# MODULE 17

# **Salary Data**

**Package used in R**

Kernlab : Used to run SVM Algorithm

**Package used in python**

Pandas : Used for data manipulation

Numpy: Used for scientific calculation

Sklearn

SVC : Used to run SVM Algorithm

train\_test\_split: Splitting of training and testing data

**Loading the data**

Loading the Salary Train and Salary Test dataset by Import Dataset option and select strings as factor

**EDA**

No NaN Data Found in the dataset.

R: Categorical Variables are converted to Factor while loading the dataset

Python: Selected first 500 observation from salary test

Categorical Variables are converted to numerical data with LabelEncoder

**Data Partitioning**

R: Using Salary train data as training data and Salary test data as testing data.

Python: Splitting the data into 80:20 Ratio. 80% Training and 20% Test Data

**SVM Modeling.**

Used different Kernal method to get the high accuracy

Model 1

Kernel Model : vanilladot

Using the Model in Test Data : **.84** Accuracy

Using the Model in Train Data: **.84** Accuracy

Model 2

Kernel Model : rbfdot

Using the Model in Test Data : **.85** Accuracy

Using the Model in Train Data: **.86** Accuracy

Model 3

Kernel Model : polydot

Using the Model in Test Data : **.84** Accuracy

Using the Model in Train Data: .**84** Accuracy

# Python

Kernal Model : Linear

Using the Model in Test Data : **.80** Accuracy

Using the Model in Train Data: **. 80** Accuracy

Kernal Model : Rbf

Using the Model in Test Data : **.80** Accuracy

Using the Model in Train Data: **. 80** Accuracy

Kernal Model : Poly

Using the Model in Test Data : **. 80** Accuracy

Using the Model in Train Data: **. 80** Accuracy

Kernal Model : sigmoid

Using the Model in Test Data : **. 77** Accuracy

Using the Model in Train Data: **. 77** Accuracy

# **Forestfires Data**

**Package used in python**

Pandas : Used for data manipulation

Numpy: Used for scientific calculation

Sklearn

SVC : Used to run SVM Algorithm

train\_test\_split: Splitting of training and testing data

**Loading the data**

Loading the forestfires dataset by Import Dataset option and select strings as factor

**EDA**

No NaN Data Found in the dataset.  
Categorical Variables are converted to Factor while loading the dataset

Selected only the Temp, RH, wind, rain, area, size\_category variables for the model

**Plot**

Scatter plot was used to find the correlation between different variables

No Correlation between wind and Temp. Points are scattered around the plot

Negative Correlation found between RH and Temp.

No Correlation between RH and Rain.

**Data Partitioning**

Forest Data is split into training and test data. First 450 observation is taken as training data and remaining data is taken as test data

**SVM Modeling.**

Used different variables to predict the and listing the models with good accuracy

Used different Kernal method to get the high accuracy

Model 1

Kernel Model : vanilladot

Variables used : temp + wind+ rain+ area

Using the Model in Test Data : **.865** Accuracy

Using the Model in Train Data: **.915** Accuracy

Model 2

Kernel Model : rbfdot

Variables used : temp + wind+ rain+ area

Using the Model in Test Data : **.766** Accuracy

Using the Model in Train Data: **.880** Accuracy

Model 3

Kernel Model : polydot

Variables used : temp + wind+ rain+ area

Using the Model in Test Data : **.850** Accuracy

Using the Model in Train Data: **.911** Accuracy

**Vanilladot Model has good accuracy compared to other model’s**

# Python

Kernal Model : Linear

Using the Model in Test Data : **.99** Accuracy

Using the Model in Train Data: **.99** Accuracy

Kernal Model : Rbf

Using the Model in Test Data : **.89** Accuracy

Using the Model in Train Data: **.90** Accuracy

Kernal Model : Poly

Using the Model in Test Data : **.93** Accuracy

Using the Model in Train Data: **.81** Accuracy

Kernal Model : sigmoid

Using the Model in Test Data : **.87** Accuracy

Using the Model in Train Data: **.86** Accuracy