CIT651 –Introduction to Machine Learning and Statistical Data Analysis

Assignment #3
(Due on: January 15, 2019 at mid-night)

Implement an algorithm that builds a 3-level decision tree to predict the outcome of the games Liverpool played in the 2017/2018 premier league season. The training data given in the file "Training_Data.xlsx" is the outcome of all games of all other teams that ended in the win of one of the two competing teams. The decision tree should predict, based on the values of the given attributes, whether the home team or the away team of the game in which Liverpool is playing will win the game. The attributes given in the file for each game are as follows:

Home Team
Away Team
HS: Home Team Shots
AS: Away Team Shots
HST: Home Team Shots on Target
AST: Away Team Shots on Target
HF: Home Team Fouls Committed
AF: Away Team Fouls Committed
HC: Home Team Corners
AC: Away Team Corners
HY: Home Team Yellow Cards
AY: Away Team Yellow Cards
HR: Home Team Red Cards
AR: Away Team Red Cards

Given that the data is numeric, you will need to discretize the data. The root node of the tree should have 2 possible values. The nodes in the two next levels should have 3 possible values. For discretization of the root node, you can do it based on whether the value is above or below the mean value of the attribute. For the nodes of the other two levels, you can discretize the data using equal spacing discretization (uniform).
Deliverables:

- Your code.

- A text file specifying the nodes at each level of the tree that were identified using your algorithm.

- A confusion matrix for the games Liverpool played given in the testing data file "Liverpool.xlsx".

Important Notes:

- **Do not use Python built-in functions computing the decision tree or computing the entropy. You can use built-in functions that compute the majority in a list of numbers, the mean, the maximum or the minimum, if needed.**

- **The decision tree that you will implement will probably not lead to leaf nodes after only 3 levels. So, to keep it at this size, specify leaf nodes after level 3 based on the majority of class labels at each node.**

- This is an individual assignment. It is not a team assignment.