Lab Manual for Data Science(Data Processing and Presentation)

By
Bikash C. Singh, Ph.D.
Dept. of ICT
Islamic University, Kushtia, Bangladesh

Code:ICE 412
Full Marks: 100
Credit: 3.0

This lab focuses on the basic concepts of data science with the intention of sharing ideas, knowledge to students so that they can gather enough knowledge and capable to design new models/approaches in this area. With this aim, students will learn how to do preprocessing steps on dataset for removing missing values, outliers etc. Moreover, this lab will also introduce the ways to find out the correlation coefficient among various features of dataset so as students can learn how to find out the most significant features of a dataset for building a machine learning model. More precisely, students are going to experiments different types of machine learning techniques such as supervised learning (e.g., NB(Naive Bayes), DT (Decision Tree), k-NN (k-nearest neighbours), SVM (Support vector machine) etc.), unsupervised learning (e.g., K-mean), semi-supervised (e.g., EM), and Regression models (e.g., Simple linear regression, Multiple linear regression, Polynomial regression etc.).

We will use *Python* programming language for coding.

Note: Practical Report Required!

For this lab, you are expected to write up a concise report on what you did. The main thing is to convey your understanding of each of the steps taken. Any questions asked during the procedure text should be answered and you should provide a summary at the end.

Report format: Flexible (MS Word doc or Latex are ok)

Length: Max 2 pages

Delivery: Printout

Deadline: 1 week after lab date

Serial No.	Experiments	
	Preprocessing steps in Dataset Do the following functions on the dataset “Loan_training.csv” and “Loan_testing.csv”.	**Last date for assignment submission**
1.	Find the missing values of each columns	**31 August, 2019**
2.	Replace the missing values of the columns named ‘LoanAmount’ by considering mean value	
3.	After doing the function mentioned in 2, drop all missing values from “Loan_training.csv” dataset.	
4.	Detect and drop all outliers using IQR (Box plot) methods from “Loan.training.csv”	
5.	Detect and drop all outliers using Z-score methods from “Loan.training.csv”	
6.	Find the Pearson’s correlation coefficient among features of “Loan_training.csv” and delete the highly correlated features from “Loan_training.csv”.	
7.	Find the Spearman’s correlation coefficient among features of “Loan_training.csv” and delete the highly correlated features from “Loan_training.csv”.	
	Build machine learning model Do the following functions on the dataset “Loan_training.csv” and “Loan_testing.csv”.	
8.	Build k-NN(k-nearest Neighbours) supervised machine learning model using Loan_training dataset. Find the accuracy and F1 Score on Loan_testing dataset.	
9.	Build DT (Decision Tree) supervised machine learning model using Loan_training dataset. Find the accuracy and F1 Score on Loan_testing dataset.	
10.	Build support vector machine(SVM) supervised machine learning model using Loan_training dataset. Find the accuracy and F1 Score on Loan_testing dataset.	
11.	Build Naive Bayes (NB) supervised machine learning model using Loan_training dataset. Find the accuracy and F1 Score on Loan_testing dataset.	
12.	Build K-mean unsupervised machine learning model using Loan_training dataset. Find the accuracy and F1 Score on Loan_testing dataset.	

13.	Build simple Linear Regression machine learning model using '*LoanAmount*', *'ApplicantIncome*' columns of Loan_training dataset. Calculate the followings values on on Loan_testing dataset i. Mean Absolute Error ii. Mean Squared Error iii. Mean Squared Error	
14.	Build multiple Linear Regression machine learning model using '*LoanAmount*', *'ApplicantIncome*' columns of Loan_training dataset. Calculate the following values on Loan_testing dataset i. Mean Absolute Error ii. Mean Squared Error iii. Root Mean Squared Error	