**DATA DESCRIPTION AND RESEARCH QUESTION**

The Framingham heart study dataset is rooted in a present study on cardiovascular disease conducted in the community of Framingham, Massachusetts. The dataset is publicly available in the Kaggle web browser and includes detailed patient information, potentially enabling researchers to explore the risk factors involved with heart disease. The Framingham heart disease dataset involves 16 variables and 4,238 rows. The dataset can be found at the following URL: https://www.kaggle.com/datasets/aasheesh200/framingham-heart-study-dataset.

METADATA

male: Categorical variable encoding participant gender as “1” for males and “0” for females.

age: Numerical variable representing participant age at the time of the medical examination.

education: Ordinal categorical variable indicating the participants' education level, with “1” representing Secondary School Graduated, “2” representing Bachelor Graduated, “3” for Post Graduated and “4” represents PHD.

currentSmoker: Categorical variable represented with “1” for those who smoke and “0” for those who do not.

CigsPerDay: Numerical variable representing number of cigarettes smoked per day.

BPmeds: Categorical variable represented with “1” for those who use medications for their blood pressure and “0” for those who do not.

prevalentStroke: Categorical variable represented with “1” for those who have prevalent strokes and “0” for those who do not.

prevalentHyp: Categorical variable represented with “1” for those who have prevalent hypertension and “0” for those who do not.

diabetes: Categorical variable represented with “1” for those who have diabetes and “0” for those who do not.

totChol: Numerical variable representing total cholesterol level in milligrams per decilitre (mg/dL).

sysBP: Numerical variable representing systolic blood pressure in millimetres of mercury (mmHg).

diaBP: Numerical variable representing diastolic blood pressure in mmHg.

BMI: Numerical variable representing body mass index, calculated as weight (kg)/height (m)^2.

heartRate: Numerical variable representing heart rate in beats per minute.

glucose: Numerical variable representing blood glucose level in mg/dL.

TenYearCHD: Categorical target variable representing the risk of the patient to have coronary heart disease in the next ten years with “1” for those at risk and “0” for the others.

Coronary Heart Disease (CHD) remains a significant global health challenge, leading to cardiovascular morbidity and mortality. Early identification of individuals at risk is crucial for effective intervention and prevention strategies. This project seeks to utilize advanced data analysis and predictive modelling techniques to attain deeper insights into factors associated with Ten Year CHD risk and to construct a suitable predictive model. Given that CHD is a complex condition, understanding the degree to which each variable influences the likelihood of impact aids in identifying individuals at elevated risk.

**DATA CLEANING**

This entire section of the project is executed from page 2 to page 23 in Appendix A

First, the csv file was read into RStudio as a data frame and stored in an object labelled fhd.df. Inspecting the variables, it was seen in the output of the code generated in page 3, that not all variables were encoded in the appropriate type as described in the metadata. Categorical variables were coerced into the factor class at the end of page 3 and the levels within them are inspected from page 4 to 5. To clarify the variables representation, the male factor was renamed gender, the “1” level was renamed male and the “0” was renamed female in page 5. Similarly, the levels within the education level from “1” to “4” were renamed as Secondary School, Graduation, Post Graduation and PHD to offer clarity in page 5. The percentage of missing values per variable was then assessed at the beginning of page 7, with the glucose variable possessing the highest percentage at 9.16 and the other variables with missing percentages being under three percent. All rows with missing values were first eliminated and upon inspection, the dataset row size reduced from 4238 to 3656, a decrease of 582 rows, and this dataframe was stored in the object fhd.df\_omit in 7. This dataset retained 86.27 percent of the original dataframe as seen in page 7 and was the first possible consideration for our cleaned dataset. Next, multiple boxplots were created from page 8 to 15 to look for outliers in each numerical variable and while outliers existed, none of them were outside the realm of possible values hence they remained unchanged. With an understanding of medical parameters, it was ensured that sysBP values were always higher than diaBP values in page 15. An alternative attempt at handling missing values was then attempted. First, the numerical variable “glucose” was addressed due to its relatively significant percentage of missingness, prompting an attempt at a slightly more rigorous approach. First a correlation between glucose values and all other numerical variables was investigated with the intent of imputing for mean/median across subsets of values at the risk of variance bias however page 16 showed that there was no correlation. Another approach considered involved comparing the differences amongst variables for observations with missing glucose values and observations with available glucose values. First, factor levels revealed that a significant difference occurs in the gender category, where females were more likely to have their glucose info withheld as seen by the significant difference in gender level ratios across page 19. The smaller differences in proportion were ignored at the risk of their significance being amplified by sample size. Next, the mean and median of all numerical variables between those whose glucose information was missing and those that were available was compared and there was no significant difference as seen from page 20 to 21. Of course, certain numerical values were missing which impacted the insight acquired along with the limitations in using mean and median as a means of comparison. Random imputation was applied using values from the female gender level due to the difference in proportion seen earlier amongst the level between the missing and available glucose subsets. In reality, certain individuals might have felt that their glucose level was too sensitive to disclose while others might have felt that theirs is too high to reveal, however each individual’s definition of "too high" may vary hence it was difficult to determine a range within which these values might lie beyond what has already been done. A major reason for selecting the risk of noise (random imputation) over variance bias (median/mean imputation) here was that when searching for relationships and correlations in the machine learning process, it was expected that a suitable model should be able to overcome the noise introduced and this could help to validate the predictive strength. Also, median and mean imputation also introduce a certain amount of noise. Assessing the impacts of this imputation approach, while it was seen from page 21 to 22 that the mean and median of the glucose levels amongst male and female were nearly identical, there was less variance amongst the glucose level for females in page 22. Considering the remaining numerical variables with missing values, the information within them was likely considered too sensitive to disclose by a number of participants, hence the assumption was that they were likely not missing at random. However, given that they represented a smaller percentage of the overall dataset, we could afford to apply median imputation as the effects of variance bias would not be as significant. Similarly, considering the small proportion of missing values for the categorical variables, mode imputation was initially considered for the BPMeds variable due to the majority of values lying in the "0" level already, as seen at the end of page 22. However, the risk here was that in reality, it was possible that those who withheld information on if they took blood pressure medication probably did as they might view it as sensitive information, which made the "1" category a possibility, hence random imputation was applied. Imputing for “education” was more complicated for numerous reasons. First there was a larger percentage of missing values (2.48) than BPMeds (1.25) as seen in page 23, secondly there were four factor levels which alone reduced the chances of guessing correctly via random imputation by half when compared to BPMeds and realistically, unlike BPMeds where the answer was strictly "1" or "0" there was the possibility that those who had withheld their info did not have any level of education or were not willing to disclose theirs, meaning we might have needed another level "Below Secondary School" or "Undisclosed". Finally, both of those assumptions could be incorrect. Here, the “undisclosed” level was created to represent the missing values, but this meant that the education level lost its ordinal property. Therefore, an identical dataset was created off fhd.df and stored in the object, fhd.df\_ord\_edu, and for this dataset, the missing values were replaced with a “Below Secondary School” level. Hence, there were three data frames going into the following section. The data frame where all rows with missing values were omitted, stored in fhd.df\_omit, the dataset where missing values were replaced and the education variable was an ordinal category, stored in fhd.df\_ord\_edu, and the data frame where missing values were replaced and the education variable was a nominal category, stored in fhd.df.

**EXPLORATORY DATA ANALYSIS**

This entire section of the project is executed from page 24 to 91 in Appendix A

The exploratory data analysis were applied to the data frames whose missing values were replaced first (both fhd.df and fhd.df\_ord\_edu were identical asides the education level). A correlation plot was created on page 27 and on page 26, only systolic and diastolic blood pressure displayed correlation which could potentially contribute to a principal component further in the analysis. A set of boxplots were produced on page 36, comparing each numerical variable to the target category and the “age” variable displayed the most significant relationship with the target variable, while sysBP, diaBP and BMI showed marginal relationships. Next, the categorical variables were all compared to the target category using hypothesis testing, but first the frequency count within each table for each independent category against target category was assessed from page 36 to 37 to ensure that all frequencies were greater than or equal to five. Once this was confirmed, chi-squared test was selected as the means of testing the hypothesis in both the fhd.df and fhd.df\_ord\_edu data frame with the exception of the education variable in the latter data frame which was ordinal. For this, Wilcoxon rank sum test was used due to the ordinal nature of the category. The results were displayed from page 37 to 38 and the only variable which displayed no dependency with target variable was “currentSmoker” with a p-value above 0.05. Next the relationship between each category and the target category was viewed with mosaic plots from page 38 to 44 and individuals without prevalentStroke were notably less likely to be at risk of TenYearCHD than those who had it. Next, principal component analysis was performed to reduce the number of numerical variables, hence the dimension of the data frame, as well as encode each numerical variables at varied weights into new principal component variables for new and possibly more useful insights. The first five principal components were selected based on the cumulative sum plot on page 45, due to the fifth principal component being the first to be positioned just above the horizontal line at 0.8. Next the loading of each original numerical variable within each of these five selected principal components were viewed in page 46 in order to understand the degree to which each variable contributed to the values of each principal component as well as the nature of the contribution (positive or negative). PC1 mostly encoded sysBP and diaBP, which displayed correlation earlier. PC2 encoded cigsPerDay and heart rate. PC3 encoded glucose levels. PC4 encoded totChol negatively and glucose levels positively and PC5 encoded cigsPerDay negatively and heart rate positively, unlike PC2. Notably, the relevance in age and BMI information were sacrificed in the process, and this was significant because as seen prior, the relationship between age and the target variable was the most notable amongst the numerical variables. Similar to the numerical variables, all PC variables were plotted in a correlation matrix at the top of page 47. PC3 and PC4 showed a bit of a trend as they both significantly encode glucose levels. They both displayed a similar relationship with PC1 and PC2, as well as PC5, where there was a bit of a negative relationship due to PC5 negatively loading glucose. A biplot was not created due to the number of observations being too large. On page 49, PC1 to PC5 were compared to the target category via a boxplot. Patients with risk of TenYearCHD had slightly higher PC1 values and slightly lower PC5 values. This relationship might be rooted in most loadings for both PC values skewing positive for PC1 or negative for PC5, as seen in the loadings earlier. There was also a marginal relationship with PC4. Perhaps using the original variables would create a more accurate model where there was a clearer separation between both levels of the target variables based on the numerical variables, especially "age", whose relevance has been diminished amongst the selected principal components. Every exploratory data analysis process executed so far was then repeated for the data frame with missing values and the results were nearly identical. The correlation matrix could be viewed on page 50, boxplots on page 61 showed slightly clearer separation of target levels. Frequency table and hypothesis tests for categorical variables were executed from page 61 to 64. Mosaic plots were displayed from page 65 to 69. Cumulative sum plot on page 70 and the PCA loadings on page 71. PCA correlation matrix on page 72 and PCA boxplots on page 74. The final stage of the exploratory data analysis involved cluster analysis application. The data frame with missing values replaced was addressed first, where k-means clustering was applied to split each observation into two clusters based on PC1 to PC5. Hierarchical clustering was avoided due to the size of the dataset. The purpose of the cluster analysis was to inspect the difference between data points pertaining to the target variable, hence two clusters were created in an attempt to reflect the separation between the two levels of the target variable. As seen in the cluster means table toward the end of page 74, the greatest difference between cluster means comes from PC1 and PC4. This is relevant as they are likely to be the best suited PC values to view the separation of target levels. The principal component boxplot can also be referred to as they were the most significant separators of the target levels in the view as well. Visualizing the cluster separation using PC1 and PC4 and comparing it with the separation in target variable level. There was a clearer separation on the x-axis for the cluster groups and a less obvious separation for the target variable, with significant overlap on the axis plane as seen in page 77 and 78. This was consistent with the relationship seen in the boxplot between the target variable and PC1. A chi-squared test was then carried out between the cluster groups and the target variable to reveal a p-value less than 0.5, which suggested that they could be dependent. This was seen on page 78. Proportion table at the beginning of page 79 suggested that most patients at risk of TenYearCHD fell into cluster 1 while most without fell in cluster 2. Regardless, both clusters were dominated by the “0” target level as seen in page 79 due to the skew in proportion of target levels. It was important to note that cluster 1 had a lower proportion of these “0” level patients than cluster 2 while possessing a higher proportion of patients who were at risk than cluster 2. While there was a relationship between the clusters and target variable there still seemed to be a significant overlap between the two levels of the target variables based on the PC clusters. Assessing the accuracy of the clusters in predicting the correct level (cluster 1 to target variable level 1 and cluster 2 to target variable level 0), the accuracy is at about 63.3 percent as seen at the end of page 79. While the PC values might have been marginally useful in predicting the accurate level, a more robust supervised learning method which utilized the categorical variables might have aided in delivering more accurate results. This cluster analysis process was then replicated for the data frame with missing rows omitted. The results were again, nearly identical with the difference in cluster groups and target levels compared through scatterplots in page 89 and 90, hypothesis testing of both categories along with the contingency tables viewed in page 90, proportion of target variable levels within each cluster viewed in page 91, and the accuracy of the cluster just below, in page 91. The cluster accuracy for this dataset was marginally better at 63.6 percent.

**MACHINE LEARNING PREDICTION**

This entire section of the project is executed from page 91 to page 113 in Appendix A

The supervised machine learning method used for this dataset was decision trees. The dataset with missing values omitted was selected as it retained a sufficient amount of information relative to the dataset with missing values cleaned, while also possessing less noise. The reduction in noise also meant that it was more likely to deliver more accurate results despite potentially overfitting the model. Additionally, the decision tree was applied twice. The first time using numerical variables "age", "sysBP", "diaBP" and "BMI" from the original dataset, while the second application utilized "PC1", "PC4" and "PC5" from the principal component analysis. The selection of these variables amongst the numerical was based on the relationships they displayed in respect with the target variable during the exploratory data analysis phase. Based on the chi-squared test in the exploratory data analysis, all categorical variables asides "currentSmoker" will be used for the decision tree as well. Beginning with the original numerical data frame, data preparation for the modelling was done by creating a 70/30 training/test set split beginning from page 92. The split was applied specifically to each level of the target variable due to the significantly greater proportion of the "0" level. After splitting each subset of the data frame, both training and test sets for the “0” and “1” levels were joined. This addressed the possibility of either training or test set failing to capture any of the “1” level in the target variable. The decision tree produced in page 93 did not require any pruning as its structure was simple with only 3 terminal nodes, all of which led to target factor level "0". Notably, no categorical variables were involved in the tree formation. The most relevant division in the data, based on the target variable is the "age" variable, separated by those who are 48 years or younger and those who are above 48, amongst the latter group there is another relevant division based on those with a Systolic blood pressure value that's 160 or below and those with a Systolic blood pressure above 160. Despite these relevant divisions, all outcomes led to the "0" level, likely due to both the majority of the observations in the dataset falling within the "0" level and significant overlaps between observations of both levels within each node of the tree. This process was repeated with consistent sampling for data preparation using the principal component variables specified earlier. The decision tree produced at the top of page 95 did not require any pruning with only 4 terminal nodes. Again, no categorical variables were involved. The most relevant variable involved in attempting to separate all observations based on target variable level is PC1 (Blood Pressure component), which mostly encodes Systolic and Diastolic BP. The split occurred between those with values below 0.934921 and those above 0.934921. Amongst those below, there was another relevant level of separation based on PC4, which positively encoded glucose levels and negatively encoded total cholesterol. The separation occurred between values above and below -0.566995. Finally, for PC1 values above 0.934921, there was a relevant separation based on PC5, which positively encoded heart rate and negatively encoded the number of cigarettes per day. The separation occurred for values above and below -0.784373. Due to the encoding of the original numerical variables during Principal Component Analysis (PCA), this decision tree gave a slightly more detailed breakdown on how the data was separated based on the levels of the target variable across more independent variables but was also more difficult to interpret due to the PC values needing clarification including interpretation of how the lessened effects of other variables within each Principal Component affected the split. This model also lost the relevance of the "age" variable. From page 96 to 97, the contingency table and accuracy of both models can be seen and by assuming that all observations fall into the “0” level, a high accuracy of 84.7 was achieved by essentially reflecting the percentage of dominant level in the target variable. This, however, did not give any significant insight into the relevance of the other variables either because they were not relevant for the categorical target variable regardless of supervised machine learning method utilized or because the decision tree model in RStudio was not robust enough to deal with the structure and spread of the data despite its general suitability for categorical target variables. In order to amend this issue, the data was balanced first via down sampling. Given that the ratio of target variable levels was nearly 1:5, the dominant "0" target level was randomized and split into 5 equal parts. Each part was separately joined to the “1” level variable for a partial balance. This approach all observations to be taken into consideration across separate models. The results of each mode were then compared. Viewing each decision tree from page 98 to 102, "age" and "sysBP" were consistently influential regardless of the sampled section. Similar to the unbalanced implementation, the same process was repeated for the PC variables and as seen in the decision trees from page 104 to 108, PC1 and PC5 dominated all the decision tree samples. Contingency tables from page 109 and 110 for non-PC and PC variables respectively, revealed that the “1” level was now predicted for as expected. Despite the added insight, these models were not suitably accurate, delivering less accuracy than the initial models. Additionally, they were likely to display even less accuracy if applied to a different portion of the original dataset as a result of severe overfitting with each model being trained on less data. The results of the model prediction for the non-PC and PC variables were displayed from page 110 to 111, and the non-PC predictions were slightly more accurate. Finally, the model was applied with an up sampled dataset for balancing in accordance with Thabtah *et al.* (2020). It was agreed across the members of the group that the up sampled dataset would be used for comparison due to the original model delivering no insight as a result of target variable level skew and the down sampled model either eliminating important observations or creating too many models for a cross comparison of numerous overfitted models which would have likely led to the report word limit being breached. For consistency across the training and test data, the up sampling and training test split executed by Abdul Nawabi in the chunk beginning from the end of page 91 for random forests needed to be established at the beginning of the supervised machine learning section for all members to use for their up sample. Also, I personally only applied the up-sampling balance to the non-PC values as the insignificance in differing levels of accuracy was not enough to justify the difficulty of interpreting the tree design. After up sampling, the tree produced at the beginning of page 112 was simple and similar to the first model. Again, the “age” variable was the most significant separator and those below 47 fell into the "0" level, while the remaining patients were likely to fall within the “1” level despite there being a notable difference between those with “sysBP” below and above 155.25. Ultimately the relevance of age here makes sense as in real life older people tend to be at risk of coronary heart disease and so the older participants are more likely to be at risk within ten years. Assessing the accuracy at the end of page 112, this model with 64.6 percent accuracy was more accurate than most of the down-sample models asides fhd\_omit\_target\_b. While both sampling methods offered more insight, they were both significantly overfitted, in the case of up-sampling the overfitting came from artificially replicated instances of the less dominant target level, hence insights might have been misleading. Without access to more info, perhaps it could be assumed that the target level variable occurs rarely in reality and the variables in the dataset along with the tree model delivered by R could not offer significant insight into what the causes might have been, and any model generated would have possessed high predictive accuracy by mirroring the target level spread. Explaining the inaccuracy of the up sampled model, the contingency table towards the end of page 112 revealed that a large number of “0” rows were predicted as ”1”. This is likely due to the amplification of the space inhabited by the repetition of “1” level instances resulting from the up-sampling working in conjunction with the fact that these two levels significantly overlap on the variable plane seen in page 113.

**HIGH PERFORMANCE COMPUTATIONAL IMPLEMENTATION**

This section was executed in appendix B and the results will be compared to Appendix A

Replicating the correlation matrix generated at the beginning of page 52 in Appendix A, it could be seen at the top of page 4 in Appendix B that the default aesthetics gave the correlation matrix a blue colour and that the variable names across the matrix had been replaced with their respective histograms, representing the counts across ranges of values within each numerical variable. Boxplots generated from page 7 to 14 of Appendix B were similar to those atop page 61 of Appendix A. The mosaic plots from page 15 to 20 of Appendix B had a default maroon and green aesthetic in comparison to the grey aesthetic of those generated from page 64 to 69 of Appendix A. The cumulative sum plots on page 70 in Appendix A and the same on page 23 in Appendix B are also similar. Finally, replicating the decision tree model applied at the beginning of the machine learning section which did not require balancing, the model produced from page 25 of Appendix B not only delivered an accuracy level of 83.3 percent, but also delivered insight as to what type of separation led to the “1” level target variable. There were seven instances where this occured based on the tree. First instance involved diaBP above 98.75 amongst BMI values less than 21.94 amongst sysBP greater than 128.75 amongst diaBP less than or equal to 107.75 amongst age less than 48.5. This tree was interpretable and went into more detail (while creating intricate subsets) in an attempt to precisely split the levels due to the superior computing power and ability to handle greater information afforded by Apache spark. Here, insights could be gathered without compromising accuracy.

**PERFORMANCE EVALUATION AND COMPARISON OF METHODS**

This section involves comparisons from page 111 to the end of Appendix A

Comparing the up-sampling accuracy of 64.6 percent as seen at the end of page 112 to the random forest up sampling of 95.7 percent seen at the end of page 120, it was seen that while both insights might have been made superficially due to overfitting, the random forest method was much more precise in handling errors. The random forest model suffered relatively less from overfitting and was able to apply different iterations of decision trees at each node of random subsamples using a random set of independent variables which was why all variables were used in generating the formula. A luxury not afforded by the default decision tree. The decision tree model might also be sub optimal due to the process by which it selected the best split at each node which might have ended the tree prematurely and cut off insights which might have been found further down a different split which was not as significant at the time of splitting. For the random forest, 500 trees were made, and the best outcome was determined via minimum error. The plot of variable importance generated at the beginning of page 120 by the random forest displayed that age remained an important variable in creating the model, which had been established with the decision tree analysis as well. SysBP was also relatively relevant within both models. Of course, the relevance of these variables were also impacted by the training data selection and could be altered slightly with a different set. The random forest method developed a better model due to its robust method of application.

Comparing to the up-sampled support vector machine (SVM) machine learning model which used a linear kernel and radial kernel, the decision tree model was less accurate in prediction, but not to a significant degree, as in page 127, it was seen that the linear and radial kernel accuracy was 69 percent and 71.3 percent respectively, about 5 to 7 percent higher than the decision tree model accuracy, which could at least suggests somewhat similar levels of suitability given the dataset. The accuracy of the linear kernel being slightly less than the radial is likely rooted in the nature of the target level spread across the variable plane which could be seen in the graph plotted earlier in page 113, where age and sysBP were used due to the fact that they were the best numerical variables for splitting the target levels as was seen in the exploratory data analysis. In the graph there was a significant overlap amongst the target levels which meant that the maximal margin classifier could not be oriented to deliver a good separation regardless of the cost parameter. Unlike the decision tree, this model took all numerical variables into account simultaneously for one flexible separation, and because the categorical variables did not create much separation themselves within the decision tree, the SVM model was not disadvantaged, leading to a somewhat similar level of accuracy.

Comparing the decision tree accuracy to the 76.8% accuracy seen towards the end of page 131 of the k-nearest neighbour (k-NN) model implemented, it was important to understand the role that up-sampling the "1" level in the data played here. A large proportion of the accuracy stemmed from the accurate prediction of the "1" level as seen in the contingency table produced in page 131. This was a result of the replication of instances by the up-sampling process, which led to the k-NN method likely overfitting itself as all replicated datapoints would share the same space on the variable plane, therefore inhabiting a large amount of the space and influencing the model’s decision making. Additionally, looking at the scatterplot generated in page 113, it was clear to see that the "1" variables already occupied a slightly more congested area within the “0” target area of the plane.

Comparing to the 66.9 percent accuracy generated by the logistic regression model as seen on page 135, the difference in accuracy of less than 3 percent was marginal and could have changed upon the use of different test and training samples, suggesting similar levels of model suitability for the dataset. The regression model predicted where each observation belonged based on each variable, and the effects of each variable had a different level of pull on the prediction which was defined by their individual relationship with the target variable. Ultimately this model was not too dissimilar to the decision tree in this instance as the decision tree model for the up-sampling took only sysBP and age into account, which based on the summary atop page 134, had 3-star significance to the logistic regression model and were amongst the lowest p-values for the variables involved.

**DISCUSSION OF THE FINDINGS**

Up-sampling significantly overfitted all models, therefore the relevance of accuracy amongst all models should be understood in the context of an up-sampling which multiplied the instances of patients with the risk of coronary heart disease in ten years by five or more. The random forest prediction was the most accurate by a significant margin at 95.7 percent due to its robustness and iteration of multiple trees at multiple nodes. Next was the k-NN model which probably benefited the most from the up-sampling balancing approach utilized across the group, with an accuracy of 76.8 percent. The remaining three methods were all less than 7 percent apart and some could have been more or less accurate than the others depending on the training and test data used. The SVM linear kernel at 69 percent was only slightly less effective than the 71.3 percent of the radial kernel due to the target level overlap within the hyperplane, and the logistic regression as well as decision tree were not too dissimilar in terms of the variables that significantly factored into their models with accuracies of 66.9 and 64.6 respectively. Finally, the efficiency of each model was not only subject to the class of each variable and relationship each independent variable had with the target variable, but also the spread of the data in each of them. Ultimately the relevance of age across the various machine learning methods makes sense as in real life older people tend to be at risk of coronary heart disease and so the older participants were more likely to be at risk within ten years. Principal component analysis’ contribution to the decision tree model applied was not significant due to the marginalization of the relevant age variable. The positives gained from combining both sysBP and diaBP were not able to offset the difficulty in interpretation introduced by the PC variables. The overlap in the dataset on the variable hyperplane rendered certain methods (support vector with linear kernel) sub optimal but a more sophisticated splitting via HPCI decision tree addressed the overlap across numerical variables of the target levels.

**REFERENCES**

Thabtah, F., Hammoud, S., Kamalov, F. and Gonsalves, A. (2020). Data imbalance in classification: Experimental evaluation. *Information Sciences*, 513, pp.429-441.

**APPENDIX**

The appendix material was submitted as a zipped/archived folder containing the following files:

1. RStudio\_CS5811\_2351044 - Appendix A

2. HPCI\_CS5811\_2351044 – Appendix B

**DATA MANAGEMENT PLAN AND AUTHORSHIP CONTRIBUTION STATEMENT**

DATA MANAGEMENT PLAN

1. Overview

Researchers: Chidubem Henry Nworah, Abdul Nawabi, Lawrence Obute Ameh and Renuka Naineni.

Project Title: Predicting the Likelihood of Coronary Heart Disease in Ten Years

Project Duration: 3 Months

Project Context: The discipline was Data Science and Analytics, the subject was Distributed Data Analysis.

1. Defining Data Resources
   1. The data came from Kaggle, the URL was <https://www.kaggle.com/datasets/aasheesh200/framingham-heart-study-dataset>
   2. The amount of data generated was 37.2 megabytes
   3. File formats were RMarkdown and Jupyter Source File converted to pdf
2. Organizing The Data
   1. RMarkdown file knitted to pdf, and Jupyter source file was converted to pdf via code and saved as RStudio\_CS5811\_2351044.pdf and HPCI\_CS5811\_2351044 respectively.
   2. The pages of each visualization, code or code output are stated in report and the Appendix included is specified at the beginning of each section if not in the body.
   3. RMarkdown and pdf for the R code and Jupyter source file and pdf for the HPCI code
3. Looking After Data
   1. Data was stored in my laptop.
   2. Several copies of the data saved on google drive and hard drive every week.
4. Sharing Data
   1. The data is publicly available.
   2. Data is available for use to everyone
5. Archiving Data
   1. R and Python files along with dataset and report were archived
6. Executing the plan
   1. All group members are responsible for executing the plan.
   2. The plan will be reviewed weekly.
   3. Data description and research question, preparation and cleaning, exploratory data analysis, machine learning prediction and high performance computational implementation

AUTHORSHIP CONTRIBUTION STATEMENT

L.O.A. acquired the data and defined the research question. C.H.N., A.N., R.N. and L.O.A. selected and justified the data cleaning methods. C.H.N. and A.N. cleaned the data. C.H.N., L.O.A. and R.N. decided on the exploratory data analysis plan. C.H.N., R.N. and A.N. decided on the principal component analysis plan. C.H.N. decided on the cluster analysis plan. C.H.N. and L.O.A implemented the exploratory data analysis plan. C.H.N. implemented the principal component analysis and cluster analysis plan. R.N. and C.H.N. interpreted the principal component analysis. C.H.N. interpreted the cluster analysis R.N and A.N designed and perfected all visualizations. A.N. suggested the balancing method. C.H.N. implemented and interpreted the decision tree model. A.N. implemented and interpreted the random forest model. L.O.A. implemented and interpreted the support vector model and R.N. implemented and interpreted the k-nearest neighbour and logistic regression model.