

Decision Tree

```
In [95]: import pandas as pd
```

```
In [96]: # data Mining
df= pd.read_csv('stu.csv')
df
```

Out[96]:

| | Company | Job | Degree | Salary |
|----|---------|----------|--------------|--------|
| 0 | Infosys | Sales | Gradute | <50000 |
| 1 | Infosys | Sales | Post Gradute | >50000 |
| 2 | Infosys | Devloper | Gradute | <50000 |
| 3 | Infosys | Devloper | Post Gradute | <50000 |
| 4 | Infosys | Manager | Gradute | >50000 |
| 5 | Infosys | Manager | Post Gradute | <50000 |
| 6 | Wipro | Sales | Gradute | <50000 |
| 7 | Wipro | Sales | Post Gradute | >50000 |
| 8 | Wipro | Devloper | Gradute | <50000 |
| 9 | Wipro | Devloper | Post Gradute | <50000 |
| 10 | Wipro | Manager | Gradute | <50000 |
| 11 | Wipro | Manager | Post Gradute | <50000 |
| 12 | TCS | Sales | Gradute | <50000 |
| 13 | TCS | Sales | Post Gradute | >50000 |
| 14 | TCS | Devloper | Gradute | <50000 |
| 15 | TCS | Devloper | Post Gradute | <50000 |
| 16 | TCS | Manager | Gradute | <50000 |
| 17 | TCS | Manager | Post Gradute | <50000 |

```
In [97]: df.head(1)
```

Out[97]:

| | Company | Job | Degree | Salary |
|---|---------|-------|---------|--------|
| 0 | Infosys | Sales | Gradute | <50000 |

```
In [98]: df.sample(1)
```

Out[98]:

| | Company | Job | Degree | Salary |
|---|---------|-------|---------|--------|
| 0 | Infosys | Sales | Gradute | <50000 |

```
In [99]: x = df.drop('Salary', axis= 'columns')
x
```

Out[99]:

| | Company | Job | Degree |
|----|---------|----------|--------------|
| 0 | Infosys | Sales | Gradute |
| 1 | Infosys | Sales | Post Gradute |
| 2 | Infosys | Devloper | Gradute |
| 3 | Infosys | Devloper | Post Gradute |
| 4 | Infosys | Manager | Gradute |
| 5 | Infosys | Manager | Post Gradute |
| 6 | Wipro | Sales | Gradute |
| 7 | Wipro | Sales | Post Gradute |
| 8 | Wipro | Devloper | Gradute |
| 9 | Wipro | Devloper | Post Gradute |
| 10 | Wipro | Manager | Gradute |
| 11 | Wipro | Manager | Post Gradute |
| 12 | TCS | Sales | Gradute |
| 13 | TCS | Sales | Post Gradute |
| 14 | TCS | Devloper | Gradute |
| 15 | TCS | Devloper | Post Gradute |
| 16 | TCS | Manager | Gradute |
| 17 | TCS | Manager | Post Gradute |

```
In [100]: y = df['Salary']  
y
```

```
Out[100]: 0    <50000  
1    >50000  
2    <50000  
3    <50000  
4    >50000  
5    <50000  
6    <50000  
7    >50000  
8    <50000  
9    <50000  
10   <50000  
11   <50000  
12   <50000  
13   >50000  
14   <50000  
15   <50000  
16   <50000  
17   <50000  
Name: Salary, dtype: object
```

```
In [101]: # Labling for transform  
from sklearn.preprocessing import LabelEncoder
```

```
In [102]: com_1 = LabelEncoder()  
job_1 = LabelEncoder()  
deg_1 = LabelEncoder()
```

```
In [103]: x['com_1'] = com_1.fit_transform(x['Company'])# indexing Elphabetic (0,2,1)
x
```

Out[103]:

| | Company | Job | Degree | com_1 |
|----|---------|----------|--------------|-------|
| 0 | Infosys | Sales | Gradute | 0 |
| 1 | Infosys | Sales | Post Gradute | 0 |
| 2 | Infosys | Devloper | Gradute | 0 |
| 3 | Infosys | Devloper | Post Gradute | 0 |
| 4 | Infosys | Manager | Gradute | 0 |
| 5 | Infosys | Manager | Post Gradute | 0 |
| 6 | Wipro | Sales | Gradute | 2 |
| 7 | Wipro | Sales | Post Gradute | 2 |
| 8 | Wipro | Devloper | Gradute | 2 |
| 9 | Wipro | Devloper | Post Gradute | 2 |
| 10 | Wipro | Manager | Gradute | 2 |
| 11 | Wipro | Manager | Post Gradute | 2 |
| 12 | TCS | Sales | Gradute | 1 |
| 13 | TCS | Sales | Post Gradute | 1 |
| 14 | TCS | Devloper | Gradute | 1 |
| 15 | TCS | Devloper | Post Gradute | 1 |
| 16 | TCS | Manager | Gradute | 1 |
| 17 | TCS | Manager | Post Gradute | 1 |

```
In [104]: x['job_1'] = job_1.fit_transform(x['Job'])
```

```
In [105]: x['deg_1'] = deg_1.fit_transform(x['Degree'])
```

```
In [106]: y.replace({'<50000':0, '>50000':1}, inplace= True)  
y
```

```
Out[106]: 0      0  
          1      1  
          2      0  
          3      0  
          4      1  
          5      0  
          6      0  
          7      1  
          8      0  
          9      0  
         10      0  
         11      0  
         12      0  
         13      1  
         14      0  
         15      0  
         16      0  
         17      0  
Name: Salary, dtype: int64
```

```
In [107]: x1 = x.drop(['Company', 'Job', 'Degree'], axis='columns')  
x1
```

Out[107]:

| | com_1 | job_1 | deg_1 |
|----|-------|-------|-------|
| 0 | 0 | 2 | 0 |
| 1 | 0 | 2 | 1 |
| 2 | 0 | 0 | 0 |
| 3 | 0 | 0 | 1 |
| 4 | 0 | 1 | 0 |
| 5 | 0 | 1 | 1 |
| 6 | 2 | 2 | 0 |
| 7 | 2 | 2 | 1 |
| 8 | 2 | 0 | 0 |
| 9 | 2 | 0 | 1 |
| 10 | 2 | 1 | 0 |
| 11 | 2 | 1 | 1 |
| 12 | 1 | 2 | 0 |
| 13 | 1 | 2 | 1 |
| 14 | 1 | 0 | 0 |
| 15 | 1 | 0 | 1 |
| 16 | 1 | 1 | 0 |
| 17 | 1 | 1 | 1 |

```
In [108]: x1.head(1)
```

Out[108]:

| | com_1 | job_1 | deg_1 |
|---|-------|-------|-------|
| 0 | 0 | 2 | 0 |

```
In [109]: x1['total']='0'  
x1
```

Out[109]:

| | com_1 | job_1 | deg_1 | total |
|----|-------|-------|-------|-------|
| 0 | 0 | 2 | 0 | 0 |
| 1 | 0 | 2 | 1 | 0 |
| 2 | 0 | 0 | 0 | 0 |
| 3 | 0 | 0 | 1 | 0 |
| 4 | 0 | 1 | 0 | 0 |
| 5 | 0 | 1 | 1 | 0 |
| 6 | 2 | 2 | 0 | 0 |
| 7 | 2 | 2 | 1 | 0 |
| 8 | 2 | 0 | 0 | 0 |
| 9 | 2 | 0 | 1 | 0 |
| 10 | 2 | 1 | 0 | 0 |
| 11 | 2 | 1 | 1 | 0 |
| 12 | 1 | 2 | 0 | 0 |
| 13 | 1 | 2 | 1 | 0 |
| 14 | 1 | 0 | 0 | 0 |
| 15 | 1 | 0 | 1 | 0 |
| 16 | 1 | 1 | 0 | 0 |
| 17 | 1 | 1 | 1 | 0 |

```
In [110]: # Decision Tree  
from sklearn import tree
```

```
In [111]: model =tree.DecisionTreeClassifier()  
model.fit(x1,y)
```

Out[111]:

▼ DecisionTreeClassifier

DecisionTreeClassifier()

```
In [112]: model.score(x1,y)
```

Out[112]: 1.0

```
In [113]: model.predict([[2,1,0,1]])
```

C:\Users\Administrator\AppData\Local\Programs\Python\Python311\Lib\site-packages\sklearn\base.py:465: UserWarning: X does not have valid feature names, but DecisionTreeClassifier was fitted with feature names
warnings.warn(

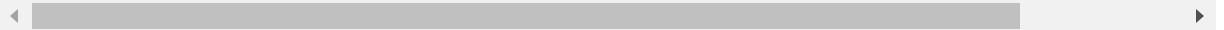
```
Out[113]: array([0], dtype=int64)
```

```
In [114]: import pandas as pd  
import matplotlib.pyplot as plt
```

```
In [115]: df = pd.read_csv('die.csv')  
df.head(1)
```

```
Out[115]:
```

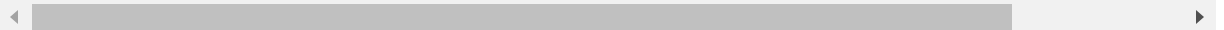
| | Pregnancies | Glucose | BloodPressure | SkinThickness | Insulin | BMI | DiabetesPedigreeFunction |
|---|-------------|---------|---------------|---------------|---------|------|--------------------------|
| 0 | 6 | 148 | 72 | 35 | 0 | 33.6 | 0.627 |



```
In [116]: df.sample()
```

```
Out[116]:
```

| | Pregnancies | Glucose | BloodPressure | SkinThickness | Insulin | BMI | DiabetesPedigreeFunction |
|----|-------------|---------|---------------|---------------|---------|------|--------------------------|
| 87 | 2 | 100 | 68 | 25 | 71 | 38.5 | 0.324 |




```
In [117]: x =df.drop('Outcome', axis='columns')
x
```

Out[117]:

| | Pregnancies | Glucose | BloodPressure | SkinThickness | Insulin | BMI | DiabetesPedigreeFuncio |
|-----|-------------|---------|---------------|---------------|---------|------|------------------------|
| 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 |
| ... | ... | ... | ... | ... | ... | ... | . |
| 763 | 10 | 101 | 76 | 48 | 180 | 32.9 | 0.17 |
| 764 | 2 | 122 | 70 | 27 | 0 | 36.8 | 0.34 |
| 765 | 5 | 121 | 72 | 23 | 112 | 26.2 | 0.24 |
| 766 | 1 | 126 | 60 | 0 | 0 | 30.1 | 0.34 |
| 767 | 1 | 93 | 70 | 31 | 0 | 30.4 | 0.31 |

768 rows × 8 columns

```
In [118]: x.sample()
```

Out[118]:

| | Pregnancies | Glucose | BloodPressure | SkinThickness | Insulin | BMI | DiabetesPedigreeFuncio |
|-----|-------------|---------|---------------|---------------|---------|------|------------------------|
| 108 | 3 | 83 | 58 | 31 | 18 | 34.3 | 0.33 |

```
In [119]: y =df['Outcome']
# Split Dataset
```

```
In [120]: feature_col= ['Pregnancies','Glucose','BloodPressure','SkinThickness','Insulin'
```

```
In [121]: from sklearn.model_selection import train_test_split
```

```
In [122]: x_train,x_test, y_train,y_test= train_test_split(x,y)# test size bydefault= 75%
```

```
In [123]: x_train.shape
```

Out[123]: (576, 8)

```
In [124]: # df.info()
```

In [125]: `x_test.shape`

Out[125]: (192, 8)

In [126]: `576/768`

Out[126]: 0.75

In [127]: `x_train`

Out[127]:

| | Pregnancies | Glucose | BloodPressure | SkinThickness | Insulin | BMI | DiabetesPedigreeFuncio |
|------------|-------------|---------|---------------|---------------|---------|------|------------------------|
| 153 | 1 | 153 | 82 | 42 | 485 | 40.6 | 0.68 |
| 519 | 6 | 129 | 90 | 7 | 326 | 19.6 | 0.58 |
| 478 | 8 | 126 | 74 | 38 | 75 | 25.9 | 0.16 |
| 176 | 6 | 85 | 78 | 0 | 0 | 31.2 | 0.38 |
| 202 | 0 | 108 | 68 | 20 | 0 | 27.3 | 0.78 |
| ... | ... | ... | ... | ... | ... | ... | . |
| 69 | 4 | 146 | 85 | 27 | 100 | 28.9 | 0.18 |
| 449 | 0 | 120 | 74 | 18 | 63 | 30.5 | 0.28 |
| 652 | 5 | 123 | 74 | 40 | 77 | 34.1 | 0.26 |
| 552 | 6 | 114 | 88 | 0 | 0 | 27.8 | 0.24 |
| 733 | 2 | 106 | 56 | 27 | 165 | 29.0 | 0.42 |

576 rows × 8 columns



In [128]: `# Model`
`# from sklearn import DecisionTreeClassifier`
`from sklearn import tree`

In [129]: `# model = DecisionTreeClassifier()`
`model = tree.DecisionTreeClassifier()`

In [130]: `model.fit(x_train, y_train)`

Out[130]: `DecisionTreeClassifier`
`DecisionTreeClassifier()`

In [131]: `predicted_value = model.predict(x_test)`

```
In [132]: model.score(x_test, y_test)
```

```
Out[132]: 0.671875
```

```
In [140]: from sklearn import metrics
```

```
In [141]: metrics.accuracy_score(y_test, predicted_value)
```

```
Out[141]: 0.671875
```

```
In [142]: # pip install pydotplus  
import pydotplus
```

```
In [143]: from sklearn.tree import export_graphviz
```

```
In [144]: # pip install six  
from six import StringIO
```

```
In [145]: # pip install Pillow  
from IPython.display import Image
```

```
In [146]: a_data = StringIO()  
import cv2  
from PIL import Image
```