Savitribai Phule Pune University
Third Year of Computer Engineering (2019 Course)
310256: Data Science and Big Data Analytics Laboratory

Home

Teaching Scheme	Credit Scheme:	Examination Scheme and Marks
Practical: 04 Hours/Week	**02**	**Term work: 50 Marks** **Practical: 25 Marks**

Companion Course: Data Science and Big Data Analytics (310251)

Course Objectives:

- To understand principles of Data Science for the analysis of real time problems
- To develop in depth understanding and implementation of the key technologies in Data Science and Big Data Analytics
- To analyze and demonstrate knowledge of statistical data analysis techniques for decision-making
- To gain practical, hands-on experience with statistics programming languages and Big Data tools

Course Outcomes:

On completion of the course, learners will be able to

CO1: Apply principles of Data Science for the analysis of real time problems
CO2: Implement data representation using statistical methods
CO3: Implement and evaluate data analytics algorithms
CO4: Perform text preprocessing
CO5: Implement data visualization techniques
CO6: Use cutting edge tools and technologies to analyze Big Data

Guidelines for Instructor's Manual

The instructor‘s manual is to be developed as a reference and hands-on resource. It should include prologue (about University/program/ institute/ department/foreword/ preface), curriculum of the course, conduction and Assessment guidelines, topics under consideration, concept, objectives, outcomes, set of typical applications/assignments/ guidelines, and references.

Guidelines for Student's Laboratory Journal

The laboratory assignments are to be submitted by student in the form of journal. Journal consists of Certificate, table of contents, and handwritten write-up of each assignment (Title, Date of Completion, Objectives, Problem Statement, Software and Hardware requirements, Assessment grade/marks and assessor's sign, Theory- Concept in brief, algorithm, flowchart, test cases, Test Data Set(if applicable), mathematical model (if applicable), conclusion/analysis. Program codes with sample output of all performed assignments are to be submitted as softcopy. As a conscious effort and little contribution towards Green IT and environment awareness, attaching printed papers as part of write-ups and program listing to journal must be avoided. Use of DVD containing students programs maintained by Laboratory In-charge is highly encouraged. For reference one or two journals may be maintained with program prints in the Laboratory.

Guidelines for Laboratory /Term Work Assessment

Continuous assessment of laboratory work should be based on overall performance of Laboratory assignments by a student. Each Laboratory assignment assessment will assign grade/marks based on parameters, such as timely completion, performance, innovation, efficient codes, punctuality and

Guidelines for Practical Examination

Problem statements must be decided jointly by the internal examiner and external examiner. During practical assessment, maximum weightage should be given to satisfactory implementation of the problem statement. Relevant questions may be asked at the time of evaluation to test the student‘s understanding of the fundamentals, effective and efficient implementation. This will encourage, transparent evaluation and fair approach, and hence will not create any uncertainty or doubt in the minds of the students. So, adhering to these principles will consummate our team efforts to the promising start of student's academics.

Guidelines for Laboratory Conduction

The instructor is expected to frame the assignments by understanding the prerequisites, technological aspects, utility and recent trends related to the topic. The assignment framing policy need to address the average students and inclusive of an element to attract and promote the intelligent students. Use of open source software is encouraged. Based on the concepts learned. Instructor may also set one assignment or mini-project that is suitable to respective branch beyond the scope of syllabus.

Set of suggested assignment list is provided in groups- A and B. Each student must perform 13 assignments (10 from group A, 3 from group B), 2 mini project from Group C

Operating System recommended :- 64-bit Open source Linux or its derivative

Programming tools recommended: - JAVA/Python/R/Scala

Virtual Laboratory:

- "Welcome to Virtual Labs - A MHRD Govt of india Initiative"
- http://cse20-iiith.vlabs.ac.in/List%20of%20Experiments.html?domain=Computer%20Science

Suggested List of Laboratory Experiments/Assignments

Assignments from all Groups (A,B,C) are compulsory.

Sr. No.	Group A : Data Science
1.	**Data Wrangling, I** Perform the following operations using Python on any open source dataset (e.g., data.csv) 1. Import all the required Python Libraries. 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). 3. Load the Dataset into pandas dataframe. 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 data frame. 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 type conversions. 6. Turn categorical variables into quantitative variables in Python. 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.
2.	**Data Wrangling II** Create an "Academic performance" dataset of students and perform the following operations using Python. 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 with them. 2. Scan all numeric variables for outliers. If there are outliers, use any of the suitable techniques to deal with them. 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 normal distribution. Reason and document your approach properly.

3.	**Descriptive Statistics - Measures of Central Tendency and variability** Perform the following operations on any open source dataset (e.g., 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. 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 categorical variable. 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.csv dataset. Provide the codes with outputs and explain everything that you do in this step.
4.	**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). 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. The objective is to predict the value of prices of the house using the given features.
5.	**Data Analytics II** 1. Implement logistic regression using Python/R to perform classification on Social_Network_Ads.csv dataset. 2. Compute Confusion matrix to find TP, FP, TN, FN, Accuracy, Error rate, Precision, Recall on the given dataset.
6.	**Data Analytics III** 1. Implement Simple Naïve Bayes classification algorithm using Python/R on iris.csv dataset. 2. Compute Confusion matrix to find TP, FP, TN, FN, Accuracy, Error rate, Precision, Recall on the given dataset.
7.	**Text Analytics** 1. Extract Sample document and apply following document preprocessing methods: Tokenization, POS Tagging, stop words removal, Stemming and Lemmatization. 2. Create representation of document by calculating Term Frequency and Inverse Document Frequency.
8.	**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. 2. Write a code to check how the price of the ticket (column name: 'fare') for each passenger is distributed by plotting a histogram.
9.	**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') 2. 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. (e.g., https://archive.ics.uci.edu/ml/datasets/Iris). Scan the dataset and give the inference as: 1. List down the features and their types (e.g., numeric, nominal) available in the dataset. 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. 4. Compare distributions and identify outliers.
	Group B- Big Data Analytics – JAVA/SCALA (Any three)
1.	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.
2.	Design a distributed application using MapReduce which processes a log file of a system.
3.	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.
4.	Write a simple program in SCALA using Apache Spark framework
	Group C- Mini Projects/ Case Study – PYTHON/R (Any TWO Mini Project)
1.	Write a case study on Global Innovation Network and Analysis (GINA). Components of analytic plan are 1. Discovery business problem framed, 2. Data, 3. Model planning analytic technique and 4. Results and Key findings.
2.	Use the following dataset and classify tweets into positive and negative tweets. https://www.kaggle.com/ruchi798/data-science-tweets
3.	Develop a movie recommendation model using the scikit-learn library in python. Refer dataset https://github.com/rashida048/Some-NLP-Projects/blob/master/movie_dataset.csv
4.	Use the following covid_vaccine_statewise.csv dataset and perform following analytics on the given dataset https://www.kaggle.com/sudalairajkumar/covid19-in-india?select=covid_vaccine_statewise.csv a. Describe the dataset b. Number of persons state wise vaccinated for first dose in India c. Number of persons state wise vaccinated for second dose in India d. Number of Males vaccinated d. Number of females vaccinated
5.	Write a case study to process data driven for Digital Marketing **OR** Health care systems with Hadoop Ecosystem components as shown. (Mandatory) ● HDFS: Hadoop Distributed File System ● YARN: Yet Another Resource Negotiator ● MapReduce: Programming based Data Processing ● Spark: In-Memory data processing ● PIG, HIVE: Query based processing of data services ● HBase: NoSQL Database (Provides real-time reads and writes) ● Mahout, Spark MLLib: (Provides analytical tools) Machine Learning algorithm libraries ● Solar, Lucene: Searching and Indexing

Reference Books :

1. Chirag Shah, "A Hands-On Introduction To Data Science", Cambridge University Press,(2020), ISBN : ISBN 978-1-108-47244-9.
2. Wes McKinney, "Python for Data Analysis", O' Reilly media, ISBN : 978-1-449-31979-3.
3. "Scikit-learn Cookbook", Trent hauk,Packt Publishing, ISBN: 9781787286382
4. R Kent Dybvig, "The Scheme Programming Language", MIT Press, ISBN 978-0-262-51298-5.
5. Jenny Kim, Benjamin Bengfort, "Data Analytics with Hadoop", OReilly Media, Inc.
6. Jake VanderPlas, "Python Data Science Handbook" https://tanthiamhuat.files.wordpress.com/2018/04/pythondatasciencehandbook.pdf
7. Gareth James, "An Introduction to Statistical Learning" https://www.ime.unicamp.br/~dias/Intoduction%20to%20Statistical%20Learning.pdf
8. Cay S Horstmann, "Scala for the Impatient", Pearson, ISBN: 978-81-317-9605-4,
9. Alvin Alexander, "Scala Cookbook", O'Reilly, SPD,ISBN: 978-93-5110-263-2

References :

- https://www.simplilearn.com/data-science-vs-big-data-vs-data-analytics-article
- https://hadoop.apache.org/docs/current/hadoop-mapreduce-client/hadoop-mapreduce-client-core/MapReduceTutorial.html
- https://www.edureka.co/blog/hadoop-ecosystem
- https://www.edureka.co/blog/mapreduce-tutorial/#mapreduce_word_count_example
- https://github.com/vasanth-mahendran/weather-data-hadoop
- https://spark.apache.org/docs/latest/quick-start.html#more-on-dataset-operations
- https://www.scala-lang.org/

MOOCs Courses link:

- https://nptel.ac.in/courses/106/106/106106212/
- https://onlinecourses.nptel.ac.in/noc21_cs33/preview
- https://nptel.ac.in/courses/106/104/106104189/
- https://onlinecourses.nptel.ac.in/noc20_cs92/preview

@The CO-PO Mapping Matrix

PO/CO	PO1	PO2	PO3	PO4	PO5	PO6	PO7	PO8	PO9	PO10	PO11	PO12
CO1	2	2	2	2	2	2	-	-	-	-	3	-
CO2	2	2	2	2	3	-	-	-	-	-	-	-
CO3	2	2	2	-	2	-	-	-	-	-	-	-
CO4	2	2	2	2	2	2	-	-	-	-	-	-
CO5	2	2	2	2	2	2	-	-	-	-	-	-
CO6	2	2	2	2	2	2	-	-	-	-	-	-
CO7	2	2	2	2	3	2	-	-	-	-	-	-
CO8	2	2	2	2	3	2	-	-	-	-	3	-