Nutan Maharashtra Vidya Prasarak Mandal's
NUTAN MAHARASHTRA INSTITUTE OF ENGINEERING AND TECHNOLOGY
Under Administrative Support - Pimpri Chinchwad Education Trust
ESTD : 1906
Approved by AICTE Accredited by NAAC Affiliated to SPPU
"Samarth Vidya Sankul", Vishnupuri, Telegaon Dabhade, Taluka Maval, District Pune - 410507
Tel. No. 02114 – 231666,777,888 E-mail : nmiettalegaon@gmail.com Web : www.nmiet.edu.in
AICTE ID - 1-8618657 AISHE ID - C-41640 DTE ID – 6310 UNIVERSITY ID - CEGP013890

DS&BDA Lab Index

Department: - Computer Engineering **Year & Semester Course Offered**: T.E.Sem-II

Expt. No.	Name of the Experiment
01	Data Wrangling I Perform the different operations using Python on any open source dataset (eg. data.csv)
02	Data Wrangling II Perform the different operations using Python on any open source dataset (eg. data.csv) like Scan all variables for missing values and inconsistencies, Scan all numeric variables for outliers., Apply data transformations on at least one of the variables. Reason and document your approach properly.
03	Basic Statistics - Measures of Central Tendencies and Variance Perform the following operations on any open source dataset (eg. data.csv) 1. Provide summary statistics (mean, median, minimum, maximum, standard deviation) for a dataset (age, income etc.) with numeric variables grouped by one of the qualitative (categorical) variable.
04	Data Analytics I Create a Linear Regression Model using Python/R to predict home prices using Boston Housing Dataset (https://www.kaggle.com/c/boston-housing). There are 506 samples and 14 feature variables in this dataset.
05	Data Analytics II 1. Implement logistic regression using Python/R to perform classification on Social_Network_Ads.csv dataset Compute Confusion matrix to find TP, FP, TN, FN, Accuracy, Error rate, Precision, Recall on the given dataset.
06	Data Analytics III 1. Implement Simple Naïve Bayes classification algorithm using Python/R on iris.csv dataset. II. Compute Confusion matrix to find TP, FP, TN, FN, Accuracy, Error rate, Precision, Recall on the given dataset.
07	Text Analytics 1. Extract Sample document and apply following document preprocessing methods: Tokenization, POS Tagging, stop words removal, Stemming and Lemmatization. Create representation of document by calculating Term Frequency and Inverse Document

	Frequency.
08	Data Visualization I 1. Use the inbuilt dataset 'titanic'. The dataset contains 891 rows and contains information about the passengers who boarded the unfortunate Titanic ship. Use the Seaborn library to see if we can find any patterns in the data. Write a code to check how the price of the ticket (column name: 'fare') for each passenger is distributed by plotting a histogram
09	Data Visualization II 1. Use the inbuilt dataset 'titanic' as used in the above problem. Plot a box plot for distribution of age with respect to each gender along with the information about whether they survived or not. (Column names : 'sex' and 'age') Write observations on the inference from the above statistics.
10	Data Visualization III Download the Iris flower dataset or any other dataset into a DataFrame. (eg https://archive.ics.uci.edu/ml/datasets/Iris). Scan the dataset and give the inference as: 1. How many features are there and what are their types (e.g., numeric, nominal)? 2. Create a histogram for each feature in the dataset to illustrate the feature distributions. 3. Create a boxplot for each feature in the dataset. Compare distributions and identify outliers
11	Write a code in JAVA for a simple WordCount application that counts the number of occurrences of each word in a given input set using the Hadoop MapReduce framework on local-standalone set-up
12	Design a distributed application using MapReduce which processes a log file of a system.
13	Locate dataset (eg. sample_weather.txt) for working on weather data which reads the text input files and finds average for temperature, dew point and wind speed.
14	Group C- Mini Projects/ Case Study – PYTHON/R