PUNE INSTITUTE OF COMPUTER TECHNOLOGY
DHANKAWADI, PUNE – 43.

## SCHEDULE OF LAB EXPERIMENTS

**Date:** 15/01/2024

**DEPARTMENT**: Computer Engineering **CLASS:** T.E

**ACADEMIC YEAR**: 2023-24 **SEMESTER:** II

**SUBJECT:**Data Science and Big Data Analytics Lab (310256)

| LAB EXPT. NO | PROBLEM STATEMENT | LAST DATE FOR COMPLETION |
|---|---|---|
| | **GROUP A** | |
| 1 | **Data Wrangling, I**<br>Perform the following operations using Python on any open-source dataset (e.g., data.csv)<br>1. Import all the required PythonLibraries.<br>2. Locate an open-source data from the web (e.g. https://www.kaggle.com). Provide a clear description of the data and its source (i.e., URL of the web site).<br>3. Load the Dataset into pandas' dataframe.<br>4. Data Preprocessing: check for missing values in the data using pandas isnull(), describe() function to get some initial statistics. Provide variable descriptions. Types of variables etc. Check the dimensions of the dataframe.<br>5. Data Formatting and Data Normalization: Summarize the types of variables by checking the data types (i.e., character, numeric, integer, factor, and logical) of the variables in the data set. If variables are not in the correct data type, apply proper typeconversions.<br>6. Turn categorical variables into quantitative variables inPython.<br>In addition to the codes and outputs, explain every operation that you do in the above steps and explain everything that you do to import/read/scrape the data set. | 29 Dec 2023 |
| 2 | **Data Wrangling II**<br>Create an "Academic performance" dataset of students and perform the following operations using Python. | 05 Jan 2024 |

| | | |
|---|---|---|
| | 1. Scan all variables for missing values and inconsistencies. If there are missing values and/or inconsistencies, use any of the suitable techniques to deal withthem.<br>2. Scan all numeric variables for outliers. If there are outliers, use any of the suitable techniques to deal withthem.<br>3. Apply data transformations on at least one of the variables. The purpose of this transformation should be one of the following reasons: to change the scale for better understanding of the variable, to convert a non-linear relation into a linear one, or to decrease the skewness and convert the distribution into a normaldistribution.<br>Reason and document your approach properly. | |
| 3 | **Descriptive Statistics - Measures of Central Tendency and variability**<br>Perform the following operations on any open-source dataset (e.g., data.csv)<br>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. For example, if your categorical variable is age groups and quantitative variable is income, then provide summary statistics of income grouped by the age groups. Create a list that contains a numeric value for each response to the categoricalvariable.<br>2. Write a Python program to display some basic statistical details like percentile, mean, standard deviation etc. of the species of 'Iris-setosa', 'Iris-versicolor' and 'Iris- versicolor' of iris.csvdataset.<br>Provide the codes with outputs and explain everything that you do in this step. | 12 Jan 2024 |
| 4 | **Data Visualization I**<br>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 thedata.<br>2. Writeacodetocheckhowthepriceoftheticket(columnname:'fare')foreachpassenger is distributed by plotting a histogram.<br>The objective is to predict the value of prices of the house using the given features. | 19 Jan 2024 |
| 5 | **Data Visualization II**<br>1. Use the inbuilt dataset 'titanic' as used in the above problem. Plot a box plot for distribution of age with respect to each | 29 Jan 2024 |

| | | |
|---|---|---|
| | gender along with the information about whether they survived or not. (Column names: 'sex' and'age') | |
| 6 | **Data Visualization III**<br>DownloadtheIrisflowerdatasetorany otherdatasetinto a DataFrame.(e.g., https://archive.ics.uci.edu/ml/datasets/Iris).<br>Scan the dataset and give the inferenceas:<br>1. List down the features and their types (e.g., numeric, nominal) available in thedataset.<br>2. Create a histogram for each feature in the dataset to illustrate the featuredistributions.<br>3. Create a box plot for each feature in thedataset.<br>Compare distributions and identifyoutliers | 05 Feb 2024 |
| 7 | **Text Analytics**<br>1. Extract Sample document and apply following document preprocessing methods:Tokenization, POS Tagging, stop words removal, Stemming andLemmatization.<br>2. Create representation of document by calculating TermFrequency and InverseDocumentFrequency. | 12 Feb 2024 |
| 8 | **Data Analytics I**<br>Create a Linear Regression Model using Python/R to predict home prices using Boston Housing Dataset.<br>(https://www.kaggle.com/c/boston-housing).<br>The Boston Housing dataset contains information about various houses in Boston through different parameters. There are 506 samples and 14 feature variables in this dataset. | 20 Feb 2024 |
| 9 | **Data Analytics II**<br>1. Implement logistic regression using Python/R to performclassification on Social_Network_Ads.csv dataset.<br>2. ComputeConfusionmatrixtofindTP,FP,TN,FN,Accuracy,Errorrate,Precision,<br>Recall on the given dataset. | 27 Feb 2024 |
| 10 | **Data Analytics III**<br>1. Implement Simple Naïve Bayes classification algorithm using Python/R on iris.csv dataset.<br>2. ComputeConfusionmatrixtofindTP,FP,TN,FN,Accuracy,Errorrate,Precision,<br>Recall on the given dataset | 4 Mar 2024 |
| | **Group B- Big Data Analytics – JAVA/SCALA** | |
| 11 | Write a code in JAVA for a simple Word Count application that counts the number of occurrences of each word in a given input set using the | 11 Mar 2024 |

| | | |
|---|---|---|
| | Hadoop Map-Reduce framework on local-standalone set-up. | |
| 12 | Locate dataset (e.g., sample_weather.txt) for working on weather data which reads the text input files and finds average for temperature, dew point and wind speed using the Hadoop Map-Reduce framework on local-standalone set-up. | 18 Mar 2024 |
| 13 | Write a simple program in SCALA using Apache Spark framework | 26 Mar 2024 |
| | **Group C- Mini Projects/ Case Study – PYTHON/R (Any TWO Mini Project)**<br>**(Students will select one mini project from 14,15,16)** | |
| 14 | Use the following dataset and classify tweets into positive and negative tweets. https://www.kaggle.com/ruchi798/data-science-tweets | 10 Apr 2024 |
| 15 | Develop a movie recommendation model using the scikit-learn library in python.<br>Refer dataset<br>https://github.com/rashida048/Some-NLP-Projects/blob/master/movie_dataset.csv | |
| 16 | Use the following covid_vaccine_statewise.csv dataset and perform following analytics on the given dataset<br>https://www.kaggle.com/sudalairajkumar/covid19-in-india?select=covid_vaccine_statewise.csv<br>a. Describe thedataset<br>b. Number of persons state wise vaccinated for first dose inIndia<br>c. Number of persons state wise vaccinated for second dose inIndia<br>d. Number of Malesvaccinated<br>d. Number of females vaccinated | |
| 17 | Write a case study to process data driven for Digital Marketing **OR** Health care systems with Hadoop Ecosystem components as shown. (Mandatory)<br>● HDFS: Hadoop Distributed FileSystem<br>● YARN: Yet Another ResourceNegotiator<br>● MapReduce: Programming based DataProcessing<br>● Spark: In-Memory data processing<br>● PIG, HIVE: Query based processing of dataservices<br>● HBase: NoSQL Database (Provides real-time reads andwrites)<br>● Mahout, Spark MLLib: (Provides analytical tools) Machine Learning algorithm libraries<br>Solar, Lucene: Searching andIndexing | 19 Apr 2024 |
| | **Question -Answer session with students about all above experiments** | **At the end of term** |

**Head of Department**
**Dr. G V Kale**

**Subject Coordinator**
**Mrs. P. P. Joshi**