A1 report

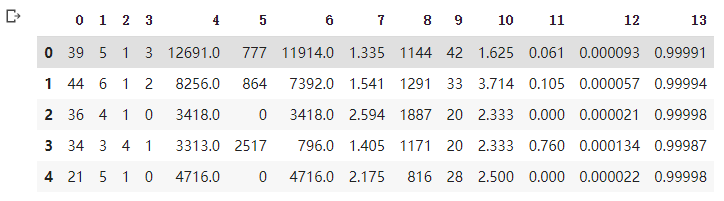
Zheng Yin

V00915261

Seng 474

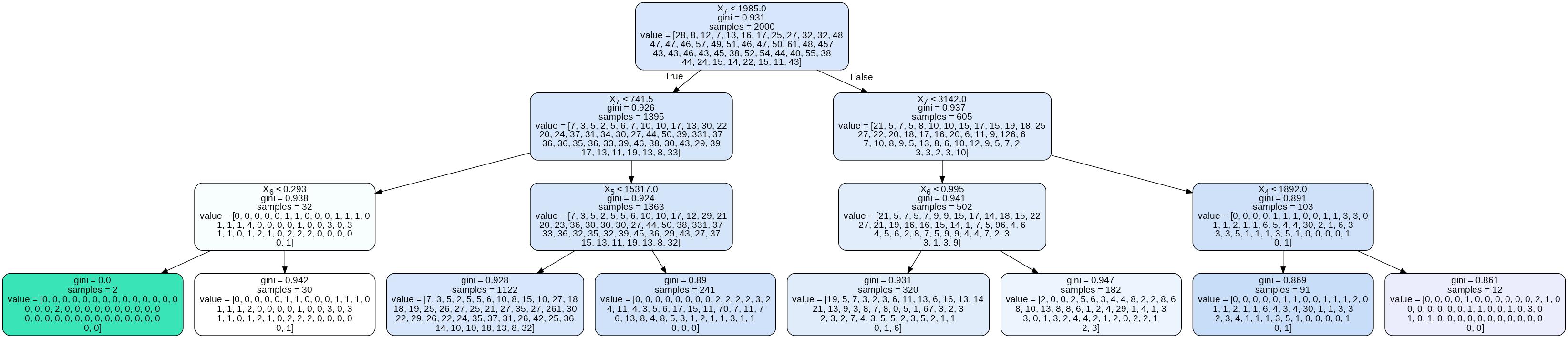
In this assignment, I choose a credit card information dataset, it includes most information about the credit card user. I choose this dataset because it is kindly relating to our group project, which is loan evaluation, there are several values that can help us to understand the data mining in business.

In this dataset, there is 5000 samples which include 14 variables, which I choose ave\_open\_to\_buy as the Y value and others are X value.

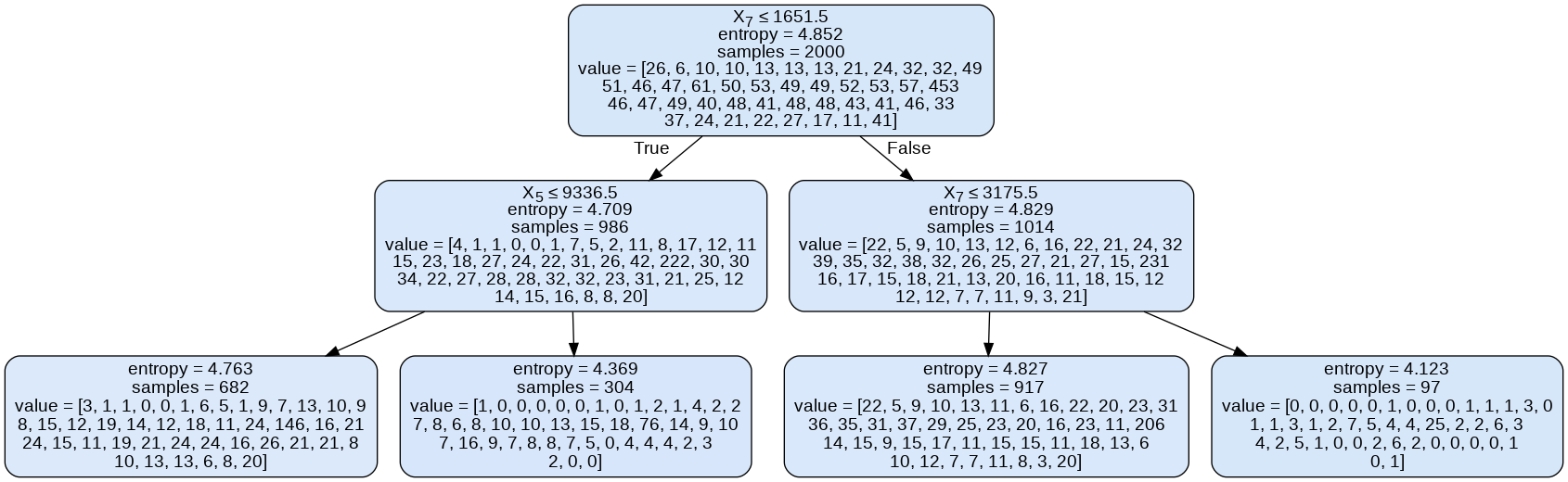


The most interesting part of this dataset is it show relationship between someone’s history in bank with their credit card limit. We can estimate someone’s time of onw their credit card by reading his other bank data.

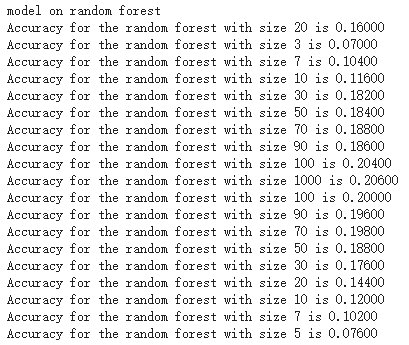
In Decision tree, the accuracy is 0.21



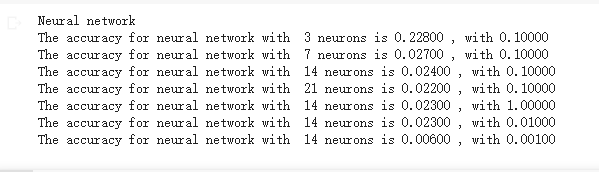
If I change the criterion to “entropy” and set max\_depth to 2, it can reach 0.23 on accuracy

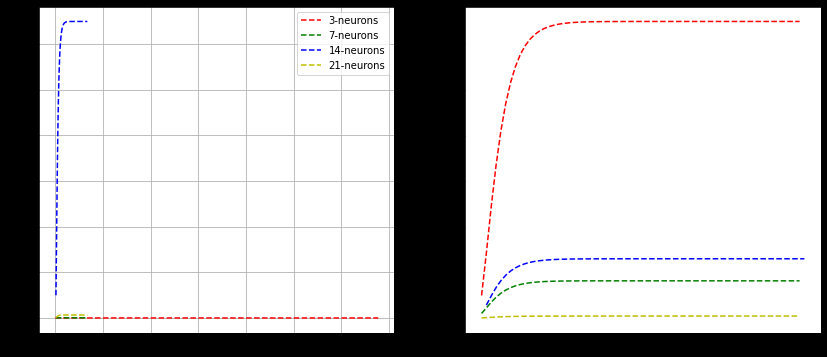


In random forest, the accuracy of about 0.2 when size of random forest with size of 100, and the accuracy is increasing by the size. Even the number of the size increased, the accuracy will not increasing above 0.2, but 0.204 is the limit of the random for



In neutral network, I changed the number of neurons which it seems like the increasing of the neurons will decreased the accuracy. In the same time, I change the alpha number in the MLPClassifier, the result is the small number of alpha can change the accuracy in the negative way.





For these three methods the data mining, look like the decision tree got the best accuracy, but it processes slow and the accuracy is not stable in several time which is a problem to me.