**Department of Computer Science and Engineering**

**Question Set 1**

**Amrita School of Engineering**

**Amrita Vishwa Vidyapeetham – Coimbatore**

III Year B.Tech. CSE V Sem

19CSE305 – Machine Learning

*Lab Evaluation* – II

**Date & Time:** 21th Oct 2021 & 03:00 to 04:00 pm

**Instructions**:

* The total mark for the evaluation is 30, of which 25 (5 marks for data preprocessing + 15 marks for building of models + 5 marks for the comparison of models) marks will be assessed through the questions asked here and for 5 marks, viva will be conducted upon the submission of solutions.
* The dataset to be used for this question set is “Census Income” in http://archive.ics.uci.edu/ml/datasets/Adult.
* The questions should be neatly worked out in Google Colab/Jupyter and it needs to be made sure that the python notebook (all models should be created in one single ipynb) is named as Roll\_Number\_Eval1.ipynb (e.g., CB.EN.U8CSE80563\_Eval1.ipynb).
* The original notebook along with its pdf exported version should be submitted as solution to the assignment that will be created on the day of evaluation.
* Grade will be based on completion of the required tasks and Accuracy (Performance) of your classifiers

1. **Data Preprocessing (5 marks)**
   1. Determine Data preprocessing methods to apply your classifier
      1. Missing values, if any, may be handled properly by using the statistical properties such as mean/median/mode.
      2. Outliers and duplicate observations, if any, need to be resolved.
      3. Feature scaling, if the feature values are not in similar unit/range.
   2. Transform the Adult Data Set with Correct Data Preprocessing Method for Your Classifier to Create a Training Set and Test Set with a Class Label.
2. **Model Building (15 marks)**
   1. Four machine learning classification models have to be created in the following manner:
      1. Model 1 – Logistic Regression
      2. Model 2 – Decision Tree (ID3)
      3. Model 3 – Random Forest
      4. Model 4 – K Nearest Neighbor
   2. The models can be created/trained/tested using the built-in utilities available in Scikit package.
   3. Use one single ipynb for all these models.
   4. Validate your result with your Test Set to compare the Accuracy of your models
   5. Display your result in Confusion Matrix and Calculate in Accuracy, Recall, Precision and F1
   6. Do 5-Fold Cross Validation (k= 5) Compare the accuracy of each test of the classifier. Your Overall Accuracy is Average of the five model accuracy from 5 runs of your classifier.
   7. Report your discussion, observation, findings on Your Result
3. **Model Comparison (5 marks)**
   1. The classification reports of the models created in the previous step have to be undergoing a comparative study.
   2. You inference about the result of the comparison need to be included in the ipynb as well as presented to the examiner who conducts viva for you.