**Lim Jie Hong**

**German Credit Data**

**Objective:**

Use C4.5 algorithm to build a decision tree for classifying between good and bad customers who borrowed from a German bank.

**Data description:**

*germancredit.txt*has 1000 cases and 20 attributes, with 70% good customers and 30% bad customers.

The attributes include Status of existing checking account, Duration of credit in months and Credit history.

**Data Preparation:**

In Handbook of Statistical Analysis & Data Mining Applications by Nisbet, Elder and Miner, it was discovered that this dataset has no outliers, missing values and other undesired features. Chi-square test was carried out for the feature selection and 9 out of 20 attributes were ranked according to their importance for the dependent variable.

As I could not find a package in R which would enable me to visualize the complete decision tree based on the output of my code, I chose to include only 3 attributes which are the most significant in the Chi-square test. This enabled me to draw the DT on a paper more easily.

**Splitting into two complimentary subsets, 70% training data and 30% test data:**

I used proportional stratified random sampling to do the split. The outputs are stored in *train\_data.csv* and *test\_data.csv*.

**Decision tree modeling with C4.5 algorithm:**

Part of the R output:

The split feature is Status\_of\_existing\_checking\_account.

The levels are: A11 A12 A13 A14

The split feature is Credit\_history.

The levels are: A30 A31 A32 A33 A34

The split feature is Duration\_in\_month and split number is 15.

This is a leaf node as no further attribute: 0 of class 1 and 4 of class 2.

This is a leaf node as no further attribute: 2 of class 1 and 3 of class 2.

The split feature is Duration\_in\_month and split number is 12.

This is a leaf node as no further attribute: 0 of class 1 and 3 of class 2.

This is a leaf node as no further attribute: 4 of class 1 and 7 of class 2.

The split feature is Duration\_in\_month and split number is 15.

This is a leaf node as no further attribute: 26 of class 1 and 17 of class 2.

This is a leaf node as no further attribute: 33 of class 1 and 42 of class 2.

The split feature is Duration\_in\_month and split number is 18.

This is a leaf node as no further attribute: 3 of class 1 and 0 of class 2.

This is a leaf node as no further attribute: 0 of class 1 and 5 of class 2.

The split feature is Duration\_in\_month and split number is 11.

This is a leaf node as no further attribute: 17 of class 1 and 0 of class 2.

This is a leaf node as no further attribute: 20 of class 1 and 13 of class 2.

The split feature is Credit\_history.

The levels are: A30 A31 A32 A33 A34

The split feature is Duration\_in\_month and split number is 20.

This is a leaf node as no further attribute: 3 of class 1 and 2 of class 2.

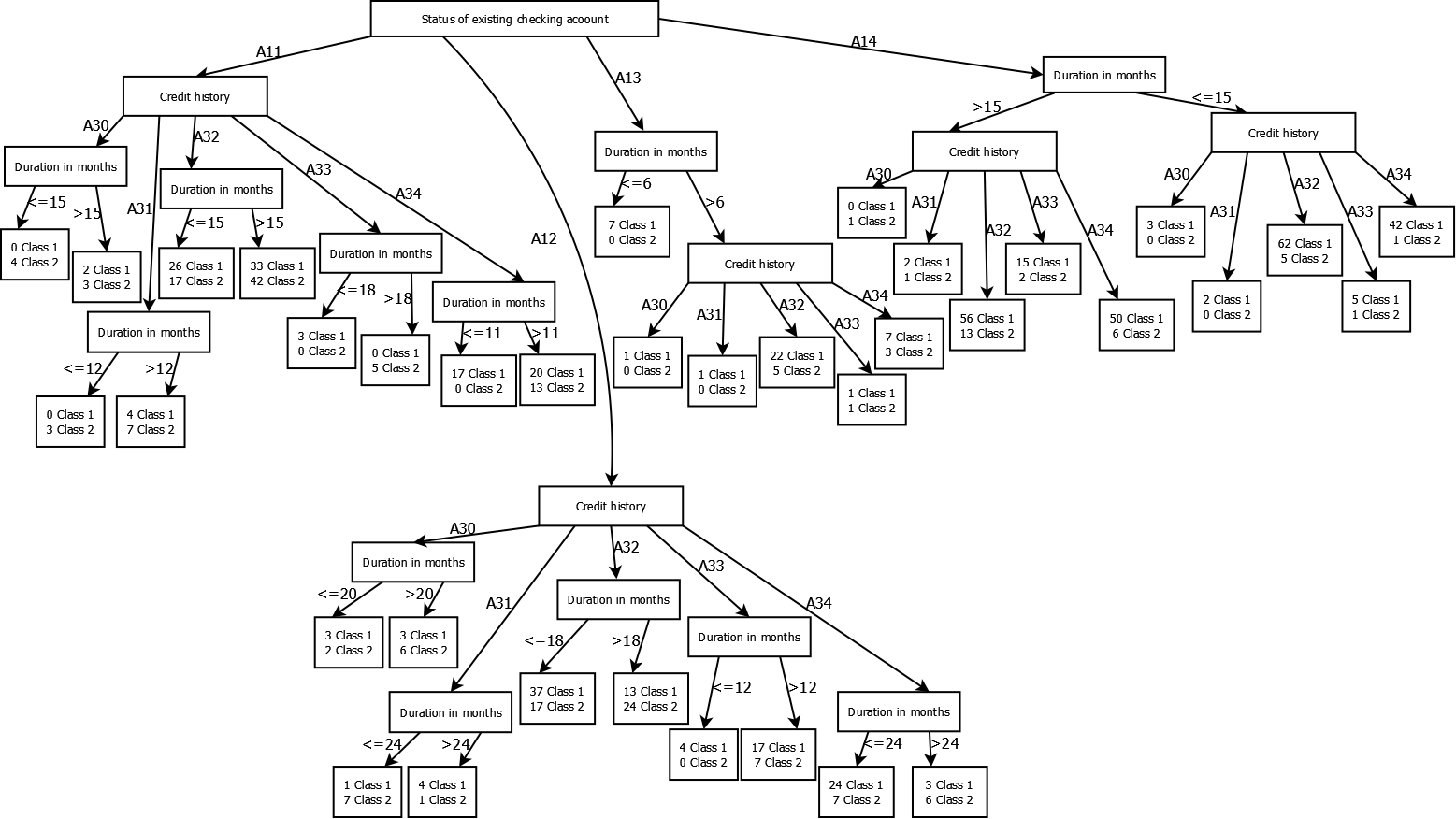
This is a leaf node as no further attribute: 3 of class 1 and 6 of class 2.

The split feature is Duration\_in\_month and split number is 24.

This is a leaf node as no further attribute: 1 of class 1 and 7 of class 2.

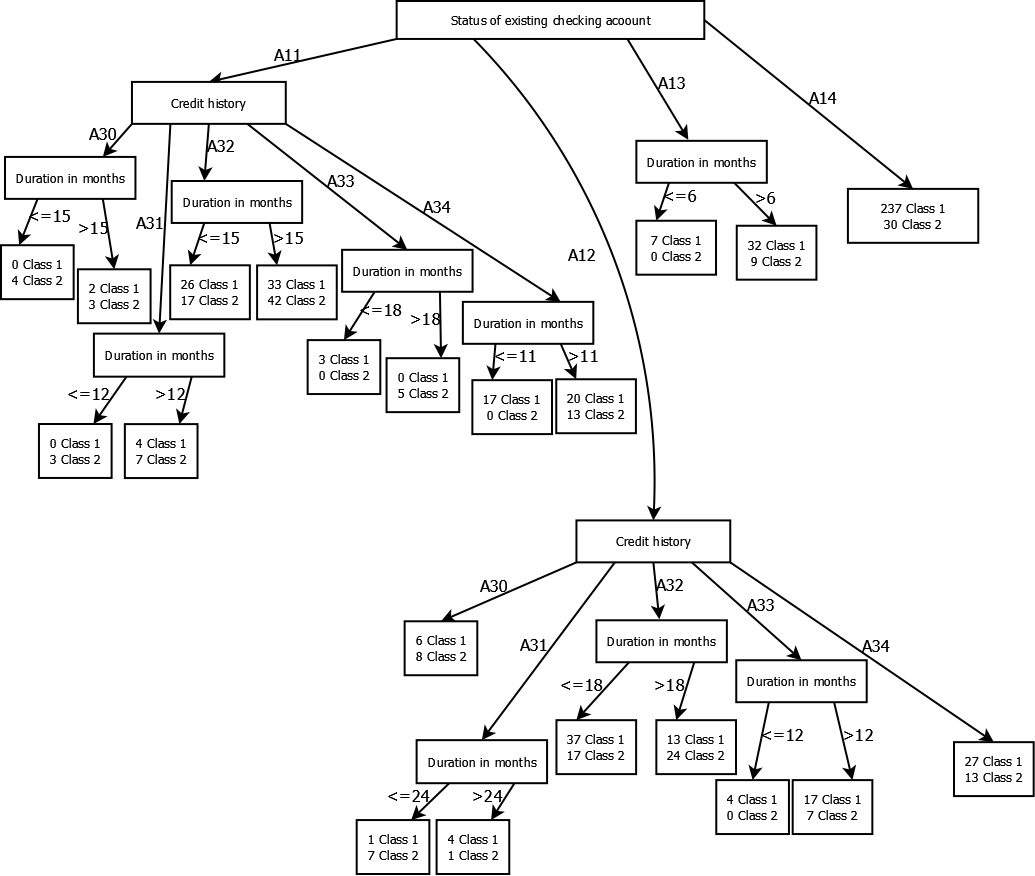
This is a leaf node as no further attribute: 4 of class 1 and 1 of class 2.

The decision tree was plotted based on R output:



**Pruning:**

After pruning, the decision tree is as follows:



**Testing:**

The predictions for the test data are as follow:

[1] 1 2 2 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 1 1 2 2 1 1 1 1 1 1 1 1 2 1 1 1

[38] 1 1 2 2 1 1 1 1 1 1 2 1 1 1 1 1 2 1 1 1 1 1 1 1 1 1 1 2 1 1 1 2 1 1 1 1 1

[75] 1 1 1 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1

[112] 2 1 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 1 1 2 1 1 1 1 1 1 1 1 1 1 1 2 1 1 1 1

[149] 1 1 1 2 1 1 1 1 1 1 1 2 1 1 1 1 1 1 1 1 1 1 1 1 2 1 1 2 1 1 1 1 1 1 1 1 1

[186] 1 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 2 1 1 2 1 2 2 2 1 1 1

[223] 2 1 1 1 2 2 2 2 2 2 2 2 2 2 2 1 2 1 1 1 1 1 1 1 1 1 2 1 2 1 2 1 2 1 2 1 2

[260] 1 2 2 2 2 2 1 2 1 1 1 2 1 1 1 1 1 1 1 1 1 2 1 1 1 2 2 2 1 1 1 2 2 1 2 2 1

[297] 2 2 2 2

The accuracy rate is 76.7%.

**Discussion:**

The choice of using only 3 attributes can be justified due to the nature of decision tree modelling. Decision tree has high variance which means it tends to be overfitting, so adding more attributes might not be useful in generalizing the model on unseen dataset. To improve its ability of generalization, ensemble methods such as random forest can be implemented. Besides that, cross-validation can be implemented to obtain more accurate testing error.