Machine Learning Report Analysis

Done by: Prisca Lim Shi Zhen

Table of contents

1. Data visualization and Unsupervised learning techniques

1.1. Data Visualization

To understand the data set from European Working Condition Survey 2016 before making assumptions and modeling prediction, data has been visualized in plots using unsupervised learning techniques: Principal Component Analysis and K-means clustering (Hierarchy too) to discover any hidden patterns that exist in the data set.

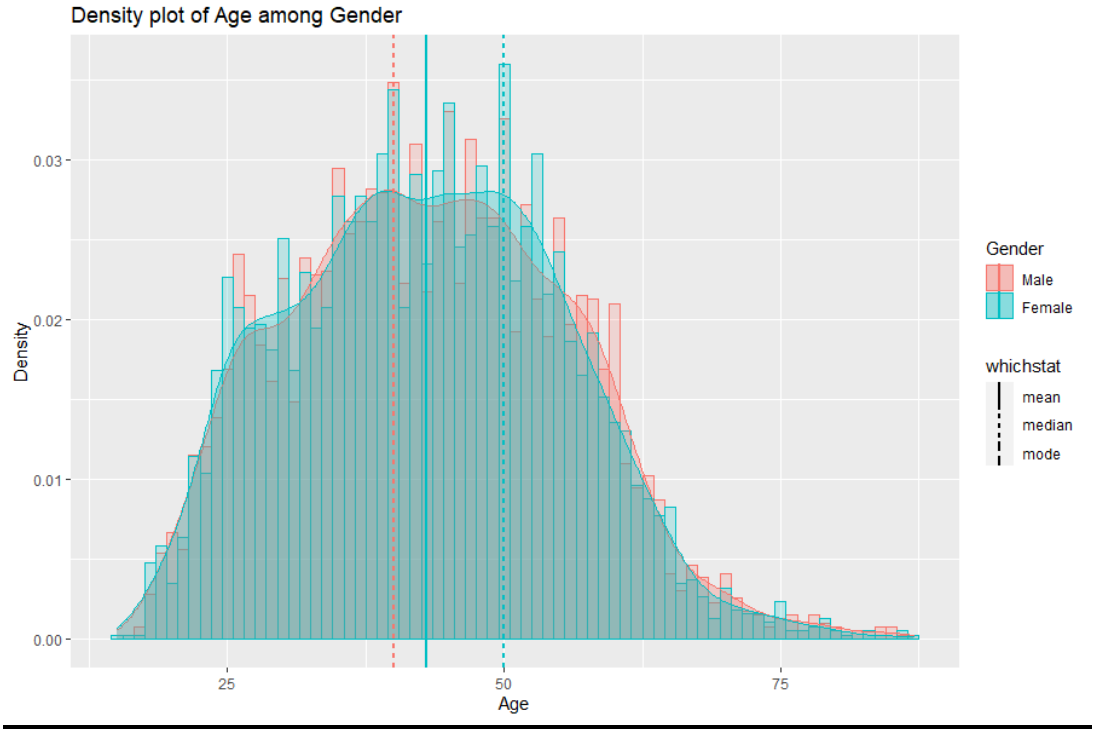

Figure 1.1.1. show the density plot of Age among Genders: Male and Female in orange and blue respectively. The solid line represents the mean, the larger dashed line represents the median, and the smaller dashed line represents the mode of Age among Genders. Note: the mean and mode for males are the same as for females.

According to figure 1.1.1, it can be observed that the distribution for male is positively skewed as the mean and mode are more than the median, indicating that there are younger males compared to females where its distribution is negatively skewed as their mean and mode is less than the median, indicating older female participants were being surveyed. The average and most frequent age in both genders being surveyed are 43, indicating that there are older and more experienced working adults participating in the survey. This may be attributed to location (rural/urban), time (peak/non-peak hours), or criteria of targeted participants of the survey.

The table below shows the questions and responses used in the survey and the data has been summarized in figure 1.1.2. below on its responses from participants per question.

Questions	**Responses**
• Q87a - I have felt cheerful and in good spirits. • Q87b - I have felt calm and relaxed. • Q87c - I have felt active and vigorous. • Q87d - I woke up feeling fresh and rested. • Q87e - My daily life has been filled with things that interest me.	1. All of the time. 2. Most of the time 3. More than half of the time 4. Less than half of the time 5. Some of the time 6. At no time

• Q90a - At my work, I feel full of energy. • Q90b - I am enthusiastic about my job. • Q90c - Time flies when I am working. • Q90f - In my opinion, I am good at my job.	1. Always. 2. Most of the time 3. Sometimes 4. Rarely 5. Never

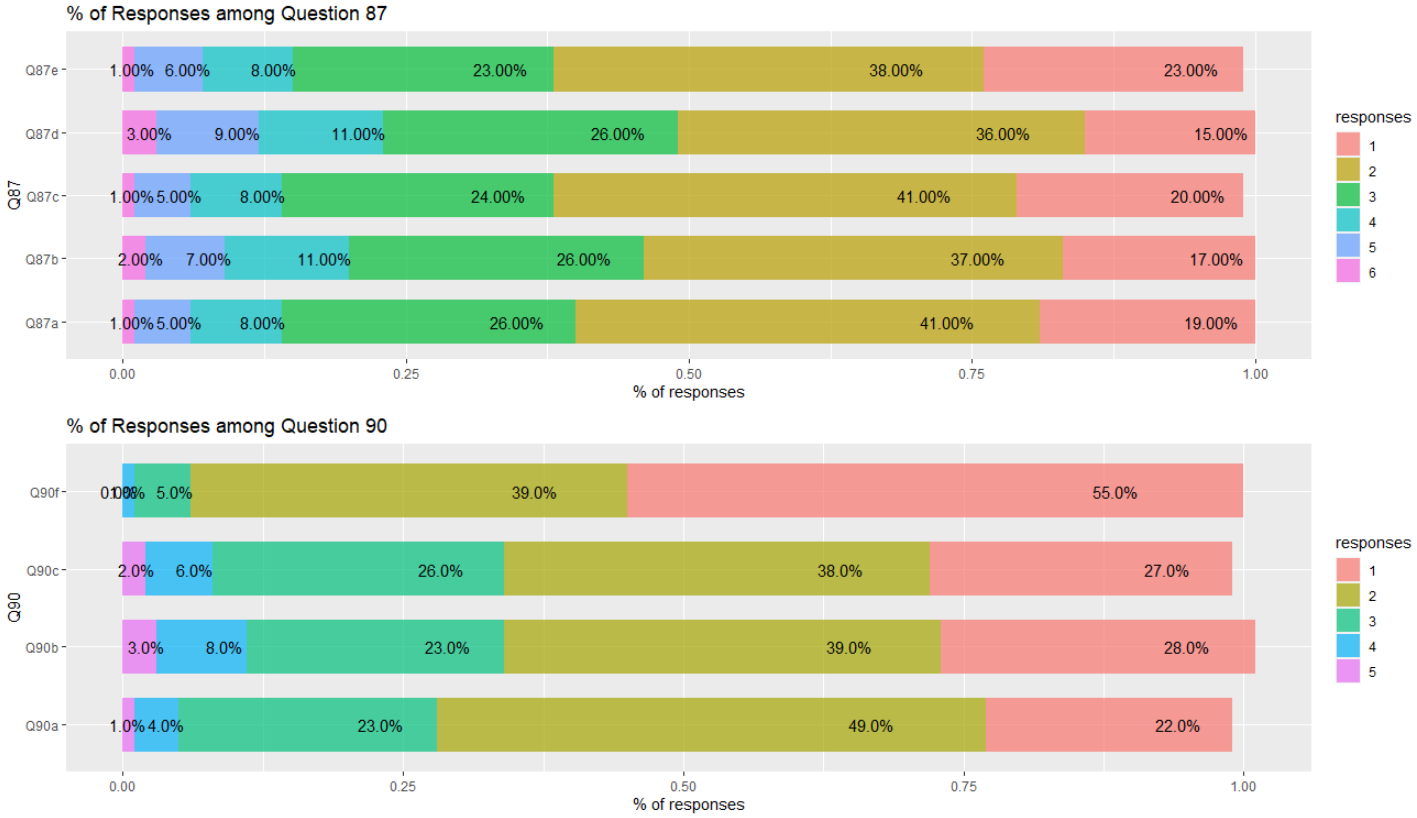

Figure 1.1.2. shows the percentage of responses coded in 1-6 among Questions 87 and 90 extracted from the survey.

According to figure 1.1.2., there is a high number of responses in 1 (all the time/always) and 2 (most of the time) for each question, indicating that workers often feel positive about their personal well-being and working attitude towards their career. Furthermore, there is a certain percentage of high responses in both questions 87 and 90 that are highly similar, suggesting a pattern of correlation among these responses.

1.2. Unsupervised Learning

Unsupervised learning techniques have been used to analyze and discover hidden patterns among the dataset gathered from the survey which allows the firm to study and understand their employees' productivity that may directly or indirectly impact the business core operation. Due to less contribution from variables: Gender and Age, it has been removed as it does not explain much in the variation in the model (see Appendix A1.1. on correlation plot and Appendix A1.2. on scree plot) thus, the final model of PCA2 is used.

In figure 1.2.1. below, it can be observed that based on the scree or elbow point, the first two principal components of PCA2 explain at least 64% of the variation of the data set gathered from the survey, indicating that these variables are sufficient for further prediction modeling. After the second component, there is only a slight increase in model explanation compared to the first two principal components and thus, resulting in a slight trade-off between accuracy

and simplicity. In figure 1.2.2. shows the importance of the contribution of each variable to the principal component model where the highest contribution in the first principal component is question 87a, 87b, 87c, 87d, 87e and the highest contribution in the second principal component are question 90b, 90c 90f. Notably, question 90a has a high contribution in both principal components 1 and 2.

To further analyze hidden patterns among these two components, data has been visualized in a biplot with K-means clustering in figure 1.2.3. for PCA2 model, where optimal cluster scree or silhouette plot (see Appendix A1.3.) is optimal at k=2 non-overlapping clusters group as it achieves the highest decrease in total within sum of squares error. It can be observed that question 90a, 90b, 90c, and 90f are highly correlated as it increases and decreases in the same direction in the second component and first component respectively, but is negatively correlated with question 80a,80b,80c,80d.80e, and 80f as it increases in opposite direction, indicating that positive working attitude towards their career is not necessarily the same as personal well-being.

Responses gathered from participants have been grouped into 2 clusters in figure 1.2.4. to show the percentage of responses corresponding to each question in each group of clusters. It can be observed that there is a higher number of overall responses between 3 to 6 in cluster 1 for each question except for question 90f and a higher number of overall responses between 1 and 2 in cluster 2 for each question, indicating that the first cluster of participants does not feel as positive as participants in cluster 2 in terms of personal well-being and working attitude towards working environment. Possibly in cluster 1, may be attributed to higher stress in the working environment such as prolonged working hours, heavy workload, or competition in job promotion among employees, causing an indirect impact on personal attitude outside the working environment to be less positive than in cluster 2. However, participants in both clusters certainly have strong skills confidence in their careers. In conclusion, the level and amount, and workload affect oppositely on their personal well-being (Similar insights mined from hierarchy in Appendix A1.4. and A1.5. below).

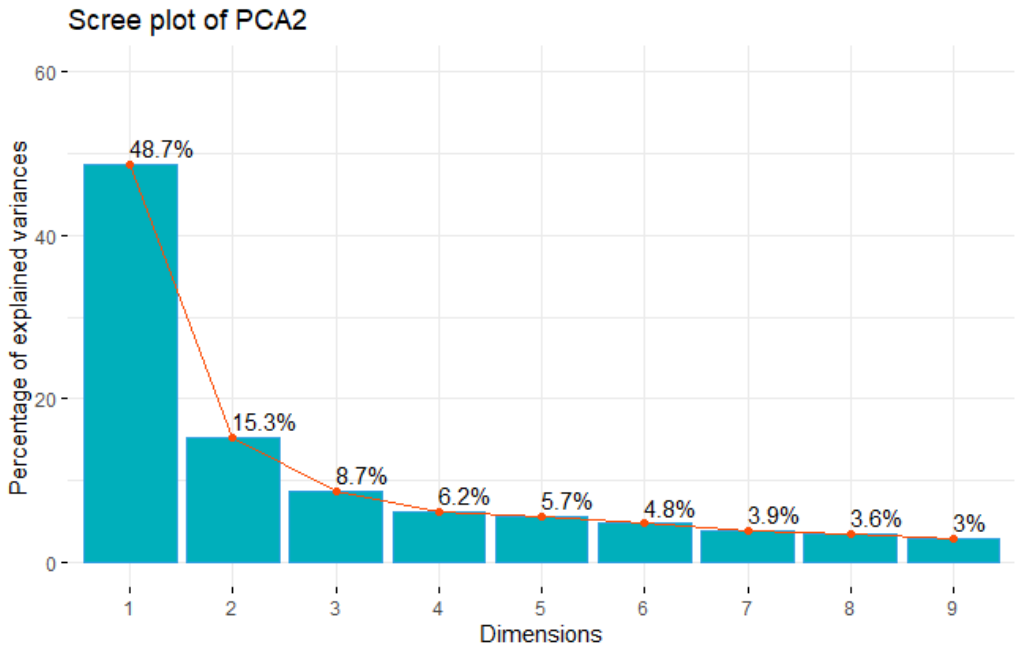

Figure 1.2.1. shows the scree plot of PCA2 where each principal component explains a certain percentage of variances of the dataset.

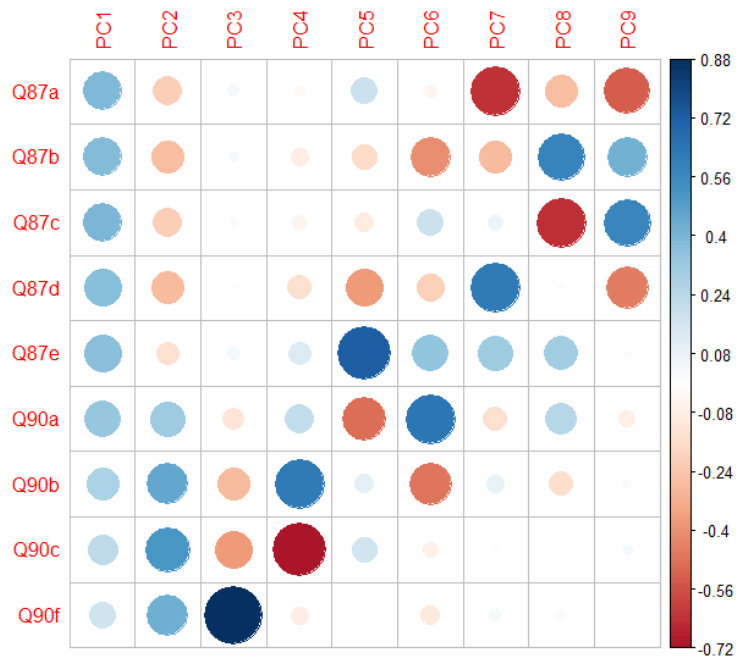

Figure 1.2.2. shows the importance contributor of each variable towards different PCA (dimension.n) models where n ranges from 1 to 9.

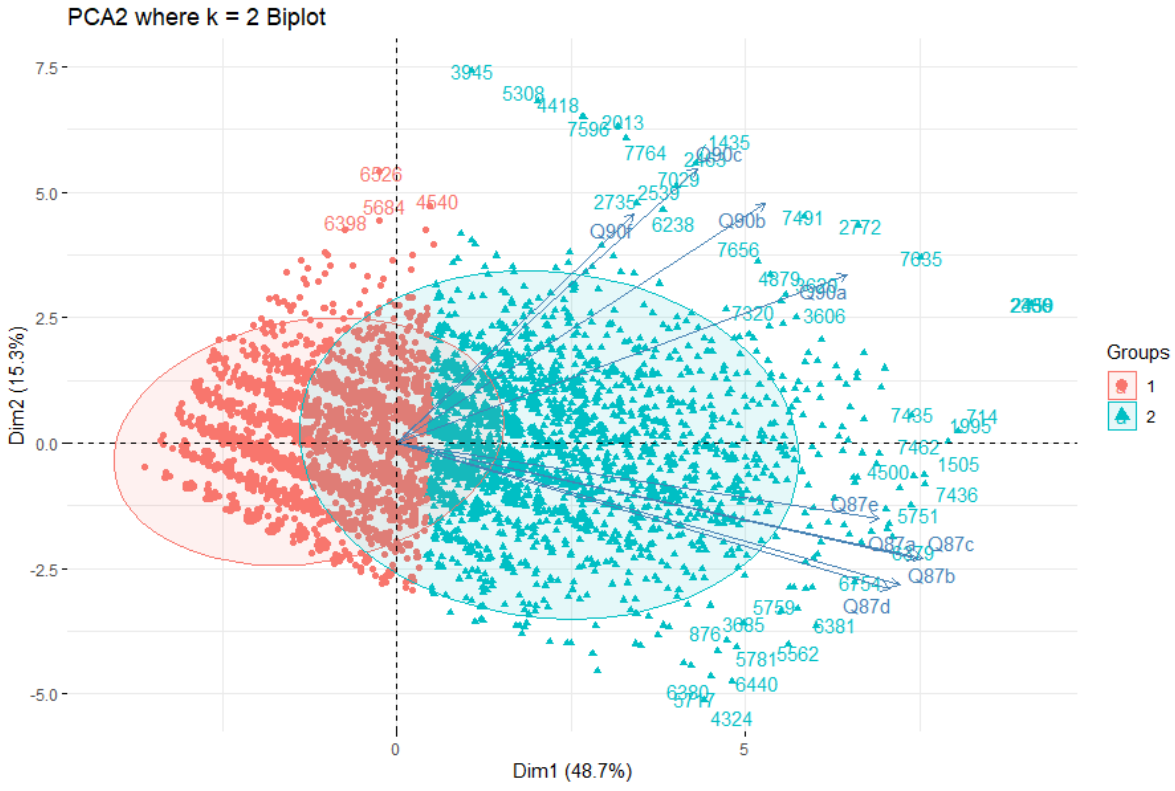

Figure 1.2.3. shows biplot of Unsupervised learning techniques – PCA analysis and K-means clustering where k=2.

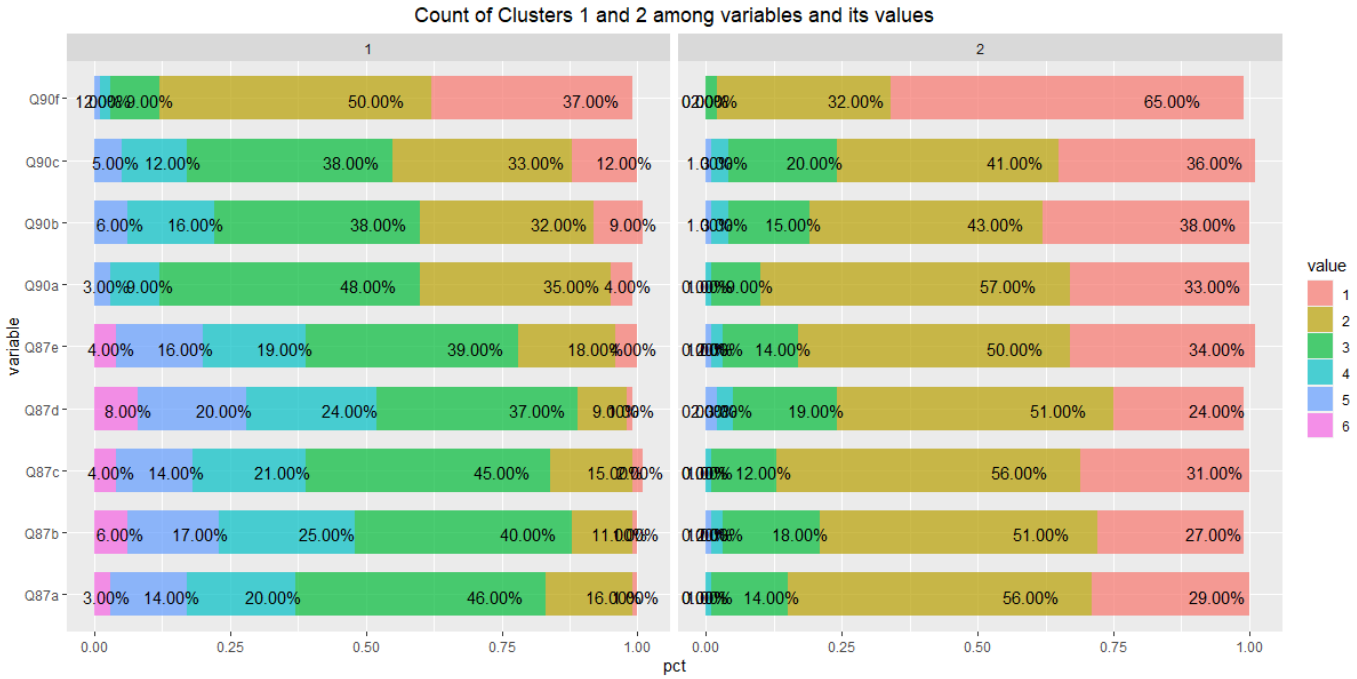

Figure 1.2.4. shows the number of clusters in clusters 1 and 2 among questions 87 and 90 its respective responses.

2. Regression model for G3 final Math and Portuguese grade

To predict the G3 final grade of Math and Portuguese without prior test grades G1 and G2, these two variables have been removed from the data set. The machine learning algorithm: Linear Regression, Ridge Regression, Lasso Regression, Classification Regression, Random Forest, and Boosting, has been used to predict regression models for G3 final grade of Math and Portuguese. For linear regression, stepwise elimination(bi-direction) has been used to include variables that provide the lowest AIC error metric score that fits well on the prediction model to minimize error from prediction against actual observation. For the remaining Regression models, it uses the full dataset as their individual algorithm structure provides optimal configuration to select and predict the best model that generates the lowest RMSE (such as lasso regression with subset feature selection or random forest with model stabilizer). Furthermore, no multicollinearity was detected in the full dataset, indicating that all the variables are independent and effective in determining changes used to predict G3 final grade and thus, able to easily discern the impact individual variables have on G3 final grade. Additionally, validation approaches: train-test split and cross-validation have been used to determine how well the predicted model on prior (seen data) dataset can predict future (unseen data) outcome of G3 final grade.

According to figure 2.1. below, it can be observed that Ridge Regression is a better prediction model as its lowest test set RMSE of 0.81 and 0.93 in predicting the G3 final grade of both Math and Portuguese respectively, indicating that the fit of the predicted G3 final grade is close to actual G3 final grade. Furthermore, none of the variables has a coefficient of 0, indicating that the full dataset does not overfit the model and is somewhat of equal importance in predicting changes to the outcome of the G3 final grade model.

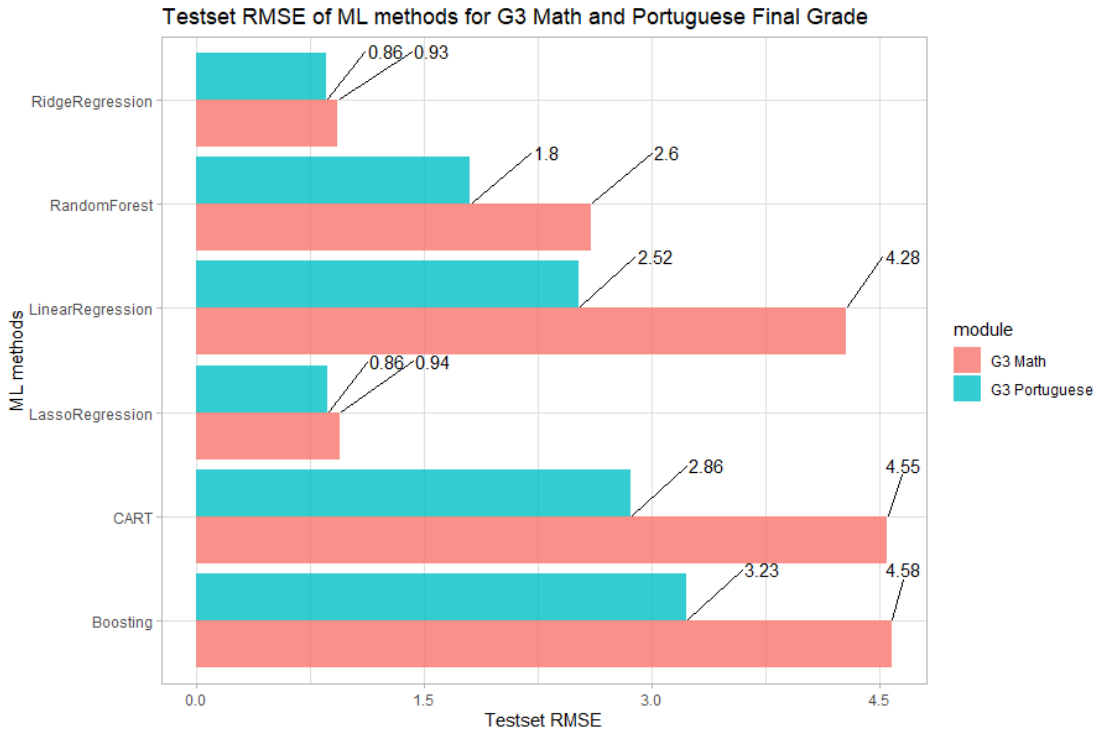

Figure 2.1. shows the Testset Root Mean Square Error (RMSE) of individual machine learning algorithms used to predict the G3 final grade of Math and Portuguese.

By analyzing the impact of individual variables have on the G3 final grade model for Math in figure 2.2., the number of past 'failures' has the highest absolute coefficient of 0.138, followed by the frequency of going out with friends (goout) with an absolute coefficient of 0.105. Due to the negative coefficient of both variables, this indicates that students who achieve a higher number of past failures while other variables are held constant tend to score poorly in the G3 final grade of Math. Similarly, students who often go out with friends while other variables are held constant tend to score poorly on the G3 final grade of Math. This may be attributed to the lesser time allocated for studying as well as lack of knowledge of math that causes lower confidence, and other contributing factors such as psychological factors: learning disabilities or behavioral issues toward learning (Kamal and Bener, 2022).

By analyzing the impact of individual variables have on the G3 final grade model for Portuguese in figure 2.2., the number of past 'failures' has the highest absolute coefficient of 0.168, followed by the type of 'school': 'GS (baseline reference) or 'MS', with an absolute coefficient of 0.162. Due to the negative coefficient of both variables, this indicates that students who achieve a higher number of past failures while other variables are held constant tend to score poorly in the G3 final grade of Portuguese. Similarly, students who attend Mousinho da Silveira (MS) school while other variables are held constant, tend to score poorly in the G3 final grade of Portuguese. This may be attributed to the lack of teaching faculty or learning facilities provided by Mousinho da Silveira (MS) and other similar reasons stated above where the lack of knowledge or other psychological factors that affect poor grade in G3 final grade for Portuguese (D'Alessandro, 2022). Thus, the number of past failures is the highest contributor that negatively impacts on G3 final grade for both Math and Portuguese.

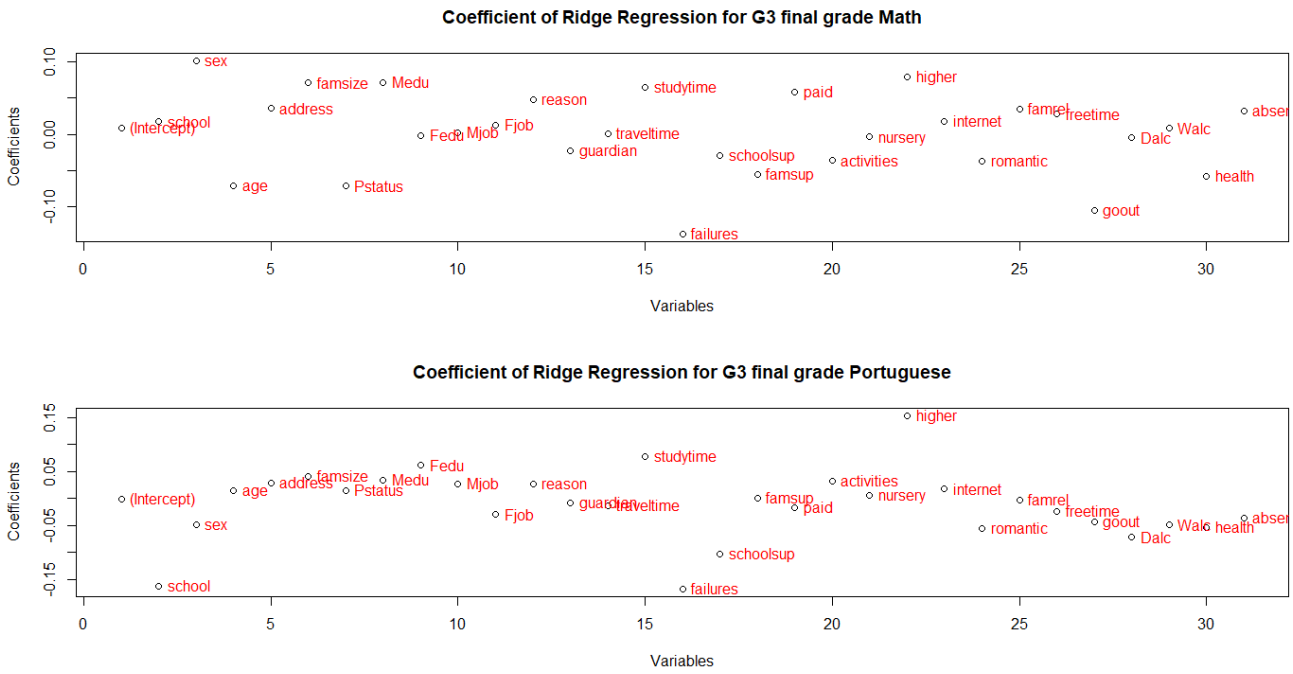

Figure 2.2. shows the coefficient of Ridge Regression for G3 final grade for both Math and Portuguese.

3. Classification model of Bank Subscription.

To predict if clients will subscribe to a term deposit with the bank, a classification model is required for a binary outcome where Yes = 1 and No = 0 (baseline reference) is used. The machine algorithm used for the classification task were Sector Vector Machine, Boosting, Random Forest, Classification Regression Tree, K-Nearest Neighbor, Lasso logistic Regression, Ridge logistic Regression, Logistic Discriminant Analysis, and Logistic Regression. Similarly, before the model evaluation mentioned in section 2 above, stepwise elimination(bi-direction) has been used for Logistic Regression and Linear Discriminant analysis to include variables that provide the lowest AIC error metric score that fits well on the prediction model to minimize accuracy error from prediction against actual observation. Thus, the final model with the lowest AIC included variables: Marital, Education, Housing, Loan, Contact, Day, Month, Duration, Campaign, Poutcome, and Family. As for the other machine learning algorithms, the full dataset has been included in the model due to its complex automated optimization such as subset features, the penalty imposed for large dataset compared or higher dimension/polynomial degree complex classification which tends to provide higher accuracy and credibility than logistic regression (Rawat, 2022). Furthermore, no multicollinearity was detected in the reduced dimension model from stepwise elimination, indicating that these selected variables are independent and effective in predicting bank subscription. Additionally, the train-test split has been used on individual machine learning algorithms to determine how well the predicted model on the prior (seen model) dataset can predict the future (unseen model) outcome for bank subscriptions. The assumption has been made on the threshold for classifying observations for a predicted model where at least 50% will be classified as “yes” and less than 50% will be “no” to bank subscription.

To determine the optimal overall accuracy for predicting bank subscription, sensitivity, specificity, positive predicted value, and negative predicted value performed by machine

learning classifiers have been taken into consideration to determine how well the predicted observations are the actual true positives and negatives among the false positives and negatives. Based on figure 3.1., it can be observed that Linear Discriminant analysis and Random Forest has an overall classification accuracy of 90.4%. However, its sensitivity and specificity differ, at 96.8% and 41% respectively for Linear Discriminant analysis and 97.7% and 34.6% respectively, for Random Forest. This indicates that Linear Discriminant analysis correctly predicts 96.8% of actual customers who will not subscribe (true positive) among the total number of customers who will not subscribe (total positive) and correctly predicts 41% of actual customers who will subscribe (true negative) among the total number of customers who will subscribe (total negative). As for Random Forest, it correctly predicts 97.7% of actual customers who will not subscribe (true positive) among the total number of customers who will not subscribe (total positive) and correctly predicts 34.6% of actual customers who will subscribe (true negative) among the total number of customers who will subscribe (total negative).

To further analyze the true positives and negatives among the predicted positives and negatives, positive (NPV) and negative predictive values (NPV) are used. The accuracy of true positives and negatives are 92.7% and 62.7% respectively for Linear Discriminant analysis and 92% and 65.9% for Random Forest. This indicates that the Linear Discriminant analysis model predicts a customer who will not subscribe to a bank has a 92.7% chance that are actual customers who will not subscribe to the bank, and the predicted customer who will subscribe to a bank has a 62.7% chance that actual customers who will subscribe to the bank. As for Random Forest, the model that predicts a customer who will not subscribe to a bank has a 92% chance that actual customers who will not subscribe to the bank and predicted customers who will subscribe to a bank has a 65.9% chance that actual customers who will subscribe to the bank.

By analyzing each variable included in the Random Forest prediction model for bank subscription, it can be observed that on the left side of the plot in figure 3.2., the ‘duration’ of the call from the last contact (in seconds (for benchmark purposes to facilitate comparison across models)) is the most important variable due to its large decrease in mean accuracy and Gini index of 86.28 and 172.89 respectively when it’s being excluded from the model, indicating that the variable contributes significantly in overall prediction accuracy and provides higher confidence in successful classification (Martinez-Taboada and Ignacio Redondo, 2022). Followed by the ‘month’ of last contract and poutcome (outcome of previous marketing campaign) which has a higher decrease in mean accuracy and Gini index of 39.16 and 78.91 for the month, and 30.65 and 37.59 for poutcome respectively, than other variables in the prediction model. However, because of the nature of the call of duration where duration (in seconds) is known after customers are being contacted, leading to a known outcome of bank subscription, but the duration of the call is unknown before being contacted as the call has yet to occur, resulting in the unknown outcome of bank subscription. Furthermore, based on the correlation shown on the right side of figure 3.2., duration is highly correlated (40.11%) with the outcome of bank subscription and thus, it is not realistically possible to determine the duration of the call and predict bank subscription accordingly as the duration of the call is either known or unknown, directly affecting the accuracy of outcome of bank subscription being either known or unknown and will be excluded in the future for more precise prediction performance.

This leads to the next following highest variable: month of last contact and outcome of previous marketing campaign (poutcome). Based on the correlation plot, the month has slightly positively correlated and the poutcome is slightly negatively correlated of 0.023 and -0.083 respectively with the outcome of bank subscription, indicating that the latter the months since

the last contact while other variables are held constant, the higher the chance that the clients will not subscribe to a term deposit, but the higher the probability of success from previous marketing campaign while other variables are held constant, the higher the overall accuracy and confidence in predicting that the customer will subscribe to the bank. For Linear Discriminant analysis, its standardized coefficient is reflected in figure 3.3. shows similar importance as the Random Forest model on the month of the last contact and poutcome of the previous marketing campaign, where the highest absolute coefficient of the last contact since March and October, and probability of the outcome of 'success' are the most significant in contributing to group classification on bank subscription ('no'/'yes').

Although both models have equal overall accuracy, their accuracy in actual and predicted true positives and negatives differs, and thus, trade-off of accuracy between true positives and negatives. Hence, depending on the bank's willingness to risk slight accuracy on true positives for true negatives or vice-versa, it must align with business objectives. For example, expanding the clients base by promoting campaign to increase the chances of subscriptions from clients who are less likely to subscribe or possibly, to identify, segment, and maintain contact with clients who are likely to subscribe to a long-term deposit as Linear Discriminant analysis model predicts better on clients who are less likely to subscribe to term deposit and Random Forest predicts better on the latter, with a slight trade-off between classifying correctly on actual clients who subscribes or does not subscribe.

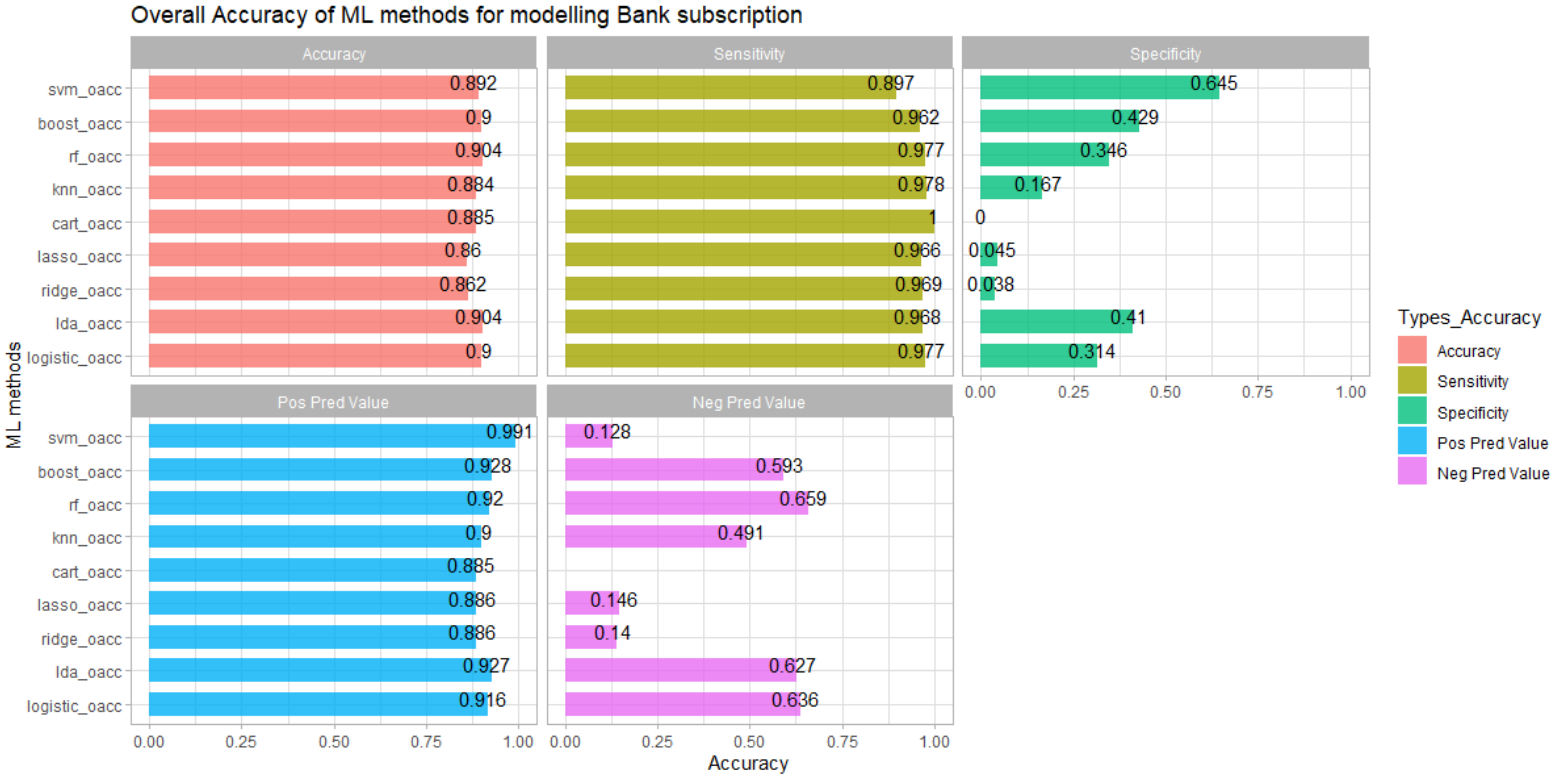

Figure 3.1. shows the overall accuracy of individual machine learning methods for bank subscriptions.

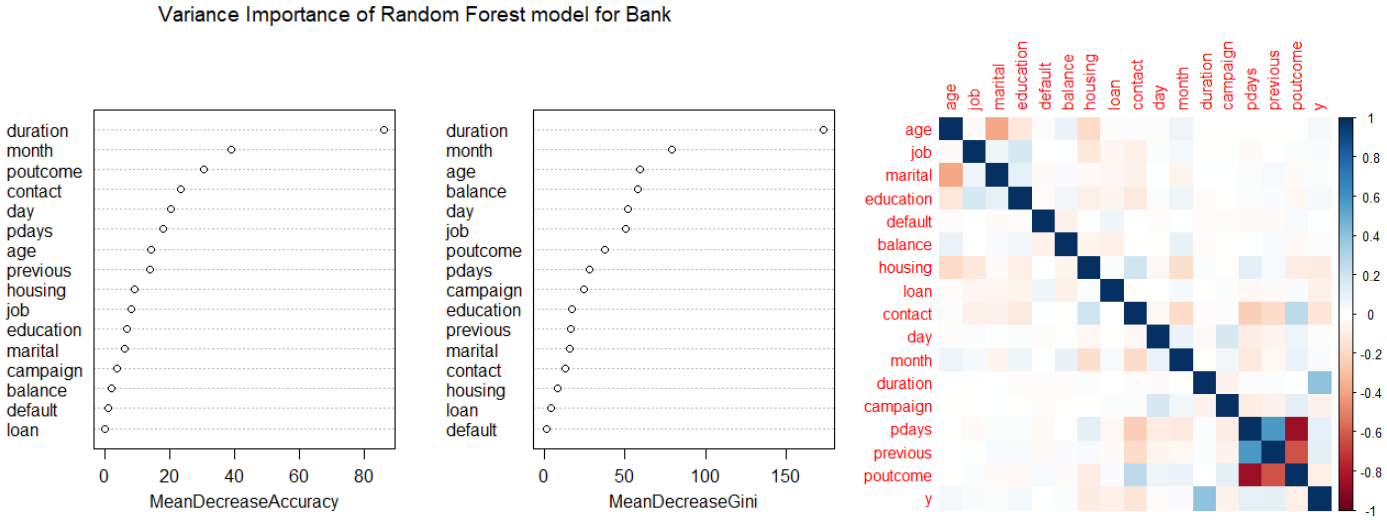

Figure 3.2. From left: shows the variance importance of the Random Forest model to determine bank subscription. From right: shows the correlation coefficient among pairs of variables in the bank data set.

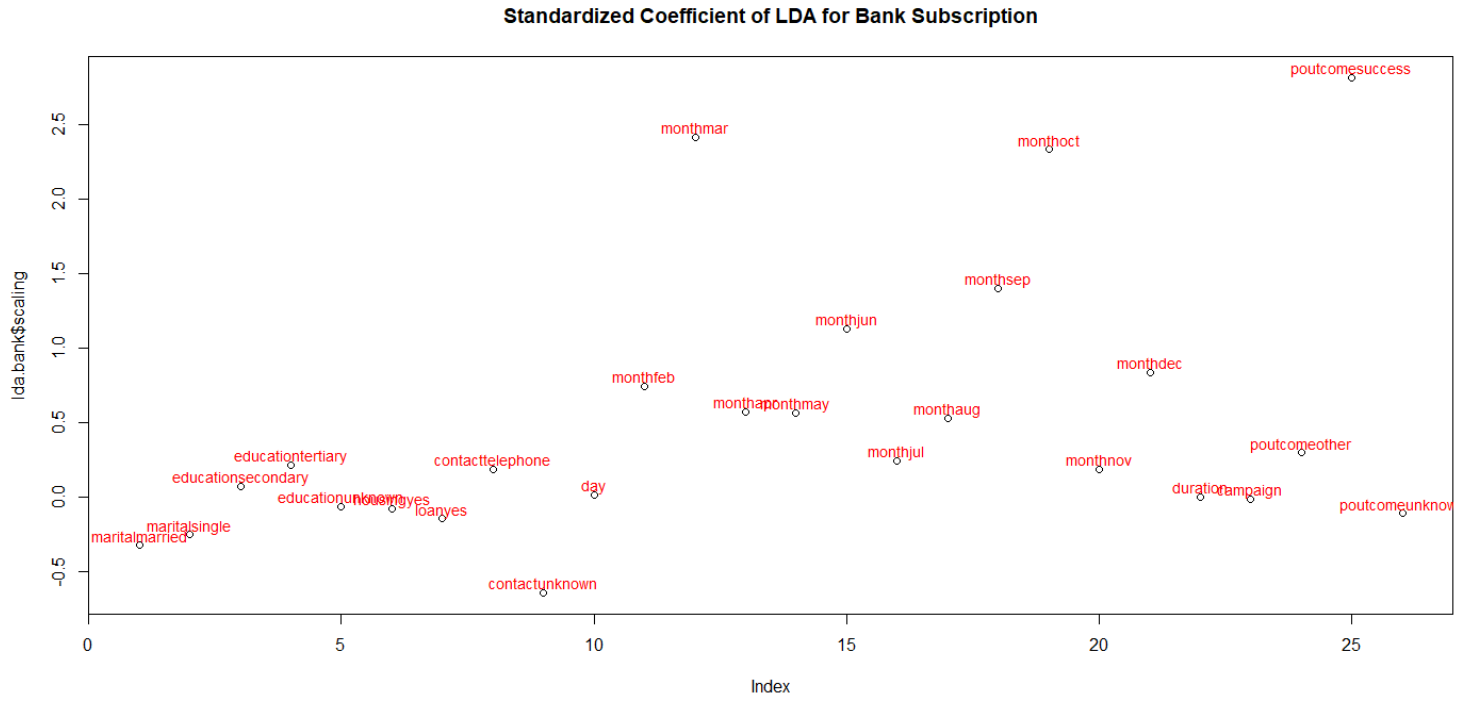

Figure 3.3. shows the standardized coefficient of linear discriminant analysis for bank subscription.

References:

Analytics Vidhya. 2022. *Skewness and Kurtosis |Shape of data: Skewness and Kurtosis*. [online] Available at: <https://www.analyticsvidhya.com/blog/2021/05/shape-of-data-skewness-and-kurtosis/#:~:text=Beginner%20Statistics,outliers%20in%20a%20given%20data.> [Accessed 21 March 2022].

SIM GE Website. 2022. *Full-time Diploma in Information Technology Course | SIM GE*. [online] Available at: <https://www.simge.edu.sg/programme/diploma-in-information-technology/> [Accessed 21 March 2022].

Sites.google.com. 2022. *Urban vs. Suburban vs. Rural Schools*. [online] Available at: <https://sites.google.com/site/urbanvssuburbanvsruralschools/> [Accessed 21 March 2022].

Hse.gov.uk. 2022. [online] Available at: <https://www.hse.gov.uk/statistics/causdis/stress.pdf> [Accessed 20 March 2022].

Kamal, M. and Bener, A., 2022. *Factors contributing to school failure among school children in very fast developing Arabian Society*. [online] PubMed Central (PMC). Available at: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3251187/> [Accessed 29 March 2022].

D'Alessandro, D., 2022. *What Causes School Failure? | Pediatric Case and Reference Article | Pediatric Education*. [online] Pediatric Education. Available at: <https://pediatriceducation.org/2018/07/30/what-causes-school-failure/> [Accessed 29 March 2022].

Qservegroup.com. 2022. *Why Overall Accuracy Isn't Sufficient? *. [online] Available at: <https://www.qservegroup.com/eu/en/i717/why-overall-accuracy-isnt-sufficientennbsp-> [Accessed 25 March 2022].

Martinez-Taboada, F. and Ignacio Redondo, J., 2022. *Variable importance plot (mean decrease accuracy and mean decrease Gini).*. [online] figshare. Available at: <https://plos.figshare.com/articles/figure/Variable_importance_plot_mean_decrease_accuracy_and_mean_decrease_Gini_/12060105/1#:~:text=The%20mean%20decrease%20in%20Gini,the%20variable%20in%20the%20model.> [Accessed 25 March 2022].

Rawat, A., 2022. *What is Support Vector Regression? | Analytics Steps*. [online] Analyticssteps.com. Available at: <https://www.analyticssteps.com/blogs/what-support-vector-regression> [Accessed 20 March 2022].

Rpubs.com. 2022. *RPubs - Discriminant Analysis*. [online] Available at: <https://rpubs.com/aaronsc32/discriminant-analysis> [Accessed 20 March 2022].

Appendix:

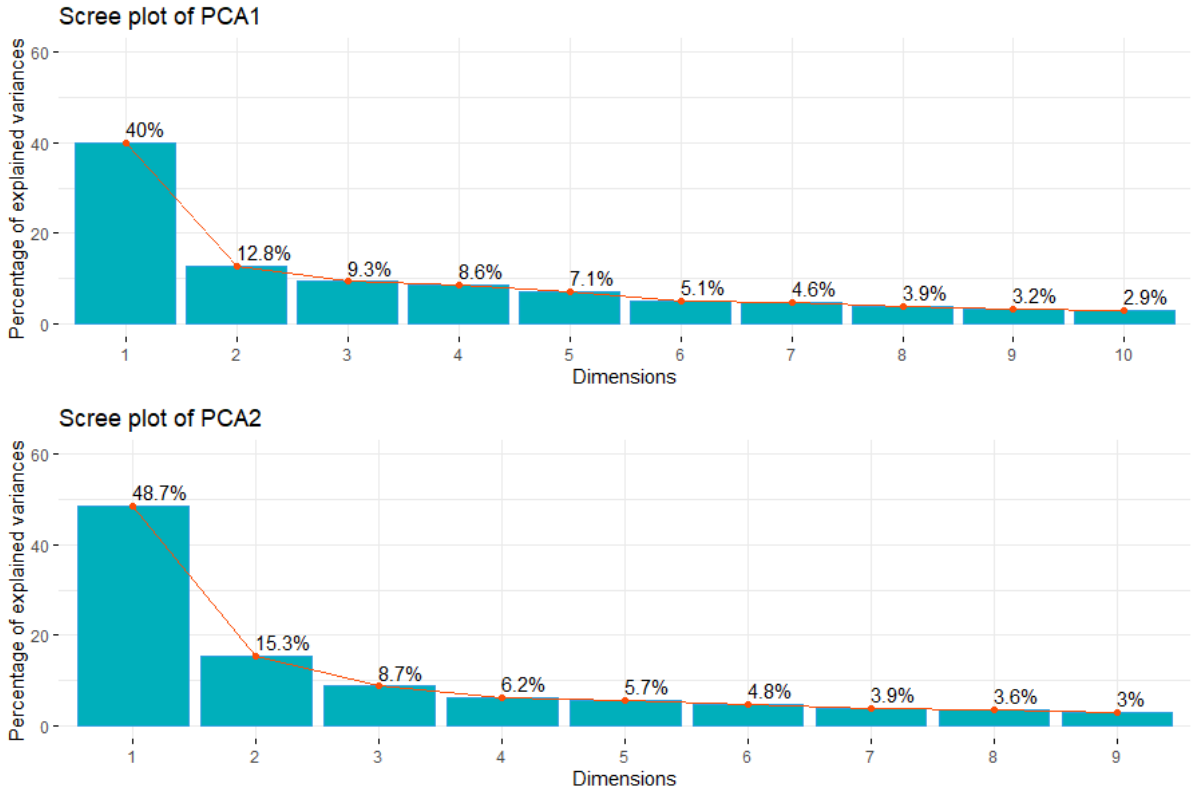

Figure A1.1. shows the comparison of the screen plot of PCA1 and PCA2 before and after removing Gender and Age respectively.

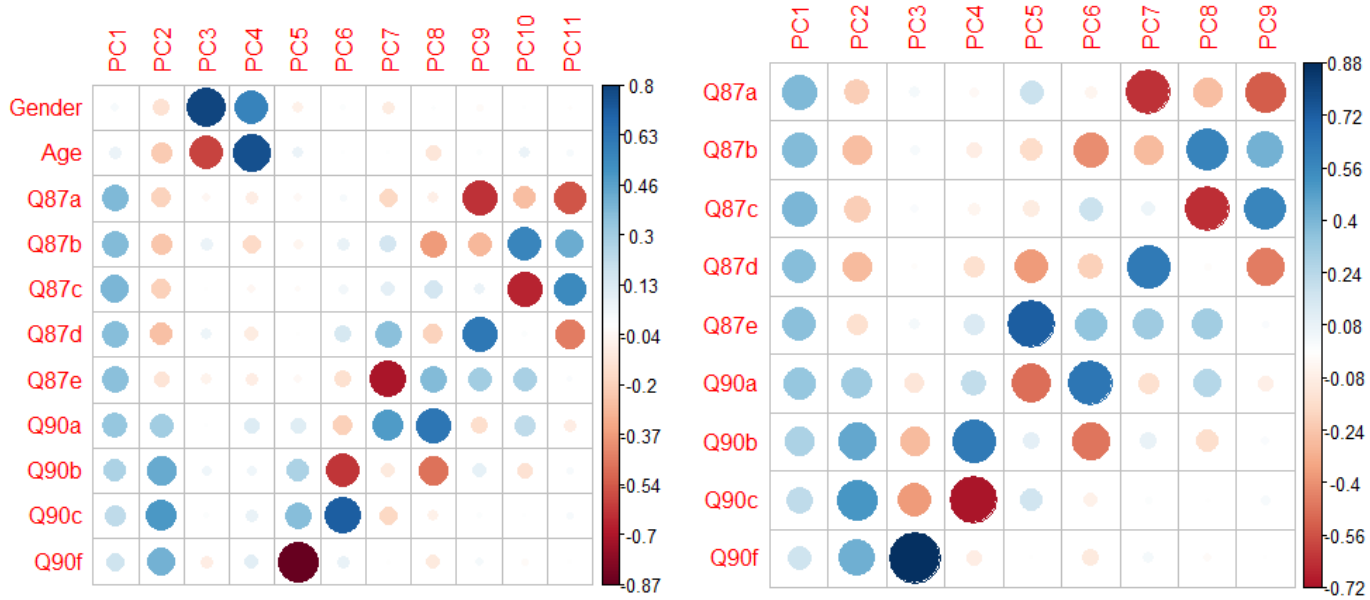

Figure A1.2. shows the contribution and its correlation of each variable in each PCA for PCA1 and PCA2 before and after removing Gender and Age variables respectively.

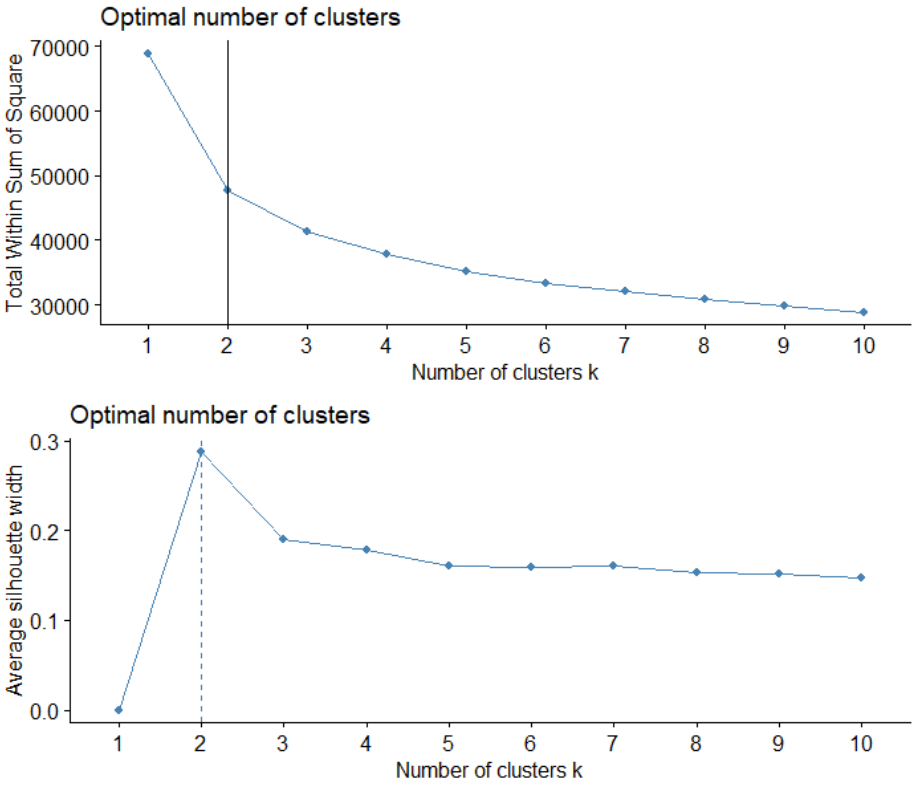

Figure A1.3. shows the scree plot and silhouette plot for optimal k-means clustering where k=2.

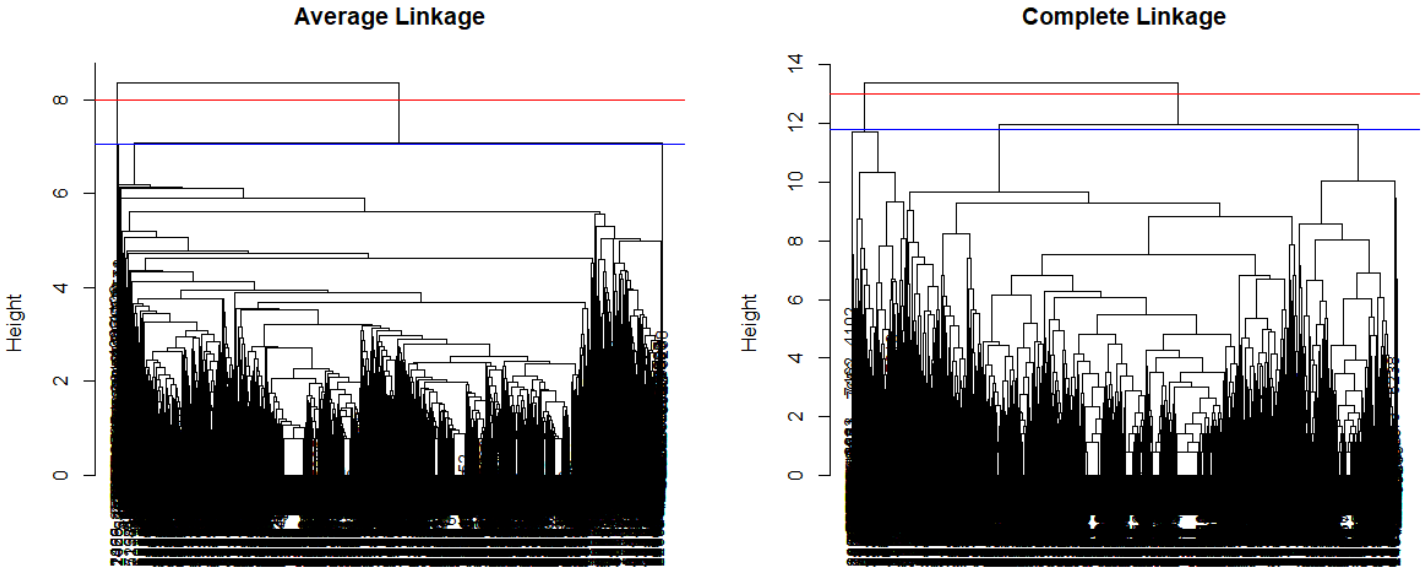

Figure A1.4. shows the comparison of Hierarchy Clustering between average and complete linkage to determine the number of clusters where on the left: cut-off height is 8 (solid red line) provides 2 clusters and cut-off height is 7.05 (solid blue line) provides 3 clusters, and on right: cut-off height is 13 (solid red line) provides 2 clusters and cut-off height is 11.8 (solid blue line) provides 3 clusters.

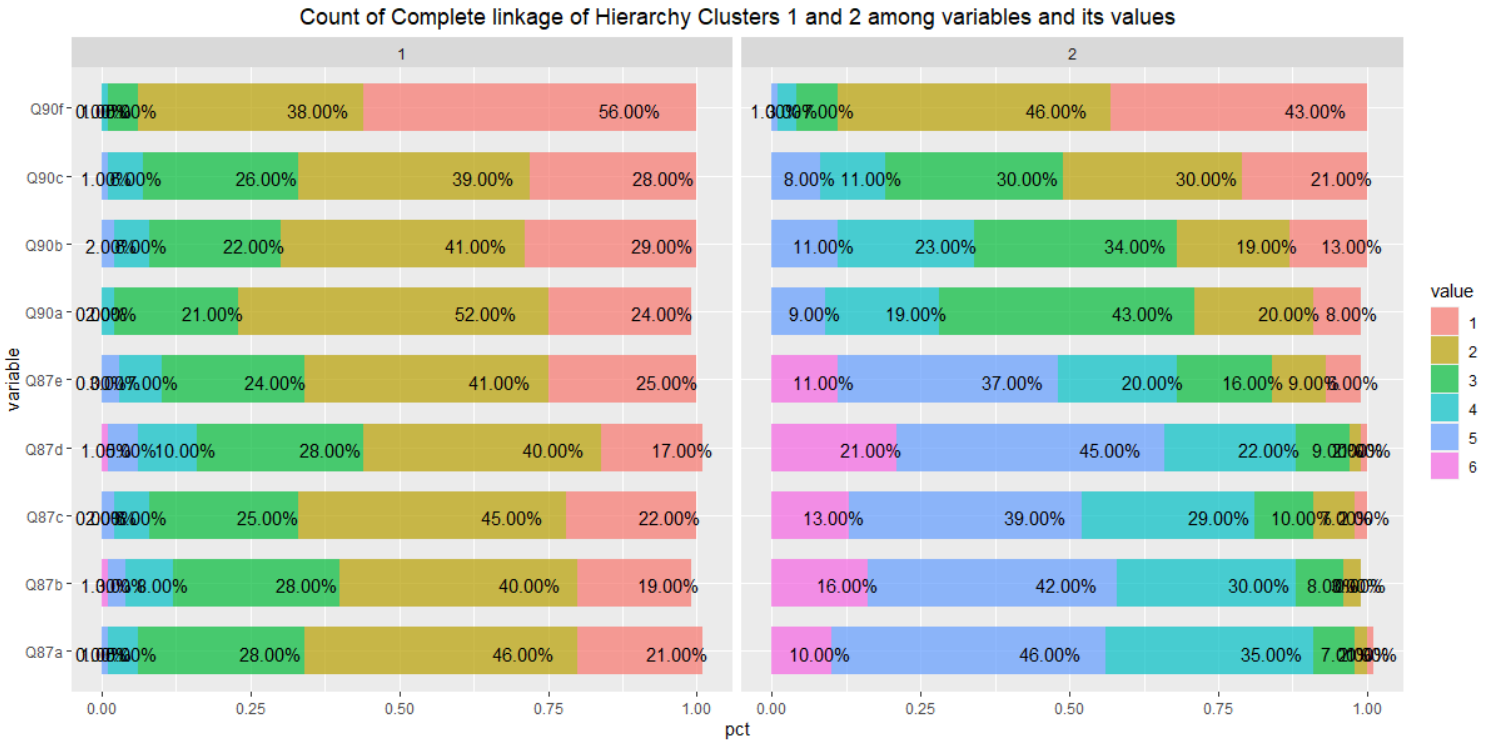

Figure A1.5. shows the count of complete linkage of Hierarchy Clusters 1 and 2 among questions 87 and 90 its respective responses.