Lab - 5

CSL2010: Introduction to Machine Learning

AY 2021-22, Semester – I

General Instructions:

1. **Prepare separate Python code files for each task and name them as task1.py and task2.py, respectively.**
2. **Also, provide your colab file link in the report. Make sure that the file is sharable.**
3. **Put both the codes in a folder named <IML_Lab5_YourRollNo>, create a zip file, and upload in google-classroom.**
4. **Submit a single report depicting the method, results, and observations for all the tasks. There is no need to add theory behind the concepts.**
5. **Clearly, mention the assumptions you have made, if any.**
6. **You are free to use any library.**
7. **Clearly, report any resources you have used while attempting the assignment.**
8. **Any submission received in another format or after the deadline will not be evaluated.**

Task 1: Decision tree classifier **Marks: 12**

Use the dataset from here and perform the following:

1. Identify the features and target from the data.
2. Look for missing values in the data, if any, and address them accordingly.
3. Find out the ordinal/nominal/categorical data, if any, and convert them into numerical equivalent.
4. Split the dataset into 70:30, 80:20, and 90:10.
5. For reproducibility, set seed = 55 throughout.
6. Use Entropy information gain for 80:20 split and Gini-index for the rest.
7. Train a decision tree classifier and report model accuracy.
8. Prepare confusion matrix and classification report.
9. Provide a graphical visualization of the tree.
10. Comment on overfitting.

Task 2: Decision tree regressor **Marks: 8**

Use the dataset from here and perform the following:

1. Identify the features and target from the data.
2. Split dataset into training and test set using the following formula:
   Let your roll number be B20XX207, and the last three digits of your roll number be S. If S is odd split ratio is 70:30, and if it is even then, the split ratio is 80:20. In the above example, S turns out to be 207, which is odd; hence split the data in 70:30.
3. Use MSE and MAE.
4. For reproducibility, set seed = 2021.
5. If the split is 70:30, set node selection strategy as ‘random’, ‘best’ otherwise.
6. Train a decision tree regressor and report model accuracy.
7. Provide a graphical visualization of the tree.
8. Prepare confusion matrix and classification report.

Note: You are required to solve tasks 1.1, 1.2, and 1.3 within the lab itself.