# MODULE 14

1. **Company Dataset**

**Package used in R**

randomForest - Used to run Random Forest

Gmodels - Used to run cross table for Decision Tree

C50 - Used to run Decision Tree

**Package used in python**

Pandas - Used for data manipulation

Numpy - Used for Mathematical calculation

Sklearn

DecisionTreeClassifier - Used to run Decision Tree Algorithm

Train\_test\_split - Used to Split the data to train and test for modeling

LabelEncoder - Typecasting the Categorical label to Numerical variable

RandomForestClassifier - Used to run Random Forest Algorithm

accuracy\_score - Calculate the accuracy of the random Forest

**Loading the data**

Loading the Company dataset in R and Python

**EDA**

No NaN Data Found in the dataset

Typecasting the sales variables to 0 or 1 according to the sales

Typecasting the ShelveLoc, Urban, US variables to numerical variables with label encoder

**Data Partitioning**

Splitting the data to X and Y. X as the predictors with all the X variables and Y as the target with the predicting variables

**Modeling and Result**

Model=DT(criterion='entropy') and test\_size=0.2

Test accuracy 0.862 and training accuracy was 1.0

**Pruning Model**

1. (criterion='entropy', max\_depth=4)

Test accuracy: 0.83 and Train accuracy : 0.91

1. (criterion='entropy', min\_samples\_split = 15)

Test accuracy: 0.75 and Train accuracy : 0.91

1. (criterion='entropy', min\_samples\_leaf = 10)

Test accuracy: 0.77 and Train accuracy : 0.91

**RANDOM FOREST**

**Data Partitioning**

Splitting the data to X and Y. X as the predictors with all the X variables and Y as the target with the predicting variables. Data is split in the ratio of Training 70 : Test 30

**Modeling and Result**

(n\_jobs=3, n\_estimators=125, criterion="entropy")

Test accuracy: 0.825 and Train accuracy : 1.0

(n\_jobs=3, n\_estimators=75, criterion="entropy",max\_depth=3)

Test accuracy: 0.791 and Train accuracy : 0.864

1. **Diabetes Dataset**

**Package used in R**

randomForest - Used to run Random Forest

Gmodels - Used to run cross table for Decision Tree

C50 - Used to run Decision Tree

**Package used in python**

Pandas - Used for data manipulation

Numpy - Used for Mathematical calculation

Sklearn

DecisionTreeClassifier - Used to run Decision Tree Algorithm

Train\_test\_split - Used to Split the data to train and test for modeling

LabelEncoder - Typecasting the Categorical label to Numerical variable

RandomForestClassifier - Used to run Random Forest Algorithm

accuracy\_score - Calculate the accuracy of the random Forest

**Loading the data**

Loading the Diabetes dataset in R and Python

**EDA**

No NaN Data Found in the dataset

Renaming the Dataset variables as all the variables as space in it

Creating dummy variables on the variables column

**Data Partitioning**

Splitting the data to X and Y. X as the predictors with all the X variables and Y as the target with the predicting variables

**Modeling and Result**

Model=DT(criterion=gini) and test\_size=0.2

Test accuracy 0.70 and training accuracy was 1.0

**Pruning Model**

1. (criterion='gini', max\_depth=4)

Test accuracy: 0.766 and Train accuracy : 0.806

1. (criterion='gini', min\_samples\_split = 5)

Test accuracy: 0.75 and Train accuracy : 0.78

1. (criterion='gini', min\_samples\_leaf = 5)

Test accuracy: 0.77 and Train accuracy : 0.91

**RANDOM FOREST**

**Data Partitioning**

Splitting the data to X and Y. X as the predictors with all the X variables and Y as the target with the predicting variables. Data is split in the ratio of Training 70 : Test 30

**Modeling and Result**

(n\_jobs=3, n\_estimators=125, criterion="entropy")

Test accuracy: 0.766 and Train accuracy : 0.998

(n\_jobs=4, n\_estimators=55, criterion="gini", max\_depth=5)

Test accuracy: 0.748 and Train accuracy : 0.865

**Fraud Dataset**

**Package used in R**

randomForest - Used to run Random Forest

Gmodels - Used to run cross table for Decision Tree

C50 - Used to run Decision Tree

**Package used in python**

Pandas - Used for data manipulation

Numpy - Used for Mathematical calculation

Sklearn

DecisionTreeClassifier - Used to run Decision Tree Algorithm

Train\_test\_split - Used to Split the data to train and test for modeling

LabelEncoder - Typecasting the Categorical label to Numerical variable

RandomForestClassifier - Used to run Random Forest Algorithm

accuracy\_score - Calculate the accuracy of the random Forest

**Loading the data**

Loading the Diabetes dataset in R and Python

**EDA**

No NaN Data Found in the dataset

Renaming the Dataset variables as all the variables as space in it

Typecasting the Taxable variables if above 30000 its converted 1 and other entries converted 0

Typecasting the categorical data of Undergrad ,Marital, Urban, to numerical data

Rearranging the variable to split the data

**Data Partitioning**

Splitting the data to X and Y. X as the predictors with all the X variables and Y as the target with the predicting variables

**Modeling and Result**

Model=DT(criterion=entropy) and test\_size=0.2

Test accuracy 0.66 and training accuracy was 1.0

**Pruning Model**

1. (criterion='entropy', max\_depth=3)

Test accuracy: 0.825 and Train accuracy : 0.787

1. (criterion='gini', min\_samples\_split = 15)

Test accuracy: .683 and Train accuracy : 0.8

1. (criterion='gini', min\_samples\_leaf = 5)

Test accuracy: 0.70 and Train accuracy : 0.785

**RANDOM FOREST**

**Data Partitioning**

Splitting the data to X and Y. X as the predictors with all the X variables and Y as the target with the predicting variables. Data is split in the ratio of Training 70 : Test 30

**Modeling and Result**

(n\_jobs=3, n\_estimators=125, criterion="gini")

Test accuracy: 1.0 and Train accuracy :1.0

(n\_jobs=4, n\_estimators=2, criterion="gini", max\_depth=5)

Test accuracy: 0.994 and Train accuracy : 1.0