Week 2 – Group Activity – Naïve Bayes

Jonathan Y Ben Avraham
2124450
Group – Jonathan and Kareem

Question 1

Part I

a. Loan_Id is an identification field, normalizing it would result in anonymisation of the data.
b. Loan_Id is not an attribute that we use during the training stage.
c. Including Loan_Id into the features for analysis may introduce bias into the decision tree.

Part II

a. Loan_ID is in ascending order, doesn't form part of the dataset to be normalised.
b. It is an identification number.
c. Loan_ID would introduce bias into the decision tree.

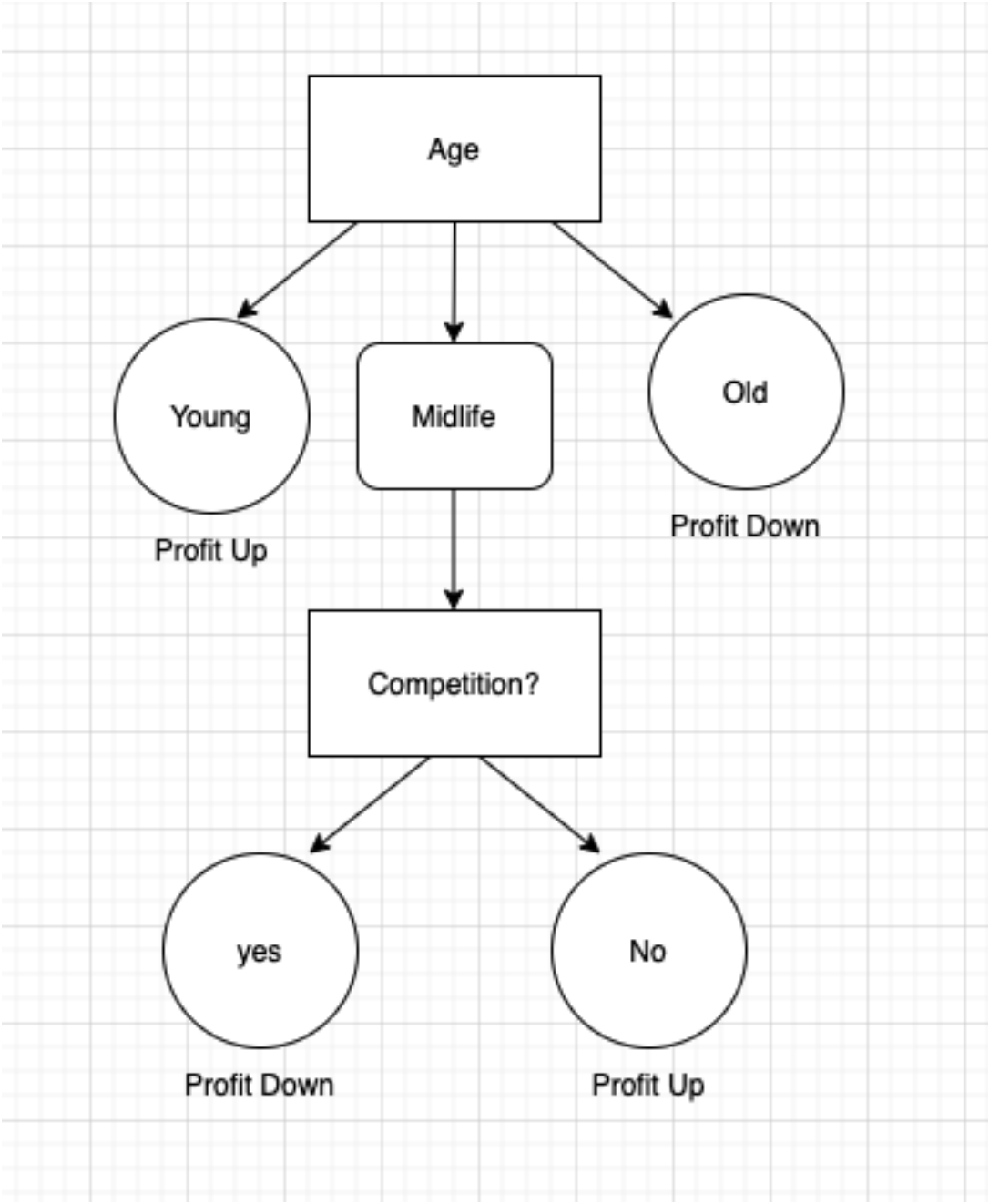
Question 2

Part I

In the dataset provided we already have a column for Profit, this attribute should be used as the target class. In the case of our dataset, the root node should be age. The reason for the root node being age is that it has greater information gain, that is there is more information to be got from this attribute than any other attributes in our dataset. Age has three values: Young, Midlife and Old. The possible values for our target class 'Profit' are Up and Down. The possible values for our attributes Competition and Type are respectively, Yes and No, and Software and Hardware.

Part II

Target class should be profit. Age should be the root node. We have concluded that company type does not have an impact on whether or not a company's profit goes up or down, however age has a direct impact on entropy or information gain, as well as competition.



Bibliography

Han, J., Pei, J., Kamber, M., 2011. Data Mining: Concepts and Techniques, 3rd Edition, Third Edition. ed.