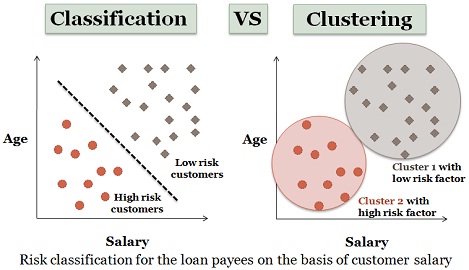
Task 1

Describe the difference between classification and clustering?

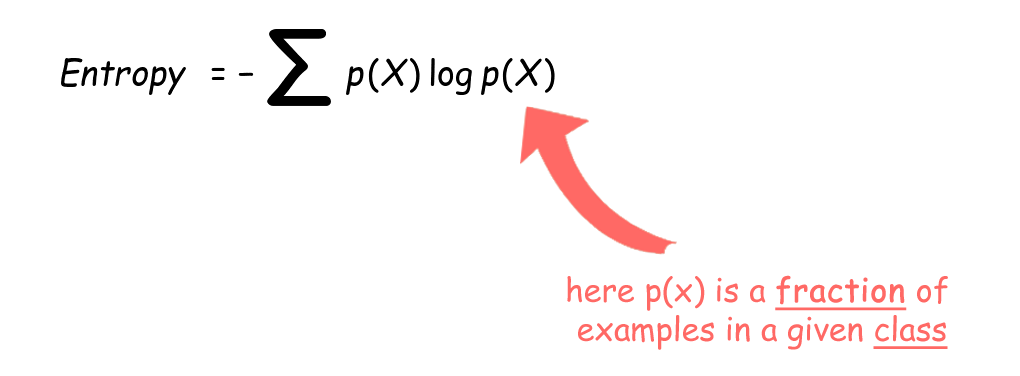
Ans: Clustering and Classification are two types of Learning Methods. Classification is supervised Learning whereas Clustering is Unsupervised Learning. In classification the predefined labels are assigned to the predictors which we tend to predict whereas in clustering labels are not defined.



Task 2

Describe what is entropy?

Entropy is measure of impurity or disorderness in the data. It also refers to the uncertainty in data.



Task 3

Describe and compare the following “feature selection measures” or called “splitting criteria”:

information gain, gain ratio, and Gini index?

Ans:

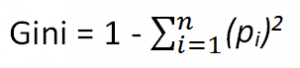
Information Gain:

Information gain refers to the quality of information given about a dataset. It can also be referred as purity of data.

Gain Ratio:

Gini Index:

Gini index or Gini impurity measures the degree of a feature variable being wrongly classified when it is randomly chosen from the rest.



Task 4

Given training instances and their attributes, construct and plot three decision trees

• ID3: information gain (a.k.a., entropy based gain)

• C4.5: gain ratio

• CART: gini index

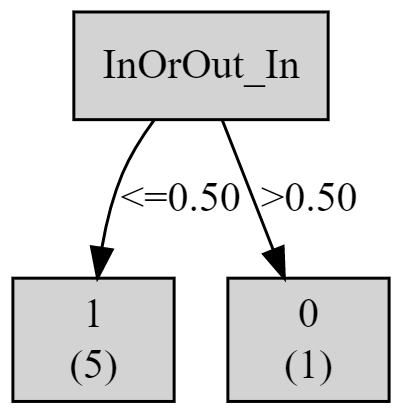
Question 1: We have

(1) 6 training instances and 6 testing instances

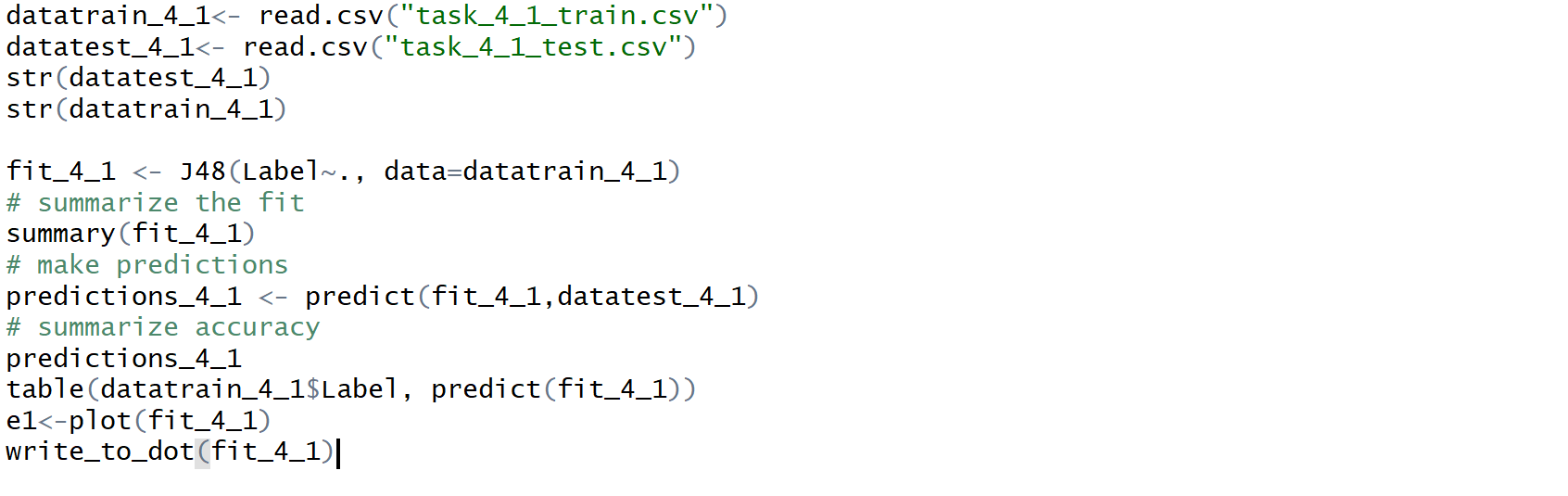
(2)3 attributes: (a) 2-value attribute (Home/Away), (b) 2-value attribute (In/Out), (c) 4-value

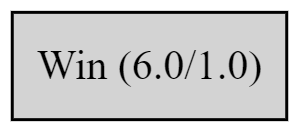
attribute (NBC/ESPN/FOX/ABC)

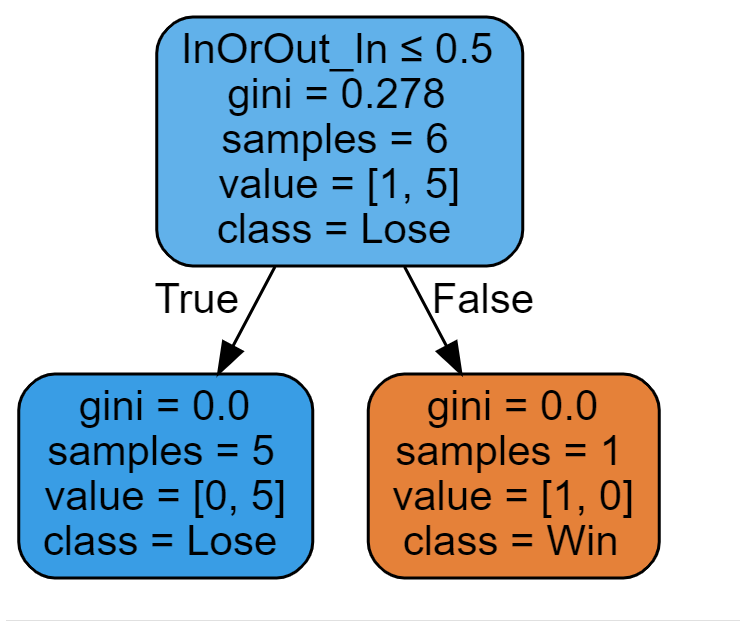
ID3:



C4.5





CART:



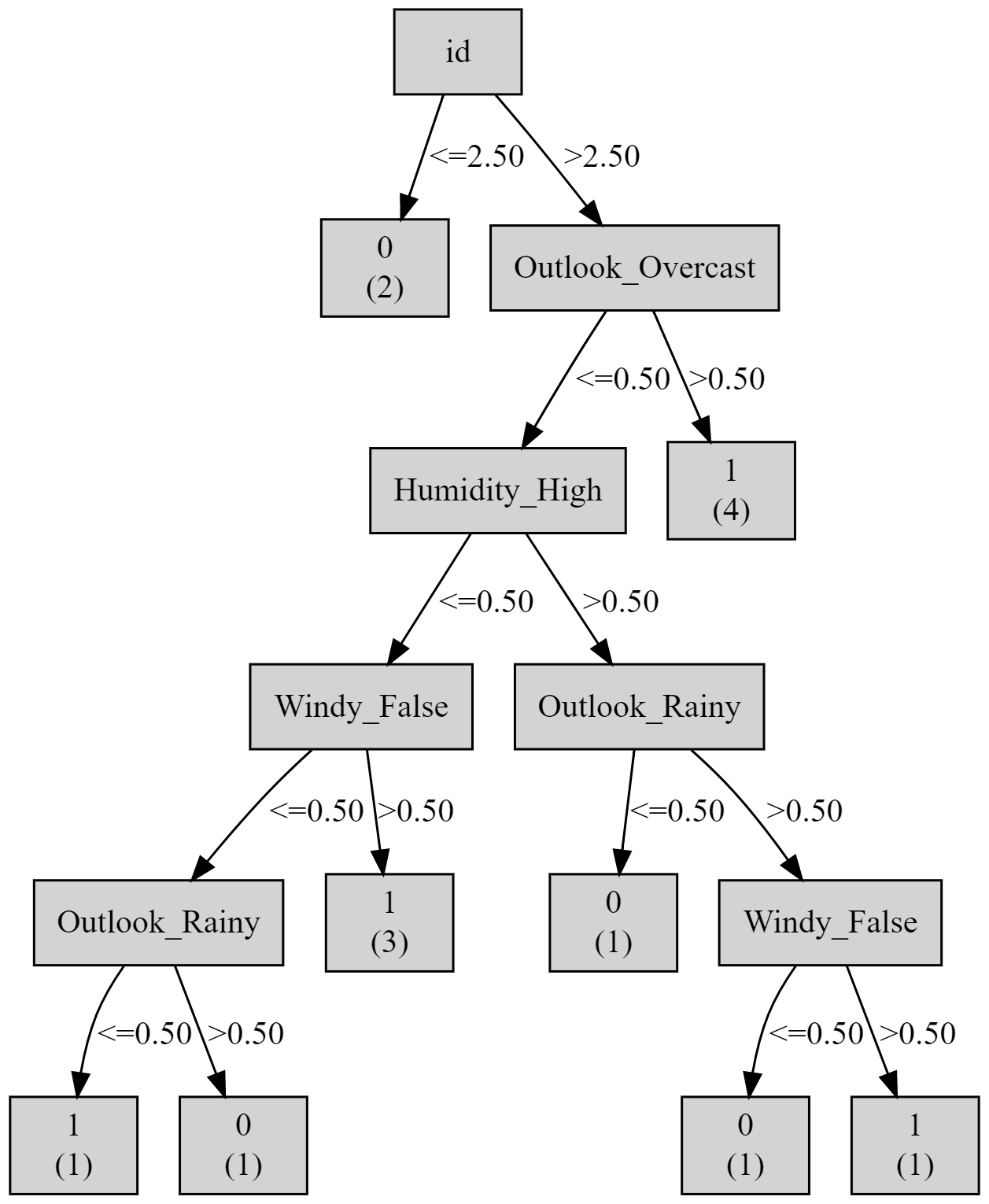
Question 2: We have

(1) 14 training instances and 1 testing instance

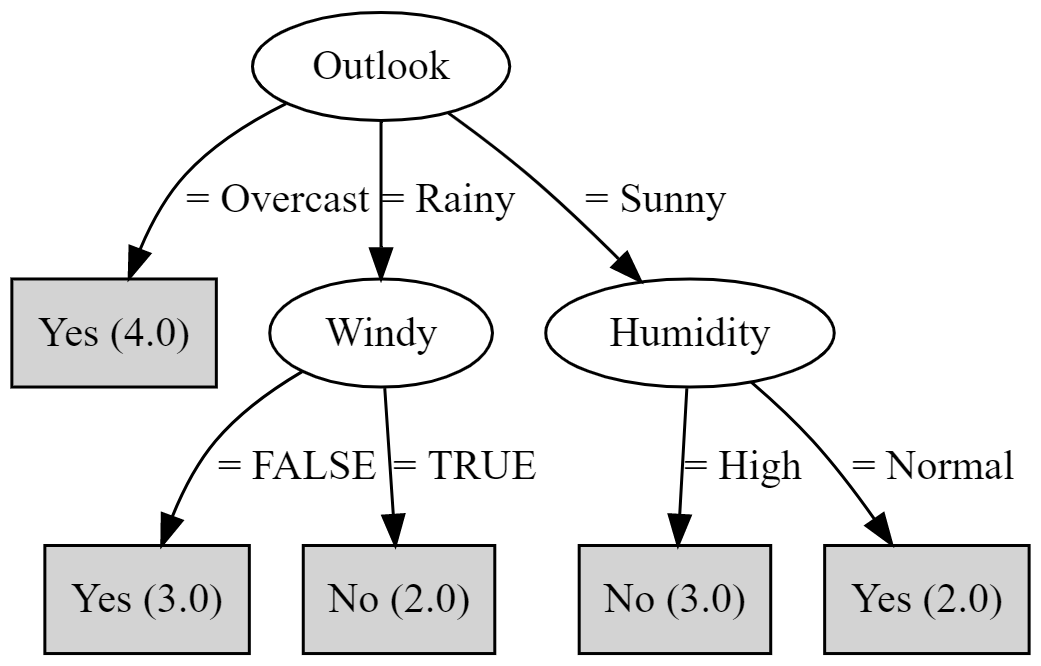
(2) 4 attributes: (a) 3-value attribute (Sunny/Overcast/Rainy), (b) 3-value attribute

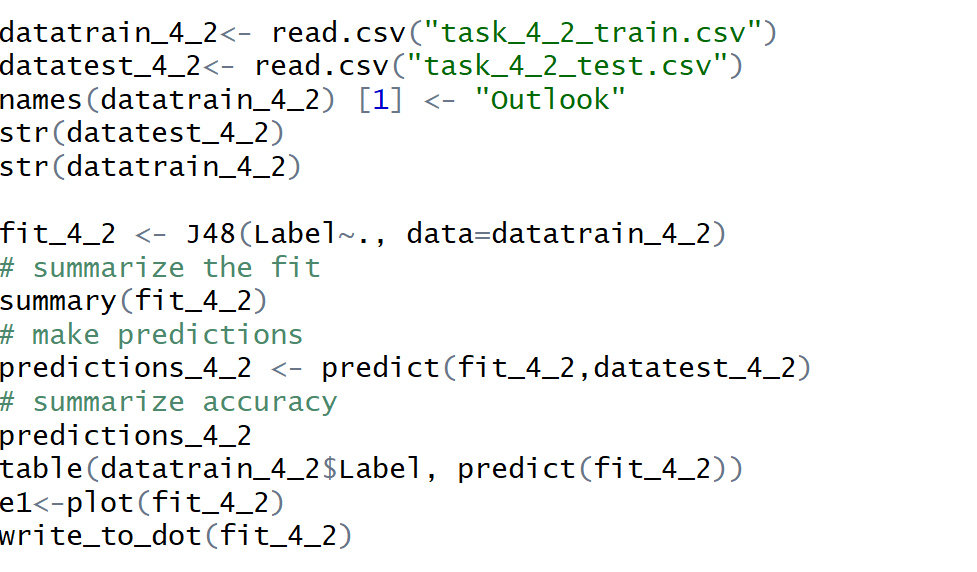
(Hot/Mild/Cool), (c) 2-value attribute (High/Normal), (d) 2-value attribute (True/False)

ID3:

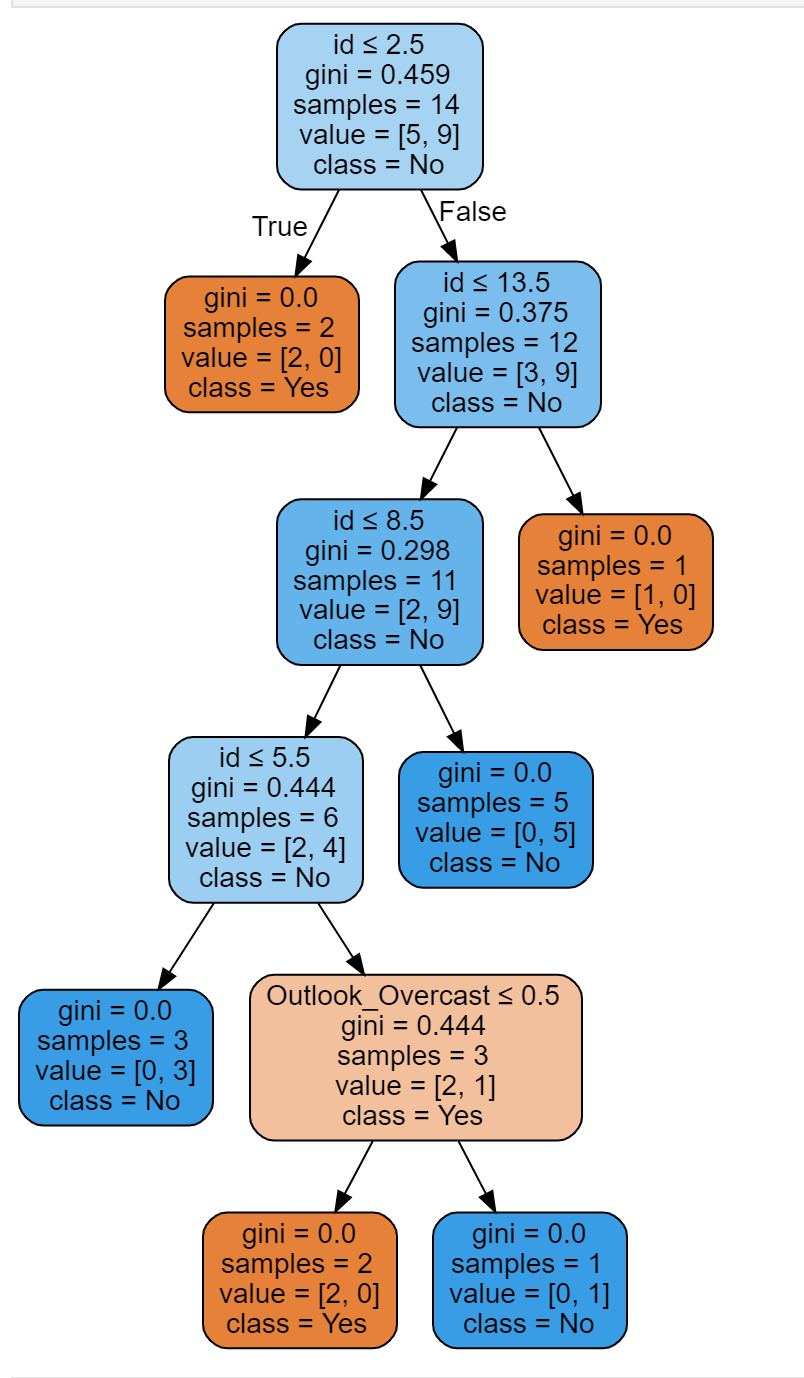


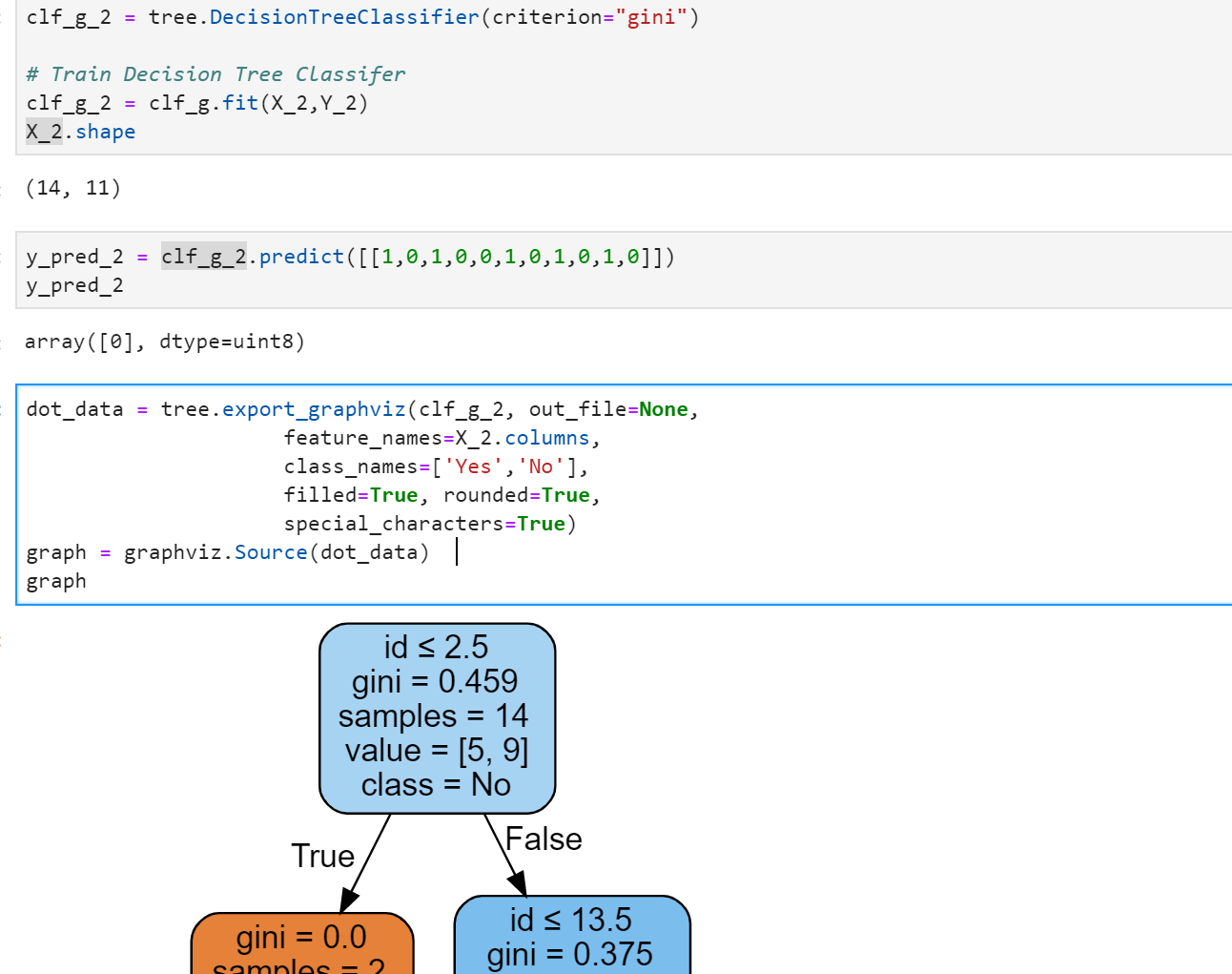
C4.5:





CART:



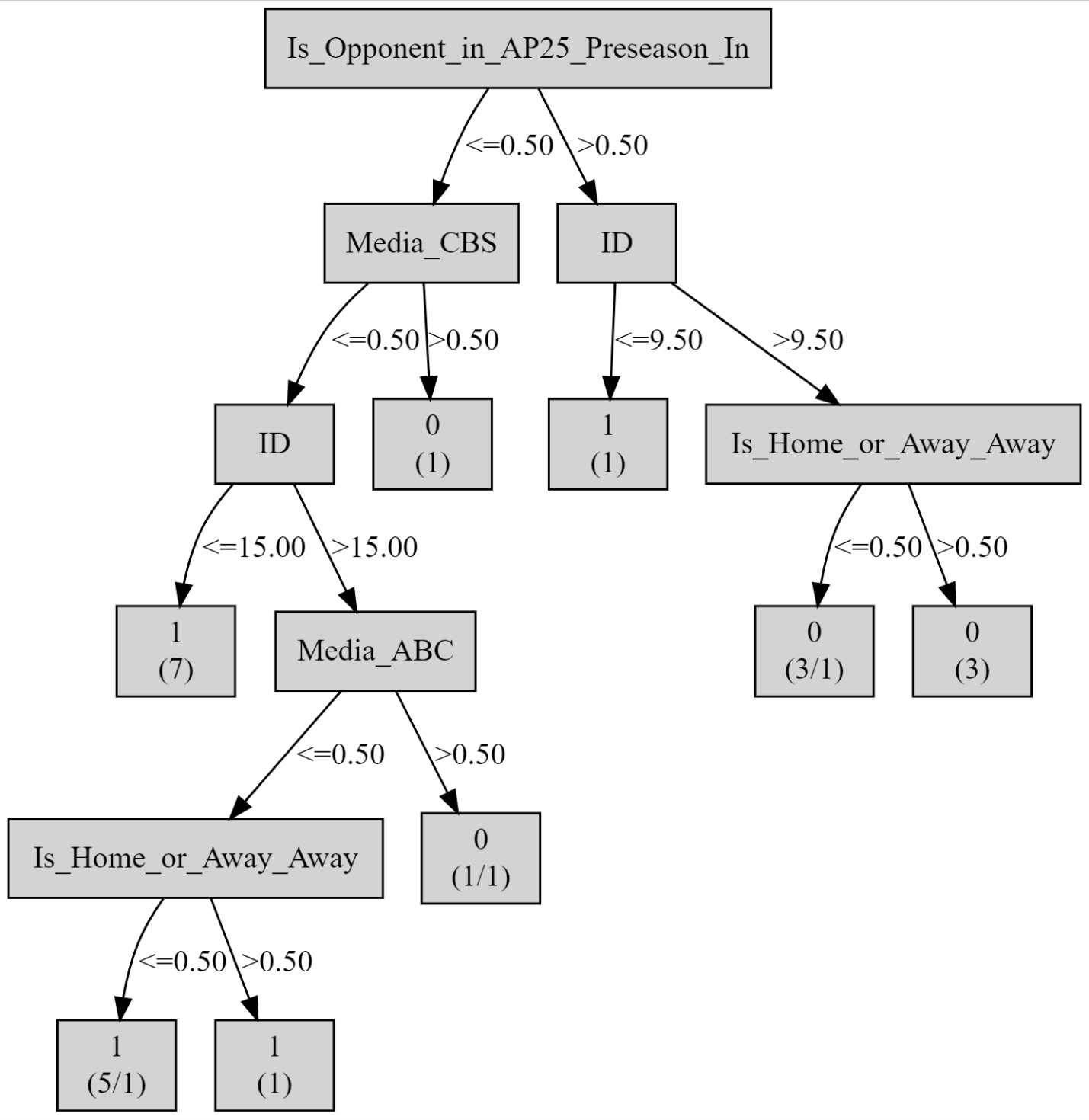


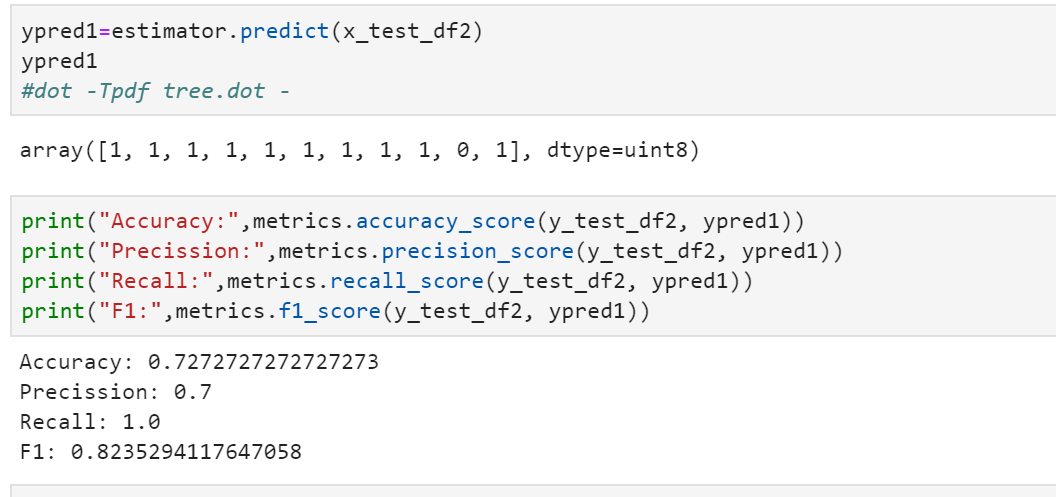
Task 5

Given a university’s football game data for the last two seasons, please construct the two

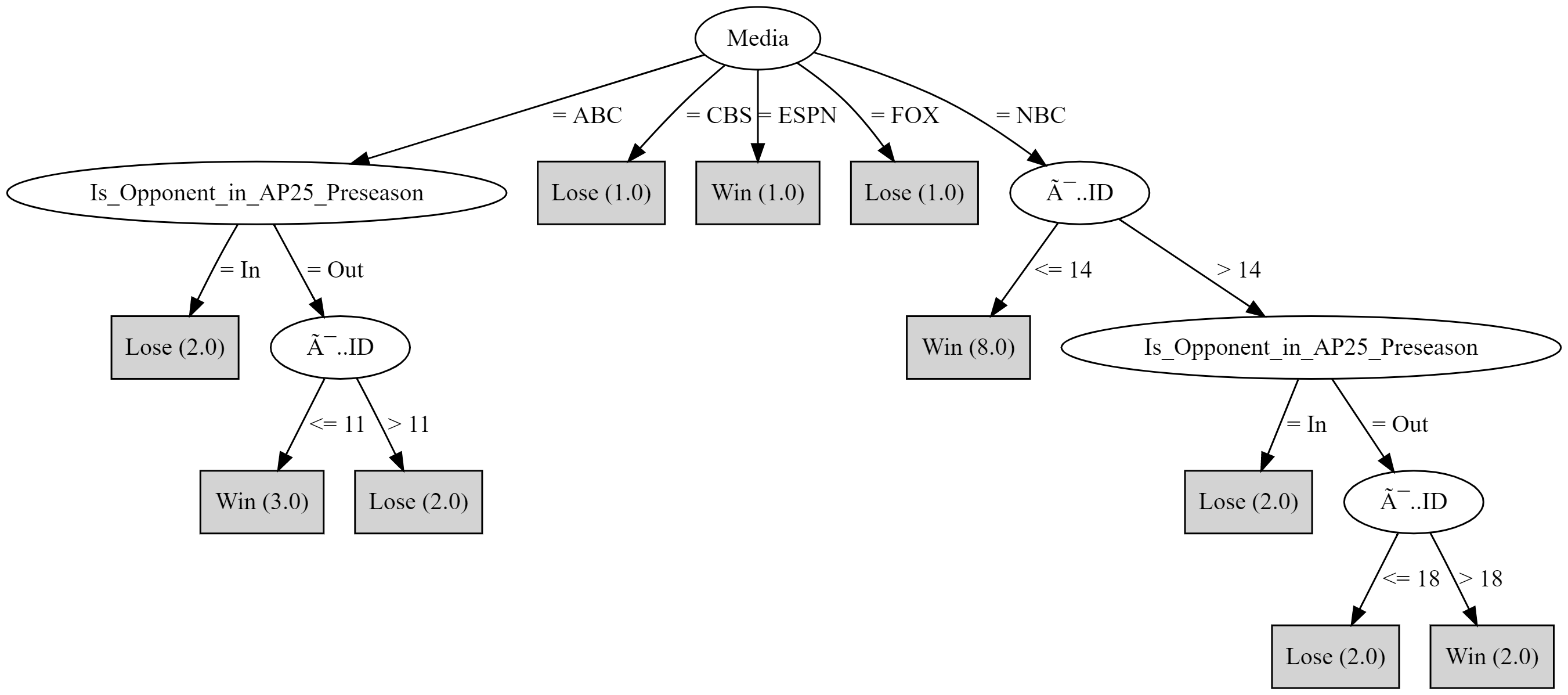
classification models ID3 and C4.5 to predict game results on games, and evaluate the model

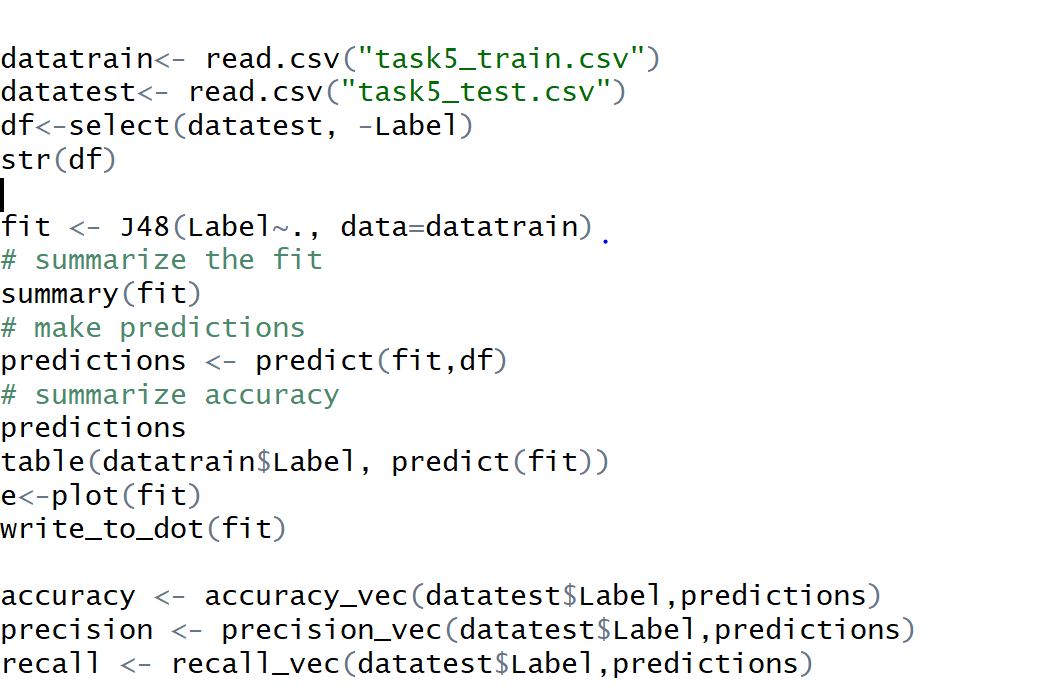
performance.

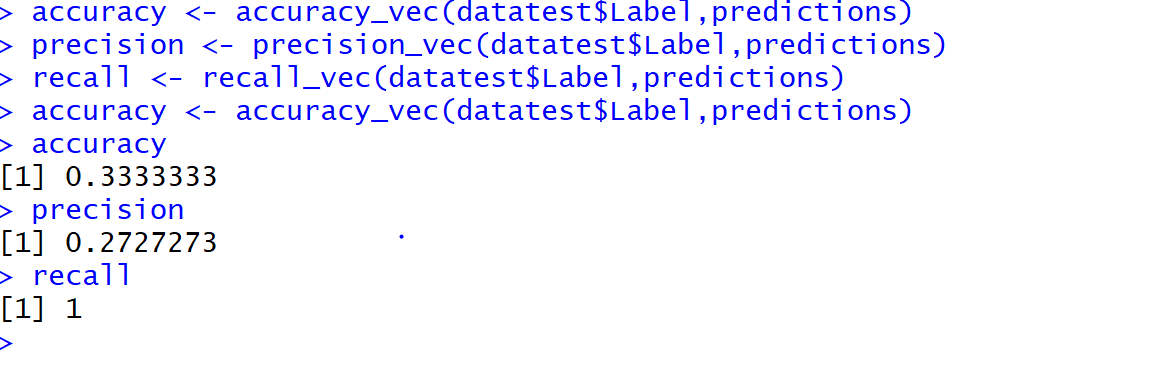




C4.5:







Question 3: which model is the best; which model performs the worst? Can you explain why?

We can infer that the accuracy for id3 was more than C4.5 for my code.

Github link for code: <https://github.com/bhuvaneshkj/ML_Assignemt3.git>

Please contact me in case of any queries, I can explain code. I m in hurry of submission