# MODULE 19

# **Salary Data**

**Package used in R**

Readr : used to load the dataset

e1071: used to run Naive Bayes

Gmodels: Used to plot Cross Table

**Package used in Python**

Pandas : Used for Data Manipulation

Numpy: Used for Scientific Calculation

Sklearn.naive\_bayes

MultinomialNB : Used for Multinomial Naive Bayes

GaussianNB : Used for Gaussian Naive Bayes

CategoricalNB : Used for Categorical Naive Bayes

BernoulliNB : Used for Bernoulli Naive Bayes

Sklearn.preprocessing

LabelEncoder : Used to Label Encoding

**Loading the data**

Loading the Salary Train and Salary Test dataset

**EDA**

No NaN Data Found in the dataset.

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

Python: Categorical Variables are converted to numerical data with LabelEncoder

**Data Partitioning**

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

**Naive Bayes Modeling**

|  |  |  |
| --- | --- | --- |
| **Model** | **Train Accuracy** | **Test Accuracy** |
| Multinomial NB | 0.772 | 0.774 |
| Gaussian NB | 0.795 | 0.794 |
| Categorical NB | 0.795 | 0.856 |
| Bernoulli NB | 0.725 | 0.728 |

**Gaussian Model has given a good accuracy on both train and test model**

1. **SMS Raw Data**

**Package used in R**

Readr : used to load the dataset

Tm: used to text cleaning

e1071: used to run Naive Bayes

Gmodels: Used to plot Cross Table

**Package used in Python**

Pandas : Used for Data Manipulation

Numpy: Used for Scientific Calculation

Sklearn.naive\_bayes

MultinomialNB : Used for Multinomial Naive Bayes

GaussianNB : Used for Gaussian Naive Bayes

BernoulliNB : Used for Bernoulli Naive Bayes

Sklearn.feature\_extraction.text

CountVectorizer : Used to create Bag Of Words (Count Matrix)

TfidfTransformer: Used to Transform the data to document-frequency.

**Loading the data**

Loading the sms\_Raw\_NB dataset

**EDA**

No NaN Data Found in the dataset.

Converting all the upper case to lower case

Removing all the numbers and only selecting the words longer than 3 letters

After Data Splitting, creating bag of words with the CountVectorizer and transforming all the train and test data.

Now transforming the CountVectorizer to TF-IDF matrix format.

**Data Partitioning**

Creating predictors and target data from the dataset and Splitting the data to X and Y. X as the predictors with all the X variables except Salary Variables and Y as the target with the predicting variables.

**Naive Bayes Modeling**

|  |  |  |
| --- | --- | --- |
| **Model** | **Train Accuracy** | **Test Accuracy** |
| Multinomial NB | 0.969 | 0.955 |
| Gaussian NB | 0.899 | 0.839 |
| Bernoulli NB | 0.981 | 0.972 |

Bernoulli Model has gave better accuracy compared to other models