Big Data Analytics

Lab Practical and date – Practical 2, Monday 27th July 2020

Name and Roll Number- Het Shah, 17BIT103

Practical Objective- Learning limitation of data analytics by applying Machine Learning Techniques on large amount of data. Write R/Python program to Read data set from any online website, excel file and CSV file and to perform

a) Linear regression and logistic regression on iris dataset.
b) K-means clustering.

Steps Involved-

We perform data scarping using python by reading data from different file format such as excel, csv and also performing data analytics on it

Background

Python is an interpreted, high-level, general-purpose programming language. Created by Guido van Rossum and first released in 1991, Python's design philosophy emphasizes code readability with its notable use of significant whitespace.

Libraries Used-

1) Pandas-pandas is a software library written for the Python programming language for data manipulation and analysis

2) SciKit-Learn-Scikit-Learn a free software machine learning library for the Python programming language. It features various classification and regression and clustering algorithms including support vector machines, random forests, gradient boosting, *k*-means and DBSCAN, and is designed to interoperate with the Python numerical and scientific libraries NumPy and SciPy.
3) Mathplotlib- Matplotlib is a plotting library for the Python programming language and its numerical mathematics extension NumPy. It provides an object-oriented API for embedding plots into applications using general-purpose GUI toolkits like Tkinter, wxPython, Qt, or GTK+.

Data-set

1) Iris DataSet-The data set consists of 50 samples from each of three species of *Iris* (*Iris setosa*, *Irisvirginica* and *Iris versicolor*). Features were measured from each sample: the length and the width of sepals and petals, in centimeters. Based on the combination of these four features, Fisher developed a linear discriminant model to distinguish the species from each other.

Algorithms

1) Linear regression- linear regression is a linear approach to modeling the relationship between a scalar response and one or more explanatory variables.
2) K-Means-k-means clustering is a method of vector quantization, originally from signal processing, that aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster.
3) Logistic Regression-In statistics, the logistic model is used to model the probability of a certain class or event existing such as pass/fail, win/lose, alive/dead or healthy/sick. This can be extended to model several classes of events such as determining whether an image contains a cat, dog, lion, etc.

Conclusion-

In this Experiment we applied various machine learning algorithms on the given IRIS dataset and plotted the results using the mathplotlib.