**Department of Computer Science and Engineering**

**Amrita School of Engineering**

**Amrita Vishwa Vidyapeetham – Coimbatore**

**III Year B.Tech.**

**Computer Science and Engineering**

**Fifth Semester**

***19CSE305 – Machine Learning Lab Evaluation – I***

***Sec-E Date:*** *26.08.2021 (AN)* ***Duration:*** *1 Hour 15 Mins*

**Set - II**

***Instructions:***

1. The total mark for the evaluation is 30, of which 24 marks will be assessed through the questions asked here and for 6 marks, viva will be conducted upon the submission of solutions.
2. The question set consists of 6 questions (there maybe sub questions under some main questions) each carries 3 to 5 marks.
3. The questions should be neatly worked out in Google Colab/Jupyter, and it needs to be made sure that the python notebook is named as Roll\_Number\_Eval1.ipynb (e.g. CB.EN.U8CSE80409\_Eval1.ipynb)
4. 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.

**Questions:**

1. **Loading the dataset (1 x 3 = 3)** 
   1. Import all required packages and create alias for them as per your interest.
   2. Load the dataset named “Drug Classification Dataset” from the url: "https://www.kaggle.com/prathamtripathi/drug-classification” to create a dataframe (df). (If you get UnicodeDecodeError while loading, pass an extra parameter as encoding='latin-1')
   3. Display
2. Size of the dataframe.
3. Number of dependent and independent variables.
4. Datatype of each variable
5. **Basic operations with dataframe (1.5 x 2 = 3)**
6. Display descriptive statistics for all the attributes in the dataset. Provide your inference in the solution document.
7. Apply appropriate label encoding techniques to convert categorical variables to numerical variables.
8. **Rescaling mechanisms (2+2=4 marks)**
   1. Apply z-score rescaling technique. – 2marks
   2. Show the data distribution of all the attributes and provide your inference (Symmetric or skewed) – 2 marks
9. **Outlier removal (2+2=4)**
10. From Q3, display the row numbers of an outlier for any attribute (2 marks)
11. Remove the outlier and provide your inference. (2 marks)
12. **Missing value (1.5 x 2 = 3)**
    1. Create an empty value for the cell 11th row for an age.
    2. Impute the missing values for that cell by mean with respect to class label.
13. **Data Visualization (1.5 x 2 = 3)**
14. Provide your inference about correlation between two variables.
15. Also remove the highly correlated attribute if any, with proper inference.
16. **Train:Test split ( 2+2= 4)**
    1. Create two different dataframe after rescaling. One dataframe should contains only the dependent variables and another dataframe with independent variable.
    2. Split the dataset into train and test with 70:30 ratio. Note: Each class should be distributed equally.

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*All The Best\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*