Subject: 23CSE301

Lab Session: 06

Notes:

1. Please read the assignment notes carefully and comply to the guidelines provided.
2. Code should be checked into GitHub and the report to TurnItIn. Main Section (Mandatory):

Please use the data associated with your own project.

Ref: Please refer to your notes on Entropy, information gain and Decision Tree.

1. https://scikit-learn.org/stable/modules/generated/sklearn.tree.plot_tree.html
2. https://www.datacamp.com/tutorial/decision-tree-classification-python
3. https://scikit-learn.org/stable/auto_examples/tree/plot_iris_dtc.html#sphx-glr-auto-examples-tree-plot-iris-dtc-py

A1. Write a function to calculate the entropy associated with your dataset. (If your dataset contains only continuous numeric data for outcome (a regression problem), employ equal width binning and divide your data into 4 bins. Each bin may be considered as a categorical data value. Write a function for equal width binning).

$$H = -\sum_i p_i(\log_2 p_i)$$

Here, p_i refers to the probability of occurrence of each outcome value.

A2. Calculate the Gini index value for your dataset.

$$Gini = 1 - \sum_j p_j^2$$

A3. Write your own module for detecting the feature / attribute for the root note of a Decision Tree. Use Information gain as the impurity measure for identifying the root node. Assume that the features are categorical or could be converted to categorical by binning.

A4. If the feature is continuous valued for A3, use equal width or frequency binning for converting the attribute to categorical valued. The binning type should be a parameter to the function built for binning. Write your own function for the binning task. The number of bins to be created should also be passed as a parameter to the function. Use function overloading to allow for usage of default parameters if no parameters are passed.

A5. Expand the above functions to built your own Decision Tree module.

A6. Draw and visualize the decision tree constructed based on your data. (Refer above provided web sources [1] & [2] for understanding and learning on how to visualize a DT).

A7. Use 2 features from your dataset for a classification problem. Visualize the decision boundary created by your DT in the vector space. (Refer above provided web source [3] & [2] for understanding and learning on how to draw decision boundary for a DT).

Report Assignment:

1. Update your understanding of your project in the introduction section of the report.
2. Study the downloaded papers & update the literature survey section of your report.
3. Expand the methodology and results sections with outcomes of this experiments & results obtained. Please discuss your observations, inferences in results & discussion section. Please conclude the report appropriately with these experiments. Consider following points for observation analysis & inferences.