

COMP3425 Assignment 2

Tianxiang Zhang

U6773547

1. Platform

- Edition Windows: 10 Pro
- Processor: Intel(R) Core (TM) i7-8650U CPU @ 1.90GHz 2.11 GHz
- Installed RAM: 8.00 GB
- System type: 64-bit operating system, x64-based processor
- R: x64 4.0.4
- Rattle: Version 5.4.0

2. Data

(a) In your own words, briefly describe the purpose and means of data collection.

Data collection is a questionnaire that includes 7 sections, which involves 5 aspects.

Section A: Political standpoint and Satisfaction with current Australian government.

Section B & C: Experiences and Social status with COVID-19.

Section D & E: Mental health and Financial status in the last few weeks.

Section F: Satisfaction and Discrimination in Australia.

Section G: Survey Feedback.

The Purpose: Training models to find the relationship between data from different aspects of the survey, and then predict people's answers to different questions.

(b) Look at the pairwise correlation amongst the numeric variables using Pearson product-moment correlation. Qualitatively describe the pairwise correlations amongst each of the variables $p_age_group_sdc$, $C3_a$, $C3_b$, $C3_c$, $C3_d$, $C3_e$, and $C3_f$. Explain what you see in terms of the meaning of the data.

Quantitative analysis:

Figure 2.b.1 is the correlation summary using the Pearson covariance and **Figure 2.b.2** is the scatter plot of correlation using Pearson. We can find that the Pearson's correlation coefficient of $C3_d$ and $C3_e$ is 0.9659, which is positive and large. The Pearson's correlation coefficient of $C3_b$ and $C3_f$ is 0.3637, $C3_a$ and $C3_b$ is 0.4448, $C3_b$ and $C3_c$ is 0.3347, which are positive and medium.

Qualitative analysis:

We can find that people have two different ways to get information about COVID-19. Firstly, traditional way, people who read newspapers and magazines ($C3_d$) are also likely to watch radio and TV ($C3_e$) to get information about COVID-19. Secondly, modern way, people who get information about COVID-19 from professional advice ($C3_b$), maybe they also get information from official government sources ($C3_a$), family or friends and social media($C3_f$).

Correlation summary using the 'Pearson' covariance.

Note that only correlations between numeric variables are reported.

	p_age_group_sdc	C3_e	C3_d	C3_c	C3_f	C3_b	C3_a
p_age_group_sdc	1.00000000	-0.043091561	-0.036172989	-0.01809889	-0.022219859	-0.022713088	-0.022713088
C3_e	-0.04309156	1.000000000	0.965934801	0.37443905	0.004880821	0.003261864	0.003261864
C3_d	-0.03617299	0.965934801	1.000000000	0.37855162	0.009458487	0.009062038	0.009062038
C3_c	-0.01809889	0.374439053	0.378551617	1.00000000	0.157825485	0.334714692	0.334714692
C3_f	-0.02221986	0.004880821	0.009458487	0.15782548	1.000000000	0.363760721	0.363760721
C3_b	-0.02271309	0.003261864	0.009062038	0.33471469	0.363760721	1.000000000	1.000000000
C3_a	-0.03282946	0.002421357	0.005352398	0.189628348	0.200830103	0.444885248	0.444885248

Figure 2.b.1

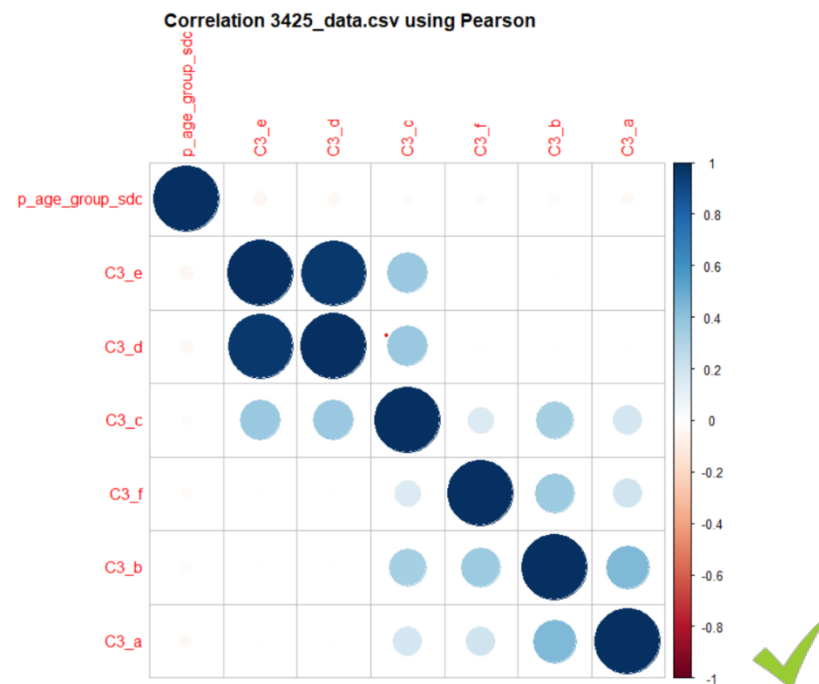


Figure 2.b.2

3. Association Mining

(a) Generate association rules, adjusting *min_support* and *min_confidence* parameters as you need. What parameters do you use? Bearing in mind we are looking for insight into what factors affect A1, find 3 interesting rules, and explain both objectively and subjectively why they are interesting.

<i>min_support</i>	<i>min_confidence</i>	<i>data (type: Categorical)</i>
0.5	0.5	A1, A5_a, B1_a, D1_a, D1_b, E1_a, E1_b, E10, E14

According to **Figure 3.a**, we can find there are 3 rules about A1 (RHS), which are **Rule 6**, **Rule 7** and **Rule 16**, the lift value of them are greater than 1.

Rule 6: E1_b = (1,2] => A1 = (1,2]

Rule 7: B1_a = (1,2], E1_b = (1,2] => A1 = (1,2]

Rule 16: B1_a = (1,2] => A1 = (1,2]

Objectively, LHS of three rules consistent of “You have been tested by a doctor or nurse for COVID-19” (***BI_a***) and “In education (not paid for by employer), even if on vacation” (***EI_b***). RHS always is “You very satisfied or satisfied with the way the country is heading” (***AI***).

Subjectively, People who have been tested for COVID-19 during the epidemic are highly satisfied with Australia, probably because they are satisfied with Australia’s medical system. People who still are students satisfied with Australia; the reason may be that their financial situation has not been affected during the COVID-19.

[6]	{TFC_EI_b=(1,2)}	=>	{TFC_AI=(1,2)}	0.5738225	0.6281606
[7]	{TFC_BI_a=(1,2),TFC_EI_b=(1,2)}	=>	{TFC_AI=(1,2)}	0.5611413	0.6279777
[8]	{TFC_AI=(1,2),TFC_BI_a=(1,2)}	=>	{TFC_EI_b=(1,2)}	0.5611413	0.9225614
[9]	{TFC_EI0=(-98,1]}	=>	{TFC_BI_a=(1,2)}	0.5113225	0.9817391
[10]	{TFC_BI_a=(1,2)}	=>	{TFC_EI0=(-98,1]}	0.5113225	0.5219602
[11]	{TFC_EI_a=(-98,1]}	=>	{TFC_BI_a=(1,2)}	0.5212862	0.9804089
[12]	{TFC_BI_a=(1,2)}	=>	{TFC_EI_a=(-98,1]}	0.5212862	0.5321313
[13]	{TFC_BI_a=(1,2)}	=>	{TFC_DI_b=[1,1]}	0.5480072	0.5594082
[14]	{TFC_DI_b=[1,1]}	=>	{TFC_BI_a=(1,2)}	0.5480072	0.9797571
[15]	{TFC_AI=(1,2)}	=>	{TFC_BI_a=(1,2)}	0.6082428	0.9795770
[16]	{TFC_BI_a=(1,2)}	=>	{TFC_AI=(1,2)}	0.6082428	0.6208969

Figure 3.a


(b) Comment on whether, in general, association mining could be a useful technique on this data.

Association mining is a useful method of data analysis on this data.

The Association rule algorithm is the most used algorithm. Association mining allows us to find the relationship between items in the data set. We can find strong rules in the data set by adjusting ***min_support*** and ***min_confidence***, and the association rule mining algorithm usually does not consider the transaction or the sequence between events.

4. Simple Classification

(a) This should be a very easy task for a learner. Why?

Opinionated is a Boolean version of *A4A5_agg*. Because there are only two types of variables in *Opinionated*, ***Ture*** and ***False***, then it can be transformed into a binary classification task to avoid a lot of complex calculations by using *A4A5_agg*. 

(b) Train each of a Linear, Decision tree, SVM and Neural Net classifier, so you have 4 classifiers. Evaluate each of these 4 classifiers, using a confusion matrix and interpreting the results in the context of the learning task.

```

Error matrix for the Linear model on 3425_data.csv [validate] (counts):

      Predicted
Actual  FALSE TRUE Error
FALSE   209   35  14.3
TRUE    46   61  43.0

Error matrix for the Linear model on 3425_data.csv [validate] (proportions):

      Predicted
Actual  FALSE TRUE Error
FALSE   59.5 10.0  14.3
TRUE   13.1 17.4  43.0

Overall error: 23.1%, Averaged class error: 28.65%

```

Figure 4.b.1 (Error matrix for the Linear model)

```

Error matrix for the Decision Tree model on 3425_data.csv [validate] (counts):

      Predicted
Actual  FALSE TRUE Error
FALSE   332    0    0
TRUE     0  141    0

Error matrix for the Decision Tree model on 3425_data.csv [validate] (proportions):

      Predicted
Actual  FALSE TRUE Error
FALSE   70.2  0.0    0
TRUE    0.0 29.8    0

Overall error: 0%, Averaged class error: 0%

```

Figure 4.b.2 (Error matrix for the Decision Tree model)

```

Error matrix for the SVM model on 3425_data.csv [validate] (counts):

      Predicted
Actual  FALSE TRUE Error
FALSE   231   13   5.3
TRUE     51   56  47.7

Error matrix for the SVM model on 3425_data.csv [validate] (proportions):

      Predicted
Actual  FALSE TRUE Error
FALSE   65.8  3.7   5.3
TRUE   14.5 16.0  47.7

Overall error: 18.2%, Averaged class error: 26.5%

```

Figure 4.b.3 (Error matrix for the SVM model)

```

Error matrix for the Neural Net model on 3425_data.csv [validate] (counts):

      Predicted
Actual  FALSE TRUE Error
FALSE   209   35  14.3
TRUE     24   83  22.4

Error matrix for the Neural Net model on 3425_data.csv [validate] (proportions):

      Predicted
Actual  FALSE TRUE Error
FALSE   59.5 10.0  14.3
TRUE     6.8 23.6  22.4

Overall error: 16.9%, Averaged class error: 18.35%

```

Figure 4.b.4 (Error matrix for the Neural Net model)

From the data:

Overall error (*Linear* > *SVM* > *Neural Net* > *Decision Tree*)

Averaged class error (*Linear* > *Decision Tree* > *Neural Net* > *Decision Tree*)

From the method and the model learnt:

The accuracy/error rate is indeed a very good and intuitive evaluation index, but sometimes a high accuracy/low error rate does not represent an algorithm. For example, predicting all situations that the third war will not happen tomorrow when we predict whether the third world war will happen tomorrow, the accuracy rate will be very high but meaningless. Therefore, it is not comprehensive to evaluate an algorithm model based on accuracy alone.

(c) Inspect the models themselves where that is possible to assist in your evaluation and to explain the performance results. Which learner(s) performed best and why?

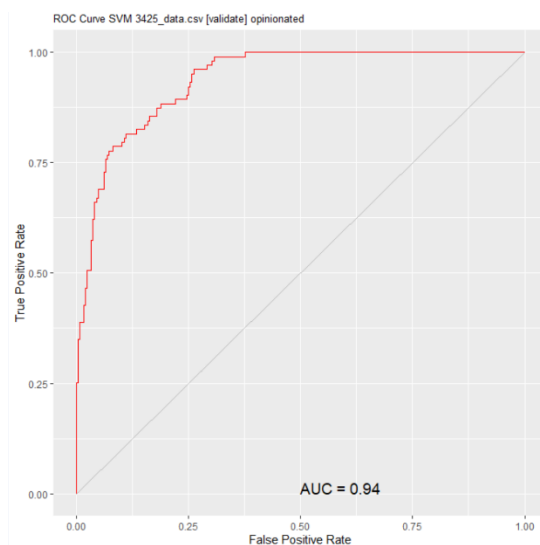


Figure 4.c.1(SVM)

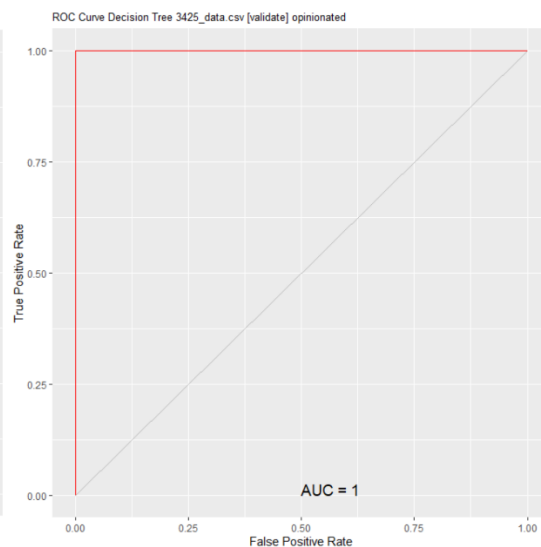


Figure 4.c.2 (Decision Tree)

Decision Tree performed best.

The overall error and averaged class error are 0 on error matrix and ACU is 1 on ROC.

5. Prediction

(a) Explain which you chose of a regression tree or neural net and justify your choice.

Regression Tree

The neural network may also need to do different preprocessing of the data to find a suitable model. It is relatively simple to process with regression trees. In other word, the neural network needs a larger data set to have better results, but the data of 3425.csv is not a large data set suitable for the neural network ✓

(b) Train your chosen model and tune by setting controllable parameters to achieve a reasonable performance. Explain what parameters you varied and how, and the values you chose finally.

<i>Input</i>	<i>Target</i>
<i>B1_a – B1_d, B4, B5, C1_a – C1_i, C3_a – C3_f</i>	<i>B6</i>

As **figure 5.b** shows, I changed *min split* from 20 to 25, *max depth* from 30 to 25, and *complexity* from 0.0100 to 0.0035.

For *min split*: if there are more than k samples in a certain node, the copy needs to be divided. The replacement variable 2 of this parameter, if there are more than 2 samples is divided into one sample. If these two samples can be divided, the budget will continue to be subdivided.

For *max depth*: we can ignore this value when there are few data or features. If the model has a large sample size and many features, it is recommended to limit this maximum depth. The specific value depends on the distribution of the data.

For *complexity*: When the complexity of the decision tree exceeds a certain level, as the complexity increases, the accuracy of the test set will decrease instead, so the established decision tree does not need to be too complicated. We can use pruning to reduce the complexity of the model, so it can reduce the risk of overfitting, thereby reducing the generalization error.

Target: B6
Algorithm: ☒ Traditional ☐ Conditional

Min Split: 25
Max Depth: 25
Priors:

Min Bucket: 7
Complexity: 0.0035
Loss Matrix:

Summary of the Decision Tree model for Classification (built using 'rpart'):

```

n= 2208

node), split, n, deviance, yval
* denotes terminal node

1) root 2208 24034.2000 3.855978
2) C1_h< -98.5 11 9622.7270 -5.545455 *
3) C1_h>=-98.5 2197 13434.3500 3.903050
6) B1_d>=1.5 785 11807.6600 3.472611
12) C3_b>=1.5 371 11095.0200 3.207547
24) C3_f< 1.5 123 10641.4600 2.707317
48) C1_h>=1.5 51 10358.1600 1.392157
96) C1_d< 1.5 15 9793.7330 -3.466667 *
97) C1_d>=1.5 36 62.7500 3.416667 *
49) C1_h< 1.5 72 132.6111 3.638889 *
25) C3_f>=1.5 248 407.5121 3.455645 *
13) C3_b< 1.5 414 663.2174 3.710145 *
7) B1_d< 1.5 1412 1400.3870 4.142351 *

Regression tree:
rpart(formula = B6 ~ ., data = crs$dataset[crs$train, c(crs$input,
crs$target)], method = "anova", model = TRUE, parms = list(split = "information"),
control = rpart.control(minsplit = 25, maxdepth = 25, cp = 0.0035,
usesurrogate = 0, maxsurrogate = 0))

```

Figure 5.b

(c) *Assess the performance of your best result using the subjective and objective evaluation appropriate for the method you chose and justify why you settled with that result.*

Objectively, I used *Score* to evaluate the model, and as **figure 5.c** shows, when *min split* and *max depth* between 25 to 30, *complexity* between 0.0030 to 0.0035, the model has the best result, which *mean* value of B6 always greater than 4. Therefore, subjectively, I believe that this is the best result base on 5.b.

	B6	rpart
Min.	:1.000	Min. :-3.467
1st Qu.:	4.000	1st Qu.: 3.710
Median :	4.000	Median : 4.142
Mean :	4.009	Mean : 3.921
3rd Qu.:	5.000	3rd Qu.: 4.142
Max. :	5.000	Max. : 4.142

Figure 5.

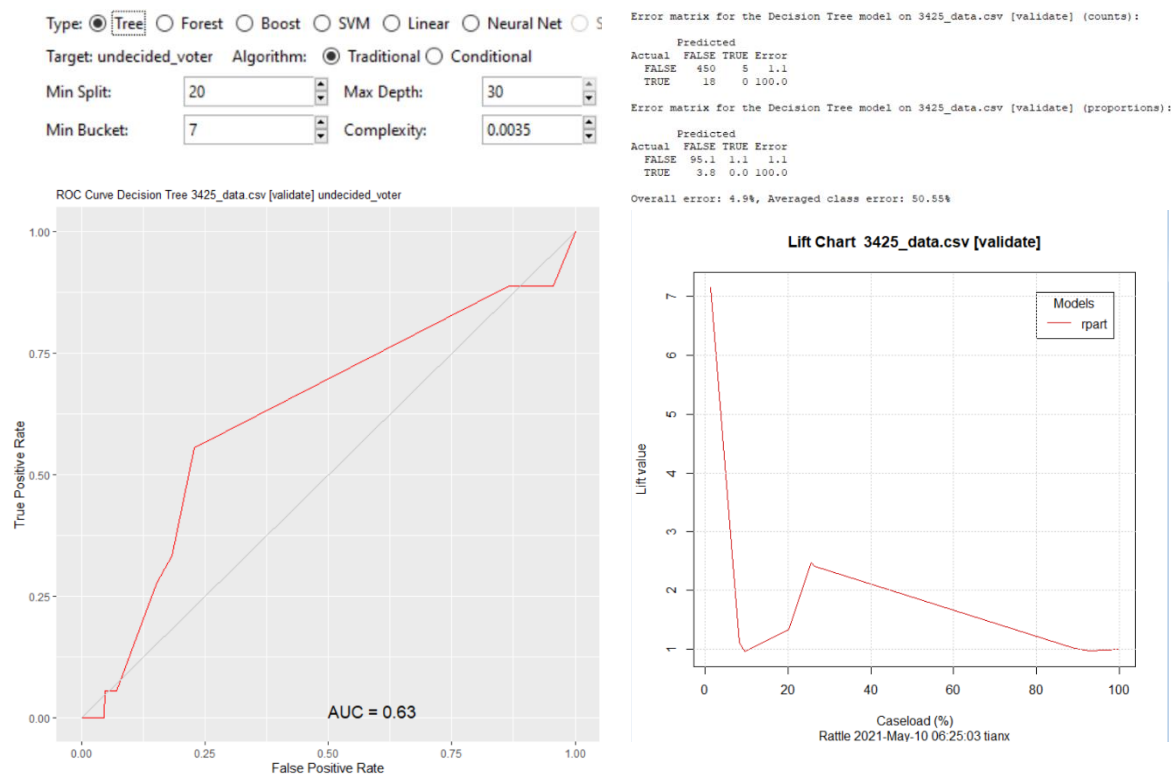
6. Complex Classification

(a) Explain how you will partition the available dataset to train and validate classification models in (b) to (d) below.

Partition: Train/Test/Validation = 80/10/10

Because the file 3425.csv is a small data set, which have 3155 observations and 62 input variables, I try to make the training set have as much data as possible to improve the accuracy of classification.

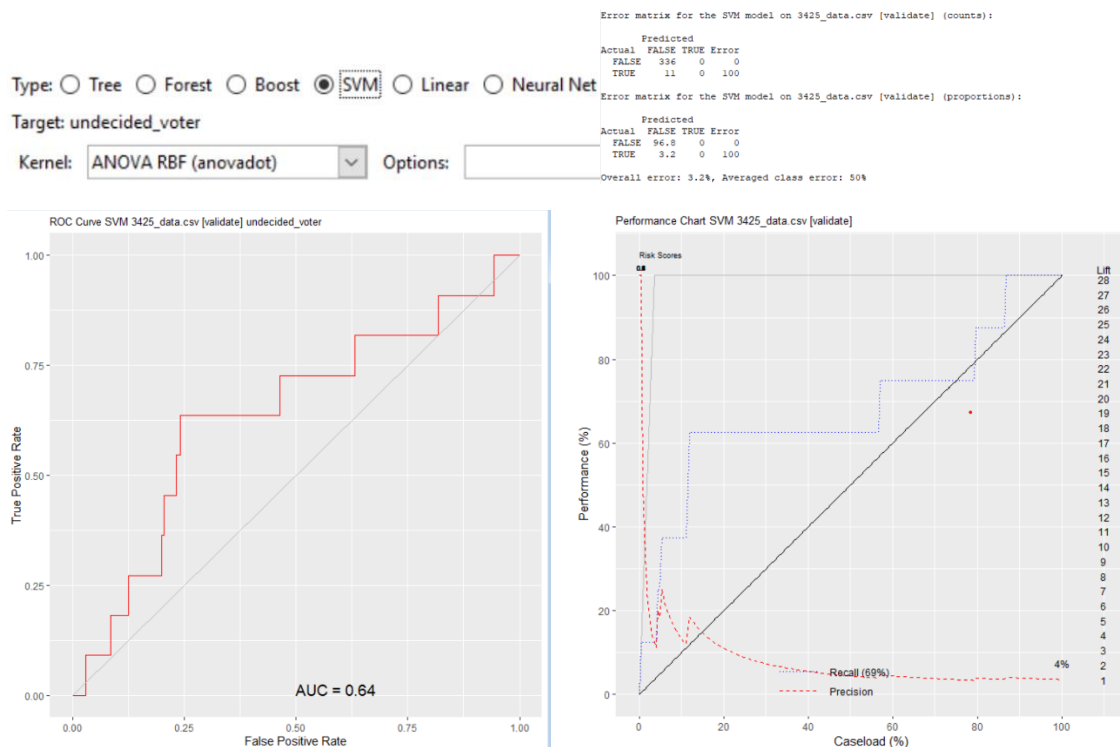
(b) Train a Decision Tree Classifier



When Complexity is 0.0035 (Defaults is 0.010), overall error is lowest, which is 4.9%.

Using **Lift** to evaluate the Decision Tree Classifier, which plots the lift against the rate of positive predictions. $(TP + FP) / TOTAL$

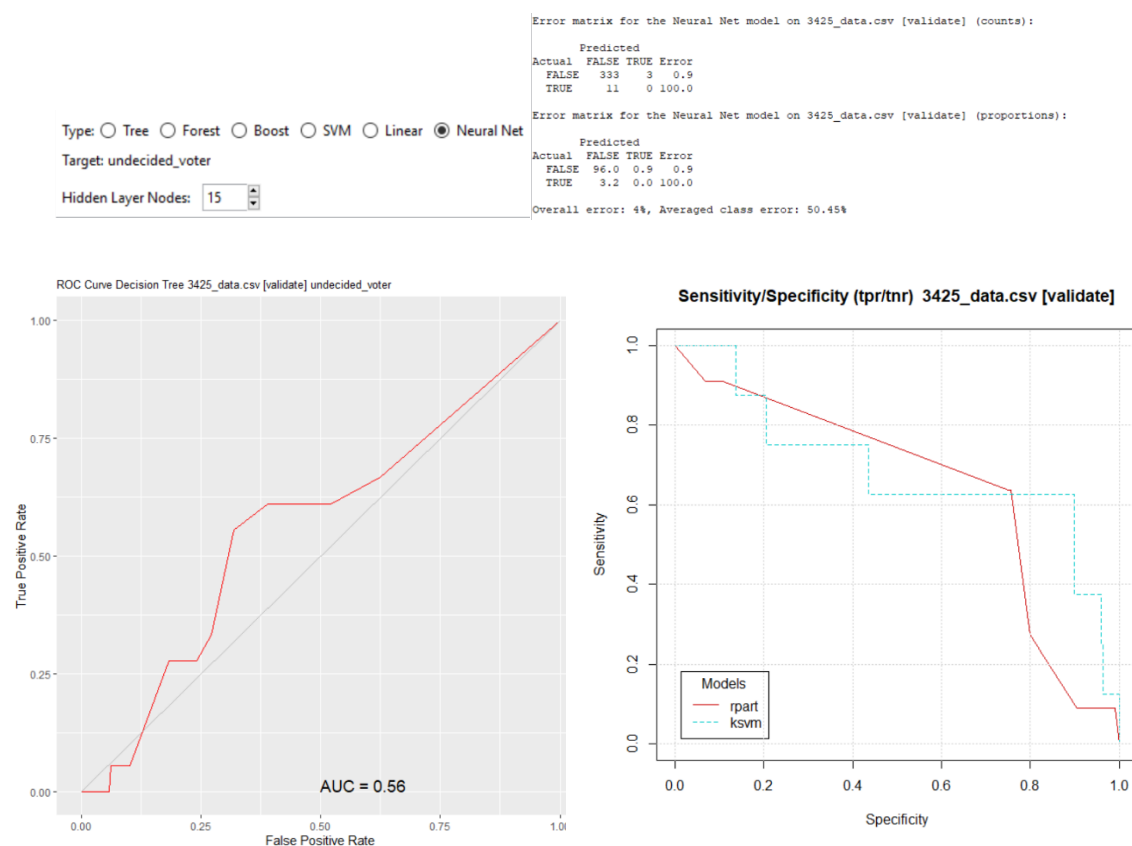
(c) Train an SVM Classifier



When Kernel is ANOVA (Defaults is Laplacian), overall error is lowest, which is 3.2%.

Using **Precision** to evaluate the SVM Classifier, the plot precision (the positive predictive value) against recall (the true positive rate).

(d) Train a Neural Net classifier



When Hidden Layer Nodes is 15 (Defaults is 10), overall error is lowest, which is 4%.

Using **Sensitivity** to evaluate Neural Net Classifier, this plots sensitivity (the true positive rate, also called recall) against the specificity (the true negative rate).

7. Clustering

(a) *Justify your choice of k as your preferred.*

As the **Figure 7.a** shows, we can find $k=5$ is the elbow point. ✓

With the number of clusters k increases, the sample division will be finer, and the degree of aggregation of each cluster will increase, and the error square sum SSE will naturally gradually become smaller. When k is less than the true number of clusters, the increase in k will be large. Increasing the degree of aggregation of each cluster, so the decline in SSE will be large, and when k reaches the true number of clusters, the return on the degree of aggregation obtained by increasing k will rapidly decrease, so the decline in SSE will also be sharply reduced. , And then tends to be flat with the increase of k value. Therefore, the k value corresponding to the elbow is the true number of clusters of the data.

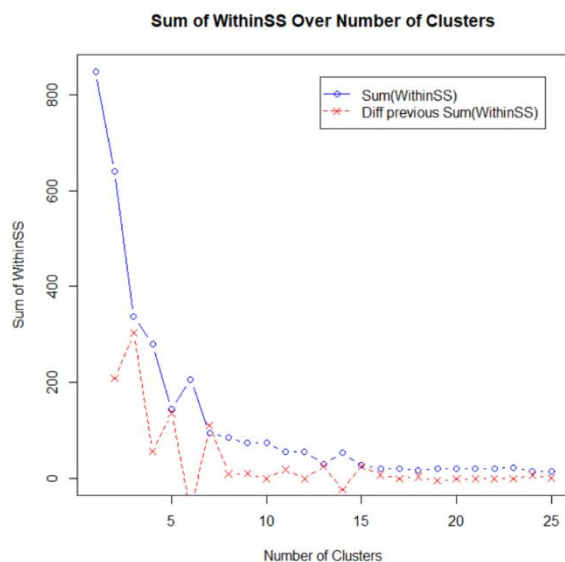


Figure 7.a

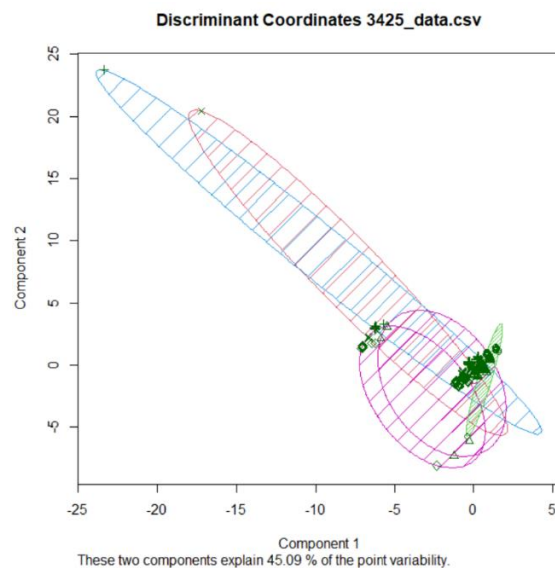


Figure 7.c

(b) *Calculate the sum of the within-cluster-sum-of-squares for your chosen model.*

```
Cluster sizes:
[1] "199 258 392 1123 405"

Data means:
      Al      Cl_a      Cl_c p_age_group_sdc p_education_sdc
0.9727445 0.9832262 0.9896117 0.5863133 0.3880942

Cluster centers:
      Al      Cl_a      Cl_c p_age_group_sdc p_education_sdc
1 0.9720935 0.9905965 0.9903975 0.2026801 1.0
2 0.9723790 0.9954333 0.9902525 0.2700258 0.5
3 0.9703041 0.9694635 0.9907052 0.8520408 1.0
4 0.9735538 0.9835042 0.9894907 0.5571386 0.0
5 0.9734148 0.9843784 0.9880944 0.8000000 0.5

Within cluster sum of squares:
[1] 7.213291 5.397939 22.261937 128.957226 16.674965
```

Figure 7.b

As the *figure 7.b* shows, the within-cluster-sum-of-squares are 7.21, 5.39, 22.26, 128.95 and 16.67. Therefore, *the sum of the within-cluster-sum-of-squares* is 180.48.

(c) Look at the cluster centres for each variable. Using this information, discuss qualitatively how each cluster differs from the others.

As *Figure 7.c* shows, we can find there 5 clusters differ from the others. However, the interesting thing is that these 5 clusters are overlapping, and I have tried $k=2$, which is the same situation. I think the reason of overlapping is that K-Means can only be locally optimal and is greatly affected by outliers.

(d) Use a scatterplot to plot (a sample of) the objects projected on to each combination of 2 variables with objects mapped to each cluster by colour. Describe what you can see as the major influences on clustering. Include the image in your answer.

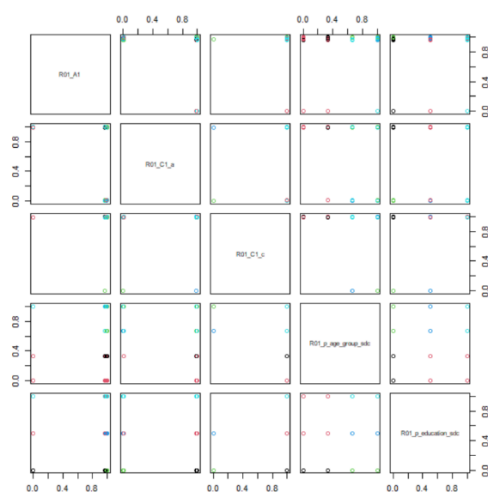


Figure 7.d

As Figure 7.d shows, we find *p_age_group_sdc* and *p_education_sdc* are the major influences on clustering.

8. Qualitative Summary

In the assignment, I used many functions on Rattle. For example, K-Means on Cluster, Association Mining on Associate, and Tree, SVM, Neural Net, etc. on Model. K-Means Clustering impressed me the most, which has a good clustering effect and fast classification. However, it slowly when I try to evaluate a Neural Net model by ROC. I still do not find a way to optimize it, but I will spend more time studying their algorithms, deeply understanding their underlying logic, and then looking for ways to optimize them so that the classification speed is faster, and the prediction accuracy is higher.

All in all, we learned and used Classification, Regression, Clustering, Association, and Neural Net, all of which are important tools to data mining. However, nowadays, more and more ways to data mining from different angles, like Web data mining. So much interesting knowledges waiting for us!