VIT®
Vellore Institute of Technology
(Deemed to be University under section 3 of UGC Act, 1956)

LAB CYCLE SHEET

Course code: PMCA507P

Course Name: Machine Learning Lab

Programme: MCA

Faculty handling the course: Dr. Parimala M and Dr. Anitha A

School of Computer Science Engineering & Information Systems

Course Outcomes:
1. Provide solution for classification and regression approaches in real-world applications
2. Gain knowledge to combine machine learning models to achieve better results
3. Choose an appropriate clustering technique to solve real world problems
4. Realize methods to reduce the dimension of the dataset used in machine learning algorithms
5. Choose a suitable machine learning model, implement and examine the performance of the chosen model for a given real world problems

Assessment-1
CO1: Provide solution for classification and regression approaches in real-world applications
Total Marks : 10 marks

The **Diabetes prediction dataset** taken from "kaggle data repository" is a collection of medical and demographic data from patients, along with their diabetes status (positive or negative). The data includes features such as age, gender, body mass index (BMI), hypertension, heart disease, smoking history, HbA1c level, and blood glucose level. This dataset can be used to build machine learning models to predict diabetes in patients based on their medical history and demographic information. This can be useful for healthcare professionals in identifying patients who may be at risk of developing diabetes and in developing personalized treatment plans. Additionally, the dataset can be used by researchers to explore the relationships between various medical and demographic factors and the likelihood of developing diabetes.

Task 1:

Import the diabetes dataset to your platform from the link given below and perform the following

https://drive.google.com/file/d/12G6B0cEBWd69QzxzM3ZGe9CyFlE7KD8m/view?usp=sharing

1. Create a data frame and display the number of samples and features with the datatype
2. Count the number of diabetes patient who never smoke
3. Display all the statistical measure about the data frame
4. Find the number of samples having missing values
5. Print the first 50 samples from the dataset

Task 2:

1. Find the number of missing samples for each feature and if there is any missing value found, then fill the values based on the type of data(continuous/discrete)
2. Remove the duplicated sample from the dataset
3. Normalize the input feature "blood_glucose_level" to the range of 0 to 1(Use the appropriate normalization type based on the requirement)

4. Map all the categorical data to ordinal data.
5. Identify the outlier range (lower bound and upper bound) for the above dataset and list the outliers.

Task 3.1

From the above dataset, consider only two attribute in the input dataset namely "blood_glucose_level" and "diabetes" for the Task3.1 and Task 3.2. Based on the blood glucose level, the person is classified under diabetic as "1" or non-diabetic as "0". Perform the following,

1. Split the given dataset into training and testing dataset
2. Assign the weights and bias using any of the approach (formula or taking random values)
3. Write your own code for building the Linear Regression model using the training dataset and predict the target class
4. Calculate the Error deviation using test dataset by applying any measures and also find the Accuracy of the model
5. Once the model is built, predict whether a person is diabetic or not for the given blood_glucose_level =155

Task 3.2

Modify the code executed in task-3.1. Step 2, 3 & 4 in the above task can be replaced with inbuilt functions. Compare the error and accuracy in both the task. If there is any deviation, write your inference and observations in the last line of code as comment line

Task 4

Using the diabetes dataset, build the multilinear regression model and perform the following

1. Split the dataset for training and testing
2. Display the intercepts/constant values calculated
3. Calculate the accuracy of the model
4. Draw the comparison graph with y and predicted y
5. Predict the person is diabetic or not for the new input feature

"Female,36,0,0,current,32.27,6.2,220"

Task 5

Using the diabetes dataset build the logistic regression model and perform the following

1. Split the dataset for training and testing
2. Build and Test the model
3. Evaluate the model using confusion matrix and other measures. Also, Calculate the accuracy of the model
4. Draw the comparison graph with y and predicted y
5. Predict the person is diabetic or not for the new input feature **"Male,80,0,0,never,22.06,9,155"**