COMP3425 Assignment 2

Created @April 15, 2024 12:18 PM

by Razeen Wasif (u7283652)

Platform

The analysis was performed on a MacOS Sonoma 14.3.1 system running the M2 Pro processor with 10 cores (4 high-efficiency + 6 high-performance) and 16GB of RAM. All of the analysis was performed directly using the R statistical software package.

Data

The data being used belongs to The Australian Data Archive's ANU Poll Dataverse repository[1]. They contain data from series of polls conducted quarterly for the purpose of assessing Australian's opinions on important and topical issues, with an emphasis on international comparisons[2].
This data in particular captures the results of a survey poll conducted in October 2023 on the topic of Australian Constitutional Referendum on the Aboriginal and Torres Strait Islander Voice to Parliament. The aim of this poll was to mainly assess the behavioural response of Australian citizens towards the Aboriginal and Torres Strait Islander Voice to Parliament and towards the Government and satisfaction with life in general. The dataset aims to broadly understand the current state of public opinion regarding these matters at hand as well as identify which demographics have particularly strong or weak opinions on said matters. Lastly the dataset can also shed insight on correlations between opinions and demographics such that any causal relationships can be verified.

Prior to conducting any form of data exploration, analysis or modelling, data cleansing was performed where required to ensure results are accurate, consistent and complete.

Correlation

The correlation between the selected variables RA2_a, RA2_b RA2_c, RA2_d and RA2_e with A1 as well as the p values were calculated in R and can be observed in Figures 1, 3 and 2.

There is almost no correlation between A1 and the RA variables. The RA questions ask the surveyors what medium they used to keep up with the Referendum whether it be a television or the newspapers and the A1 question was about the surveyors satisfaction with living in Australia and where the country is heading. These two variables should not have any correlation with each other and its easily evident in Figure 3. Their views on living here has no effect whatsoever on which way they choose to keep up with the Referendum. Amongst the RA variables however there can be seen some interesting pairs of correlations. RA2_a and RA2_b are quite strongly correlated suggesting many of the citizens primarily watched the TV as well as read the newspaper to pay attention to the

Referendum and may use both sources to consume news in general. RA2_e is quite weakly positively correlated with the others showing not many citizens used the social media over the others to keep up with the Referendum. This could mean not many use social media as a means of watching/reading the news. There is a moderately positive correlation for radio and online news with TV and newspaper but in general the TV and newspaper are the main sources people used to keep up with the Referendum. However, correlation does not always imply causation. There could be other factors involved such as demographic characteristics and or lifestyle preferences.

Association Mining: factors affecting satisfaction with the country's future

Before conducting the association mining in R, missing values were handled by replacing them accordingly. For Z_ANCESTRY_2, missing values were replaced with 9999 which is the refused category and same with the other attributes which were replaced with -99, the refused category for those attributes. This was done under the assumption that the surveyed individual either unintentionally skipped the question or did not wish to answer that question and didn't bother by inputting -99 themselves. Regardless the large negative value should still help maintain strong correlation between variables. All numeric attributes were also categorized into factors for the apriori algorithm. Lastly the "ORDER" and "VERB" columns were removed for all remaining tasks as they didn't provide any information due to them being restricted.

The min_support threshold was 0.25, confidence was 0.53 and minlen was 2. The reason for a low threshold was due to the fact that more than 53% of the records for A1 were 2. This led to the data being highly skewed towards a single variable and as such it would be difficult to determine whether the association is particularly strong or the input values is just highly frequent. For this reason, it seemed appropriate to set the min_support threshold to 0.25 which would allow for less frequently occurring but potentially more relevant associations. Lowering the confidence would also ensure that these associations would not always be required to exist.

The generated rules were filtered by lift greater than 1.5 such that rules with strong association were returned. They were further filtered by removing redundant rules for simplicity and removing statistically non-significant rules so that only meaningful patterns could be focused on. The first 25 rules are included in the appendix as Figure 4 and 5 for reference.

1. $\{\mathrm{E11a}{=}2\} \Rightarrow \{\mathrm{A1}{=}2\}$

   support count: 1107; support: 0.2624; confidence: 0.5836; lift: 1.0947

   This association suggests that people who are satisfied with the direction Australia is heading in and are satisfied with their lives in Australia typically make decent income or are "coping on present income" suggesting they're getting by without too much struggling. The lift suggests a positive association and the confidence shows 58% of all people who responded as being satisfied with the country makes decent enough

income. Understandably having money would heavily factor into how one feels about their satisfaction with the country and how it's heading and those who are struggling on low income would be unsatisfied.

2. $\{\mathrm{A4\_a}{=}2, \mathrm{E1\_c}{=}2, \mathrm{E1\_d}{=}2, \mathrm{p\_atsi}{=}4\} \Rightarrow \{\mathrm{A1}{=}2\}$

   support count: 1067; support: 0.2529; confidence: 0.6888; lift: 1.2922

   This association shows that in general, People who are statisfied with the way the country is heading are not unemployed, not of Aboriginal or Torres Strait Islander origin and have "quite a lot of confidence" in the Federal Government of Canberra. This association occurs in over a quarter of all observed records and has the highest confidence. Of all the people that chose "satisfied" as the response to A1, 68% have "quite a lot of confidence" in the federal government, have jobs and are not of Aboriginal or Torrest Strait Islander descent. The high lift also suggest that there is a positive correlation between the choices of these variables. Confidence in the federal government is inarguably a valuable factor in determining the satisfaction of Australian citizens and being employed would also aid in that. But this only seems to be the case for non-Aboriginal and Torrest Strait Islander people.

3. $\{\mathrm{A4\_c}{=}2\} \Rightarrow \{\mathrm{A1}{=}2\}$

   support count: 1273; support: 0.3017; confidence: 0.6592; lift: 1.2367

   This association suggests that people who are satisfied with how the country is going also have "quite a lot of confidence" in their local State/Territory Government. This association overall has high support, confidence and lift and is particularly interesting because it shows that having confidence in the Federal Government is a more relevant factor in determining satisfaction of people's lives in Australia rather than the State Government.

For this data in particular, despite uncovering pretty interesting associations, association mining is not quite a useful data mining technique to explore this dataset. There are many variables which adds to the complexity when determining interesting associations. The dataset also generates a large amount of rules, reaching millions, which can not only be challenging to manage and interpret given redundant or statistically insignificant rules but also requires a significant amount of computational resources. Having singular highly frequent values also naturally means that associations including those variables will have a high support and confidence making them interesting despite them potentially not having any real correlation. Each variable also has multiple responses thereby increasing the total number of associations making it arduous to interpret the associations between variables.

Classification: How people voted in the Referendum

Including both attributes would introduce redundancy into the classfication models. Redundant features can lead to overfitting and make the model less interpretable. Further more if the information from the target variable is included in the features used to predict

it (since RB1 is identical), the performance estimates will be overly optimistic. Additionally removing the redundant RB1 will improve computational efficiency for a large dataset such as this by reducing the dimensionality of the dataset and the cost of training the model.

Linear Classifier Confusion Matrix:

	Actual	
Predicted	FALSE	TRUE
FALSE	416	84
TRUE	89	539

Decision Tree classifier Confusion Matrix:

	Actual	
Predicted	FALSE	TRUE
FALSE	452	81
TRUE	53	542

SVM classifier Confusion Matrix:

	Actual	
Predicted	FALSE	TRUE
FALSE	437	92
TRUE	68	531

Neural Net classifier Confusion Matrix:

	Actual	
Predicted	FALSE	TRUE
FALSE	184	321
TRUE	82	541

The Linear classifier model has a high metrics all around. The classifier correctly identifies a large proportion (82.38%) of individuals who voted yes. The high specificity (85.62%) indicates the classifier correctly identifies a large proportion of individuals who voted no. In general, the model performs well overall in correctly classifying both yes and no votes with an accuracy of 84.66%.

Having a look at figure 6 which shows the most significant attributes for classifying how people voted in the logistic regression model, individuals with higher values of the significant predictor variables are more likely to vote yes while those with lower values are less likely to vote yes.

The neural net classifier had the worst predictive performance out of all the models but it's best to keep in mind that the neural network model was used with the default parameters without any changes. The results achieved by a neural network classifier are influenced by various factors, including the architecture of the network and the choice of activation functions. The model has a high sensitivity of 86.83% showing it correctly identifies a large proportion of individuals who voted yes. The model however has very low specificity of 36.43% indicating the model does not perform well in correctly identifying individuals who voted no. This may suggest that the classifier tends to be more biased towards predicting yes votes. It's overall performance was moderate with a accuracy of 64.27%. The interpretability of a neural network is also poor relative to a classifier such as logistic regression or decision trees due to its black box nature therefore is harder to assess how it determined prediction classes.

In the same way, SVM classifier is also hard to interpret compared to simpler models. However, it's perfomance was much better than the neural net's. It had relatively high metrics across the board and provided good predictive performance with a sensitivity of

0.8653, specificity of 0.8523 and accuracy of 0.8582. SVM's are generally effective for binary classification tasks, moreso when there is a clear margin of separation between classes, whereas neural networks might struggle. SVM's are also less prone to overfitting in high-dimensional spaces.

But the model with the best predictive performance was the Decision tree. It had a sensitivy of 89.50%, specificity of 87.00% and accuracy of 88.12%. Decision trees are inherently interpretable because they represent decision rules in a tree-like structure. Decision trees can also capture non-linear relationships better between features and the target. They automatically select the most informative features for splitting, ignoring irrelevant ones making them more robust to noise. This property helps prevent overfitting and improves generalization performance, especially in datasets with high dimensionality.

Regression: Population Demographic

A Major benefit of the decision tree for a large feature space is it's model complexity. They have a simple and intuitive structure, making them easy to interpret and faster to implement compared to neural networks which have higher computational complexity. In a neural network, all input variables will be assigned weights, even if they don't significantly determine the target variable and this may cause the neural net to overfit the data. Whereas decision trees inherently handle feature selection and automatically selects the most informative features during the tree-building process. They explicitly evaluate the predictive power of each feature and select the best split points based on criterias such as reduction in variance for regression. Extreme values of irrelevant explanatory variables will not be able to influence the prediction outcome in a decision tree. For these reasons, The decision tree model was used to predict weight_final_ref variable.

The tree was optimised using a grid to tune the minsplit, maxdepth, minbucket and cost complexity parameters. The grid was iterated over to create multiple models with varying parameters in order to find parameter values of the best performing model. The model with the best performance had the following parameter values cost complexity=0.00435, minsplit=15, maxdepth=5, and minbucket=3. The model was then pruned using `rpart`'s prune function with a cost complexity of 0.02 because the xerror stopped decreasing significantly after the 5th split which corresponded to a cp of 0.02. The model returned a RMSE of 0.0957, R-Squared value of 0.5763 and MAE of 0.0630 suggesting strong predictive performance.

More Complex Classification

SVM classifier Confusion Matrix:

	Actual	
Pred	FALSE	TRUE
FALSE	445	93
TRUE	60	530

Neural Net classifier Confusion Matrix:

	Actual	
Pred	FALSE	TRUE
FALSE	452	58
TRUE	53	565

Decision Tree classifier Confusion Matrix:

	Actual	
Pred	FALSE	TRUE
FALSE	461	84
TRUE	44	539

The decision tree and SVM were partitioned into 70/15/15 (training/validation/testing). Since the neural net is very computationally expensive, it was partitioned into 40/30/30 to hyperparameter tune and find the optimal parameter values. Once that was done, the data was re-partitioned into 70/15/15 and the model was trained using the optimal parameters.

Same as for the regression tree, the classification tree was optimised accordingly and returned the following optimal parameter values: cp=1e-10, minsplit=30, maxdepth=4, minbucket=5. The tree performed extremely well achieving the best performance with an accuracy of 88.66%, sensitivity of 91.29%, specificity of 86.52%, F1-score of 87.81% and ROC-AUC score of 0.921.

The SVM model's radial kernel turned out to be better performing for this particular dataset. It must be due to the non-linear relationships between the features and target in the data as well as the high dimensionality of the data. The cost and gamma parameters were tuned to optimise the svm model, settling with a cost of 5. The model performed very well almsot on par with the decision tree with an accuracy of 88.03%, sensitivity of 90.69%, specificity of 85.87% and F1-score of 87.16%. The model had a ROC-AUC of 0.956 displaying really strong performance.

The neural net was optimised by tuning the hidden layers and number of neurons as well as the threshold parameter. The best model had a threshold of 0.05 and 2 hidden layers, 5 neurons in the first layer and 4 neurons in the second layer. The model obtained an accuracy of 71.22%, sensitivity of 87.34% and specificity of 84.65%.

Clustering

To find the optimal k number of clusters, The elbow method and silhouette method were utilised from the `factoextra` library. The results shown in figures 15 and 16 show that according to the elbow method k=4 is optimal whereas the silhouette method shows k=5 is optimal. I decided to go with the lower k=4 for a few reasons. With lower number of clusters, the results can be more simple to interpret. There is less risk of overfitting the model to noise or specific characteristics of the dataset. Each cluster also had substantial observations meaning the number of partitions were not excessive. While the total within cluster sum of the squares was slightly lower with 5 clusters, the difference wasn't too significant.

The sum of the within cluster sum of squares (wcss) with 4 clusters was 72.6336. This value represents the compactness of clusters, indicating how close the data points are to their respective centroids. It is a evaluation metric used to assess how effectively the

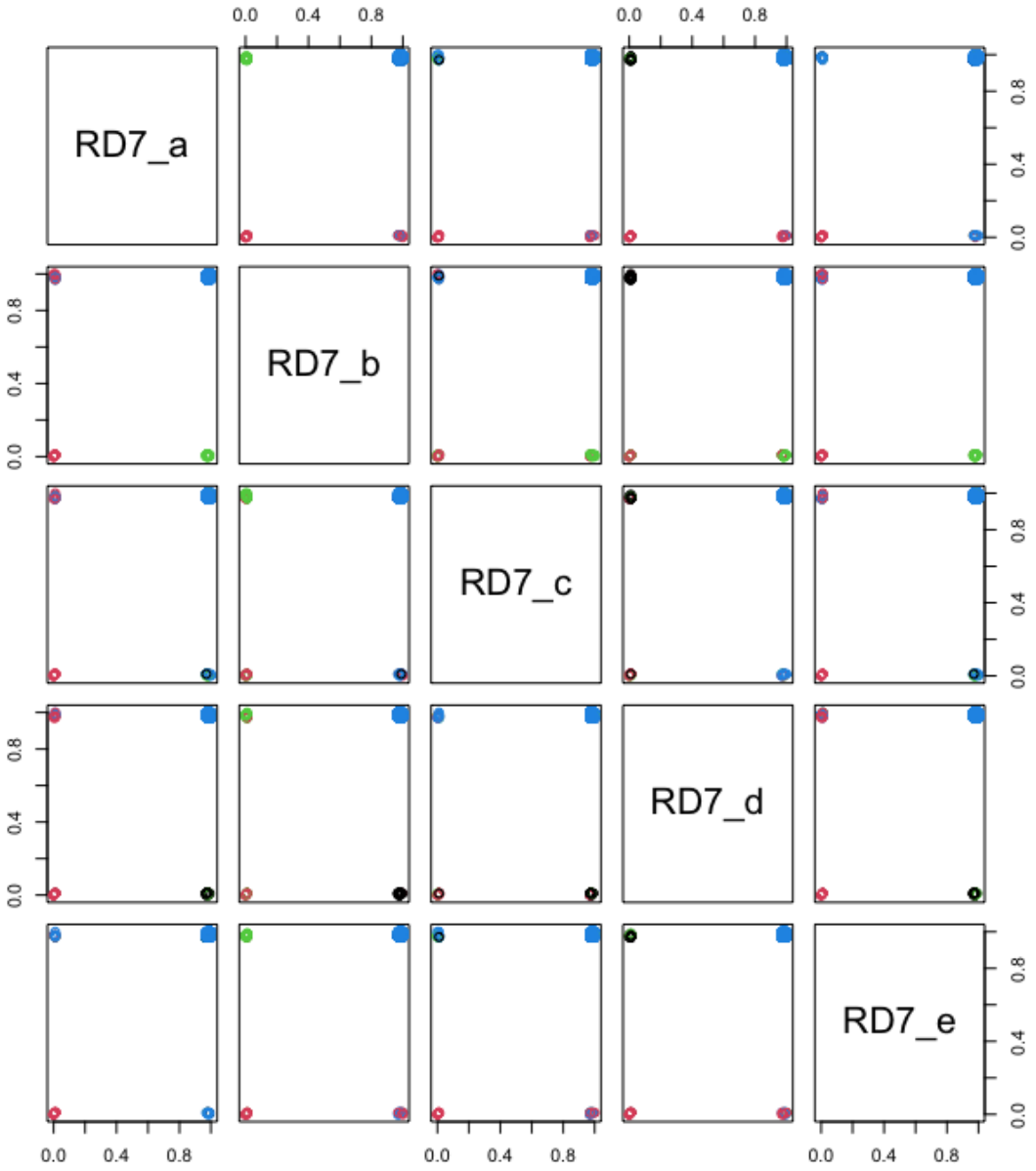
clustering algorithm has grouped similar data points together. A lower wcss is generally considered ideal as it indicates more compact and well-separated clusters. The wcss value reduces as the number of clusters (k) increases because there will be less observations in each cluster and each observation will be closer to the cluster mean. Which is why increasing k to reduce wcss isn't ideal as having excessive numbers of clusters opposes the main objective of building a model that groups observations with similar characteristics together without overfitting.

Having a look at figure 18, cluster 1 has relatively low values across all the variables. The respondents in this cluster tend to show lower agreement with the variable's statements compared to the other clusters. They are less likely to agree with Australia's multiculturism, importance of government support for reconciliation and the need for Indigenous representation.

Cluster 2 has high values for all variables. The respondents in this cluster strongly agree with statements across the board expressing strong feelings and support for multiculturism, government support for reconciliation, indigenous representation and are against the notion that Indigenous people can improve their situation by trying harder.

Cluster 3 has mixed responses, with high agreement on some variables while low agreement on others. Respondents in this cluster strongly agree with the statement Australia is better with many diverse racial groups, they agree the First nation's people should have a voice in matters affecting them but they also agree that Indigenous people can improve their situation by trying harder. This group however seems to disagree that the federal government should help improve reconciliation.

Cluster 4 has high values for all variables across the board showing strong agreement will every statement.



The scatterplot shows the variables RD7_b and RD7_a are strong influencers of the clustering as they show points belonging to clusters that are not observed in the scatterplots containing the other variables.

Qualitative Summary of Findings

Several different data mining techniques and machine learning techniques for classification, regression and clustering were implemented to explore and analyse this dataset. Given the large number of explanatory variables, the decision tree algorithm proved to be a very useful and efficient resource for classification and regression. Aside from being computationally inexpensive compared to other techniques such as neural networks, the decision tree had the best performance and was ideal for this dataset as there is no requirement to conduct extensive data exploration or pre-sort the variables

that are important to predicting the target attribute. Furthermore hyperparameter tuning a regression or classification tree is much faster and less expensive compared to other models so the results can be optimised relatively quicker.

References:

[1] Biddle, Nicholas; McAllister, Ian, 2023, "ANU Poll 57/Australian Constitutional Referendum Survey (ACRS) (October 2023): Aboriginal and Torres Strait Islander Voice to Parliament", doi:10.26193/13NPGQ, ADA Dataverse, V4

[2] ANU Poll Dataverse. Available at: https://dataverse.ada.edu.au/dataverse.xhtml?alias=anupoll (Accessed: 15 April 2024).

Appendix:

```
                A1         RA2_a        RA2_b         RA2_c        RA2_d
A1     1.0000e+00 -0.009796482 -0.01305899 4.790185e-05 -0.02009077 -
RA2_a -9.7964e-03  1.000000000  0.87869233 5.920731e-01  0.46239748
RA2_b -1.3058e-02  0.878692335  1.00000000 5.919558e-01  0.44945002
RA2_c  4.7901e-05  0.592073084  0.59195579 1.000000e+00  0.30197672
RA2_d -2.0090e-02  0.462397480  0.44945002 3.019767e-01  1.00000000
RA2_e -8.5314e-03  0.290851253  0.29244590 3.817772e-01  0.31186000
```

```
          A1 RA2_a RA2_b RA2_c RA2_d RA2_e
A1     1.00 -0.01 -0.01  0.00 -0.02 -0.01
RA2_a -0.01  1.00  0.88  0.59  0.46  0.29
RA2_b -0.01  0.88  1.00  0.59  0.45  0.29
RA2_c  0.00  0.59  0.59  1.00  0.30  0.38
RA2_d -0.02  0.46  0.45  0.30  1.00  0.31
RA2_e -0.01  0.29  0.29  0.38  0.31  1.00
```

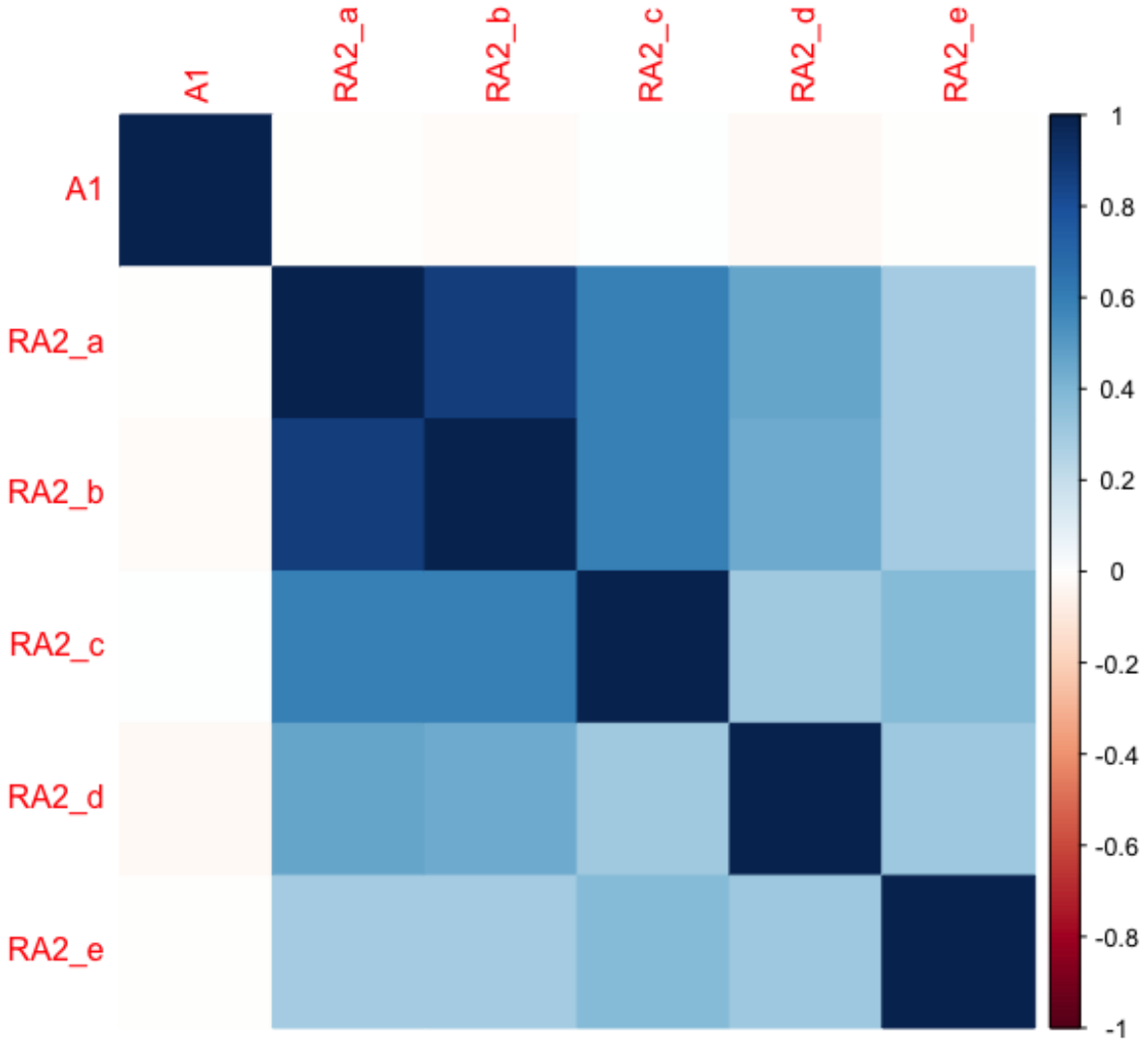


Figure [3]. Correlation Matrix of A1, RA2_a, RA2_b, RA2_c, RA2_d, RA2_e

```
lhs                                      rhs    support confidence cover
{A4_a=2, E1_c=2, E1_d=2, p_atsi=4} => {A1=2} 0.25290 0.6888315  0.367
{A4_a=2, E1_c=2, E1_d=2}           => {A1=2} 0.25669 0.6884933  0.372
{A4_a=2, E1_d=2}                   => {A1=2} 0.26238 0.6884328  0.381
{A4_a=2, E1_c=2, p_atsi=4}         => {A1=2} 0.26072 0.6883605  0.378
{A4_a=2, E1_d=2, E1_e=2}           => {A1=2} 0.25124 0.6883117  0.365
{A4_a=2, E1_d=2, p_atsi=4}         => {A1=2} 0.25788 0.6881720  0.374
{A4_a=2, E1_c=2}                   => {A1=2} 0.26451 0.6880395  0.384
```

```
{A4_a=2, E1_c=2, E1_e=2}           => {A1=2} 0.25290 0.6879433  0.367
{Mode=1, A4_a=2, E1_c=2, E1_d=2}   => {A1=2} 0.25314 0.6872587  0.368
{Mode=1, A4_a=2, E1_d=2}           => {A1=2} 0.25882 0.6872247  0.376
{Mode=1, A4_a=2, E1_c=2, p_atsi=4} => {A1=2} 0.25716 0.6871438  0.374
{Mode=1, A4_a=2, E1_d=2, p_atsi=4} => {A1=2} 0.25432 0.6869398  0.370
{Mode=1, A4_a=2, E1_c=2, E1_e=2}   => {A1=2} 0.25005 0.6868490  0.364
{Mode=1, A4_a=2, E1_c=2}           => {A1=2} 0.26096 0.6868372  0.379
{A4_a=2}                           => {A1=2} 0.27233 0.6867902  0.396
```

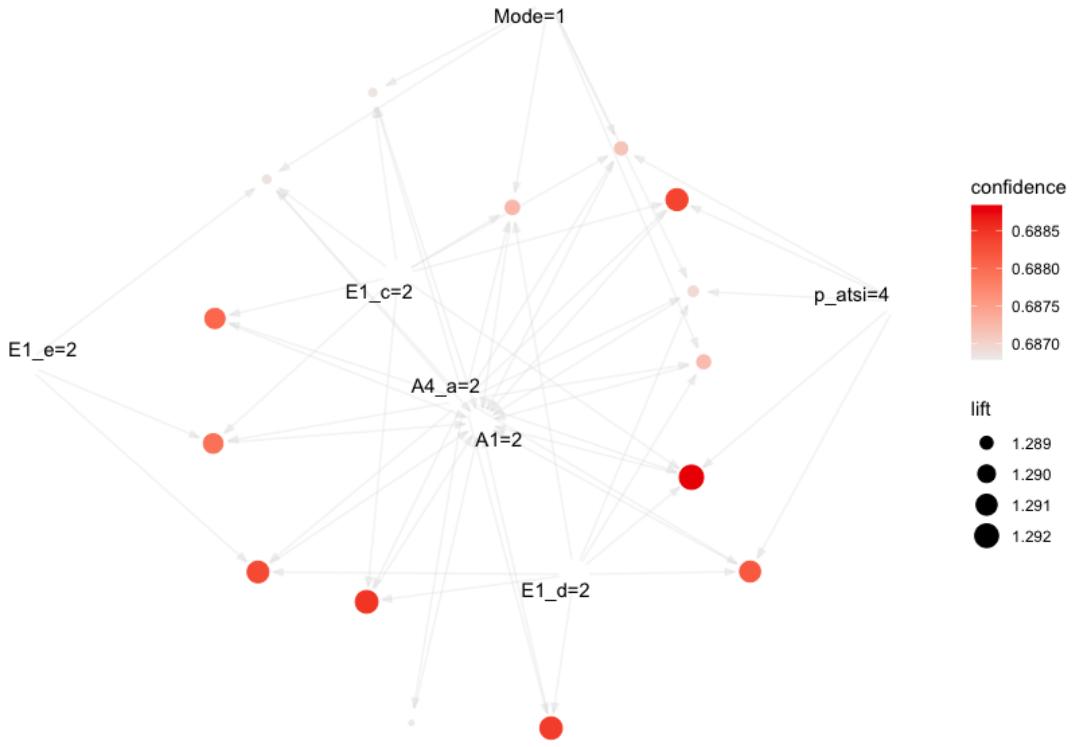


Figure [5]. graph of top 15 association rules for A1 sorted by confidence

```
                Estimate Std. Error z value Pr(>|z|)
(Intercept)      2.991e+00  1.217e+00   2.458 0.013969 *
A1              -1.762e-01  7.146e-02  -2.466 0.013672 *
RA4             -5.492e-01  1.110e-01  -4.947 7.54e-07 ***
RB5             -1.878e-01  3.406e-02  -5.513 3.52e-08 ***
RB12_a           9.695e-02  3.094e-02   3.134 0.001726 **
RB12_b          -8.773e-02  3.122e-02  -2.810 0.004948 **
RB12_d           1.029e-01  2.416e-02   4.261 2.03e-05 ***
RC5_a           -2.868e-01  3.114e-02  -9.210  < 2e-16 ***
```

```
RC5_c             -1.065e-01  2.136e-02  -4.988 6.09e-07 ***
RC5_d              2.727e-01  3.030e-02   9.000  < 2e-16 ***
RC6_a              1.148e-01  2.207e-02   5.204 1.95e-07 ***
RC6_b             -1.474e-01  2.207e-02  -6.679 2.41e-11 ***
RD1_h             -1.781e-01  3.432e-02  -5.190 2.10e-07 ***
RD1_i             -1.059e-01  3.541e-02  -2.991 0.002782 **
RD1_k              1.136e-01  3.813e-02   2.980 0.002886 **
RD2_a             -1.275e-01  3.861e-02  -3.302 0.000960 ***
RD4_b             -1.618e-01  2.921e-02  -5.539 3.04e-08 ***
RD4_d              1.971e-01  2.884e-02   6.833 8.33e-12 ***
RD6               -4.513e-02  1.322e-02  -3.413 0.000643 ***
p_atsi             1.444e-01  3.849e-02   3.752 0.000176 ***
weight_final_ref -2.649e-04  2.681e-05  -9.882  < 2e-16 ***
---Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

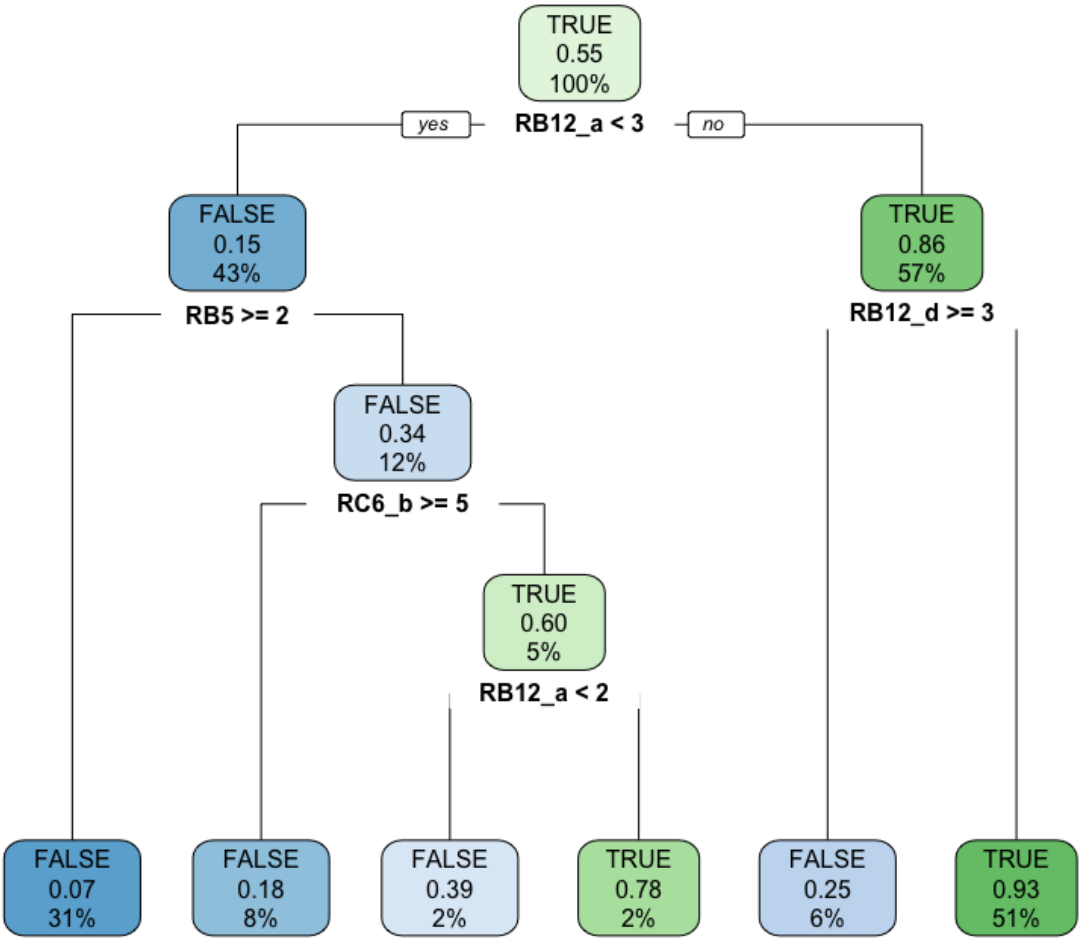


Figure [7]. Classification Tree for question 4

```
RD4_b       RD4_d       RB12_d      RC5_a       RD1_h       RC5_
1.764916    1.754769    1.179980    1.166127    1.156737    1.00
RC6_b       RB12_a      RB5         RD1_e       RD2_a       RC7_
9.507e-01   9.432e-01   8.599e-01   7.420e-01   7.359e-01   7.23
```

```
  minsplit maxdepth           cp minbucket     error
1       15        5 0.0043547483         3 0.4467056
2        8        5 0.0006487686         2 0.4474330
3       11        5 0.0049380900         8 0.4477362
4       14       15 0.0046094490         2 0.4484248
5       20        5 0.0022558438         7 0.4493242
```

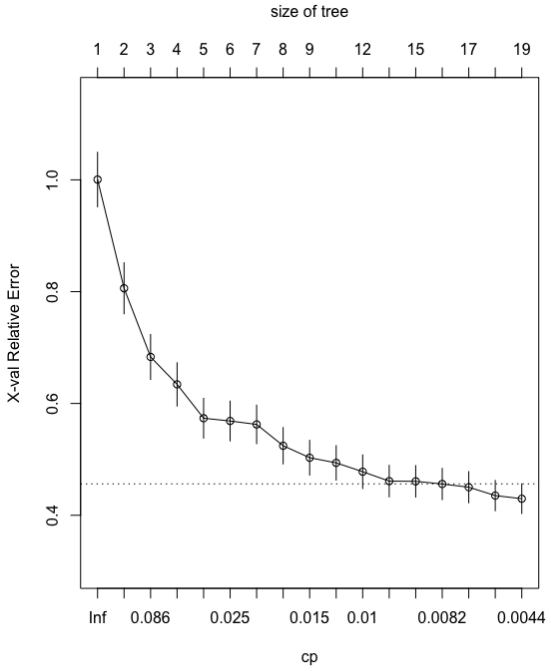


Figure [10]. Complexity parameter plot against Cross-validated error rate

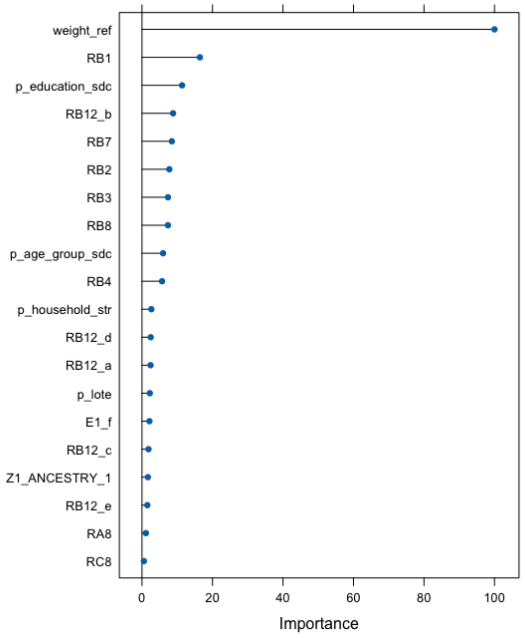


Figure [11]. important features in predicting weight_final_ref

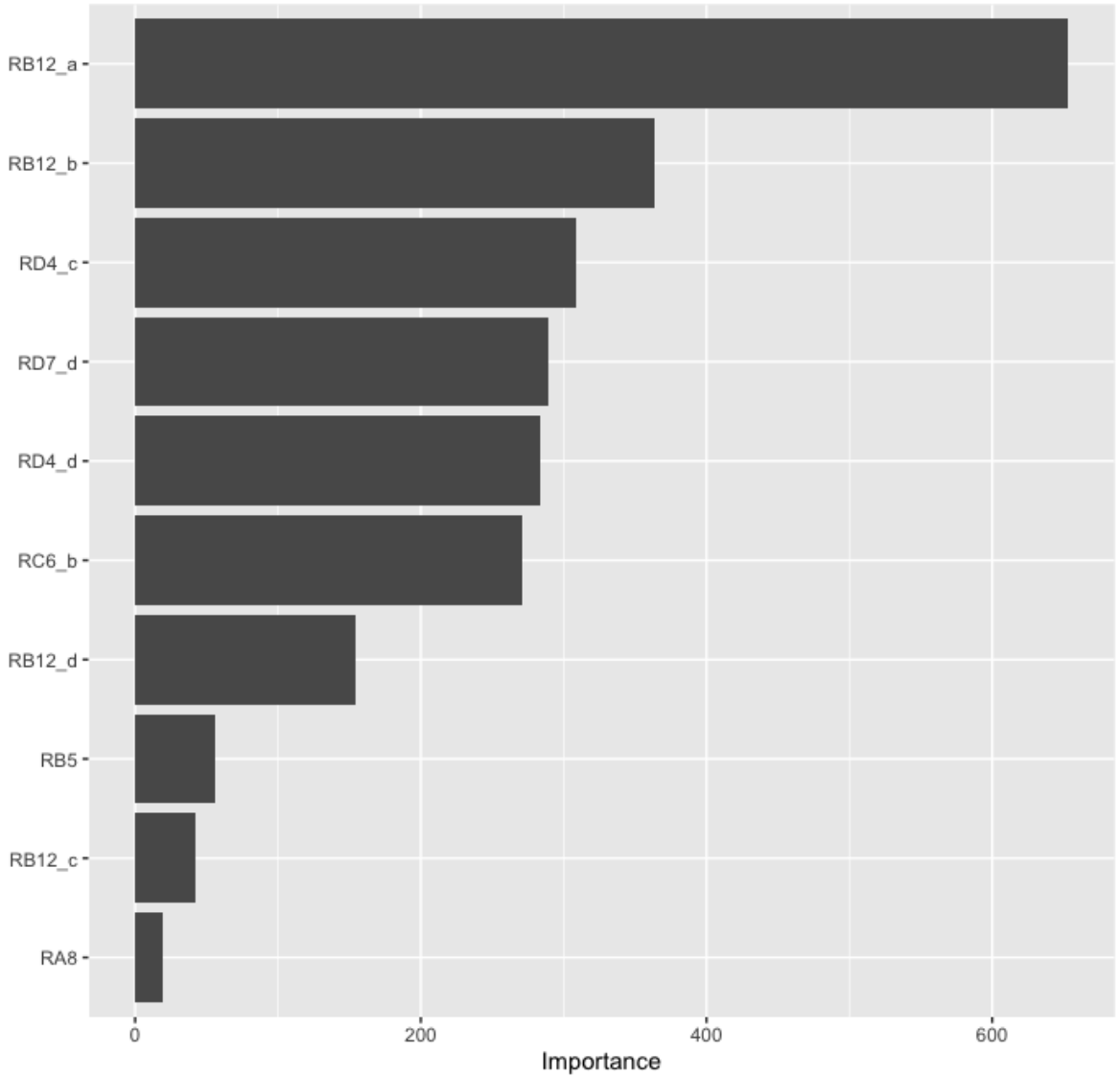


Figure [12]. Important features in predicting Ass2_voted

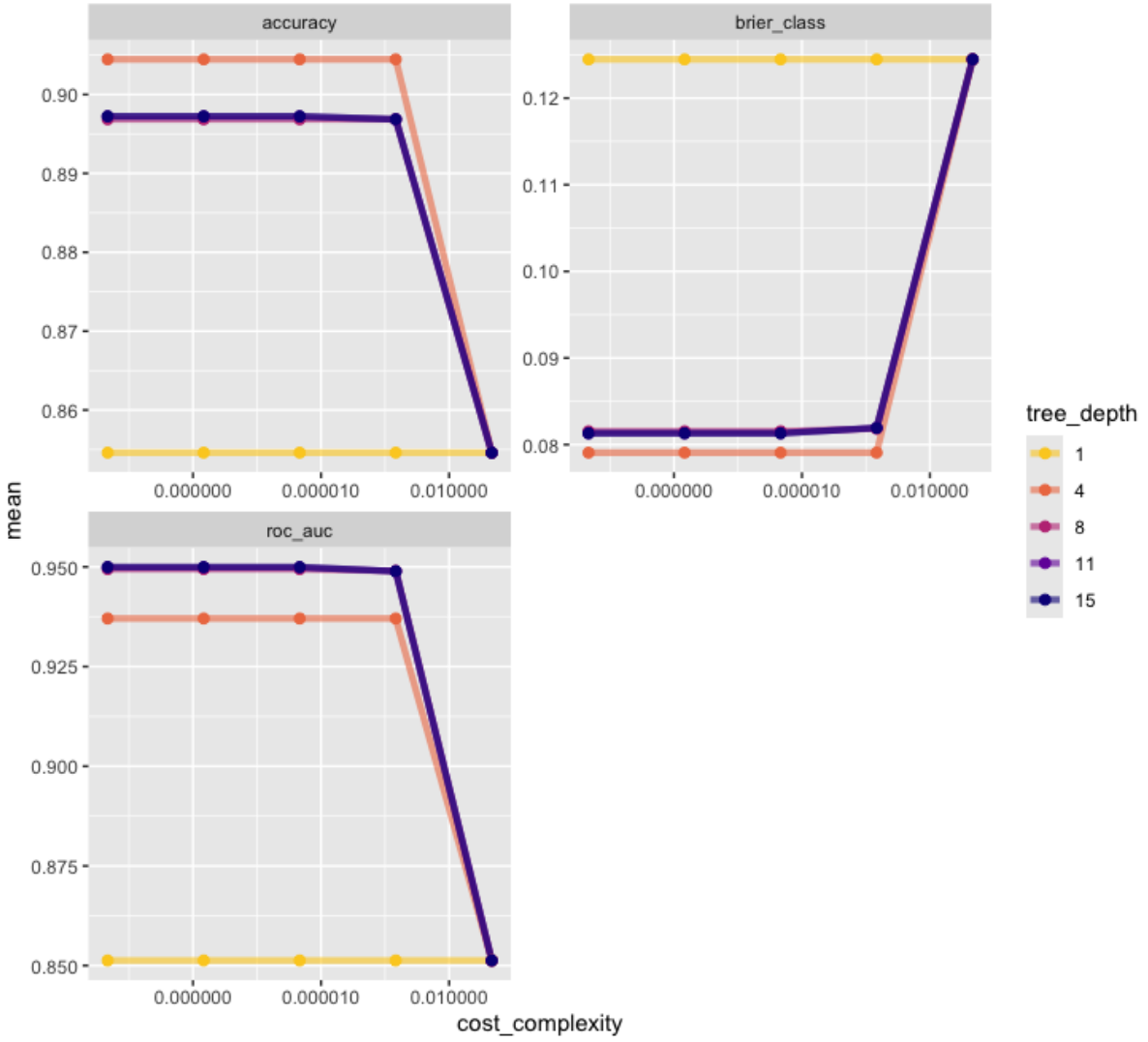


Figure [13]. Optimal tree depth and cost complexity for classification tree

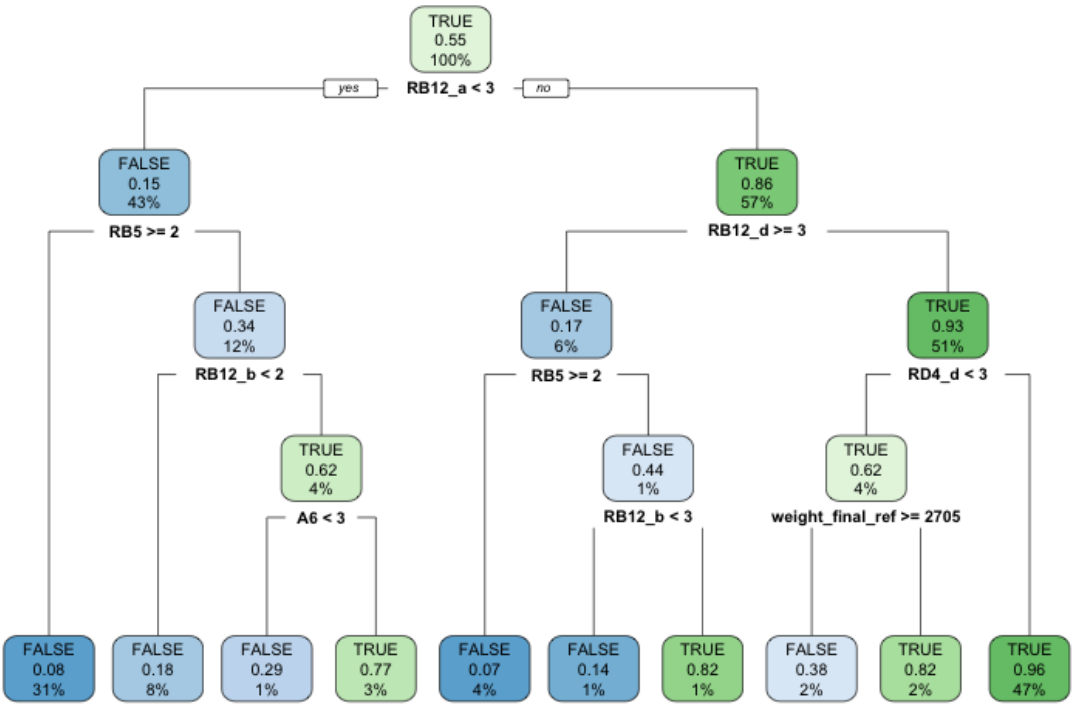
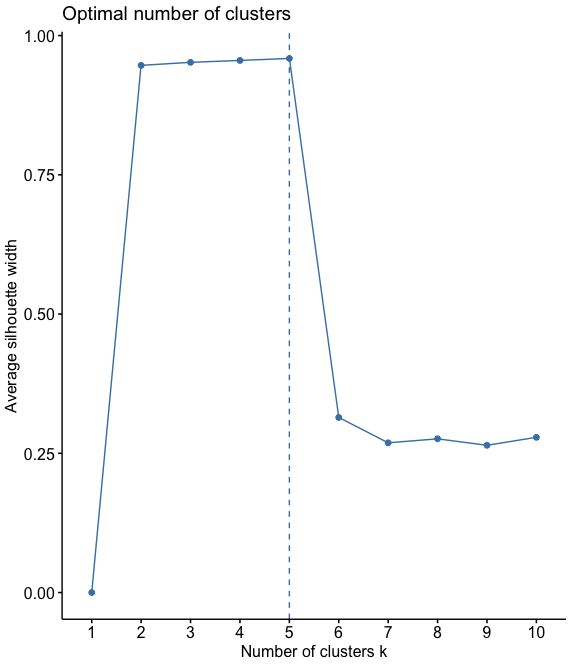
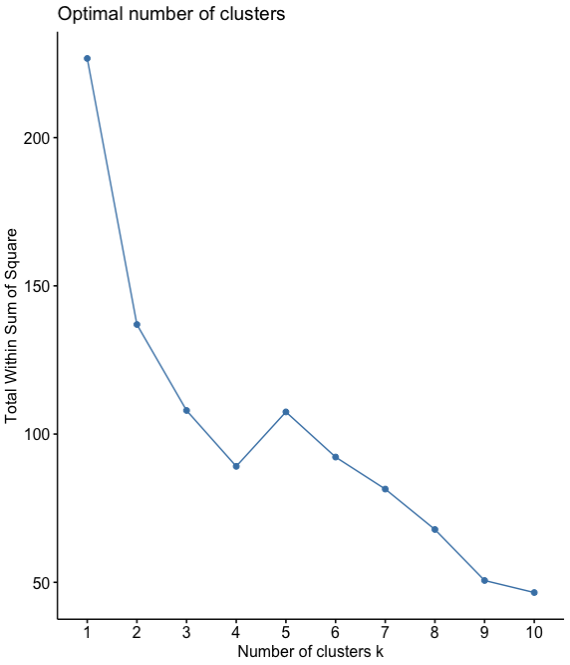


Figure [14]. Optimised classification tree



Figure[15]. Silhouette method to find optimal number of clusters for Question 7



Figure[16]. Elbow method to find optimal number of clusters for Question 7

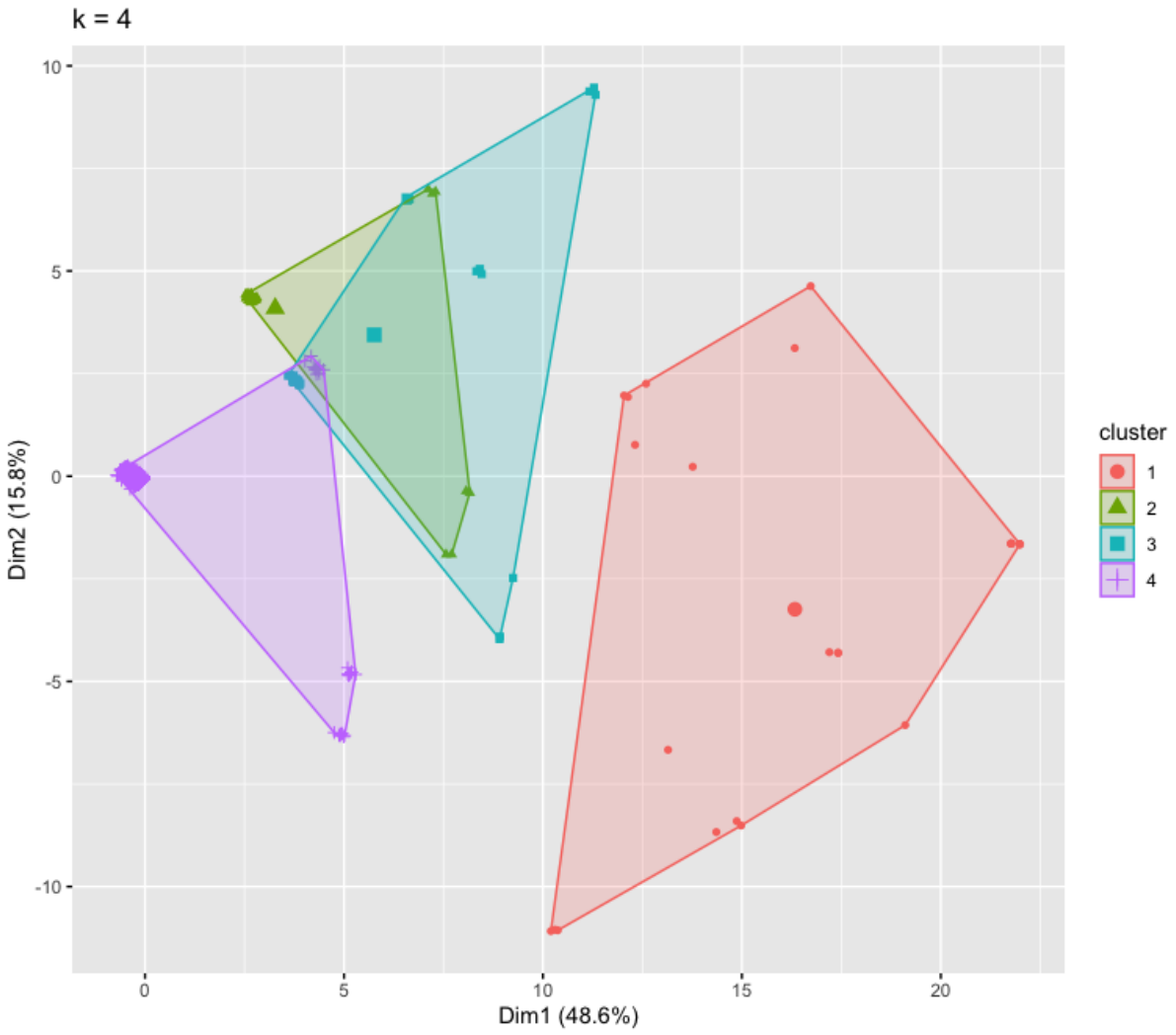


Figure [17]. Clustering of RD7_a,RD7_b,RD7_c,RD7_d,RD7_e

Cluster	RD7_a	RD7_b	RD7_c	RD7_d	RD7_e
<int>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	1	0.0843	0.320	0.395	0.318	0.201
2	2	0.943	0.981	0.925	0.00883	0.941
3	3	0.954	0.00787	0.796	0.745	0.928
4	4	0.977	0.980	0.975	0.987	0.976

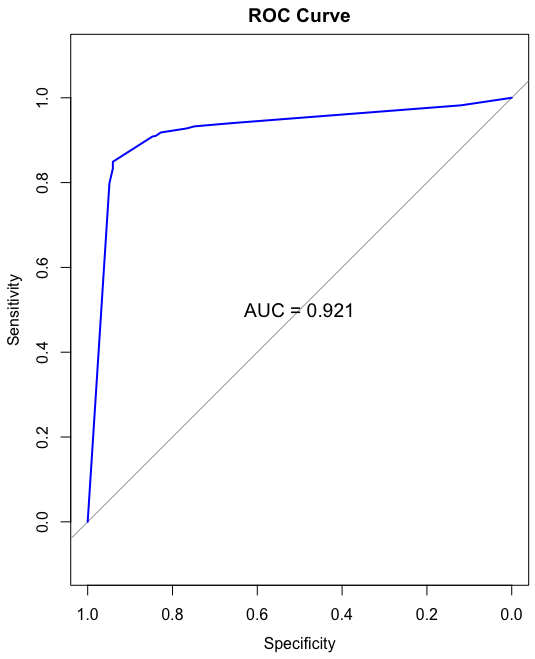


Figure [19]. ROC plot of classification tree

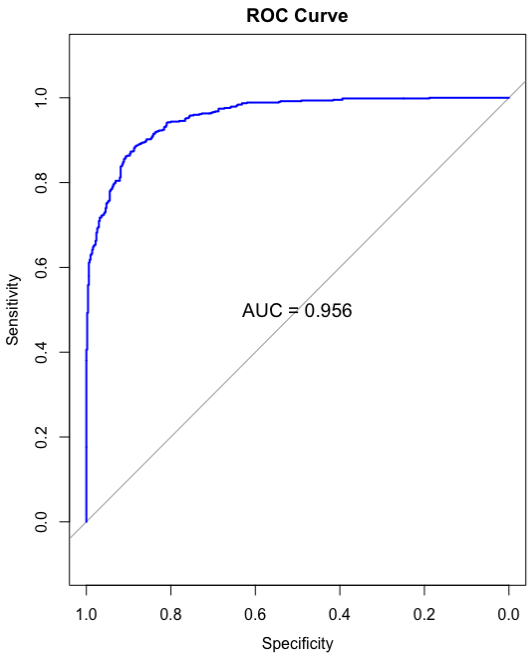


Figure [20]. ROC plot of SVM model