COMP 3425 DATA MINING

Assignment 2

Han Zhang

U6541559

1.

All the experiments are done on the lab machine in CSIT building, ANU. The operating system is 64-bit Ubuntu (Linux). The memory is 15.3 GiB and the CPU is Intel Core i7-7700 CPU @ 3.60GHz. For task 2,4,5,6,7 rattle is directly used. The implementation of association rule in task 3 is based on platform R by putting specific command.

2.

(a)

The purpose of data collection is to gather the answers to some questions from a group of people in order to draw some conclusions and make some predictions in the future. In this case, the collected data is the personal feedback about the experiences under the effect of the Coronavirus, the attitudes about the sexual harassment and assault [1]. The collection of experiences and views of participants is useful for the analysis of circumstances of Australians under the effect of COVID-19. The mean for data collection in this case is by online survey.

(b) The correlation table and correlation plot for the seven attributes are directly generated in rattle (as shown in appendix). There are a total number of 42 pairs of correlations, which are generally divided into three groups by correlation values.

Significantly correlated (absolute value of correlation > 0.6)

Variable d and e have a correlation value of 0.966 which are strongly positive correlated. This relationship indicates the participants who prefer getting information from newspaper and magazine are highly likely to get information from radio and TV, which is reasonable as these two ways are both traditional medias and similar.

Moderately correlated (0.3 < absolute value of correlation $\leq$ 0.6)

There are five pairs among these range which are a&b, c&d, c&e, b&c, b&f. All of these have the positive correlation value and thus positive relationship.

It indicates that the people getting information from official government source (a) are relatively likely to get information from professional advice (b) as these two ways are both professional and official. The people getting information from professional advice (b) or newspaper and magazine (d) or radio and TV (e) are relatively willing to communicate with family and friends (c). The people getting information from professional advice are relatively likely to get information from social media.

Not correlated (absolute value of correlation $\leq$ 0.3)

The rest of the 36 pairs are not correlated so they have no relationships. Obviously, the variable age_group_sdc have no relationship with others which means the way they acquire information is independent from the age. It is because different people rely on different sources and it is resulted from their characteristics rather than age. Meanwhile, some channels have almost no relationship with others that is mainly from the huge difference in their nature. For example, the government source is official, and the radio and TV are unofficial at all.

3.

(a)

The distribution of A1 is significantly unbalanced as most of the participants voted being

satisfied and few of them voting being dissatisfied. When generating the association rules, the parameter adjustment should be considered carefully. The minimum support number can easily filter out the unsatisfaction group. Consideration should be taken separately for satisfaction and dissatisfaction.

Minimum support = 0.59; Minimum Confidence =0.59; Minimum Length=2 ;

The set of parameters is for generating rules whose RHS is {A1=2}. In the R platform, the rule's appearance is set by command. The proportion for being satisfied (selection 2) is 62%. The minimum support for the rule whose RHS is {A1=2} should be less than 0.62.

Minimum support = 0.08; Minimum Confidence =0.85; Minimum Length=2 ;

The set of parameters is for generating rules whose RHS is {A1=4}. The proportion for being dissatisfied (selection 4) is 15%. The minimum support for the rule whose RHS is {A1=4} should be less than 0.15.

The selection for minimum confidence is determined by trying and displaying to keep the balance between the number of rules and the confidence level. The minimum length is set as 2 in default.

The selected three rules and interpretation are as follows.

Rule 1: { C1_c =1} => { A1=2}. Confidence: 61.3%

The first rule indicates reducing socialising in public spaces makes people feel satisfactory. Objectively, the confidence is 61.3%, which is persuasive. Subjectively, this rule is reasonable and interesting because limit activities in public space can increase sense of safety and lead to satisfactory.

Rule 2: { B1_b =2} => { A1=2}. Confidence: 61.4%

The second rule indicates having no contact with someone who had infected of COVID-19 makes people feel satisfactory. Objectively, the confidence is 61.4%, which is persuasive. Surjectively, this rule is reasonable and interesting because having no contact with people being infected lower the possibility getting infected and makes people keep safe. People are satisfied as their safety are not under threat of virus.

Rule 3: {A4_c =3, A4_d=2,A5_d=5,E_14=2} => { A1=4}. Confidence: 81.2%

The third rule indicates having not much confidence on government, having some confidence on GP, having no trust to politicians, coping in current income makes people dissatisfactory. Objectively, the confidence is 81.2%, which is very persuasive. Subjectively, it is quite common to being dissatisfactory because of the untrust to the external environment such as government and political situation and financial pressure can lead to more stress. It is interesting as it can gives some inspirations for the government for improvement.

(b)

Association mining is a useful tool in data mining as it can help to discover the associations between multiple variables. The interesting patterns and useful features can be summarised and presented by association mining, which can help certain organisations for evaluation, development and optimisation. For example, in the above scenario, the association mining

reveals the causality of people's satisfaction about their life in Australia with respect to multiple external factors. This can help the Australian government to optimise their future actions in governing the country and improving the life of Australian residents.

4.

(a)

It is a very simple classification task for a learner because the target variable opinionated is generated by the value of A4A5_agg. There exists the rule between the predictor variables and target variables. It is easy for a classifier such as decision tree to learn the rule. The performance of classifier will be nearly perfect as well. Commonly, the classification task has the ground truth label which is collected along with other attributes rather than calculated from other attributes. Therefore, the task is very simple with existing rule.

(b)

The confusion matrixes for these four classifiers are in appendix. For this scenario, TP means correctly classified as Opinionated class, TF means correctly classified as Not-Opinionated class, FP means wrongly classified as Opinionated class, and FN means wrongly classified as Not-Opinionated class. The confusion matrix reflects the ability of each classifier and detects which type of error the classifier tends to make.

Linear:

When training data has a proportion of 20%, the number of FP is 92 and the number of FN is 74. The average error is 20.05%. When training data has a proportion of 70%, the number of FP is 37 and the number of FN is 41. The average error is 27.05%.

With more training data, the performance of linear classifier becomes worse. Without the normalisation of the data, the classifier will easily be affected by the extreme data. The linear classifier will become overfitting with more data and have worse performance.

Tree:

When training data has a proportion of 20%, the number of FP and FN are 0. When training data has a proportion of 70%, the number of FP and FN are 0. The performance is always perfect with error rate 0% regardless of the size of training data. This is mainly because that the tree has already extracted the rule from limited training data. The tree classifier is not sensitive to outliers. The performance is always perfect and cannot be improved any more from more training data.

SVM:

When training data has a proportion of 20%, the number of FP is 41 and the number of FN is 121. The average error is 24.55%. When training data has a proportion of 70%, the number of FP is 13 and the number of FN is 36. The average error is 20.35%. The performance becomes better with more training data. This indicates that SVM classifier is underfitting with less training data and will have more prediction ability with more training data.

Neural Net:

When training data has a proportion of 20%, the number of FP is 93 and the number of FN is 109. The average error is 26.35%. When training data has a proportion of 70%, the number of FP is 37 and the number of FN is 20. The average error is 17.1%. Neural network has better classification ability with more data. With less data, neural network tends to overfit to the training data and be sensitive to outliers.

(c)

Among the four classifiers, the tree classifier is the best classifier for this classification task. Objectively, the tree classifier has 100% accuracy in all data which means it can classify all the data into right class. Although other classifiers can improve the performance by parameter tunning, they are less likely to reach an accuracy of 100%. Therefore, the tree classifier behaves best among all of the classifiers. Meanwhile, the target variable Opinionated is generated by rules and decision tree is a classifier that classifying by rules. It is more likely for the decision trees to extract rules from the data and thus get perfect classification results.

5.

(a)

I choose a decision tree as a classifier by comparing the natures of these two classifiers. The resulting model generated by regression tree is easy to visualize and friendly for non-experts to understand. Before, generating the decision tree, it is unnecessary to do the pre-process because the algorithm will not be affected by the scaling of data and missing values severely. Therefore, regression tree is less sensitive to data quality and less expensive for analysis. However, one major problem for regression tree is that it tends to overfit and has poor generalisation ability. When the data quality is high without large number of missing values and outliers, neural network will have high capacity for prediction task by adjusting the parameters and topology, thus it is more likely to get a better result than using regression tree. However, in this scenario, the data is of low quality because some participants did not answer all the questions. Therefore, it is more suitable to use regression tree for this task due to the nature of data and high interpretative ability.

(b)

As discussed above, regression tree tends to be overfitting, so it is significant to take control of the complexity of the tree. Pre-pruning is a feasible solution to stop the construction of tree before the tree is fully expanded. There are various parameters can adjust the complexity of tree, such as max depth, complexity, min split, and min backet.

Adjustment Process:

1. Change max depth from 30 to 5. Record optimal value.
2. Vary complexity from 1 to 0.0001. Record optimal value.
3. Vary min depth from 30 to 5. Record optimal value
4. Vary min backet from 10 to 5. Record optimal value
5. Try combination of different optimal parameters. Record error rate
6. Make small adjustment on the individual optimal parameter in last step.

7. Continue step 5-6.

Part of the adjustment of parameters is shown below (Step 5 & 6).

Model	max depth	complexity	min split	min backet	Testing error
-1	30	1	30	10	80%
0 (default)	30	0.01	20	7	77.4%
1	30	0.001	10	7	53.9%
2	25	0.001	10	7	53.9%
3	20	0.0001	8	5	62.5%
4	**20**	**0.0001**	**10**	**7**	**53.5%**
5	10	0.0001	10	7	54.8%
6	5	0.0001	5	5	76.9%

The default structure of the tree is over complex which leads to the model fitting the training set very well but fit the testing set with the error rate of 77.4%. When increase the values of parameters, the error rate becomes bigger with 80%. Therefore, the general idea is to simplify the structure of tree by modifying the parameters. By adjusting each of the parameters from the default values to smaller values, the error rate decreases gradually. However, when the error rate reaches a local minimum in model 4, the error rate begins to increase if continue on lowering the values of parameters.
The optimal max depth value is 20, the optimal complexity value is 0.0001, the min split is 10 and the min backet value is 7.

(c)
The best performance has an error rate of 36.7 % in training set and 53.5% in testing data. Objectively, the performance is relatively good compared to other classifiers. As neural network constructed by Pytorch has an error rate of 0% in training data and 90.7% in testing data. The neural network has much more serious problem of overfitting than regression tree. Meanwhile, the modification of parameters has covered a relatively large range. The trend of error rate is that the error rate will firstly decrease and finally increase when adjusting the parameters from large values to small values, and I tried to find the local minimum of error rate. Subjectively, I am quite satisfied with the result because the data is collected from real world. It is common to be very dirty that composed of some outliers. Using the dirty data directly without sufficient inspection and data cleaning work will unavoidably lead to some errors. Therefore, I believe, more effort should be put on the data cleaning work rather than tunning parameters, so I settled with this result.

6.

(a) In part b to d, the dataset is partitioned for training, testing, and testing in three different proportion sets.

- Training 20%; Testing 40%; Validation 40%
- Training 40%; Testing 30%; Validation 30%

- Training 70%; Testing 15%; Validation 15%

(b) The parameter setting for the optimal and default decision tree is as follows.

Optimal Min split: 25; Max depth: 20; Min bucket 6; Complexity 0.001

Default. Min split: 20; Max depth: 30; Min bucket 7; Complexity 0.01

The parameter of min split, max depth, min bucket, and complexity are varied from the default parameter. The error rate is shown below.

Error rate	20/40/40		40/30/30		70/15/15	
	Training	Testing	Training	Testing	Training	Testing
Optimal	50%	50%	43.85%	49%	43.5%	47.25%
Default	50%	50%	43.95%	50.1%	50%	50%

From the statistic of error rate in the table, there is no difference of the performance between the optimal parameters and the default parameters when 20% of the data are used for training. The performance of the optimal decision tree model is gradually improved when the size of the training data increases. To further prove the superiority of the optimal model over the default model, the ROC curves for setting with optimal parameters and default parameters when the proportion of training set is 70% are shown below.

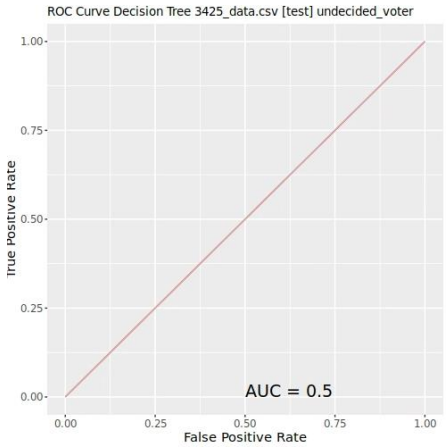


Fig 6.1 ROC for default

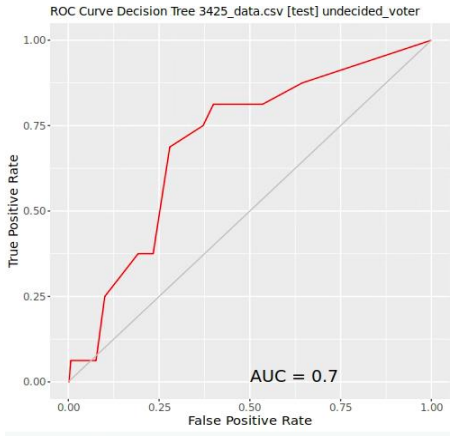


Fig 6.2 ROC for optimal

The area under the ROC curve can be used to evaluate the prediction model and the value of AUC represents the area. The value is usually between 0.5 and 1. When the value of AUC equals to 0.5, the model has poor ability for classification. When the value of AUC is larger or equal to 0.7, the model has some extent of ability for classification. In this scenario, the adjustment of the parameters improves the performance of the classifier to some extent.

(c)

Normalisation technique serves a key role in the performance of the SVM classifier, therefore before training the classifier, all the numerical attributes are normalised into range 0 to 1.

The parameter setting for the optimal and default SVM is as follows.

Optimal Kernel: ANOVA dot

Default. Kernel: RBF dot

The error rate is shown below.

Error rate	20/40/40		40/30/30		70/15/15	
	Training	Testing	Training	Testing	Training	Testing
Optimal	43.4%	49%	49%	49%	50%	49.2%
Default	43.75%	50%	50%	50%	48.3%	50%

From the statistic of error rate in the table, the performance of the SVM classifiers with two different parameter settings do not vary a lot in the error matrix. The performance of the classifier will become worse with more training data.

The ROC curves for setting with optimal parameters and default parameters when the proportion of training set is 20% are shown below.

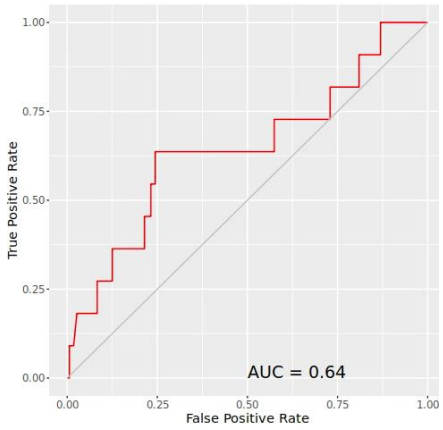


Fig 7.1 ROC for default

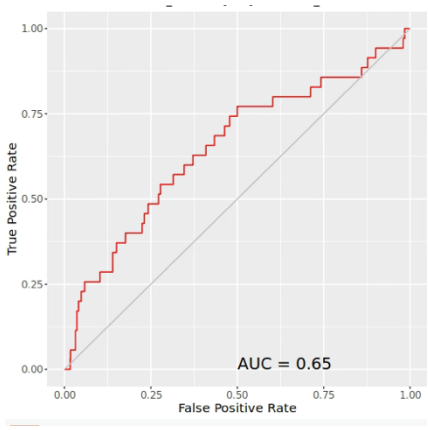


Fig 7.2 ROC for optimal

From the AUC value shown above, the performance of the optimal SVM is slight better than that of the default SVM which is consistent with the error rate table.

(d).

Before training the classifier, all the missing values were deleted, or the ROC plot cannot be generated. For the neural network classifier, the optimal parameter is the same as the default classifier which is hidden layer nodes equals to 10. When trying the number of hidden layer nodes, the error rate is higher than the default situation, so the optimal one is the best. The error rate is shown below.

Error rate	20/40/40		40/30/30		70/15/15	
	Training	Testing	Training	Testing	Training	Testing
Optimal & Default	0%	47.7%	13.55%	50.75%	20.55%	51.6%

The performance of the neural network classifier will become worse when the size of training data become bigger. In this situation, when training size is 20% of the entire data set, the network has the lowest error rate. The ROC curve is shown below.

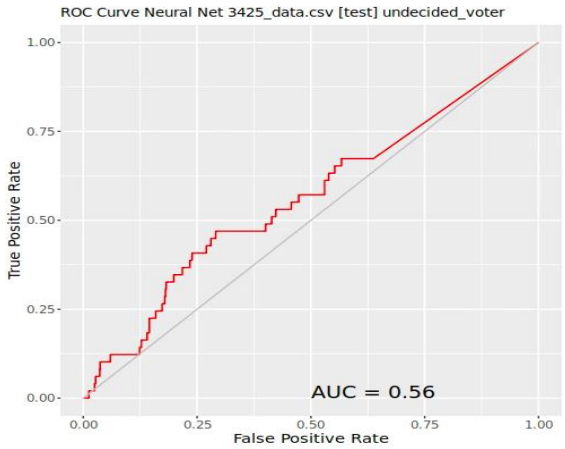


Fig 8.1 ROC for default & optimal

AUC value is 0.56, which indicates the neural network classifier has little value for classification task even for the optimal one. Thus, neural network is not suitable for the task.

7. The chosen variables are C1_i and C1_h.

(a)

The number of clusters is defined as 5. The default number of clusters is defined as $\sqrt{n/2}$ which is approximately 40 for the data set. The within cluster sum of squares measures the squared average Euclidean distance of all the points within a cluster to the cluster centroid. The smaller the value is, the closer that all the points away from their centroids and the clustering will be more reasonable. However, when the cluster number is 40, the sum of squares tends to be closed to 0 and it makes no sense to have so many clusters. The value on the y axis of the elbow plot is the value of 'within-cluster-sum-of-squares'. The determined number of clusters should be the point where the y value decreases significantly. When k equals 2, the SSE value is over 500 and decreases significantly to approximately 200 when k equals 5. Therefore, the number of clusters should be 5.

(b)

The sum of the within-cluster-sum-of-squares is 272 when k equals 5. As discussed above, the sum value can represent how far away for the points from their centroids. It is useful and interesting because the value can be used to determine the choice of number of clusters. On the one hand, when the value is relatively small, the clustering model can be considered as a good model that each point is compact to its centroid. On the other hand, when the value is extremely small, it is possible that the model does not make any sense. As when choosing the number of centroids that is equal to the size of data, the sum of within-cluster-sum-of-squares is always 0. Therefore, when determine the number of clusters or evaluate the goodness of the model, the analyst cannot completely rely on the quantity of the value. Some other measures need to be considered as well so it is quite interesting.

(c)

The cluster centres are as follows.

```
Cluster centers:

     R01_A1  R01_C1_h    R01_C1_i R01_p_age_group_sdc R01_p_education_sdc
1 0.9737699 0.9915958 0.994485221           0.6141202           0.0000000
2 0.9759455 0.9932023 0.993793409           0.0000000           0.0000000
3 0.9729590 0.9912537 0.995621156           0.3899153           0.6408776
4 0.9718011 0.9904716 0.995601152           0.8996623           0.8487699
5 0.9769700 0.5100161 0.005526134           0.6666667           0.4767442
```

Fig 9.1 ROC for default

Generally, the centroids have very coordinates in some dimensions and different coordinates in other dimensions. For example, cluster 1 and 2 have same coordinates in dimension A1, C1_h, C1_i, and education_sdc. In dimension age_group_sdc, they vary a lot. The Euclidean distance between them in the hyperplane is 0.61, which is relatively small. The table shows the distance of centroids. According to the value of distance, they are divided into two groups. When distance is bigger or equal to 0.7, the distance is small. Otherwise, the distance is big.

Pair	1&2	1&3	1&4	1&5	2&3	2&4	2&5	3&4	3&5	4&5
distance	0.61	0.68	0.89	1.2	0.75	1.23	1.45	0.55	1.15	1.18
conclusion	small	small	big	big	big	big	big	small	big	big

To conclude, the centroid of cluster 2 is far away from other centroids. The centroid of cluster 5 is far away from others as well. The centroid of cluster 4 is far away from centroids of cluster 1 and cluster 2. Other pair-wise centroids are relatively closed to each other.

(d)

The scatter plot for each combination of 2 variables are as follows. The original points are in a hyperplane of 5 dimensions. It is impossible to visualise each point in the hyperplane. The scatter plot takes 2 variables as a combination to visualise the distribution of the points in the 2-D plane (X: variable 1; Y: variable 2). There are 10 different scatter plots in total that take each pair of variables. In most of the variable pair-wise plots, it is hard to find the boundary that can partition each cluster because some points with different colours (in different clusters) even overlap. There is only one plot that has clear boundary for points with different colours which is the scatter plot taking education_sdc as X variable and age_group_sdc as Y variable.

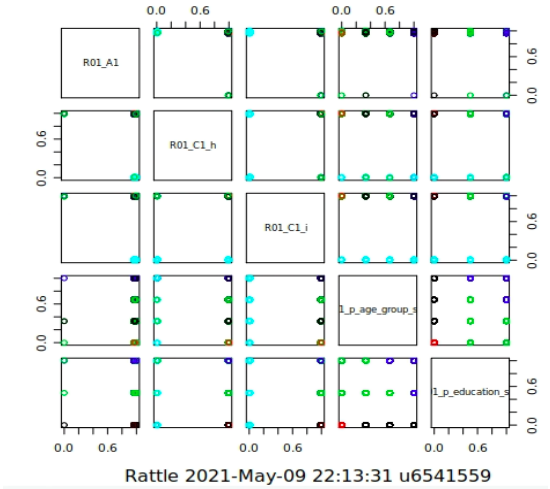


Fig 10.1 ROC for default

This phenomena hints that the major variables that plays key role in clustering is the education_sdc and age_group_sdc. The entity with similar features with respect to the two variables can be clearly clustered together and with different features with respect to the two variables are partitioned clearly. Therefore, the values of these two variables can be considered as the major influence on the clustering.

8.

For the classification task using this data, the tree classifier is the most suitable technique comparably. The most important reason is that tree optimiser can lead to the optimal result, which is noticeably better than other classifiers including linear classifier, neural network, and SVM classifier. The data is collected from the online survey so the data quality cannot be guaranteed. Large number of missing values and outliers will have large influence on the predictive ability of some classifiers. For example, SVM is sensitive to missing values. The neural network is sensitive to some extreme values. For linear classifier, it is both sensitive to missing values and outliers. The nature of data restricts the classification abilities for these classifiers. However, the algorithm for decision tree decides that it will not be affected by the low quality of data and can lead to a relatively better result than other classifiers.

For the computational complexity, the linear classifier has the simplest structure and is the least expensive one. For decision tree, the complexity is directly determined by the parameter setting and reasonable adjustment will keep a balance between the prediction ability and the computational complexity. However, one major problem for the decision tree is the overfitting. Although the parameter tunning can avoid the over-complex structure before training, it is better to perform the post-pruning technique to remove the insignificant branches.

In general, the decision tree is the most suitable technique using this data. In the future work, more efforts can be put on summarising the conclusion drawn before for decision making. For example, the association rule indicates the connection between variables and the clustering identifies entities that are similar to each other. Prediction can predict the possible values of missing or future data. The summarisation of the conclusions drawn from each task can be used to help certain organisations for evaluation, development, and optimisation. In this case, it can help the Australian government to optimise their future actions in governing the country and improving the life of Australian residents.

Reference

[1] N. Biddle, “Participant Information Sheet”, ANU. [Online]. Available: https://www.srcentre.com.au/anuethics/Information_sheet_ANUPoll_April_2021.pdf

Appendix

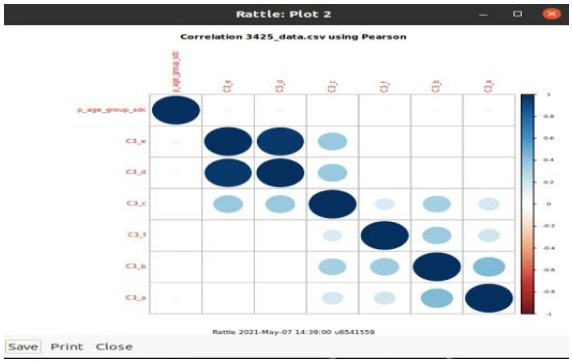
2(b) correlation table and correlation graph

```
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
p_age_group_sdc      1.00000000 -0.043091561 -0.036172989 -0.01809889
C3_e                -0.04309156  1.000000000  0.965934801  0.37443905
C3_d                -0.03617299  0.965934801  1.000000000  0.37855162
C3_c                -0.01809889  0.374439053  0.378551617  1.00000000
C3_f                -0.02221986  0.004880821  0.009458487  0.15782548
C3_b                -0.02271309  0.003261864  0.009062038  0.33471469
C3_a                -0.03282946  0.002421357  0.005352398  0.18962835
                         C3_f         C3_b         C3_a
p_age_group_sdc -0.022219859 -0.022713688 -0.032829466
C3_e             0.004880821  0.003261864  0.002421357
C3_d             0.009458487  0.009062038  0.005352398
C3_c             0.157825485  0.334714692  0.189628348
C3_f             1.000000000  0.363760721  0.200830103
C3_b             0.363760721  1.000000000  0.444885248
C3_a             0.200830103  0.444885248  1.000000000

Rattle timestamp: 2021-05-07 14:41:37 u6541559
```



4.(b) The confusion matrix for 4 classifiers

	Predicted		
Actual	FALSE	TRUE	Error
FALSE	219	25	10.2
TRUE	34	69	33.0

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

       Predicted
Actual  FALSE TRUE Error
  FALSE   564   92  14.0
  TRUE     74  209  26.1

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

       Predicted
Actual  FALSE TRUE Error
  FALSE  60.1  9.8  14.0
  TRUE    7.9 22.3  26.1

Overall error: 17.6%, Averaged class error: 20.05%
```

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

       Predicted
Actual  FALSE TRUE Error
  FALSE   220   37  14.4
  TRUE     41   60  40.6

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

       Predicted
Actual  FALSE TRUE Error
  FALSE  61.5 10.3  14.4
  TRUE   11.5 16.8  40.6

Overall error: 21.7%, Averaged class error: 27.5%
```

```
Error matrix for the SVM model on 3425_data.csv [test] (coun

       Predicted
Actual  FALSE TRUE Error
  FALSE   615   41   6.2
  TRUE    121  162  42.8

Error matrix for the SVM model on 3425_data.csv [test] (prop

       Predicted
Actual  FALSE TRUE Error
  FALSE  65.5  4.4   6.3
  TRUE   12.9 17.3  42.8

Overall error: 17.2%, Averaged class error: 24.55%

Rattle timestamp: 2021-05-07 21:13:42 u6541559
============================================================
```

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

       Predicted
Actual  FALSE TRUE Error
  FALSE   244   13   5.1
  TRUE     36   65  35.6

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

       Predicted
Actual  FALSE TRUE Error
  FALSE  68.2  3.6   5.1
  TRUE   10.1 18.2  35.6

Overall error: 13.6%, Averaged class error: 20.35%

Rattle timestamp: 2021-05-07 21:16:02 u6541559
```

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

       Predicted
Actual  FALSE TRUE Error
  FALSE   563   93  14.2
  TRUE    109  174  38.5

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

       Predicted
Actual  FALSE TRUE Error
  FALSE  60.0  9.9  14.2
  TRUE   11.6 18.5  38.5

Overall error: 21.5%, Averaged class error: 26.35%

Rattle timestamp: 2021-05-07 21:30:44 u6541559
```

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

       Predicted
Actual  FALSE TRUE Error
  FALSE   220   37  14.4
  TRUE     20   81  19.8

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

       Predicted
Actual  FALSE TRUE Error
  FALSE  61.5 10.3  14.4
  TRUE    5.6 22.6  19.8

Overall error: 15.9%, Averaged class error: 17.1%
```