SNJB's Late Sau K.B. Jain COE, Chandwad	
Department of Computer Engineering	
Academic Year: 2021-22 Class:TE Sem: I	
Name of the Subject: Data Science and Big Data Analytics Laboratory Subject Code: 310251	
Assignment No 1	1. Locate open **Iris dataset from the URL** csv_url = 'https://archive.ics.uci.edu/ml/machine-learning-databases/iris/iris.data' 2. Perform the following operation on dataset a) Display total no of rows and column b) Display type of each column c) Sort the data in descending order , by considering column sepal.length d) Slice the data: 11 to 20 rows, and only two columns, sepal.length and Species e) rename the column Species to Type
Assignment No 2	Consider the given **dataset StudentsPerformanceTest1** 1. Check that is there any missing values in dataframe as a whole 2. is there any missing values across each column 3. count of missing values across each column 4. count row wise missing value 5. count of missing values of a gender column. 6. groupby count of missing values of a column , consider column gender and score 7. replace the missing value of score column with average value of the column
Assignment No 3	1. Locate open **Iris dataset** from the URL csv_url = 'https://archive.ics.uci.edu/ml/machine-learning-databases/iris/iris.data' 2. Perform the label encoding , by considering Species as target variable.
Assignment No 4	1. Locate open **Iris dataset** from the URL csv_url = 'https://archive.ics.uci.edu/ml/machine-learning-databases/iris/iris.data' 2. Perform the One Hot encoding , by considering Species as target variable.
Assignment No 5	1. Locate open **Iris dataset** from the URL csv_url = 'https://archive.ics.uci.edu/ml/machine-learning-databases/iris/iris.data' 2. Perform the Dummy Variable Encoding , by considering Species as target variable.
Assignment No 6	1. Creation of Dataset " StudentsPerformance" using Microsoft Excel. **The features of dataset are**: Math_Score, Reading_Score, Writing_Score, Placement_Score, Club_Join_Date **Range of Values:** Math_Score [60-80], Reading_Score[75-,95], ,Writing_Score [60,80], Placement_Score[75-100], Club_Join_Date [2018-2021]. **Response VAriable :** Placement Score **In 20% data, fill the impurities.**
Assignment No 7	1. Load the **Academic Performance** dataset in data frame object. 2. Check null values in the dataset. 3. Check missing values in dataset and replace the null values with standard null value NaN 4. Replace the missing value of Math Score with Mean Value 5. Replace the missing value of Reading Score with standard deviation 6. Replace the missing value of place with common value "Nashik"
Assignment No 8	1. Load the **Academic Performance** dataset in data frame object. 2. Check null values in the dataset. 3. Count the number of null values in complete data set (Hint: eplace the null values with standard null value NaN) 4. Dropping rows with at least 1 null value 5. Dropping rows if all values in that row are missing 6. Dropping columns with at least 1 null value. 7. Dropping Rows with at least 1 null value in CSV file
Assignment No 9	1. Load the **demo dataset** in dataframe object df 2. Detect the outlier using BoxPlot. 3. Handle the outlier using Quantile based flooring and capping (Hint: the outlier is capped at a certain value above the 90th percentile value or floored at a factor below the 10th percentile value)
Assignment No 10	1. Load the **demo dataset** in dataframe object df 2. Detect the outlier using ScatterPlot 3. Handle the outlier using Quantile based flooring and capping (Hint: the outlier is capped at a certain value above the 90th percentile value or floored at a factor below the 10th percentile value)
Assignment No 11	1. Load the **demo dataset** in dataframe object df 2. Detect the outlier using Z-score 3. replace the outliers with the median value.
Assignment No 12	1. Load the **demo dataset** in dataframe object df 2. Detect the outlier using Inter Quantile Range(IQR) 3. remove the outliers from the dataset.
Assignment No 13	1. Load the **MallCustomer dataset** in dataframe object df 2. Display summary statistics (mean, median, minimum, maximum, standard deviation) for a dataset for each column 3. Display Measures of Dispersion (Mean Absolute Deviation, Variance, Standard Deviation, Range, Quartiles, Skewness) 4. if your categorical variable is age groupsa nd quantitative variable is income, then provide summary statistics (minimum and maximum) of income grouped by the age groups.
Assignment No 14	Create a Linear Regression Model using Python to predict home prices using **Boston Housing Dataset**. The objective is to predict the value of prices of the house using the given features.

Assignment No 15	1. Implement logistic regression using Python 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..
Assignment No 16	1. Implement Simple N**aïve 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.
Assignment No 17	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.
Assignment No 18	1. Use the inbuilt **dataset 'titanic'**. 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.
Assignment No 19	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.
Assignment No 20	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 box plot for each feature in the dataset. 4. Compare distributions and identify outliers.
Assignment No 21	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 Hadoop Map-Reduce framework
Assignment No 22	Design a distributed application using MapReduce which processes a log file of a system
Assignment No 23	Write a Scala program to find a number is zero, positive or negative and write a scala program to print your name
Assignment No 24	Write a Scala Program to find largest number among two numbers and write a scala program to print your name