

Lab - 2

PRML
AY 2020-21, Trimester - III

March 15, 2021

Deadline: March 20, 2021, 11:59 PM

Q1: A csv file has been provided to you at this [link](#). The dataset represents the mood of a student to go to class depending on the weather at IIT Jodhpur. We have been accustomed to online classes so this is to give you a feeling of attending classes in the post-COVID scenario. A Colab Notebook is [attached](#) for your reference about the stepwise procedure to solve the exercise. The marks distribution according to the tasks are as follows:

- i) Preprocessing the data. (5 Marks)
- ii) Cross-validation over the data. (5 Marks)
- iii) Training the final model after cross-validation (5 Marks)
- iv) Perform decision tree classification and calculate the prediction accuracy for the test data. (5 Marks)
- v) Plot the decision tree and the decision surface. (5 Marks)

Q2: In the previous case, the nodes are split based on entropy/gini impurity. The following [dataset](#) contains real-valued data, and the description of the dataset is available [here](#). The column to be predicted is '**Upper 95% Confidence Interval for Trend**' i.e. the last column present in the dataset using other columns as features. The marks distribution according to the tasks are as follows:

- i) Preprocessing the data. (5 Marks)
- ii) Cross-validation over the data. (5 Marks)
- iii) Training the final model after cross-validation (5 Marks)
- iv) Perform decision tree regression and calculate the squared error between the predicted and the ground-truth values for the test data. (5 Marks)
- v) Plot the decision tree and the decision surface. (5 Marks)

Please submit the necessary codes (Notebook) containing your output, and a PDF explaining and analyzing (e.g., what design choices would lead to better prediction) the steps for all the five parts in both the questions along with necessary plots/figures.

Note: No submission will be accepted after the final deadline.