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-2
CO2: Gain knowledge to combine machine learning models to achieve better results
Total Marks : 10 marks

Note: For the entire task, follow the Machine Learning cycle starting from loading the dataset, splitting the dataset, training the model, testing the model, Error/accuracy measures, validating with new dataset. Wherever required visualize the model using any type of graph.

Task 2.1

1. Construct a decision tree for diabetes dataset given in Assessment-1 using Information gain.
2. Train the model and display the classification tree. Explain the decision tree in text cell
3. Evaluate the model using the test dataset
4. Print the accuracy of the model and confusion matrix for the model built
5. Predict the person is diabetic or not for the new input feature
   "Male, 80, 0, 0, never, 22.06, 9, 155"

Task 2.2

Implement the SVM algorithm on the diabetes dataset with 42 samples given below.

(i) Convert the target class into binary class which has two class labels namely '0' and '1'. Transform the value '0.5' to '0'

(ii) Draw the graph with linearly separable line that separates the two distinct classes

(iii) Predict the target class value for "HbA1c=5 and blood_glucose_level-100"

(iv) Justify the statement. Model yields 100% accuracy but it fails to predict the target class for new dataset

HbA1c_level	blood_glucose_level	diabetes
6.6	140	0
6.6	80	0
5.7	158	0
5	155	0
4.8	155	0

6.6	85	0
5.7	85	0
4.8	145	0
5	100	0.5
6.1	85	0.5
6	100	0.5
5	130	0.5
5.8	200	0.5
5	160	0.5
6.6	126	0.5
6.1	200	0.5
5.7	158	0.5
5.7	80	0.5
3.5	159	0.5
6	90	0
5	90	0
4	159	0
6.5	200	1
6.5	200	1
5.7	260	1
9	160	1
7	159	1
9	159	1
8.8	155	1
8.2	126	1
6.2	220	1
7.5	300	1
6.2	159	1
6.8	280	1
6.5	280	1
9	155	1
7.5	155	1
7	300	1
6.6	280	1
9	140	1
7.5	155	1
6.5	300	1

Task 2.3

1. Consider the preprocessed data from Task 2.2
2. Apply kNN to predict whether a person is diabetic/non-diabetic for the following input features for “HbA1c=5 and blood_glucose_level-100”
3. k value is considered as 7
4. Use the Euclidean distance to find the neighbourhood points

Task 2.4

Consider the models implemented in Task 2.2(M1) and Task 2.3(M2). Also apply linear regression model (M3) for the dataset given in Task 2.2. Apply ensemble approach on these three models and compare the results from all the three models (M1,M2,M3) with ensemble technique.