Fr. C. Rodrigues Institute of Technology

Computer Engineering Department

Machine Learning Lab

Teacher Incharge: Mrs. Smita Dange

Fr. C. Rodrigues Institute of Technology
Computer Engineering Department

LAB MANUAL

Course Code and Name: CSL70011 Machine Learning   Academic Year - SH2022

Name of the Faculty: Ms. Smita Dange   Branch/SEM – COMP (VII)

No	List of Practical	CO	Level	PI	PO
1	Dataset selection and Performing exploratory data analysis on selected dataset	CO1	Apply and Analyze		
2	To implement Linear Regression and Logistic Regression	CO2, CO3	Apply and Analyze		
3	To implement multivariate Linear Regression.	CO2, CO3	Apply and Analyze		
4	To implement CART	CO2, CO3	Apply and Analyze		
5	To implement SVM	CO2, CO3	Apply and Analyze		
6	To implement Graph Based Clustering	CO2, CO3	Apply and Analyze		
7	To implement DB Scan	CO2, CO3	Apply and Analyze		
8	To implement Ensemble learning (bagging/boosting)	CO2, CO3	Apply and Analyze		
9	To implement PCA	CO2, CO3	Apply and Analyze		
10	To implement LDA	CO2, CO3	Apply and Analyze		
11	Compare the all implemented models	CO4	Analyze		
12	Mini Project	CO1,CO2,CO3,CO4	Apply and Analyze		

Course Lab Outcomes

At the end of the course, the learner will be able to:

CSL70011.1 To perform data exploratory analysis of a given dataset.

CSL70011.2 To develop machine learning model

CSL70011.3 To evaluate the built machine learning model

CSL70011.4 To identify the best machine learning model for the given dataset.

MAPPING OF COURSE LAB OUTCOMES WITH PROGRAM OUTCOMES

COs	PO1	PO2	PO3	PO4	PO5	PO6	PO7	PO8	PO9	PO10	PO11	PO12
CO1	3	3			3				2			2
CO2	3	3			3				2			2
CO3	3	3			3				2			2
CO4	3	3			3				2			2

MAPPING OF COURSE LAB OUTCOMES WITH PROGRAM SPECIFIC OUTCOMES

COs	PSO1	PSO2
CO1	3	2
CO2	3	2
CO3	3	2
CO4	3	2
CO5	3	2
CO6	3	2

Fr. C. Rodrigues Institute of Technology
Computer Engineering Department

List of Assessment tools

Sr.No	Assessment tools
1.	Practical
2.	Mini Project

Mapping Tools and COs

CO-ID	Tool1	Tool2
CSL704.1	Practical-1	Mini Project
CSL704.2	Practical-2	Practical-3
CSL704.3	Practical-4	Practical-5
CSL704.4	Practical-11	Mini Project

Fr. C. Rodrigues Institute of Technology
Computer Engineering Department

LAB MANUAL

Practical 1	Dataset selection and Performing exploratory data analysis on selected dataset
Lab outcome	LO1: Learner will be able to perform data exploratory analysis of given dataset.
Technology Used	Python
Theory	Exploratory Data Analysis (EDA) is a pre-processing step to understand the data. There are numerous methods and steps in performing EDA, however, most of them are specific, focusing on either visualization or distribution, and are incomplete.

Methodology	1. Select the topic
	2. Download two different datasets for the specific topic.
	3. Perform Exploratory analysis of given dataset.
	  i) **Import and get to know the data**
	  ii) **Data Cleaning**
	    a) Check the data type
	    b) Check for the data characters' mistakes
	    c) Check for missing values and replace them
	    d) Check for duplicate rows
	    e) Statistics summary
	    f) Outliers and how to remove them
	3. Distributions and Relationship
	  a) Categorical variable distribution
	  b) Continuous variable distribution
	  c) Relationship between categorical and continuous variables
	4. Automated EDA using pandas profiling report
	
	Web Reference: https://towardsdatascience.com/exploratory-data-analysis-on-heart-disease-uci-data-set-ae129e47b323

Output	Program Code with output
Conclusion	Justify why EDA process is require and how it helps to build good machine learning model.

Practical 2	To implement Linear Regression and Logistic Regression
Lab outcome	LO2: Learner will be able to develop machine learning model LO3: Learner will be able to evaluate the built machine learning model.
Technology Used	Python

Teacher Incharge: Mrs. Smita Dange

Theory	**Linear Regression** is a machine learning algorithm based on **supervised learning**. It performs a **regression task**. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting. Different regression models differ based on – the kind of relationship between dependent and independent variables they are considering, and the number of independent variables getting used. **Logistic regression** is a process of modeling the probability of a discrete outcome given an input variable. The most common logistic regression models a binary outcome; something that can take two values such as true/false, yes/no, and so on. Multinomial logistic regression can model scenarios where there are more than two possible discrete outcomes. Logistic regression is a useful analysis method for classification problems, where you are trying to determine if a new sample fits best into a category. As aspects of cyber security are classification problems, such as attack detection, logistic regression is a useful analytic technique.
Methodology	- Write python code to deploy linear regression and logistic regression - Evaluate the model by calculating the errors, accuracy, precision, recall, f1-score
Output	Program Code with output
Conclusion	Compare linear regression and logistic regression and justify for the given problem statement which one is more suitable.

Practical 3	To implement multivariate Linear Regression.
Lab outcome	LO2: Learner will be able to develop machine learning model LO3 Learner will be able to evaluate the built machine learning model.
Technology Used	Python
Theory	Multivariate regression is a technique used to measure the degree to which the various independent variable and various dependent variables are linearly related to each other. The relation is said to be linear due to the correlation between the variables. Once the multivariate regression is applied to the dataset, this method is then used to predict the behavior of the response variable based on its corresponding predictor variables. Multivariate regression is commonly used as a supervised algorithm in machine learning, a model to predict the behavior of dependent variables and multiple independent variables.
Methodology	- Write python code to deploy linear regression and logistic regression

Teacher Incharge: Mrs. Smita Dange

	- Evaluate the model by calculating the errors, accuracy, precision, recall, f1-score
Output	Program Code with output
Conclusion	Compare with simple linear regression and justify for the given problem statement which one is more suitable.

Practical 4	To implement CART.
Lab outcome	LO2: Learner will be able to develop machine learning model LO3 Learner will be able to evaluate the built machine learning model.
Technology Used	Python
Theory	In the decision tree, the nodes are split into subnodes on the basis of a threshold value of an attribute. The CART algorithm does that by searching for the best homogeneity for the subnodes, with the help of the Gini Index criterion. The root node is taken as the training set and is split into two by considering the best attribute and threshold value. Further, the subsets are also split using the same logic. This continues till the last pure sub-set is found in the tree or the maximum number of leaves possible in that growing tree. This is also known as Tree Pruning.
Methodology	- Write python code to deploy CART - Evaluate the model by calculating the errors, accuracy, precision, recall, f1-score
Output	Program Code with output
Conclusion	Discuss the model's efficiency and gini index value.

Practical 5	To implement SVM
Lab outcome	LO2: Learner will be able to develop machine learning model LO3 Learner will be able to evaluate the built machine learning model.
Technology Used	Python
Theory	Support Vector Machine, abbreviated as SVM can be used for both regression and classification tasks. But, it is widely used in classification objectives. Support vector machine is highly preferred by many as it produces significant accuracy with less computation power. The objective of the support vector machine algorithm is to find a hyperplane in an N-dimensional space (N — the number of features) that distinctly classifies the data points.

Methodology	- Write python code to SVM - Evaluate the model by calculating the errors, accuracy, precision, recall, f1-score
Output	Program Code with output
Conclusion	Discuss the efficiency of the SVM. Mentioned about the training time required for SVM.

Practical 6	To implement Graph based clustering
Lab outcome	LO2: Learner will be able to develop machine learning model LO3 Learner will be able to evaluate the built machine learning model.
Technology Used	Python
Theory	**Clustering** is the task of dividing the population or data points into a number of groups such that data points in the same groups are more similar to other data points in the same group and dissimilar to the data points in other groups. It is basically a collection of objects on the basis of similarity and dissimilarity between them. Using MST graph based clustering is deployed. First build the MST and the delete the branch having higher weight/ or inconsistent branch or crossing defined threshold value.
Methodology	- Write python code to deploy Graph based clustering (using MST) - Use all three methods to delete the branch
Output	Program Code with output
Conclusion	Discuss the efficiency of the model. Compare the different ways of deleting the branch. Compare with k-mean

Practical 7	To implement DBSCAN.
Lab outcome	LO2: Learner will be able to develop machine learning model LO3 Learner will be able to evaluate the built machine learning model.
Technology Used	Python
Theory	Fundamentally, all clustering methods use the same approach i.e. first we calculate similarities and then we use it to cluster the data points into groups or batches. Here we will focus on **Density-based spatial clustering of applications with noise** (DBSCAN) clustering method. Clusters are dense regions in the data space, separated by regions of the lower density of points. The *DBSCAN algorithm* is based on this intuitive notion of "clusters" and "noise". The key idea is that for each point of a cluster, the neighborhood of a given radius has to contain at least a minimum number of points.
Methodology	- Write python code to deploy DBSCAN
Output	Program Code with output
Conclusion	Discuss the efficiency of the model. Compare with k-mean

Practical 8	To implement Ensemble learning (bagging/boosting)
Lab outcome	LO2: Learner will be able to develop machine learning model LO3 Learner will be able to evaluate the built machine learning model.
Technology Used	Python
Theory	
Methodology	- Write python code to deploy linear regression and logistic regression -
Output	Program Code with output
Conclusion	Discuss the efficiency of the model.

Practical 9	To implement PCA
Lab outcome	LO2: Learner will be able to develop machine learning model LO3 Learner will be able to evaluate the built machine learning model.
Technology Used	Python
Theory	
Methodology	- Write python code to deploy linear regression and logistic regression - Evaluate the model by calculating the errors, accuracy, precision, recall, f1-score
Output	Program Code with output
Conclusion	Discuss the efficiency of the model.

Practical 10	To implement LDA.
Lab outcome	LO2: Learner will be able to develop machine learning model LO3 Learner will be able to evaluate the built machine learning model.
Technology Used	Python
Theory	
Methodology	- Write python code to deploy LDA -
Output	Program Code with output
Conclusion	Discuss the efficiency of the model.

Practical 11	To compare all the model
Lab outcome	LO2: Learner will be able to develop machine learning model LO3 Learner will be able to evaluate the built machine learning model.
Technology Used	Python
Theory	
Methodology	- Write python code to deploy linear regression and logistic regression -
Output	Program Code with output
Conclusion	

Mini Project

Mini Project is a group activity. This activity helps the learner to work in team, develop team spirit and leadership quality. Under mini project students will able to identify the best machine learning model for specific problem.

Mapping with CO: CO1, CO2,CO3,CO4

CO4: To identify the best machine learning model for the given dataset.

Step 1: Forming a group: Form a group of five students.

Step 2: Analyze all the models build for given dataset.

Step 3: Identify the best model with justification.

Step 4: Develop front end for the same.

Note: Refer the template to submit Mini Project Report

Subject In-Charge

(Ms. Smita Dange)