Question 1

The given dataset represents the data of 1000 past credit applicants from the German Credit, which is described by 30 variables.

By importing the data into Rapid miner and performing initial Exploratory Data Analysis, we obtain the proportion of the ‘Good’ to the ‘Bad’ cases to be 700 and 300 respectively.

Question 2

100% Training Data criterion:

Gain ratio, depth 25, confidence 0.02, threshold 0.68

Accuracy: 75

Gain ratio, depth 22, confidence 0.01, threshold 0.68

Accuracy: 72.9

Gain ratio, depth 17, confidence 0.01, threshold 0.6

Accuracy: 76.3

Gain ratio, depth 18, confidence 0.01, threshold 0.6

Accuracy: 77.10, Misclassification cost: 68.5

**Gain ratio, depth 18, confidence 0.01, threshold 0.6, prepruning: min gain – 0.01, min leaf size – 4, min size for split – 4, no of prepruning alternatives - 3**

**Accuracy: 72.8, Misclassification cost: 74**

Amount, Duration, Age, History, Chk\_Acct, Num\_Dependents, Used Car, Real Estate, Install-rate, Employment, Co\_applicant, New\_Car, Retraining

Question 3

Part a

Gain\_ratio, depth 10, confidence 0.01, threshold 0.68, prepruning: min gain – 0.03, min leaf size – 5, min size for split – 5, no of prepruning alternatives - 4

Accuracy: 69.8, Misclassification cost: 103

Gini\_Index, depth 18, confidence 0.01, threshold 0.7, prepruning: min gain – 0.02, min leaf size – 10, min size for split – 12, no of prepruning alternatives - 5

Accuracy: 71.6, Misclassification cost: 86

50-50 Split

**Gini\_Index, depth 10, confidence 0.01, threshold 0.72, prepruning: min gain – 0.05, min leaf size – 13, min size for split – 5, no of prepruning alternatives - 5**

**Accuracy Training: 77.6, Training Misclassification: 65.6, Accuracy: 73, Misclassification cost: 85.4**

Part b

70-30 Split

**Gini\_Index, depth 10, confidence 0.01, threshold 0.72, prepruning: min gain – 0.085, min leaf size – 8, min size for split – 5, no of prepruning alternatives - 5**

**Accuracy Training: 70.29, Training Misclassification: 67.429, Accuracy: 69.67, Misclassification cost: 81**

80-20 Split

**Gini Index, depth 10, confidence 0.01, threshold 0.72, prepruning: min gain – 0.075, min leaf size – 8, min size for split – 5, no of prepruning alternatives - 5**

**Accuracy Training: 74, Training Misclassification: 58.5, Accuracy: 71.5, Misclassification cost: 78.5**

**Part c**

**W-J48 - C-.25, M-10, Checked-RSA, Accuracy\_Tra- 72.4, Tra\_cost-63.6, Test Acc-71.46, Cost-76.4**

**CART- S-1.0, M-2, N-5, C-1, All unchecked, Acc\_tra-77, Tra\_cost 89.6, Test\_acc-75.2, Cost-82.2**

**Part D**

1995- Tra\_acc-71, Cost\_tra-55.3, Test\_acc-70, Cost-70

1992- Tra\_acc-75.6, Cost\_tra-67.6, Test\_acc-71.4, Cost-84.6

2000- Tra\_acc-73.6, Cost\_tra-69.6, Test\_acc-71.2, Cost-72.8

2200- Tra\_acc-74.2, Cost\_tra-71.4, Test\_acc-73.4, Cost-72.2

2800- Tra\_acc-73.6, Cost\_tra-57.6, Test\_acc-70.6, Cost-67

Question 4-

**Threshold 0.5**

**Training Misclassification: 84.75 48600 $**

**Test Misclassification cost: 103 11800 $**

**Threshold 0.6**

**Training Misclassification: 64.125 66900 $**

**Test Misclassification cost: 83.5 16300 $**

**Threshold 0.7**

**Training Misclassification: 58.5 78000 $**

**Test Misclassification cost: 78.5 18500 $**

**Threshold 0.8**

**Training Misclassification: 55.125 102,900 $**

**Test Misclassification cost: 74.5 24100 $**

**Question 6: 80:20 .5 thr**

**Gini index , depth 10, confidence 0.01, threshold 0.5, prepruning: min gain – 0.075, min leaf size – 8, min size for split – 5, no of prepruning alternatives - 5**

**Accuracy Training: 75.75, Training Misclassification: 84.75, Accuracy: 72.5, Misclassification cost: 83.5**

**50-50**

**Gini Index, depth 10, confidence 0.01, threshold 0.5, prepruning: min gain – 0.07, min leaf size – 10, min size for split – 5, no of prepruning alternatives - 5**

**Accuracy Training: 74, Training Misclassification: 65.6, Accuracy: 73, Misclassification cost: 85.4**