```
Today Agenda
- DecisionTree Regressor
- Random Forest
```

DecisionTree Regressor

In [1]:

```
# 1. Read the Data
import pandas as pd
companies_data = pd.read_csv("https://raw.githubusercontent.com/AP-State-Skill-Developr
                             Datasets/master/Regression/1000_Companies.csv")
companies_data.head()
```

Out[1]:

	R&D Spend	Administration	Marketing Spend	State	Profit
0	165349.20	136897.80	471784.10	New York	192261.83
1	162597.70	151377.59	443898.53	California	191792.06
2	153441.51	101145.55	407934.54	Florida	191050.39
3	144372.41	118671.85	383199.62	New York	182901.99
4	142107.34	91391.77	366168.42	Florida	166187.94

In [2]:

```
# 2. Check the any null or process the data
companies_data.isnull().sum()
```

Out[2]:

```
R&D Spend          0
Administration     0
Marketing Spend    0
State              0
Profit             0
dtype: int64
```

In [6]:

```
companies_data.dtypes
```

Out[6]:

```
R&D Spend          float64
Administration     float64
Marketing Spend    float64
State               object
Profit             float64
dtype: object
```

In [7]:

```
companies_data['State'].unique()
```

Out[7]:

```
array(['New York', 'California', 'Florida'], dtype=object)
```

In [8]:

```
companies_data.shape
```

Out[8]:

```
(1000, 5)
```

In [9]:

```
state1 = pd.get_dummies(companies_data['State'])
state1
```

...

In [10]:

```
from sklearn.preprocessing import LabelEncoder
lab = LabelEncoder()
state2 = lab.fit_transform(companies_data['State'])
print(state2)
```

...

In [11]:

```
companies_data['State_tran']=state2
companies_data
```

Out[11]:

	R&D Spend	Administration	Marketing Spend	State	Profit	State_tran
0	165349.20	136897.8000	471784.10000	New York	192261.83000	2
1	162597.70	151377.5900	443898.53000	California	191792.06000	0
2	153441.51	101145.5500	407934.54000	Florida	191050.39000	1
3	144372.41	118671.8500	383199.62000	New York	182901.99000	2
4	142107.34	91391.7700	366168.42000	Florida	166187.94000	1
5	131876.90	99814.7100	362861.36000	New York	156991.12000	2
6	134615.46	147198.8700	127716.82000	California	156122.51000	0
7	130298.13	145530.0600	323876.68000	Florida	155752.60000	1
8	120542.52	148718.9500	311613.29000	New York	152211.77000	2
9	123334.88	108679.1700	304981.62000	California	149759.96000	0
10	101913.08	110594.1100	229160.95000	Florida	146121.95000	1
11	100671.96	91790.6100	249744.55000	California	144259.40000	0
12	93863.75	127320.3800	249839.44000	Florida	141585.52000	1
13	91992.39	135495.0700	252664.93000	California	134307.35000	0
14	119943.24	156547.4200	256512.92000	Florida	132602.65000	1
15	114523.61	122616.8400	261776.23000	New York	129917.04000	2
16	78013.11	121597.5500	264346.06000	California	126992.93000	0
17	94657.16	145077.5800	282574.31000	New York	125370.37000	2
18	91749.16	114175.7900	294919.57000	Florida	124266.90000	1
19	86419.70	153514.1100	0.00000	New York	122776.86000	2
20	76253.86	113867.3000	298664.47000	California	118474.03000	0
21	78389.47	153773.4300	299737.29000	New York	111313.02000	2
22	73994.56	122782.7500	303319.26000	Florida	110352.25000	1
23	67532.53	105751.0300	304768.73000	Florida	108733.99000	1
24	77044.01	99281.3400	140574.81000	New York	108552.04000	2
25	64664.71	139553.1600	137962.62000	California	107404.34000	0
26	75328.87	144135.9800	134050.07000	Florida	105733.54000	1
27	72107.60	127864.5500	353183.81000	New York	105008.31000	2
28	66051.52	182645.5600	118148.20000	Florida	103282.38000	1
29	65605.48	153032.0600	107138.38000	New York	101004.64000	2
...	...	...	...	...	...	...
970	13856.00	112503.4128	95514.22902	Florida	60869.96038	1
971	71829.00	121065.1295	207373.29080	New York	110395.79400	2
972	131154.00	129826.5157	321841.04030	Florida	161076.62960	1

	R&D Spend	Administration	Marketing Spend	State	Profit	State_tran
973	68679.00	120599.9232	201295.35720	New York	107704.77620	2
974	108056.00	126415.2979	277273.38630	California	141344.20750	0
975	140149.00	131154.9383	339196.91740	Florida	168760.98050	1
976	56850.00	118852.9626	178471.26940	California	97599.36358	0
977	47438.00	117462.9555	160310.78970	New York	89558.77320	2
978	58867.00	119150.8423	182363.07640	Florida	99322.46927	1
979	12914.00	112364.2939	93696.63744	California	60065.21791	0
980	62574.00	119698.3089	189515.74310	New York	102489.32740	2
981	53106.00	118300.0316	171247.21120	California	94400.89669	0
982	123537.00	128701.6025	307144.01800	California	154569.49220	0
983	48901.00	117679.0180	163133.65220	Florida	90808.60147	1
984	105143.00	125985.0928	271652.74480	California	138855.65680	0
985	63615.00	119852.0486	191524.35540	New York	103378.64470	2
986	100405.00	125285.3634	262510.76090	California	134808.02420	0
987	41289.00	116554.8432	148446.27740	New York	84305.73556	2
988	39970.00	116360.0473	145901.26330	Florida	83178.92524	1
989	43532.00	116886.0996	152774.15210	Florida	86221.91110	1
990	136133.00	130561.8371	331448.03440	California	165330.14630	0
991	131106.00	129819.4269	321748.42420	New York	161035.62360	2
992	105127.00	125982.7298	271621.87280	Florida	138841.98810	1
993	46798.00	117368.4374	159075.90800	California	89012.02672	0
994	97209.00	124813.3635	256344.07010	New York	132077.70900	2
995	54135.00	118451.9990	173232.66950	California	95279.96251	0
996	134970.00	130390.0800	329204.02280	California	164336.60550	0
997	100275.47	241926.3100	227142.82000	California	413956.48000	0
998	128456.23	321652.1400	281692.32000	California	333962.19000	0
999	161181.72	270939.8600	295442.17000	New York	476485.43000	2

1000 rows × 6 columns

In [12]:

```
# seperate the input and output labels
x = companies_data[['R&D Spend','Administration','Marketing Spend','State_tran']]
x
```

Out[12]:

	R&D Spend	Administration	Marketing Spend	State_tran
0	165349.20	136897.8000	471784.10000	2
1	162597.70	151377.5900	443898.53000	0
2	153441.51	101145.5500	407934.54000	1
3	144372.41	118671.8500	383199.62000	2
4	142107.34	91391.7700	366168.42000	1
5	131876.90	99814.7100	362861.36000	2
6	134615.46	147198.8700	127716.82000	0
7	130298.13	145530.0600	323876.68000	1
8	120542.52	148718.9500	311613.29000	2
9	123334.88	108679.1700	304981.62000	0
10	101913.08	110594.1100	229160.95000	1
11	100671.96	91790.6100	249744.55000	0
12	93863.75	127320.3800	249839.44000	1
13	91992.39	135495.0700	252664.93000	0
14	119943.24	156547.4200	256512.92000	1
15	114523.61	122616.8400	261776.23000	2
16	78013.11	121597.5500	264346.06000	0
17	94657.16	145077.5800	282574.31000	2
18	91749.16	114175.7900	294919.57000	1
19	86419.70	153514.1100	0.00000	2
20	76253.86	113867.3000	298664.47000	0
21	78389.47	153773.4300	299737.29000	2
22	73994.56	122782.7500	303319.26000	1
23	67532.53	105751.0300	304768.73000	1
24	77044.01	99281.3400	140574.81000	2
25	64664.71	139553.1600	137962.62000	0
26	75328.87	144135.9800	134050.07000	1
27	72107.60	127864.5500	353183.81000	2
28	66051.52	182645.5600	118148.20000	1
29	65605.48	153032.0600	107138.38000	2
...	...	...	...	...
970	13856.00	112503.4128	95514.22902	1
971	71829.00	121065.1295	207373.29080	2

	R&D Spend	Administration	Marketing Spend	State_tran
972	131154.00	129826.5157	321841.04030	1
973	68679.00	120599.9232	201295.35720	2
974	108056.00	126415.2979	277273.38630	0
975	140149.00	131154.9383	339196.91740	1
976	56850.00	118852.9626	178471.26940	0
977	47438.00	117462.9555	160310.78970	2
978	58867.00	119150.8423	182363.07640	1
979	12914.00	112364.2939	93696.63744	0
980	62574.00	119698.3089	189515.74310	2
981	53106.00	118300.0316	171247.21120	0
982	123537.00	128701.6025	307144.01800	0
983	48901.00	117679.0180	163133.65220	1
984	105143.00	125985.0928	271652.74480	0
985	63615.00	119852.0486	191524.35540	2
986	100405.00	125285.3634	262510.76090	0
987	41289.00	116554.8432	148446.27740	2
988	39970.00	116360.0473	145901.26330	1
989	43532.00	116886.0996	152774.15210	1
990	136133.00	130561.8371	331448.03440	0
991	131106.00	129819.4269	321748.42420	2
992	105127.00	125982.7298	271621.87280	1
993	46798.00	117368.4374	159075.90800	0
994	97209.00	124813.3635	256344.07010	2
995	54135.00	118451.9990	173232.66950	0
996	134970.00	130390.0800	329204.02280	0
997	100275.47	241926.3100	227142.82000	0
998	128456.23	321652.1400	281692.32000	0
999	161181.72	270939.8600	295442.17000	2

1000 rows × 4 columns

In [13]:

```
y = companies_data['Profit']
y
```

...

In [14]:

```
from sklearn.model_selection import train_test_split
x_tr,x_te,y_tr,y_te = train_test_split(x,y,test_size=0.3,random_state=1)
```

In [15]:

```
from sklearn.tree import DecisionTreeRegressor
tree = DecisionTreeRegressor()
```

In [16]:

```
#train the model
tree.fit(x_tr,y_tr)
```

Out[16]:

```
DecisionTreeRegressor(criterion='mse', max_depth=None, max_features=None,
           max_leaf_nodes=None, min_impurity_decrease=0.0,
           min_impurity_split=None, min_samples_leaf=1,
           min_samples_split=2, min_weight_fraction_leaf=0.0,
           presort=False, random_state=None, splitter='best')
```

In [17]:

```
pred = tree.predict(x_te)
```

In [20]:

```
tree.score(x_tr,y_tr)
```

Out[20]:

```
1.0
```

In [21]:

```
from sklearn.metrics import r2_score
r2_score(y_te,pred)
```

Out[21]:

```
0.9895909666229064
```

accuracy_score,confussion these classification models purpose

In [22]:

```
tree.predict([[165349.20,136897.80,471784.10,2]])
```

Out[22]:

```
array([192261.83])
```

In [23]:

```
tree.predict([[10000,15000,50000,2]])
```

Out[23]:

```
array([51003.74933])
```

Random Forest

It is ensemble model(Combination some models).It is one of the very full algorithm.It is also used for both Classification and Regression.

In [24]:

```
import pandas as pd
diabets = pd.read_csv("https://raw.githubusercontent.com/AP-State-Skill-Development-Co
                      Datasets/master/Classification/diabetes.csv")
diabets.head()
```

Out[24]:

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunctio
0	6	148	72	35	0	33.6	0.62
1	1	85	66	29	0	26.6	0.35
2	8	183	64	0	0	23.3	0.67
3	1	89	66	23	94	28.1	0.16
4	0	137	40	35	168	43.1	2.28

In [25]:

```
# any null values
diabets.isna().sum()
```

Out[25]:

```
Pregnancies                 0
Glucose                     0
BloodPressure               0
SkinThickness               0
Insulin                     0
BMI                         0
DiabetesPedigreeFunction    0
Age                         0
Outcome                     0
dtype: int64
```

In [27]:

```
x = diabets.drop('Outcome',axis=1)
x.head()
```

Out[27]:

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunctio
0	6	148	72	35	0	33.6	0.62
1	1	85	66	29	0	26.6	0.35
2	8	183	64	0	0	23.3	0.67
3	1	89	66	23	94	28.1	0.16
4	0	137	40	35	168	43.1	2.28

In [28]:

```
y = diabets['Outcome']
y.head()
```

Out[28]:

```
0    1
1    0
2    1
3    0
4    1
Name: Outcome, dtype: int64
```

In [29]:

```
diabets.shape
```

Out[29]:

```
(768, 9)
```

In [30]:

```
from sklearn.model_selection import train_test_split
x_train,x_test,y_train,y_test=train_test_split(x,y,random_state=1,test_size=0.25)

```

In [31]:

```
from sklearn.ensemble import RandomForestClassifier
rand = RandomForestClassifier()
```

In [32]:

```
rand.fit(x_train,y_train)
```

```
C:\Users\RANGA\Anaconda3\lib\site-packages\sklearn\ensemble\forest.py:246: F
utureWarning: The default value of n_estimators will change from 10 in versi
on 0.20 to 100 in 0.22.
  "10 in version 0.20 to 100 in 0.22.", FutureWarning)
```

Out[32]:

```
RandomForestClassifier(bootstrap=True, class_weight=None, criterion='gini',
            max_depth=None, max_features='auto', max_leaf_nodes=None,
            min_impurity_decrease=0.0, min_impurity_split=None,
            min_samples_leaf=1, min_samples_split=2,
            min_weight_fraction_leaf=0.0, n_estimators=10, n_jobs=None,
            oob_score=False, random_state=None, verbose=0,
            warm_start=False)
```

In [33]:

```
y_pred = rand.predict(x_test)
```

In [34]:

```
from sklearn.metrics import accuracy_score,confusion_matrix
accuracy_score(y_test,y_pred)
```

Out[34]:

```
0.7916666666666666
```

In [35]:

```
confusion_matrix(y_test,y_pred)
```

Out[35]:

```
array([[114,   9],
       [ 31,  38]], dtype=int64)
```

1. spliting the data changing (30 and 70)
2. take original data
3. try take check or change algorithm

In [ ]:

```

```