler yapabilir; buradaki örnekte dummy değişken binominal [iki ihtimalden biri] olduğu için sadece bir kolonu almamız yeterli olacaktır [kadın değilse erkektir]

Fakat bazı durumlarda dummy variable ülke gibi değişkenleri içerisinde tutacaksa 2den fazla ülke olabilir : örn[tr,us,fr] o zaman kukla verinin tamamını tabloya eklemeliyiz.

p-value (olasılık değeri)

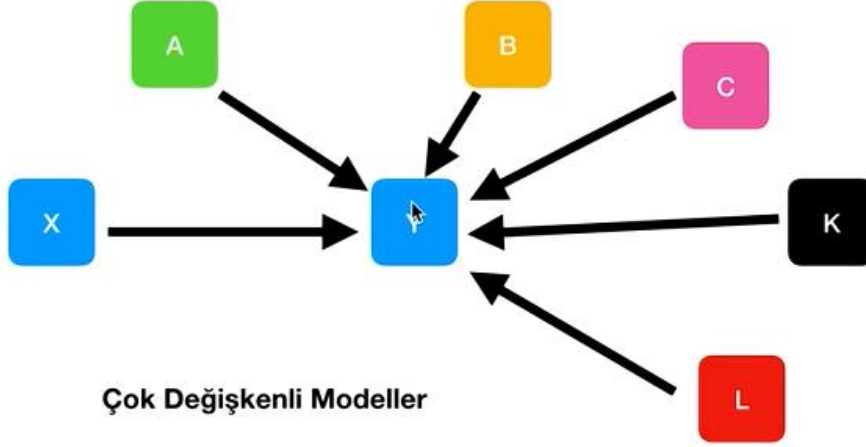
- H0: null hypothesis : Farksızlık hipotezi, sıfır hipotezi, boş hipotez
- H1: Alternatif hipotez
- p-değeri: olasılık değeri (genelde 0.05 alınır)
- P-değeri küçüldükçe H0 hatalı olma ihtimali artar
- P-değeri büyüldükçe H1 hatalı olma ihtimali artar

H0:null hypothesis: her kutuda 100 tane kurabiye vardır

H1:alternatif: her kutuda 100 tane kurabiye yoktur

p-değeri : genelde 0.05 bu örnekte de 0.05 alındığında 5 kutuda 100 tane kurabiye yoksa hipotezin yanlış olma olasılığı artar. H1'in doğru olma olasılığı artar

Değişken Seçimi



Burada önemli olan hangi bağımsız değişken bağımlı değişkeni daha fazla, daha az veya hiç etkilemiyor bunun analizi.

Hangi bağımsız değişkenleri seçmeliyiz ?

Farklı Yaklaşımlar

- Bütün Değişkenleri Dahil Etmek
- Geriye doğru eleme (Backward Elimination)
- İleri seçim (Forward Selection)
- İki Yönlü eleme (bidirectional elimination)
- Skor Karşılaştırması (Score Comparison)



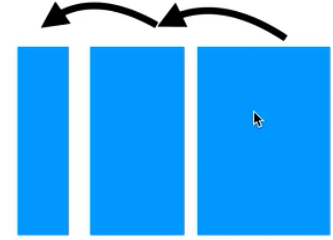
Adım adım karşılaştırma
Stepwise

Bütün Değişkenler

- Şayet değişken seçimi (selection) yapıldıysa ve değişkenlerden eminsek
- Zorunluluk varsa (örn. Bankadaki kredi skorları için geliştirilen modelin başarısının ölçülmesi)
- Keşif için (diğer 4 yöntemi kullanmadan önce bir ön fikir elde etmek için)

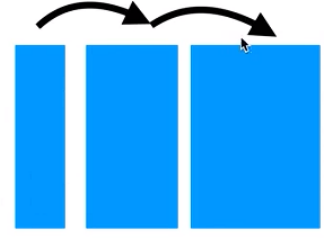
Geriye Eleme (Backward Elimination)

1. Significance Level (SL) seçilir (genelde 0.05)
2. Bütün değişkenler kullanılarak bir model inşa edilir.
3. En yüksek p-value değerine sahip olan değişken ele alınır ve şayet $P > SL$ ise 4. adıma, değilse son adıma (6. adım) gidilir
4. Bu aşamada, 3. adımda seçilen ve en yüksek p-değerine sahip değişken sistemden kaldırılır
5. Makine öğrenmesi güncellenir ve 3. adıma geri dönülür.
6. Makine öğrenmesi sonlandırılır.



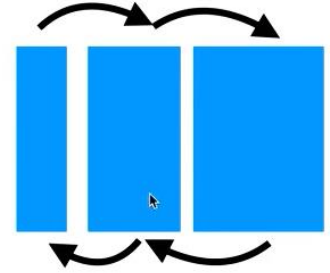
[literatürde sıkça kullanılır] ilk bütün değişkenler alınır eleye eleye gidilir

İleriye Seçim (Forward Selection)



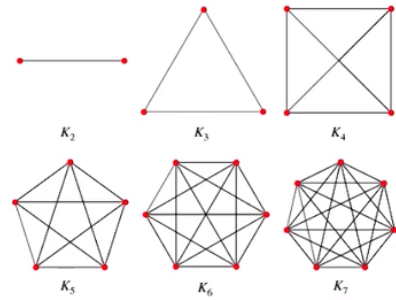
1. Significance Level (SL) seçilir (genelde 0.05)
2. Bütün değişkenler kullanılarak bir model inşa edilir.
3. En **düşük** p-value değerine sahip olan değişken ele alınır
4. Bu aşamada, 3. adımda seçilen değişken sabit tutularak yeni bir değişken daha seçilir ve sisteme eklenir
5. Makine öğrenmesi güncellenir ve 3. adıma geri dönülür, şayet en düşük p-değere sahip değişken için $p < SL$ şartı sağlanıyorsa 3. Adıma dönülür. Sağlanmıyorsa biter (6. Adıma geçilir)
6. Makine öğrenmesi sonlandırılır.

Çift Yönlü Eleme (Bidirectional Elimination)



1. Significance Level (SL) seçilir (genelde 0.05)
2. Bütün değişkenler kullanılarak bir model inşa edilir.
3. En **düşük** p-value değerine sahip olan değişken ele alınır
4. Bu aşamada, 3. adımda seçilen değişken sabit tutularak diğer bütün değişkenler sisteme dahil edilir ve en düşük p değerine sahip olan sistem de kalır
5. SL değerinin altında olan değişkenler sistemde kalır ve eski değişkenlerden hiçbirisi sistemden çıkarılamaz.
6. Makine öğrenmesi sonlandırılır.

Bütün Yöntemler



1. Başarı kriteri belirlenir.
2. Bütün olası regresyon modelleri inşa edilir (ikili seçim olur)
3. Başta belirlenen kriteri (1. adım) en iyi sağlayan yöntem seçilir
4. Makine öğrenmesi sonlandırılır.

Geriye eleme(backward elimination)

```
# Geri Eleme (Backward Elimination)

import numpy as np
import statsmodels.api as sm # stat-> istatistik

X = np.append(arr=np.ones((22,1)).astype(int), values=data, axis=1) # axis = 1 -> kolon olarak eklemek
# burada birlerden oluşan bir kolon ekliyoruz, denklemdeki çarpan değeri 1[etkisiz eleman]
# y=b0 +b1x1 +b2x2.....+e. aslında b0'i ekliyoruz

X_list = data.iloc[:,[0,1,2,3,4,5]].values # list olarak tanımlıyoruz ki
# analiz yapıldığında hangisinin p değeri yüksek görelim ve onu çıkaralım
X_list = np.array(X_list, dtype=float) # np arraya dönüştürmek

model = sm.OLS(genderDF.iloc[:,0:1],X_list).fit()
print(model.summary())
# bazı p değerleri 0.05'in üstünde çıktı öncelikle en büyük cıkanı çıkartıp tekrar deneme yanılma
# yöntemiyle devam edilebilir, p değerleri dustukce sonuca yaklasiyo sayilabiliriz ;

# X_list = data.iloc[:,[3,4,]].values gibi
```

	coef	std err	t	P> t	[0.025	0.975]
x1	2.2338	1.174	1.903	0.075	-0.254	4.722
x2	2.2461	1.075	2.089	0.053	-0.033	4.525
x3	1.8514	1.122	1.651	0.118	-0.526	4.229
x4	-0.0204	0.010	-2.098	0.052	-0.041	0.000
x5	0.0308	0.008	3.682	0.002	0.013	0.048
x6	-0.0077	0.010	-0.813	0.428	-0.028	0.012

Omnibus: 0.140 Durbin-Watson: 1.516
Prob(Omnibus): 0.932 Jarque-Bera (JB): 0.212
Skew: -0.158 Prob(JB): 0.899
Kurtosis: 2.637 Cond. No. 4.53e+03

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified
[2] The condition number is large, 4.53e+03. This might indicate that there are strong multicollinearity or other numerical problems.

In [55]:

POLYNOMİAL REGRESSION

Polinomal Regresyon (Polynomial Regression)

Çoklu Doğrusal Regresyon

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon.$$

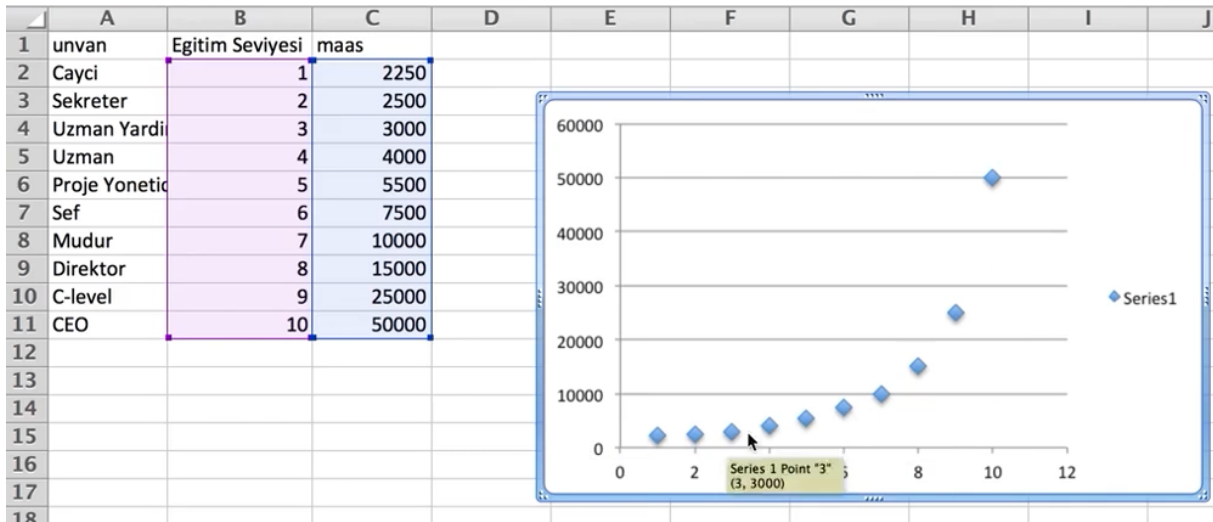
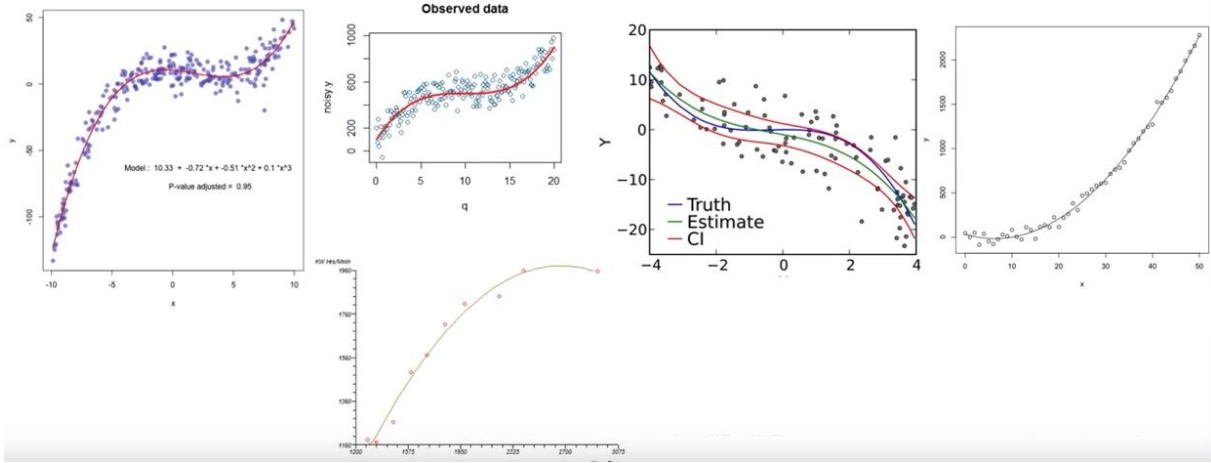
Polinomal Regresyon

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \dots + \beta_h X^h + \epsilon,$$

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \beta_{12} x_1 x_2 + \epsilon$$

$$\text{Boy} = a + b (\text{kilo}) + c (\text{yaş}) + d (\text{ayakkabı no}) + e$$

Polinomal Regresyon (Polynomial Regression)



```
# linear regression
lin_reg = LinearRegression().fit(X,Y)

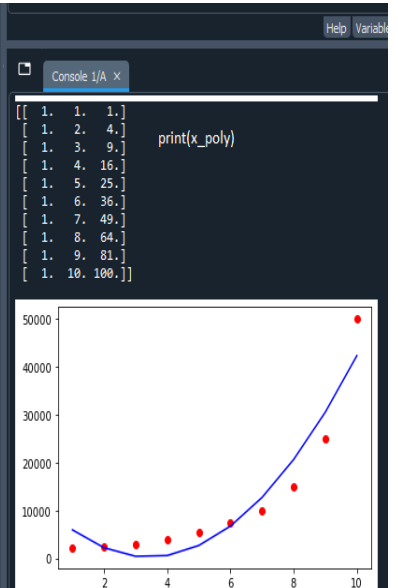
plt.scatter(X,Y, color='red')
plt.plot(X,lin_reg.predict(X),color='blue')
plt.show()

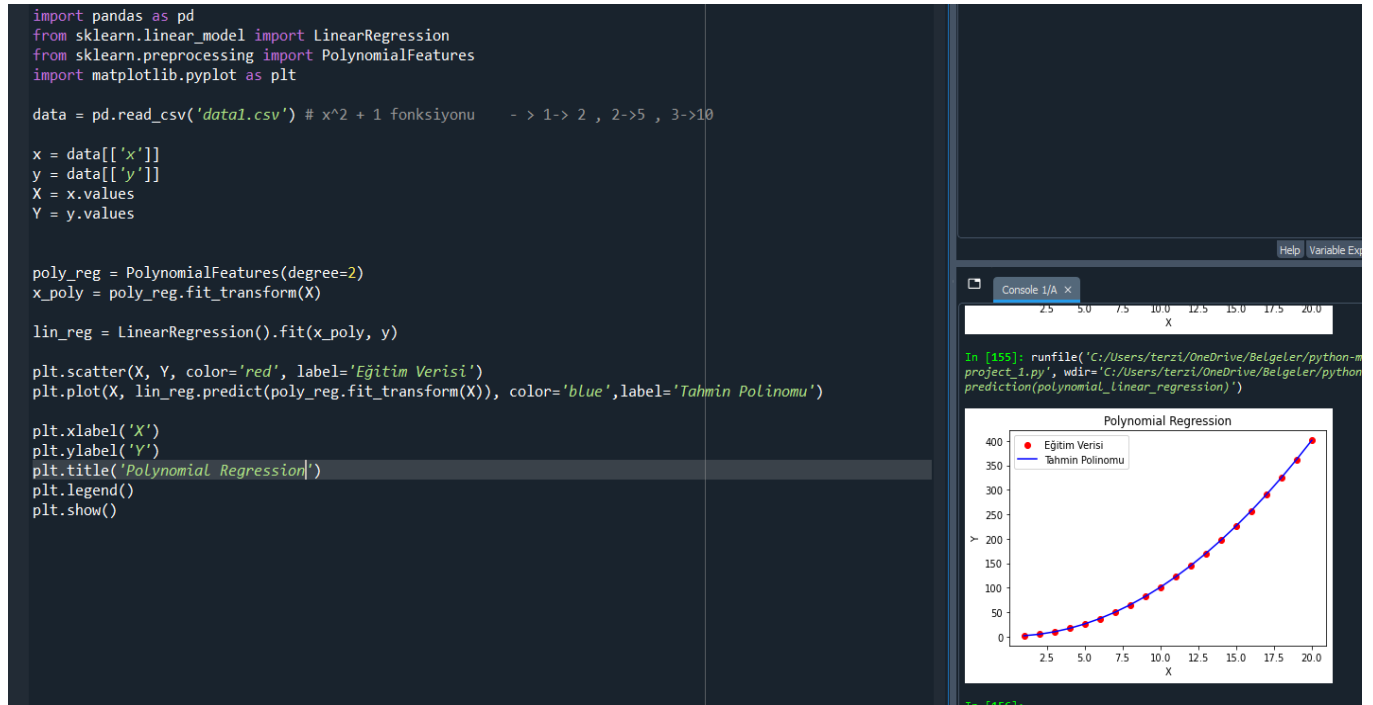
# polynomial regression
from sklearn.preprocessing import PolynomialFeatures

poly_reg = PolynomialFeatures(degree = 2) #2. dereceden polynomial obje olustur
x_poly = poly_reg.fit_transform(X) # X degerini polinomal dunyaya cevirdi
print(x_poly)

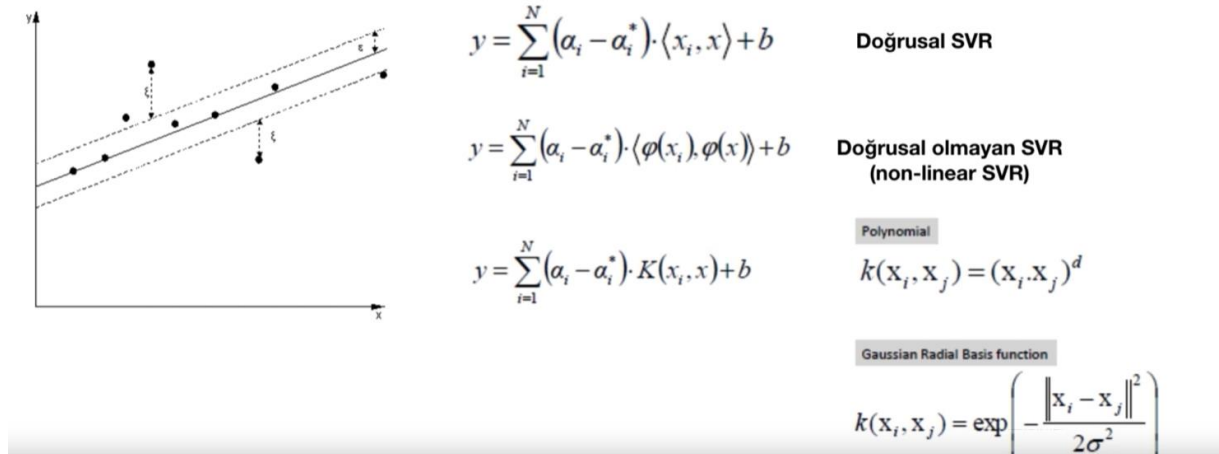
lin_reg2 = lin_reg = LinearRegression().fit(x_poly,Y) # 3 tane kolonun carpanlarini ogreniyi
#x^0 x^1 x^2

plt.scatter(X,Y, color='red')
plt.plot(X,lin_reg2.predict(poly_reg.fit_transform(X)),color='blue')
```





Destek Vektör Regresyonu (Support Vector Regression)

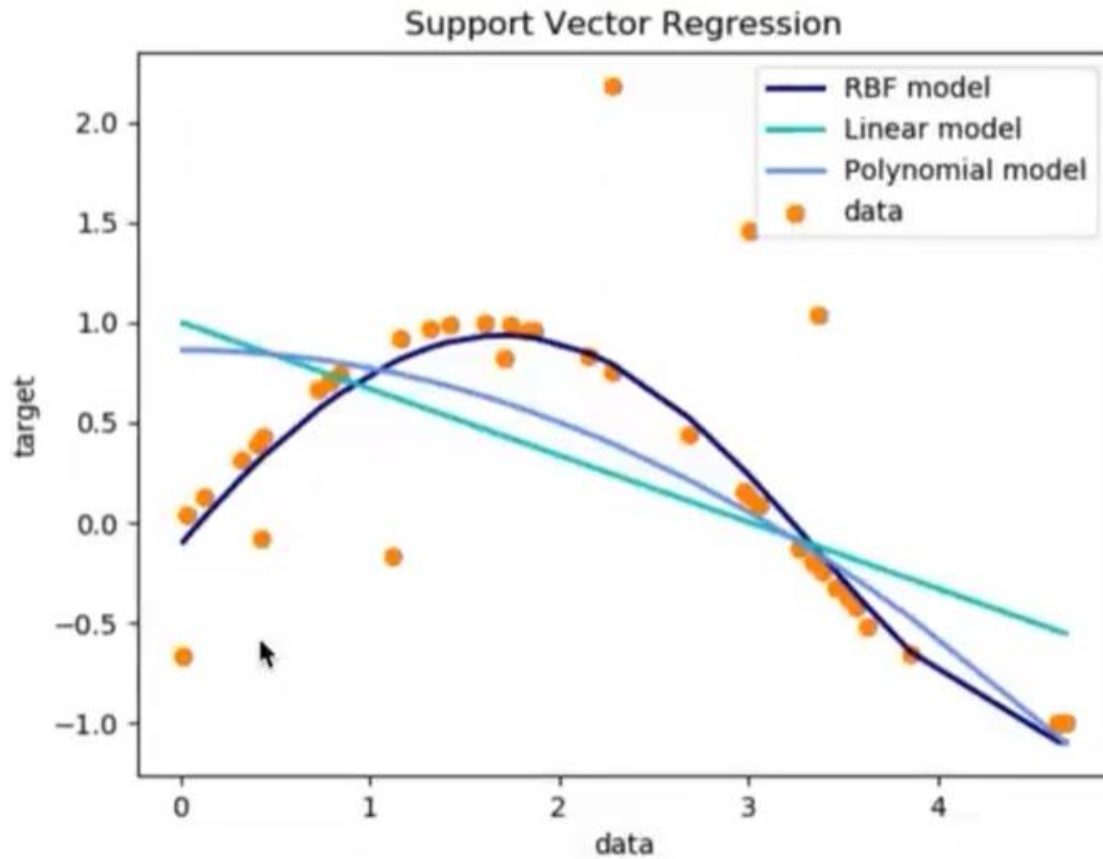


Support vector regression margin değerlerini tanımlıyor ve bu margin değerlerine giren maksimum noktayı elde edebileceği minimum margin değerine sahip fonksiyonu almayı amaçlıyor

Birden fazla doğru çizilebiliyorsa minimum margin değerine sahip aynı noktaları içine alabilecek değeri elde etmeye çalışıyor

Farklı fonksiyonlar kullanılabilir;

Doğrusal , doğrusal olmayan[polynomial, rbf(radial basis function)] , exp



```
svr_rbf = SVR(kernel='rbf', C=1e3, gamma=0.1)
svr_lin = SVR(kernel='linear', C=1e3)
svr_poly = SVR(kernel='poly', C=1e3, degree=2)
```

```
# support vector regression scaler ile kullanmamız gerekli veriler üzerindeki marjinal verilere duyarlı

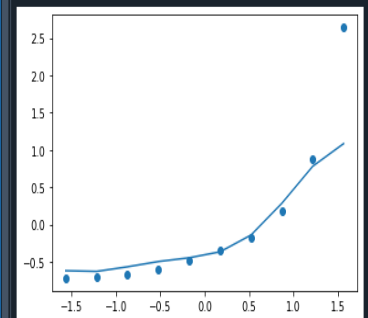
x_sc = StandardScaler().fit_transform(x)
y_sc = StandardScaler().fit_transform(y)
plt.show()

# SVR      svm -> support vector machine
from sklearn.svm import SVR

svr_reg = SVR(kernel="rbf")
svr_reg.fit(x_sc, y_sc)

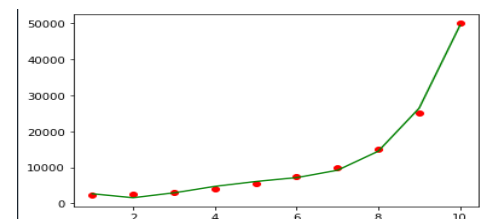
plt.scatter(x_sc, y_sc)
plt.plot(x_sc, svr_reg.predict(x_sc))
plt.show()
```

was passed when a 1d array was expected. Please change the
y = column_or_id(y, warn=True)



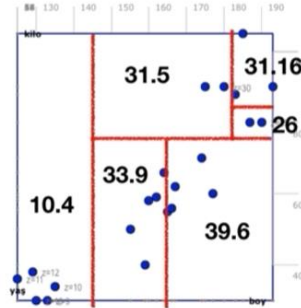
In [31]:

(rbf)



(polynomial regression)

Karar Ağacı ile Tahmin (Decision Tree)



örnek, boy = 180, kilo = 80

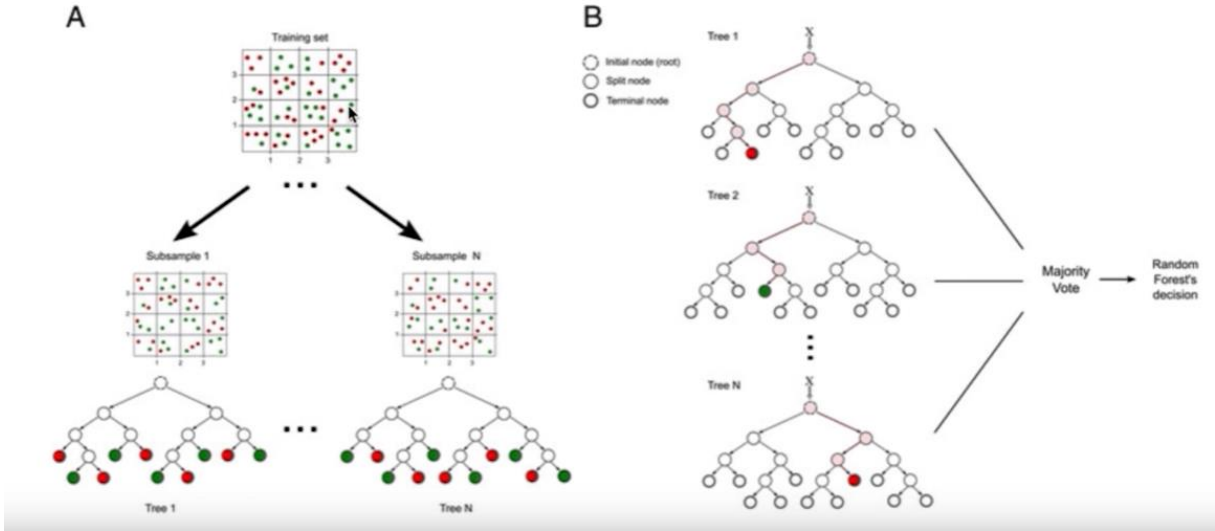
Karar ağaçları genelde sınıflandırma için kullanılır fakat tahmin için de kullanılabilir.

Hassaslığı arttırmak için karar ağacını daha fazla dallandırmalıyız



5.6 – 6 – 6.5 degerlerini aynı değerde tahmin etti

Rassal Orman



Rassal ağaç (random forest) : birden fazla decision tree'nin aynı problem aynı veri kümesi üzerinde çözilmesi ve daha sonra problemin çözümünde hep birlikte kullanılmasına dayanıyor.

Olukça güçlü bir yöntemdir

Bu duruma *Majority voted learning* - > çoğunluğun oyu ismi verilir

Veri kümesini küçük parçalara bölüyor birden fazla decision tree'yi öğreniyor

Bu şekilde birden fazla tahmin algoritması geliyor ve çoğunluğun oyuna(majority voted) göre *sınıflandırma* yapılıyor

Tahminde ise verilen kararların (majority voted) *ortalaması* sonuç oluyor

Karar ağaçlarında verinin artması durumunda başarının düşmesi ile ilgili bir sonuç var

Bazı durumlarda verinin artması sonucunda karar ağacına etkisi ;

- Sonuçların yanlış çıkması
- Overfitting (aşırı öğrenme) -> ezberlemeye gidebilir
- Ağacın çok fazla dallanması durumunda hesaplama zamanı sıkıntısı çıkartabilir



Ensemble Learning (Kollektif Öğrenme)

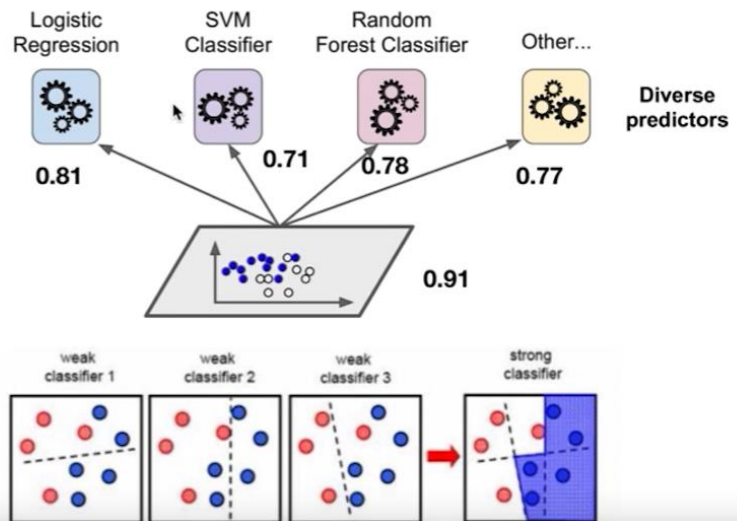
Birden fazla sınıflandırma veya birden fazla tahmin algoritması aynı anda kullanılarak daha başarılı bir sonuç çıkartılabilir

Buna [Ensemble Learning \(kollektif öğrenme\)](#) denir

Random forest ile decision tree'nin aynı anda kullanılması ensemble learning'e örnektir

Ensemble Learning Kollektif Öğrenme

- Boosting
- Bagging
 - Random Forest
- AdaBoost
- Stacking
- Blending
- MAVL



Tahmin Algoritmalarının Değerlendirilmesi (Evaluation of Predictions):

R2 Hesaplama;

- R2 (R-Square , R-Kare) Yöntemi

$$\text{TOPLA (gerçek - tahmin)}^2$$

$$\text{Hata Karaleri Toplamı} = \text{Topla}(y_i - y'_i)^2$$

$$\text{TOPLA (gerçek - tahminORT)}^2$$

$$\text{Ortalama Farkların Toplamı} = \text{Topla}(y_i - y_{\text{ort}})^2$$

$$R^2 = 1 - \frac{\text{HKT}}{\text{OFT}}$$

OFT = Olabilecek En Kötü Tahmin

R2 değeri 1'e yaklaştıkça iyi sonuç 0'a yaklaştıkça kötü sonuç negatif değerler çok kötü sonuç

boy	kilo	yas	tahmin	başarı	= (C2-D2)*(C2-D2)	= (C2-D\$24)*(C2-D\$24)
130	30	10	12	10	4	355.836777
125	36	11	11	11	0	319.109504
135	34	10	10	10	0	355.836777
133	30	9	10	9	1	394.56405
129	38	12	11	12	1	284.382231
180	90	30	35	30	25	1.29132231
190	80	25	26	25	1	14.927686
175	90	35	47	35	144	37.6549587
177	60	22	22	22	0	47.1095041
185	105	33	33	33	0	17.1095041
165	55	27	27	27	0	3.4731405
155	50	44	44	44	0	229.109504
160	58	28.5	27	28.5	2.25	0.1322314
162	59	41	40	41	1	147.291322
167	62	55	50	55	25	683.109504
174	70	47	47	47	0	328.927686
193	90	28.5	29	28.5	0.25	0.1322314
187	80	27	22	27	25	3.4731405
183	88	28	33	28	25	0.74586777
159	40	29	22	29	49	0.01859504
164	66	32	33	32	1	9.83677686
166	56	42	44	42	4	172.56405
			28.8636364		308.5	3406.63636



$$308.5 \quad 3406.63636 \quad =1-F24/G24$$

Düzeltilmiş R² (Adjusted R²) Yöntemi:

R² üzerinden bir model değerlendirilirken eklenen yeni değişken sisteme asla *negatif* etki yapmaz, R² değerini *asla düşürmez* bu sebepten dolayı R² bazı durumlarda *yetersiz ve yanıltıcı* kalabilir

Yöntemleri Karşılaştırmak

- R² (R-Square , R-Kare) Yöntemi
- Düzeltilmiş R² (Adjusted R²) Yöntemi

Doğrusal Regresyon

$$y = a_0x_0 + a_1x_1 + b$$
$$y = a_0x_0 + a_1x_1 + a_2x_2 + b$$

n = eleman sayısı
p = değişken sayısı

$$\text{Hata Karaleri Toplamı} = \text{Topla}(y_i - y'_i)^2$$

$$\text{Ortalama Farkların Toplamı} = \text{Topla}(y_i - y_{\text{ort}})^2$$

$$R^2 = 1 - \frac{\text{HKT}}{\text{OFT}}$$

$$\text{Düzeltilmiş } R^2 = 1 - (1 - R^2) \frac{n - 1}{n - p - 1}$$

```
>>> from sklearn.metrics import r2_score
>>> y_true = [3, -0.5, 2, 7]
>>> y_pred = [2.5, 0.0, 2, 8]
>>> r2_score(y_true, y_pred)
0.948...
```

-Desicion tree regresyon modelinde R² SCORE kullanmak yanıltıcı olabilir.