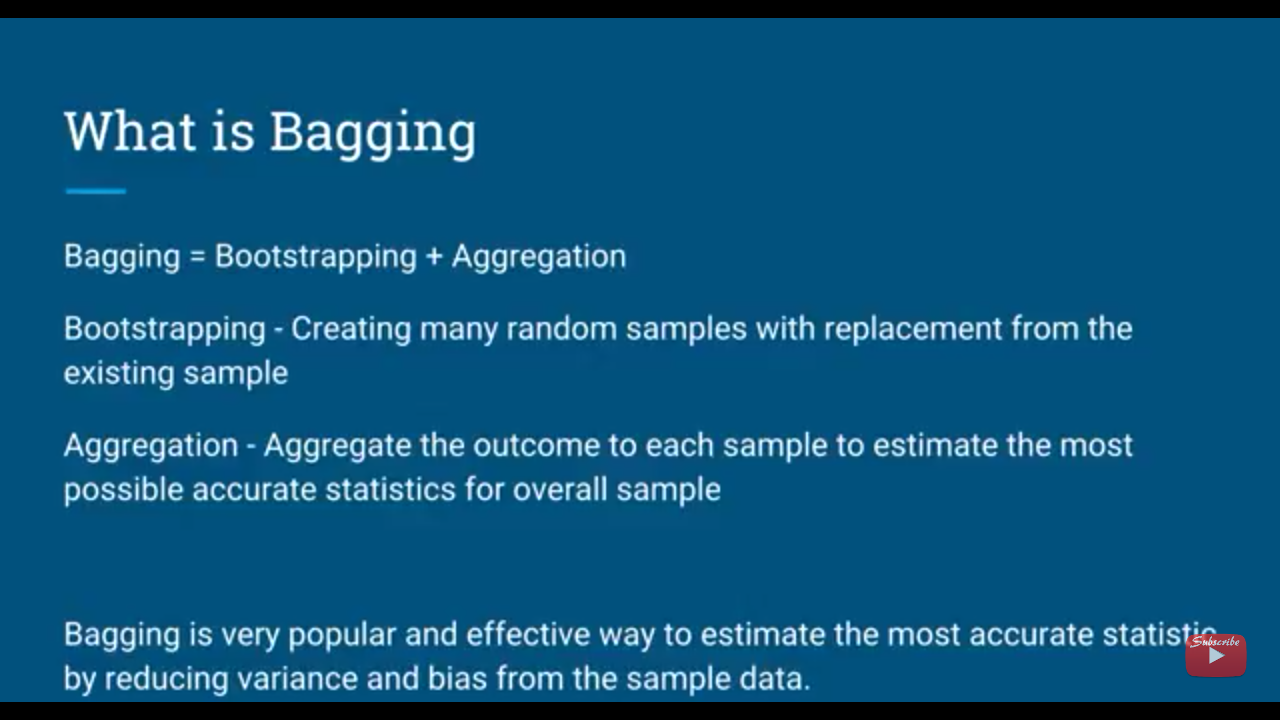
**Random Forest Regression**

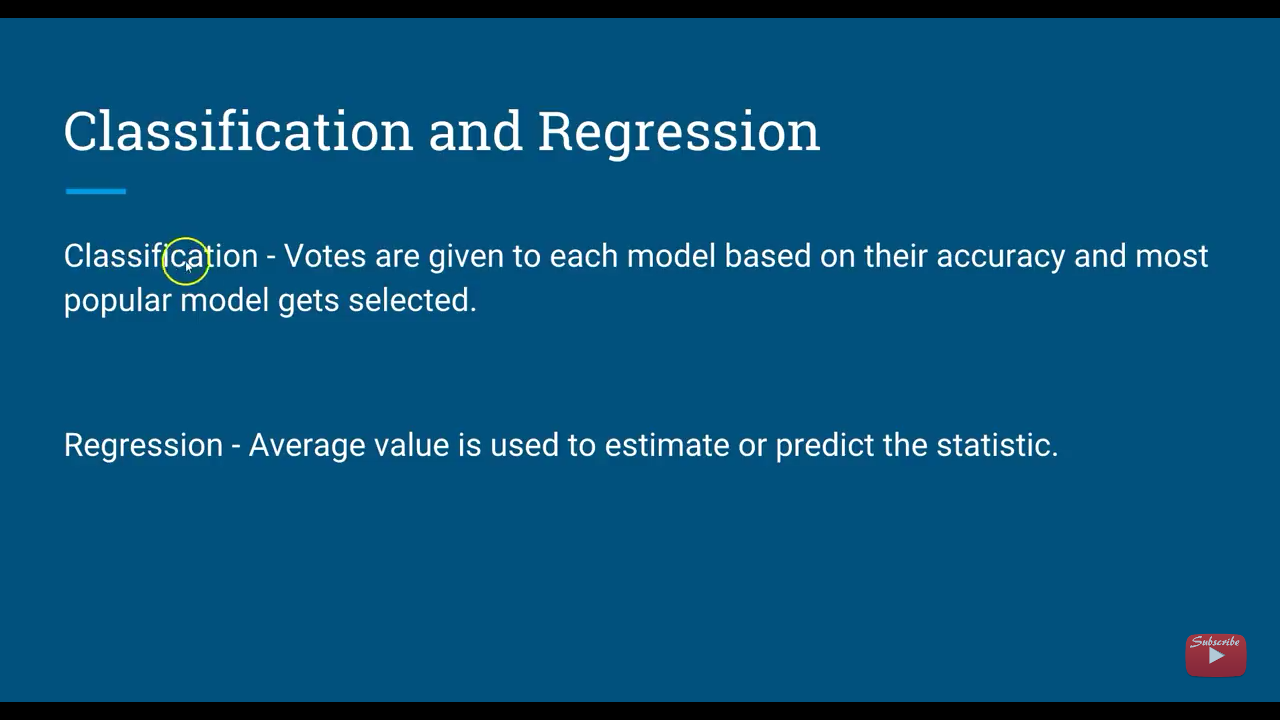
Random forest algorithm is a supervised classification algorithm. As the name suggest, this algorithm creates the forest with a number of trees.

In general, the more trees in the forest the more robust the forest looks like. In the same way in the random forest classifier, the higher the number of trees in the forest gives the high accuracy results.

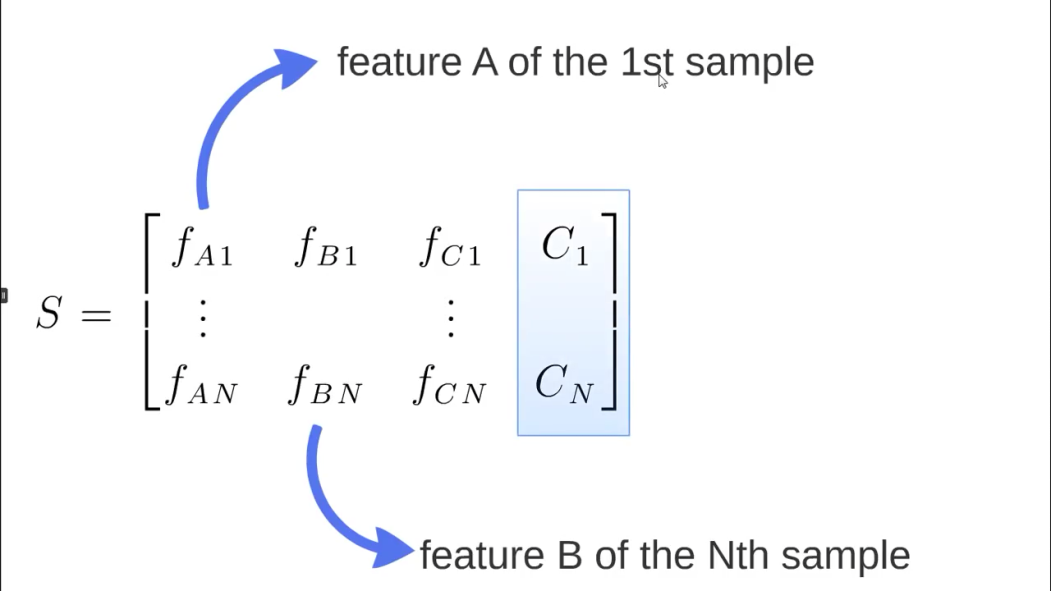
Random forest gets training set and divided it by “Bagging = Bootstrap Aggregating” which is algorithm to increase accuracy by prevent over fitting and decrease variance. It starts to divide data set to 60% as unique decision tree and 30% as overlapping data.

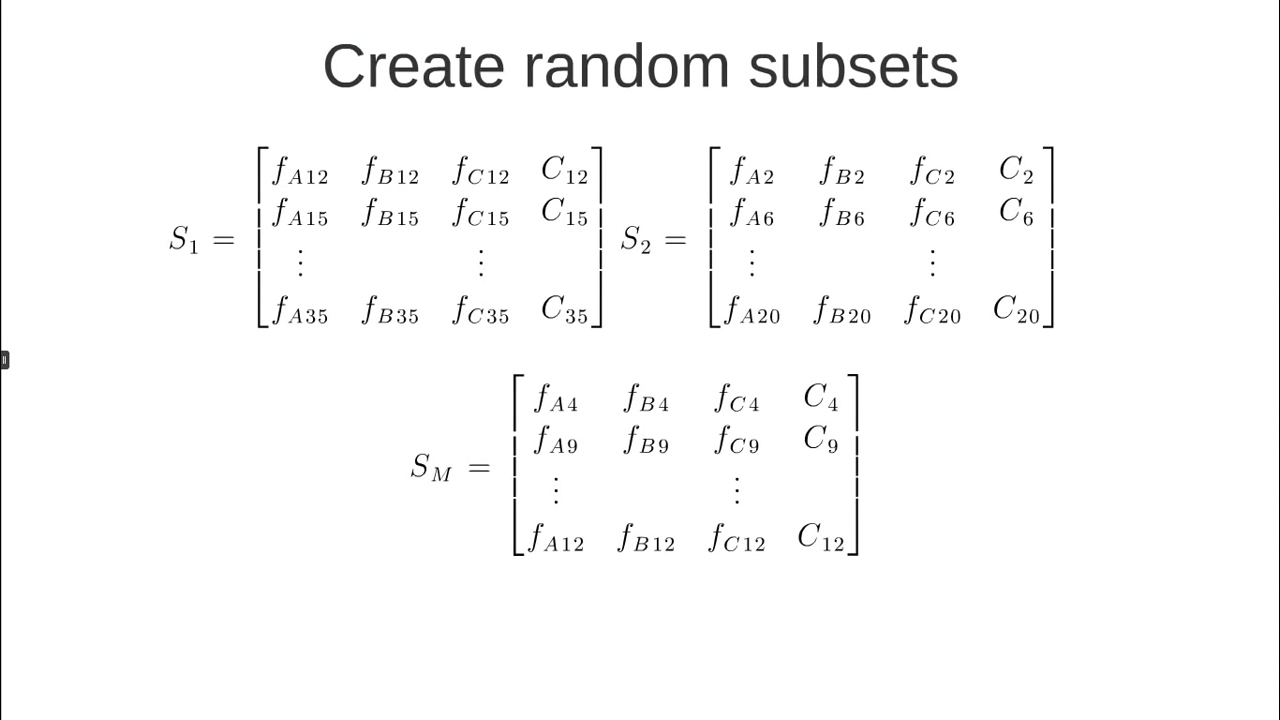
**Bagging-**

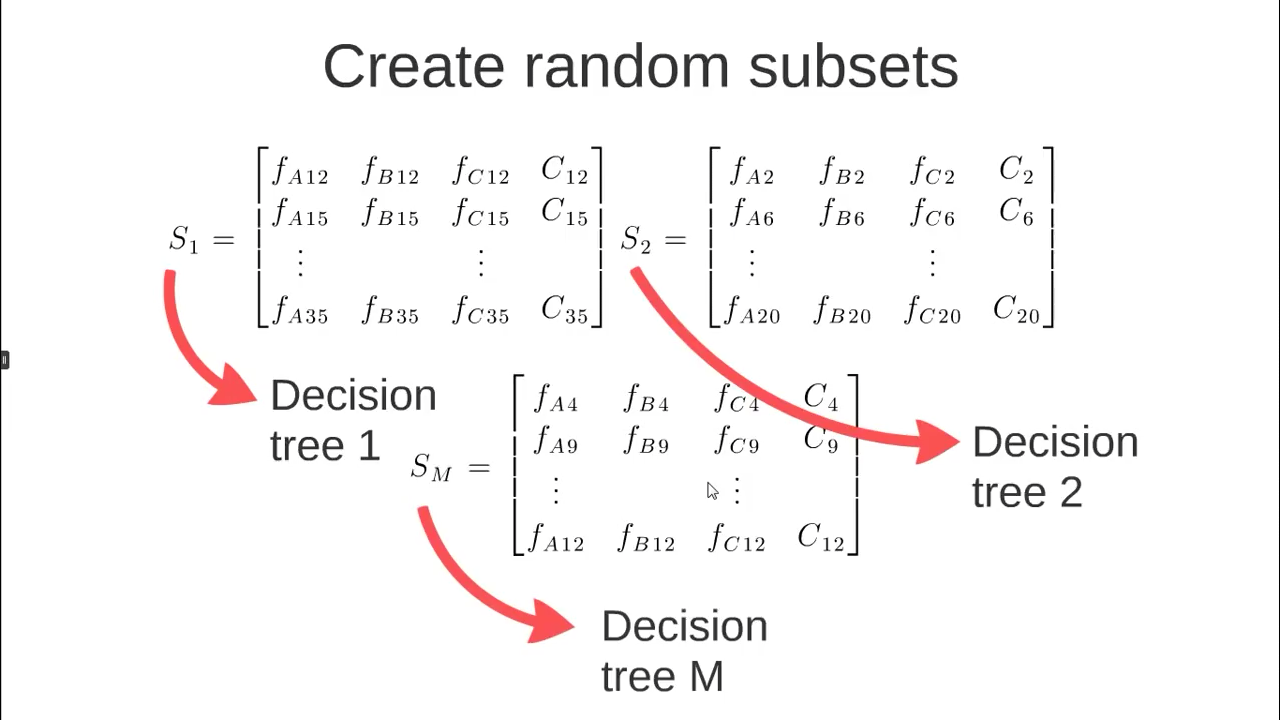




**Formulae:**







**Decode Complex Algorithm**

**Step: 1**

**Salary bands :**

Band 1 : Below $40,000

Band 2: $40,000 – 150,000

Band 3: More than $150,000

**Decision Tree 1 of variable Age**-

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| variable Age | | | | |
|  | Salary Band | 1 | 2 | 3 |
| |  | | --- | | Age | | | Below 18 | 20% | 80% | 60% |
| 19-27 | 50% | 25% | 90% |
| 28-40 | 20% | 90% | 50% |
| 40-55 | 80% | 75% | 75% |
| Morethan 55 | 30% | 55% | 90% |

**Decision Tree 2 of variable Gender**-

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Variable Gender | | | | |
|  | Salary Band | 1 | 2 | 3 |
| Gender | Male | 50% | 30% | 90% |
| Female | 80% | 20% | 70% |

**Decision Tree 3 of variable Education**-

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Variable Education | | | | |
|  | Salary Band | 1 | 2 | 3 |
| Education | <=HighSchool | 20% | 70% | 20% |
| Diploma | 50% | 90% | 50% |
| Bachelors | 20% | 50% | 20% |
| Masters | 20% | 50% | 20% |

**Decision Tree 4 of variable Residence**-

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Variable Residence | | | |  |
| Residence | Salary Band | 1 | 2 | 3 |
| Metro | 30% | 80% | 50% |
| Nonmetro | 30% | 25% | 90% |

**Decision Tree 5 of variable Industry-**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Variable Industry | | | |  |
| Industry | Salary Band | 1 | 2 | 3 |
| Finance | 20% | 90% | 50% |
| Manufacturing | 80% | 75% | 75% |
| Others | 30% | 55% | 90% |

Now let us find the random Forest of below Cart

**1. Age : 28 years , 2, Gender : Female , 3. Highest Educational Qualification : Masters,**

**4. Industry : Manufacturing, 5. Residence : Metro**

**RandomForest-(Clustering of all Decision Trees)**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| CART | Band | 1 | 2 | 3 |
| Age | 28-40 | 20% | 90% | 50% |
| Gender | Female | 80% | 20% | 70% |
| Education | masters | 20% | 50% | 20% |
| Industry | Manufacturing | 80% | 75% | 75% |
| Residence | Metro | 30% | 80% | 50% |
| Final Probability(Average of all decision Tree) | | 46% | 63% | 53% |

**Conclusion-**

The final probability is the average of the probability in the same salary bands in different CART models. From this analysis, that there is 46% chance of this individual falling in class 1 (less than $40,000) and around 63% chance of the individual falling in class 2 and 53% under class 3 respectively.

Applications:

**1.Banking:**

In the banking sector, random forest algorithm widely used in two main application. These are for finding the loyal customer and finding the fraud customers.

**2.Medicine**

In medicine field, random forest algorithm is used identify the correct combination of the components to validate the medicine. Random forest algorithm also helpful for identifying the disease by analyzing the patient’s medical records.

**3.Stock Market**

In the stock market, random forest algorithm used to identify the stock behavior as well as the expected loss or profit by purchasing the particular stock.

**4.E-commerce**

In e-commerce, the random forest used only in the small segment of the recommendation engine for identifying the likely hood of customer liking the recommend products base on the similar kinds of customers.

Python code:

import numpy as np

import matplotlib.pyplot as plt

import pandas as pd

#read the dataset

df = pd.read\_csv('C:/pycharm/salary\_position.csv')

X = df.iloc[:,1:2].values

y = df.iloc[:,2].values

##Decision tree regression

from sklearn.ensemble import RandomForestRegressor

regressor = RandomForestRegressor(n\_estimators= 10,random\_state=0)

regressor.fit(X,y)

##prediction a new result

y\_pred = regressor.predict(6.5)

##plottiing the simpole regression

X\_grid = np.arange(min(X),max(X),0.1)

X\_grid = X\_grid.reshape((len(X\_grid),1))

plt.scatter(X,y,color = 'red')

plt.plot(X\_grid,regressor.predict(X\_grid),color = 'blue')

plt.title('Random forests regression')

plt.title('salary\_position')

plt.xlabel('position')

plt.ylabel('salary')

plt.show()

output:

