PUNE INSTITUTE OF COMPUTER TECHNOLOGY

DHANKAWADI, PUNE –43

LIST OF LAB ASSIGNMENTS

ACADEMIC YEAR: 2020- 2021

Honours* in Data Science	Date: 09/12/2020
Class: T.E (Honours* in Data Science)	Semester: I
Subject: Data Science and Visualization Lab	Examination scheme: TW-50

Data Science and Visualization Lab

Data Science and Visualization Lab assignment List

Sr.No	Problem Statement	
A1	Download the Iris flower dataset or any other dataset into a DataFrame. (eg https://archive.ics.uci.edu/ml/datasets/Iris). Use Python and Perform following – a) How many features are there and what are their types (e.g., numeric, nominal)? b) Compute and display summary statistics for each feature available in the dataset. (eg. minimum value, maximum value, mean, range, standard deviation, variance and percentiles c) Data Visualization-Create a histogram for each feature in the dataset to illustrate the feature distributions. Plot each histogram. d) Create a boxplot for each feature in the dataset. All of the boxplots should be combined into a single plot. Compare distributions and identify outliers.	
A2	Access an open source dataset “Titanic”. Apply pre-processing techniques on the raw dataset.	
A3	Build training and testing dataset of assignment 2 to predict the probability of a survival of a person based on gender, age and passenger-class. Use Naive Bayes classification algorithm to predict the class of passanger.	
A4	Download Abalone dataset. (URL: http://archive.ics.uci.edu/ml/datasets/Abalone)	

Data set has total 8 Number of Attributes.

Sr.No.	Name	Data Type	Measurement Unit	Description
1.	Sex	nominal	--	M, F, and I (infant)
2.	Length	continuous	mm	Longest shell measurement
3.	Diameter	continuous	mm	perpendicular to length
4.	Height	continuous	mm	with meat in shell
5.	Whole weight	continuous	grams	whole abalone
6.	Shucked weight	continuous	grams	weight of meat
7.	Viscera weight	continuous	grams	gut weight (after bleeding)
8.	Shell weight	continuous	grams	after being dried
9.	Rings	integer	--	(age/class of abalone)

a) Load the data from data file
b) Explore the shape of dataset
c) Summarize the properties in the training dataset.
d) Check the dataset for any missing values, impute the missing values and also print out the correlation matrix.
e) Split data into train, test sets
f) Predict the age of abalone from physical measurements using linear regression.
g) Plot scatterplot of real data points and regression line.
h) Display the coefficients & intercept ,accuracy score,Mean Squared Error (MSE) and RMSE.

A5

Use the dataset in assignment 4 (Abalone dataset).
a) Load the data from data file
b) Explore the shape of dataset
c) Summarize the properties in the training dataset. Write findings from column description.

d) Check the dataset for any missing values, impute the missing values and also print out the correlation matrix.
e) Split data into train, test sets
f) Predict ring class as classification problem using Naive Bayes

	and Decision Tree Classifier g) Calculate the accuracy score of the two models for both training and test data set. h) Display confusion matrix i) Display the classification report j) Compare the two models based on accuracy score and classification report and give your reasoning on which is the best model in this case.	
A6	Use Netflix Movies and TV Shows dataset from Kaggle and perform following operations : 1. Make a visualization showing the total number of movies watched by children 2. Make a visualization showing the total number of standup comedies 3. Make a visualization showing most watched shows. 4. Make a visualization showing highest rated show Make a dashboard (DASHBOARD A) containing all of these above visualizations.	