An Analysis of The Framingham Heart Study Dataset

Jon-Paul Boyd
School of Computer Science and Informatics
De Montfort University
United Kingdom
1st May 2018

Abstract This report presents a thorough analysis of the Framingham heart dataset within SAS Enterprise Miner. Based on input from data exploration techniques that included distribution, significance, association and clustering analysis, the main risk factors and cohorts are identified. Dataset split by gender and classified with neural network models deliver best performance in predicting mortality by cardiovascular disease. Recommendations include lifetime observation of participants, improving quality of data collection, targeting healthcare guidance and treatment, and a two-stage approach to operationalising the classifiers.

Keywords Cardiovascular Disease, Framingham, Blood Pressure, Supervised Machine Learning, Classifiers, Data Mining, SAS Enterprise Miner

1. Introduction

All diseases affecting the heart or blood vessels, including coronary heart disease (clogged arteries) fall under the category Cardiovascular Disease (C.V.D.). According to [1] "*Coronary heart disease continues to be a leading cause of morbidity and mortality among adults in Europe and North America*". Data from the Framingham Heart study, started in 1948 and observing 5209 participants aged 28 to 62 at entry into the trial [2], will be analysed. The goal of the study is to use data mining methods to identify the relationship between C.V.D. and health indicators including blood pressure rates and cholesterol level, then evaluate the findings to formulate recommendations for continued proactive health guidance and monitoring. "*Elevated blood pressure is the most important risk factor for death and disability worldwide, affecting more than one billion individuals and causing an estimated 9. 4 million deaths every year*" [3], so any decrease in C.V.D. risk factors not only improves individual wellbeing and increases life expectancy, there is potential for significant reduction in associated healthcare costs.

A range of risk prediction classifiers will be developed to determine the likelihood of mortality as a function of provided health indicators. By accurately assessing the risk profiles of the study participants, those individuals and groups most at risk can be identified, helping formulate action with health guidelines for publication through appropriate channels and recommendation of treatment strategies. An assumption is made that those dataset observations with binary target variable "*Status*" level "*Dead*" died due to C.V.D., and those with level "*Alive*" have no current C.V.D. symptoms nor prior C.V.D. events.

This paper will adopt an evaluation methodology similar to the author's previous research assessing a range of binary classifiers [4], this time using SAS Enterprise Miner V14.1 (S.E.M.) in place of Python, with the paper organized as follows. *Section 2* details initial dataset analysis, focusing on variable distribution. *Section 3* goes further, analysing variable significance and association strength with target "*Status*", and grouping common characteristics through clustering. *Section 4* covers data preparation strategy. *Section 5* provides an overview of the binary classifier models assessed, with *Section 6* outlining the design of the evaluation framework. *Section 7* covers the modelling flow in S.E.M., assembly of the classifiers and efforts to tune for maximum performance. *Section 8* provides a final classifier evaluation, while *Section 9* briefly discusses classification thresholds. Delivery of final conclusions and recommendations are delivered in *Section 10*, which can be considered an executive summary.

2. Initial Data Analysis

The available dataset has 5209 observations and 13 variables, with basic structure information as provided to the author in *Appendix A*. The SAS Enterprise Miner nodes *StatExplore*, *Graph Explore* and *MultiPlot* were used in initial statistical, distribution and correlation analysis which follows.

All observations have "***Status***", a categorical and nominally scaled variable, with two distinct non-missing levels, and therefore binary. It identifies whether the anonymous study participant is alive or dead, with minority class "*Dead*" representing 38.22% (1991 participants), and "*Alive*" 61.78% (3218). "*Status*" is set as the target response variable to facilitate correlation between death due to C.V.D. and its associated risk factors. In other words, "*Status*" is the response or dependant variable as it's of interest to identify which independent variables the two levels of "*Status*" depend on. "*Status*" is sorted in descending order, placing emphasis on level "*Dead*" and factors contributing to death by C.V.D.

Interval variable "***Death_age***" gives age in years at time of death, present for all observations with variable "*Status*" level "*Dead*". Analysis shows a min. of 36, max. of 93 and mean of 70.536. With standard deviation (SD) of 10.559, negatively skewed (-0.316) and with negative kurtosis (-0.365), the distribution is flatter. Further considering the min. of 36 and mean of 70 this highlights the values are dispersed and C.V.D. not a disease suffered only by the elderly. Of all variables it is the one with the highest proportion of missing data, easily explained as systematically null in observations with status "*Alive*" (61.78%). Due to number of missing values this variable will be rejected from further analysis.

Binary variable "***Sex***" represents gender with all observations identified with level "*Female*" (55.15%, freq. 2873) or "*Male*" (44.85%. freq. 2336), therefore slightly biased with a larger female population.

Interval variable "***Diastolic***" represents artery blood pressure at heart rest between beats. UoM is assumed mmHg (millimetres of mercury). Min. is 50, max. 160 and mean 85.358. SD is 12.973 with a large positive spread out from mean giving a positive skew (0.875) and heavy-tail (Fig. 1). 52 outliers record diastolic rates above +3SD, the second most skewed variable after "*Systolic*", with both "*Status*" levels having longer upper whiskers than lower whiskers (Fig. 2). Kurtosis of 1.854 signals distribution a little peaked around the mean. There are no missing values. "*Diastolic*" is generally higher for the "*Dead*" level than "*Alive*" as evidenced by the higher inter-quartile range for "*Dead*", and the median further from the mean for "*Alive*" with a generally lower diastolic rate (Fig. 2.). Mean for "*Dead*" is 89.52 and lower for "*Alive*" with 82.77. "*Dead*" has a higher max. whisker (max. excluding outliers) of 125, in comparison with "*Alive*" with 112. Both groups have outliers although "*Dead*" has more extreme values.

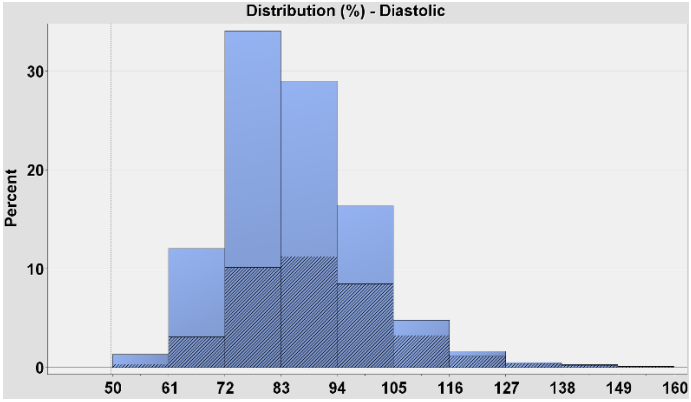

Fig. 1. Diastolic (%) histogram

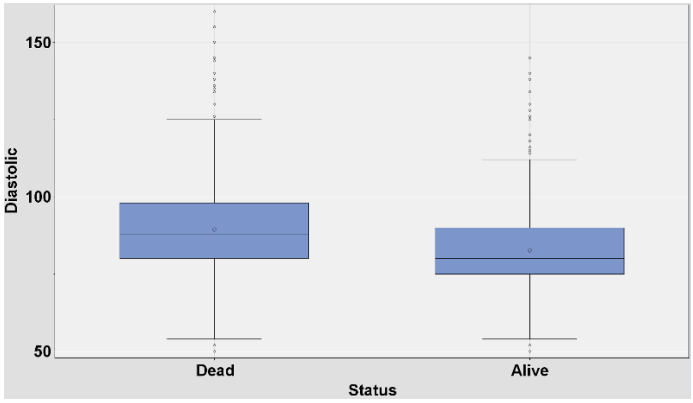

Fig. 2. Diastolic value boxplot

Interval variable "***Systolic***" represents artery blood pressure when the heart is beating, UoM assumed mmHg. Over all observations a min. of 82, max. of 300 and mean of 136.909 are recorded. With the highest kurtosis of all variables (4.228), the most skewed (1.487), and a large SD of 23.739, the data signals a tail heavy, positively skewed distribution with a large variance (Fig. 3). A significant 83 outliers

(1.59%) are above +3SD (normal distribution 0.3%). There are no missing values. Systolic rates are generally higher for the "*Dead*" level with higher inter-quartile, with max. whiskers longer than min., and greater distance between mean and median, another indication of skewness (Fig. 4). Mean for "*Dead*" is 146.16 compared with 131.18 for "*Alive*". Inter-quartile for "*Dead*" is comparatively taller than for "*Alive*", suggesting a greater spread of systolic rates for the target level "*Dead*". Both show outliers although "*Dead*" has more extreme and dispersed values. Only "*Alive*" has outliers below min. whisker. According to [5], "*Elevated systolic blood pressure (SBP) is a well-established risk factor for the development of cardiovascular disease (CVD)*".

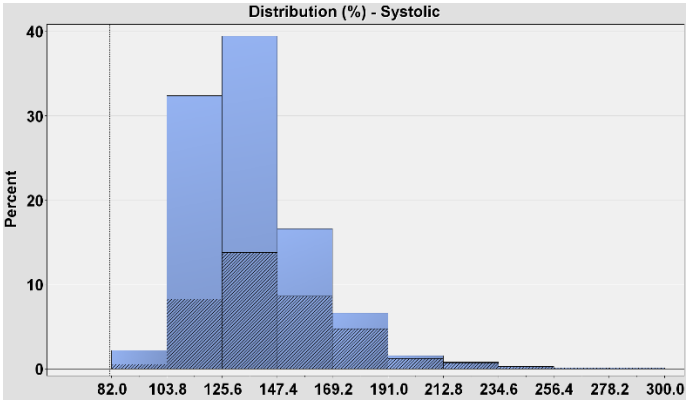

Fig. 3. Systolic (%) histogram

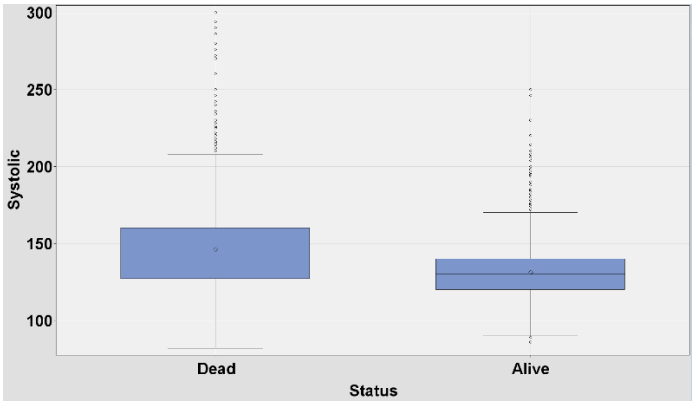

Fig. 4. Systolic value boxplot

Interval variable "***Height***" records participant height in inches, with a min. of 51.5, max. of 76.5, and mean of 64.813. A small SD of 3.582, and lowest skewness (0.177) and kurtosis of -0.396, it is the most normally distributed variable (Fig. 5), this assessment further supported by a median of 65 and so extremely close to mean. 6 observations (0.115%) are missing height data, 5 of those associated with dead participants. Both target levels have similar distribution as indicated by comparable boxplots, with means of 64.98 ("*Dead*") and 64.70 ("*Alive*"), although it's noted "*Dead*" is the only target level with outliers below min. whisker (Fig. 6).

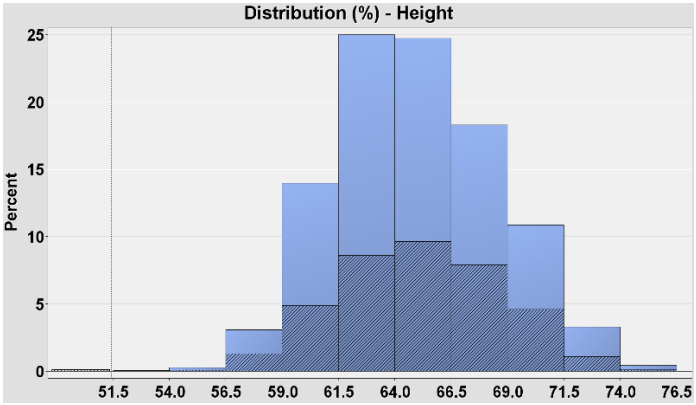

Fig. 5. Height (%) histogram

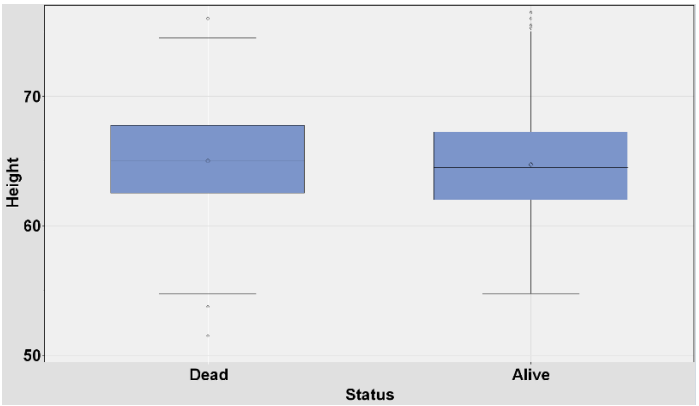

Fig. 6. Height value boxplot

Interval variable "***Weight***", measured in kg, gives min. of 67, max of 300 and mean of 153.086. SD is 28.915. 26 outliers (0.5% of observed with weight) above +3SD with avg. weight of 259kg result in a tail heavy, positively skewed distribution (0.555) (Fig. 7). There are no values for 0.115% of population (6 observations), split equally between "*Alive*" and "*Dead*". Mean for "*Dead*" is higher than for "*Alive*", with 158.27 vs 149.88. Inter-quartile for "*Dead*" is slightly taller than for "*Alive*", suggesting a greater value dispersion, with a higher max. whisker. This is supported by a taller min. and max. whiskers for "*Dead*" which suggests greater weight variance in death by C.V.D. (Fig. 8). Similar to variable "*Height*", "*Dead*" is the only group with outliers below min. whisker. Both groups have extreme upper outliers including highest 300kg, although those above max. whisker are more clustered in the "*Alive*" level.

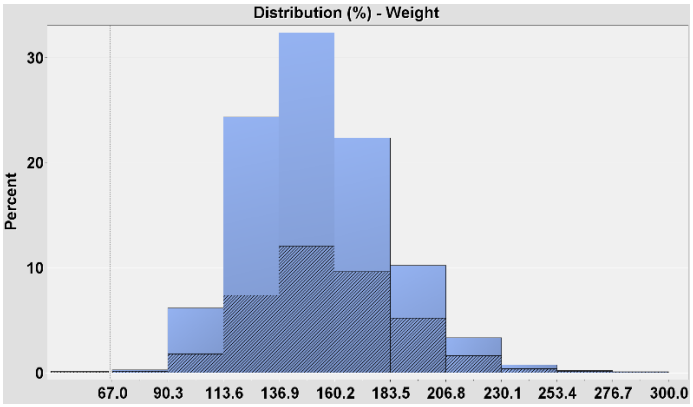

Fig. 7. Weight (%) histogram

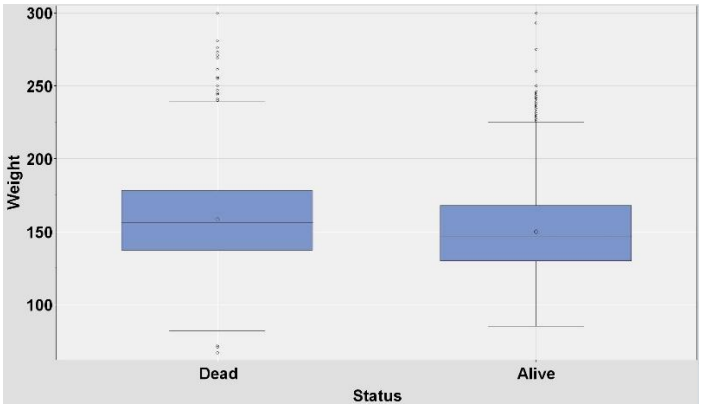

Fig. 8. Weight value boxplot

Variable "***Smoking***" represents number of cigarettes smoked per week, with values 0, 1, 5, 10, 15, 20, 25, 30, 35, 40, 45, 50, 55 and 60. 48.01% are non-smokers, the 2nd largest group smoking 20 per week (17.68%), and 3rd 5 per week (8.95%) (Fig. 9). 16 observations (0.691%) have missing values. Both levels for target "*Status*" have a min. whisker and first quartile of 0 (Fig. 10), emphasising the dominance of non-smokers in the population. The "*Dead*" group has a median of 5 and mean of 10.86 compared to median of 0 and mean of 8.44 for "*Alive*", suggesting proportionally more smokers in the target "*Dead*".

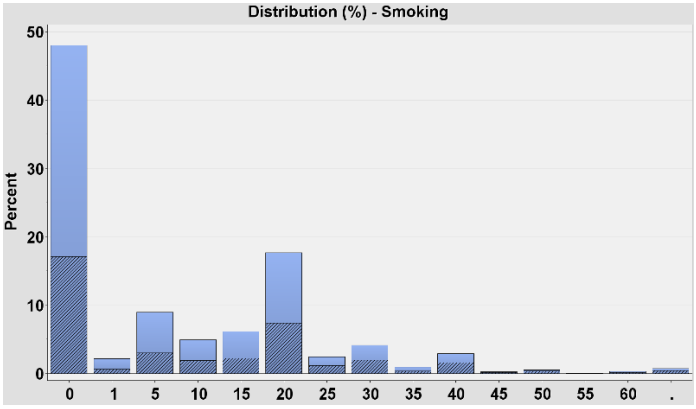

Fig. 9. Smoking (%) histogram

Fig. 10. Smoking value boxplot

Interval variable "***Cholesterol***" shows min. 96, max. 568 and mean 227.417. SD is 44.93. 32 participants above +3SD record significant outliers (avg. 405), supporting observed tail-heavy, skewed positive distribution (0.816) (Fig. 11). A 2nd highest kurtosis of 2.103 indicates some clustering around mean. Values are missing for 2.918% (152 observations), ranking 2nd in nullity value frequency after "*Death_age*". Fig. 12 shows a slightly higher inter-quartile for "*Dead*" suggesting generally higher cholesterol levels. "*Dead*" shows longer min. and max. whiskers, indicating a greater variance, in addition to greater dispersity and extremity of outliers above max. whisker, in contrast to shorter whiskers and outliers for "*Alive*" tightly clustered closer to max., suggesting less variance. Mean for "*Dead*" is 236.31 compared with 221.95 for "*Alive*".

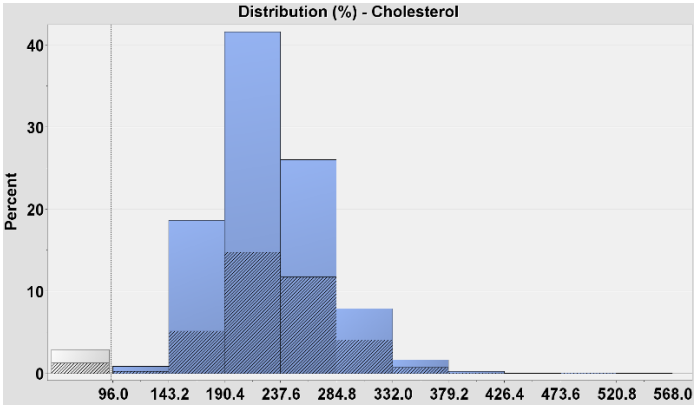

Fig. 11. Cholesterol (%) histogram

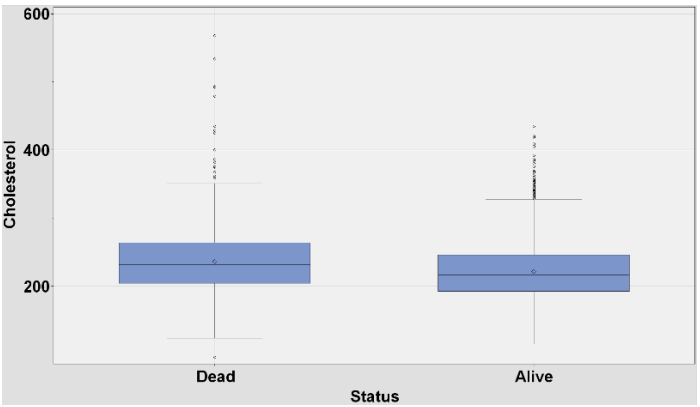

Fig. 12. Cholesterol value boxplot

Nominal variable "***Smoking_status***" has the following 5 levels – "*Non-smoker*" (48.01%), "*Light (1-5)*" (11.12%), "*Moderate (6-15)*" (11.06%), "*Heavy (16-25)*" (20.08%) and "*Very Heavy (>25)*" (9.04%). This is per week according to dataset specification given in *Appendix A*. Considering the US National Institute of Health defines light smokers as smoking less 39 cigarettes per week [6], it is reasonable to conclude the dataset figures actually represent daily consumption. This is further supported by "*Heavy smokers were classified as those who smoked, on average, 25 or more cigarettes per day, moderate smokers as 15–24 cigarettes per day, and light smokers less than 15 cigarettes per day*" [7]. As per variable "*Smoking*", 16 observations (0.691%) having missing values.

Nominal variable "***Bpressure_status***" categorises blood pressure with 3 levels – "*Normal*", "*Optimal*" and "*High*". There are no missing values. Correlation analysis of variables "*Systolic*" and "*Dystolic*" with "*Bpressure_status*" determines ranges as follows: "*Normal*" ("*Systolic*" 101-140, "*Dystolic*" 54-88), "*High*" ("*Systolic*" 112-300, "*Dystolic*" 52-160) and "*Optimal*" ("*Systolic*" 82-118, "*Dystolic*" 50-78). These ranges do not align with latest medical opinion according to the American Heart Association [8] (*Appendix B* Fig. 36), which for example indicates normal pressure when systolic less than 120mmHg and not 140mmHg. This supports scepticism over the quality of rules categorising blood pressure status at time of study data collection, or perhaps highlights further refinement in modern medical guidelines.

Nominal variable "***Weight_status***" categorises participant weight with 3 levels – "*Underweight*" (range determined 67kg-150kg, 3.47%), "*Normal*" (92kg-186kg, 28.26%) and "*Overweight*" (104kg-300kg, 68.15%). Note these categories overlap by weight range, likely inferred taking into account participant height. Going by this weight status the population is dominated by the overweight. This variable logs no value for 0.115% or 6 observations, split equally between alive and dead participants. It is bound with variable "*Weight*" as the same 6 observations are absent of both "*Weight*" and "*Weight_status*" value. Let's consider observation 15 which represents a male participant weighing 155kg (24.4 stone), height of 69 inches (5' 7"), and "*Weight_status*" of "*Normal*". According to National Health Service (NHS) UK guidelines [9] (*Appendix B* Fig. 37) this participant would be classified as very obese. This further supports the view that at least 3 existing categorical variables in this dataset are unreliable and misaligned with today's recommendations, and therefore must be treated with caution or rejected.

"***Chol_status***" is a categorical variable with categories "*Desirable*" (range determined cholesterol 96-199, 26.97% of population), "*Borderline*" (cholesterol 200-239, 35.73%), and "*High*" (cholesterol 240-568, 34.38%), with the same 152 observations missing cholesterol categorisation as they are without cholesterol measure. Recommendations on cholesterol levels from the Texas Heart Institute suggest "*In general, you want to have a cholesterol level below 200 mg/dL. Between 200 mg/dL and 239 mg/dL, your cholesterol level is elevated or borderline-high and should be lowered if you can. With a level of 240 mg/dL or above, your cholesterol level is high, and there is a need for action*" [10]. The categorical levels in the dataset as defined by the cholesterol ranges align with these recommendations.

Table 1 presents interval variables mean by "*Sex*" and target level "*Status*". The proportion of alive and dead are more evenly balanced for males than females. Both blood pressure rate components – systolic and diastolic – have a higher mean for the dead group, backed by "*28% of CHD events in men and 29% in women were attributable to blood pressure levels that exceeded high normal (130/85)*" [1]. The cholesterol mean for the alive group is roughly the same for both genders, but higher for the females in the dead group than males. Weight is only significantly higher in the dead group for females, while height differences between status are insignificant. Only the dead male group shows higher smoking. There are no univariate or time-series variables. Note hatched area on histograms highlights data associated with target level "*Dead*". Boxplots plotted using lattice with X axis target "*Status*" and Y the independent variable. See *Appendix C* for more graphs.

Table 1 Interval Variable Characteristics (mean) by Sex and Status

Characteristic	Female		Male	
	Alive (No C.V.D.)	*Dead (C.V.D.)*	*Alive (No C.V.D.)*	*Dead (C.V.D.)*
Frequency	1977	896	1241	1095
Percent	68.81%	31.19%	53.13%	46.88%
Systolic	131.12	149.59	131.27	143.35
Diastolic	82.16	90.12	83.75	89.04
Cholesterol	221.96	243.19	221.94	230.76
Weight	138.97	146.72	167.23	167.73
Height	62.70	62.27	67.88	67.20
Smoking	5.31	5.61	13.43	15.17

3. Secondary Data Analysis

Building upon the detailed examination of variable distribution it is now appropriate to explore the predictive value of each independent variable by determining how significant the relationship is between it and target variable "*Status*". 5209 participants have been observed. Are they "*Alive*" or "*Dead*" simply due to chance or do any of the input variables contribute?

3.1 Significance with Chi-Square

The Chi-Square (x^2) test, first investigated by Karl Pearson in 1900 [11] can be used to examine this variable relationship. According to [12], the Chi-Square "*is a non-parametric tool designed to analyse group differences when the dependent variable is measured at a nominal level*" and "*the Chi-square provide considerable information about how each of the groups performed in the study. This richness of detail allows the researcher to understand the results and thus to derive more detailed information from this statistic than from many others*".

The purpose of the test is to accept or reject our null hypothesis being "*no statistically significant difference between observed and expected variable frequencies*". In other words, using an example, "*the relationship between observed systolic rate and "Status" (Alive/Dead) is not statistically significant*". With target variable "*Status*" having two levels, "*Alive*" and "*Dead*", there is one degree of freedom in outcome, and referencing the statistical table for x^2distribution gives us a critical value of 3.841 [13]. If the calculated x^2 for each variable relationship with target variable exceeds 3.841 the null hypothesis can be rejected as there is 95% confidence the variable relationship is statistically significant.

The S.E.M. node *StatExplore* can provide a x^2 plot of relationship strength between input variables and target variable "*Status*" (Fig. 13). To calculate x^2 for continuous interval variables like "*Systolic*" they must first be binned categorically by setting the "*Chi-Square Statistics*" group property "*Interval Variables"* to "*Yes*" and "*Number of Bins"* to default of 5, with the plot showing sum total x^2 for binned categorical variables. Variable "*Death_age*" is not included as rejected in the previous *File Import* node.

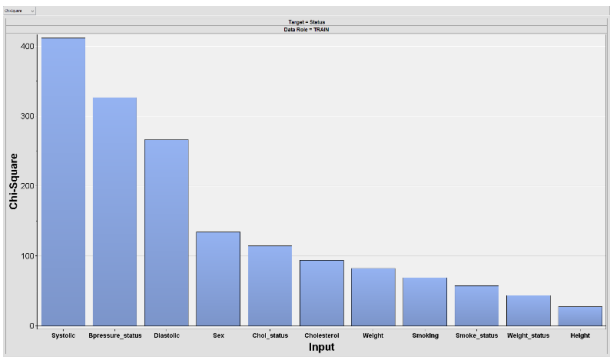

Fig. 13. Chi-Squared

Target	Input	Target: Formatted Value	Input: Formatted Value	Frequency Count	Chi-Square
Status	Systolic	Alive	125.6 -169.2	1752	1.424418
Status	Systolic	Alive	169.2 -212.8	116	81.80537
Status	Systolic	Alive	212.8 -256.4	8	20.44546
Status	Systolic	Alive	LOW-125.6	1342	47.57264
Status	Systolic	Dead	125.6 -169.2	1166	2.302248
Status	Systolic	Dead	169.2 -212.8	309	132.2198
Status	Systolic	Dead	212.8 -256.4	48	33.04545
Status	Systolic	Dead	256.4 - HIGH	10	9.984963
Status	Systolic	Dead	LOW-125.6	458	76.89039

Fig. 14. StatExplore node Cell Chi-Square view variable "Systolic" per bin

The x^2 plot (Fig. 13) suggests variable "*Systolic*" is the most important variable as it has the highest total x^2 value and hence strongest relationship with target variable "*Status*". With a x^2 of 411.8685 the null hypothesis can be rejected with 95% confidence. x^2 can be calculated with $\frac{(O-E)^2}{E}$ where O is the observed frequency and E the expected frequency. To obtain estimated frequencies we first derive observed rates, calculated as totals for target level divided by total observations:

"*Alive*" rate = 3218 / 5209 = 0.6177769245536571

"*Dead*" rate = 1991 / 5209 = 0.3822230754463429

The estimated status frequencies are then calculated by multiplying the observed row total for bin by rate for "Status" level, with bin "LOW – 125.6" as an example:

LOW 125.6 "*Alive*" = 1800 * 0.6177769245536571 = **1111.998464196583**

LOW 125.6 "*Dead*" = 1800 * 0.3822230754463429 = **688.0015358034172**

Having derived estimated frequencies x^2 can now be calculated, using $\frac{(O-E)^2}{E}$ with bin "*LOW - 125.6*" "*Alive*" as an example. Table 2 gives values used in calculation.

SQRT(1342 - 1111.998464196583) / 1111.998464196583 = 52900.70647193051 / 1111.998464196583 = **47.57264**

TABLE 2 Chi-Squared for Systolic with target Status

Systolic	**Observed Status Frequencies**			**Estimated Status Frequencies**			**Chi-Squared**	
Bin	***Alive***	***Dead***	***Total***	***Alive***	***Dead***	***Total***	***Alive***	***Dead***
LOW - 125.6	1342	458	1800	1111.998464196583	688.0015358034172	1800	47.57264	76.89039
125.6 – 169.2	1752	1166	2918	1802.673065847571	1115.326934152429	2918	1.424418	2.302248
169.2 – 212.8	116	309	425	262.5551929353043	162.4448070646957	425	81.80537	132.2198
212.8 – 256.4	8	48	56	34.5955077750048	21.4044922249952	56	20.44546	33.04545
256.4 – HIGH	0	10	10	6.177769245536571	3.822230754463429	10	6.17777	9.984963
Totals	3218	1991	5209	3218	1991	5209	157.425658	254.442851
Rate	0.6177769245536571	0.3822230754463429						
Total Systolic Chi-Squared							411.868509 *	

* Rounded to 6 decimals

The *StatExplore* node also provides the individual Chi-Squares and frequency counts per input variable by navigating to *View* > *Summary Statistics* > *Cell Chi-Squares* (Fig. 14), which is particularly interesting as the statistic is given for both target levels "*Alive*" and "*Dead*" per individual bin. This suggests that systolic values binned to category "*212.8 – 256.4*" have more significance (33.045) with dead study participants than those alive (20.445). Bin "*125.6 – 169.2*" for "*Alive*" and "*Dead*" target levels are statistically insignificant with x^2 values of 1.4244 and 2.3022 respectively, both below the critical value of 3.841 and thus 95% sure the null hypothesis can be accepted.

Several observations are drawn from the full x^2 value table given in *Appendix D*, split at critical value. For level "*Dead*", systolic bin "*169.2 – 212.8*" has a higher x^2 (132.21) than bin "*LOW - 125.6*" (76.89) as the difference between observed and estimated frequency counts is greater. Bin "*LOW - 125.6*" remains statistically significant to target level "*Dead*" given frequency count and difference between estimated (688) and observed (458), but also suggests that despite a low systolic rate there are other factors are contributing to C.V.D. Systolic bin "*256.4 – HIGH*" with x^2 of 9.98 is statically significant given the frequency count is only 10 observations, greater than the estimated 3.82 and all "*Dead*". Despite more females in the study (55.15%), dead males have greater x^2 (45.75) than dead females (37.20), suggesting males are at greater risk. A high "*Bpressure_status*" has a significantly higher x^2 for "*Dead*" (107.70) than "*Optimal*" (57.39) and "*Normal*" (36.59), and greater than for "*Alive*" with high blood pressure (66.64), implying the higher the blood pressure the greater the risk of death from C.V.D. Likewise, a

"*Chol_status*" of "*High*" gives 35.29 for "*Dead*" compared with 21.83 for "*Alive*". Non-smokers dying from C.V.D. is statistically significant with x^2 of 4.41, less than for very heavy smokers (18.03).

3.2 Association with Cramer's V

The author of [12] suggests "*The Chi-square is a significance statistic, and should be followed with a strength statistic. The Cramer's V is the most common strength test used to test the data when a significant Chi-square result has been obtained*". Cramer's V is based on Chi-Square and published by Harald Cramer in 1946 [14]. It is a measure of association, with output ranging between 0 and 1. The formula for Cramer's V is $\sqrt{\frac{x^2}{n}}$, where x^2 is Chi-Square and n the total number of observations. For total "*Systolic*" it's calculated as $\sqrt{\frac{411.868509}{5209}}$, resulting in 0.281 as shown leftmost bar Fig. 15. Cramer's V calculated values can be interpreted in accordance with the scale given in Table 3 [15, p. 219]. "*Systolic*" (1st, 0.281), "*Bpressure_status*" (2nd, 0.250) and "*Diastolic*" (3rd, 0.226) have a moderate association with target variable "*Status*". "*Sex*" (4th, 0.160), "*Chol_status*" (5th, 0.148), "*Cholesterol*" (6th, 0.134), "*Weight*" (7th, 0.125), "*Smoking*" (8th, 0.114) and "*Smoke_status*" (9th, 0.105) have a weak association. "*Weight_status*" (10th, 0.091) and "*Height*" (11th, 0.072) have negligible association.

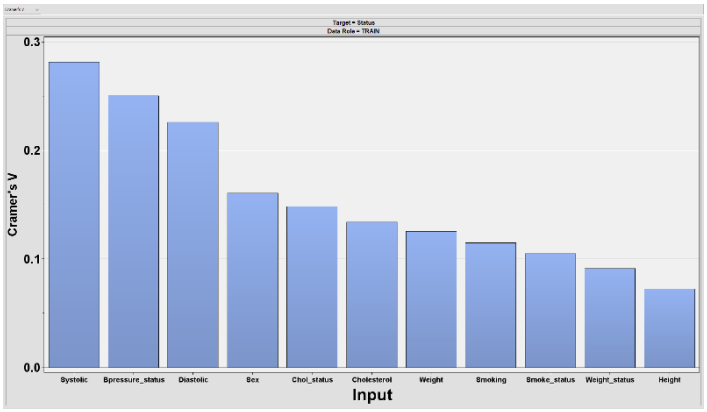

Fig. 15 Cramer's V independent input variable vs target dependent Status

Cramer's V Measure	***Interpretation***
0 and under 0.1	Negligible association
0.1 and under .20	Weak association
.20 and under .40	Moderate association
.40 and under .60	Relatively strong association
.60 and under .80	Strong association
.80 to 1.00	Very strong association

Table 3 Cramer's V Interpretation

Having assessed the significance and association strength of input variables to target "*Status*" the findings further support distribution observations made earlier. An example is "*Systolic*", shown to be the most skewed with boxplot for the "*Dead*" target level taller and higher than for "*Alive*", and Cramer's V demonstrating the strongest association of all input variables. Conversely "*Height*", the most normally distributed of variables with comparable boxplots also returns the weakest Cramer V association. An observation to be verified during classification model evaluation is the difference in strength between interval variable(s) and their associated categorical representation. One example is categorical variable "*Bpressure_status*" which sits between the interval variables "*Systolic*" and "*Diastolic*" from which it is derived. Another is "*Weight_status*" as a derived category from "*Weight*" (weak) and "*Height*" (negligible) intervals, likely negatively impacted by the low predictive value of "*Height*". These categorical variables may introduce noise and classifier testing will include comparison benchmarking with strategies that use and reject them.

3.3 Variable Correlation with Scatter Plots

Correlation, or the relationship between variables, is explored with scatter plots drawn with the S.E.M. *Graph Explore* node. The plots illustrate the form, direction and strength of the relationship, and highlight outliers. Fig. 16 plots the two blood-pressure related measurements, "*Systolic*" on the X axis versus "*Diastolic*" on the Y, coloured blue if the participant is dead and red if alive. The two variables have a moderate and reasonably bunched, positively linear relationship. Grouping by "*Status*" with colour indicates greater risk to participants with death by C.V.D. more common as the values of these two associated variables rises. Many outliers can be seen outside of the general linear form, especially

within the upper right quadrant of the plot, the majority representing dead participants with known high systolic and diastolic rates. Given this risk indicator, any later filtering or imputation of systolic and diastolic outliers must be very carefully considered both for model overfitting concerns or important signals lost.

"*Systolic*" versus "*Height*" is plotted in Fig. 17, showing a weak, non-linear relationship. Following the Y axis there is no discernible pattern with regard to death influenced by "*Height*" for a given "Systolic". Outcome without doubt is dominated by "*Systolic*", with occurrences of death increasing as "*Systolic*" rate climbs, regardless of whether "*Height*" is 54" or 72". There are some outlier exceptions – one with "*Height*" of 51.5" and "*Systolic*" of 126, and several around the "*Systolic*" 175-200 range and "*Height*" around -3SD(54") or +2SD(72"). Imputation of "*Height*" outliers should be considered, especially as this variable was determined weakest of all in earlier Cramer's V testing.

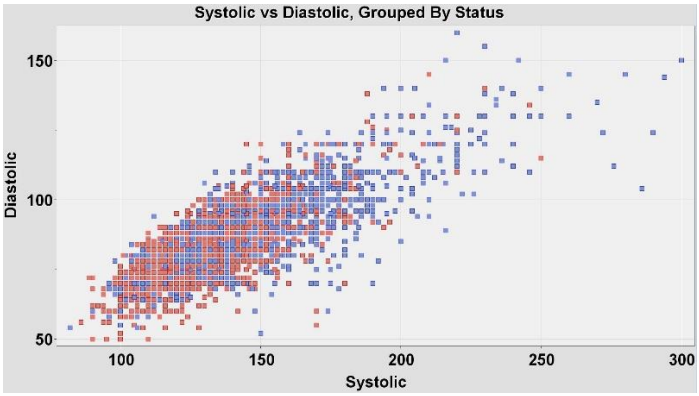

Fig. 16 Systolic vs Diastolic, grouped by Status (Blue=Dead, Red=Alive)

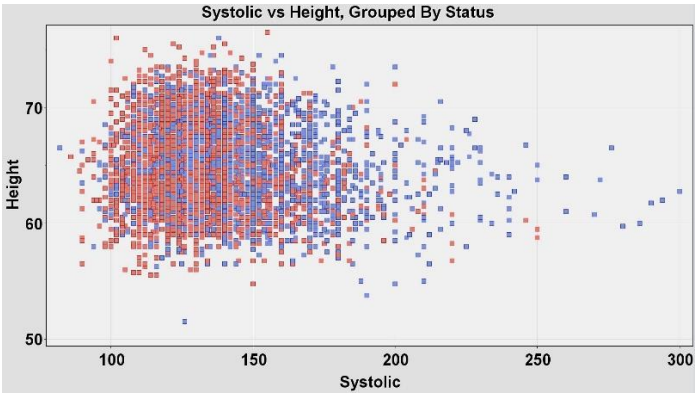

Fig. 17 Systolic vs Height, grouped by Status (Blue=Dead, Red=Alive)

Brief analysis for "*Systolic*" vs "*Cholesterol*", and "*Systolic*" vs "*Weight*" is given in *Appendix E*.

3.4 Grouping with Clustering

Clustering is an unsupervised learning method that groups data by looking for distinct patterns and does not look to predict the class of target variable "*Status*". This study performs clustering following distribution and significance analysis to see whether similar variable dispersal, associations and strengths are used to group participants by common characteristics. It is done prior to supervised modelling as discovery here may help direct assembly of best possible classifier algorithms by influencing which input variables are used. Generally, with clustering, the aim is to maximize both inter-cluster homogeneity (variance within cluster) and intra-cluster separation (distance between clusters). In addition, clustering should make sense according to expert insight and background research on the problem domain, but also according to ground truth from empirical analysis of distribution and significance. The following criteria form the basis of cluster evaluation:

- **Cluster variance** – assessed with avg. RMSSTD (Root Mean Square Standard Deviation)
- **Cluster separation** – assessed with avg. distance to nearest cluster segment (DNCS)
- **Variable importance** – cluster variable importance comparison with Cramer's V significance
- **Expert knowledge** – does the segmentation make sense?

The avg. RMSSTD should be as small as possible, as it indicates similarity of observations and compactness of a cluster. Average DNCS should be as large as possible for good separation between groups. DNCS can however decrease as more clusters are introduced to the space.

The S.E.M. *Cluster* node is modified with "*Selection Criterion > Preliminary Maximum"* = 150 (number clusters preliminary pass), resolving an issue where the algorithm was not creating the actual final maximum cluster number specified. It is assumed this increase helped deal with the large combined

variance in dataset observation characteristics. Property "*Missing Values > Interval Variables"* was also explicitly set to "*Mean*" as there is no interest in additionally segmenting participants by missing "*Cholesterol*" or "*Weight*" values. After exhausting testing a subset of input variables marked "*Yes*" were used (Fig. 18) in the *Cluster* node, with "*Height*" rejected (Cramer's V negligible association), and "*Smoking"* changed from nominal to interval in the *File Import* node. The process for cluster evaluation follows, results summarised in Table 4:

1. Execute the *Cluster* node, varying the *Final Maximum* value from 2 to 10
2. Capture variable importance
3. For each segment log RMSSTD and DNCS
4. Calculate average RMSSTD and DNCS

Name	Use
Bpressure_statu	No
Chol_status	No
Cholesterol	Yes
Death_age	No
Diastolic	Yes
Height	No
Sex	Yes
Smoke_status	No
Smoking	Yes
Systolic	Yes
Weight	Yes
Weight_status	No

Fig. 18 Cluster variables

Final Max.	***RMSSTD (Avg)***	***DNCS (Avg)***	***Order of Variable Importance***
2	0.8376	2.3272	systolic, diastolic, weight, cholesterol, sex, smoking
3	0.7526	2.5787	diastolic, weight, sex, smoking, systolic, cholesterol
4	0.7091	2.3028	diastolic, weight, sex, smoking, systolic, cholesterol
5	0.6758	2.1486	diastolic, weight, smoking, systolic, sex, cholesterol
6	0.6543	2.1249	diastolic, systolic, weight, smoking, sex, cholesterol
7	0.6678	1.9613	diastolic, weight, smoking, systolic, sex, cholesterol
8	**0.6267**	**2.0740**	**diastolic, systolic, weight, smoking, sex, cholesterol**
9	0.6280	1.9849	diastolic, systolic, weight, smoking, sex, cholesterol
10	0.6225	1.9734	diastolic, weight, systolic, smoking, sex, cholesterol

Table 4 Cluster variation, distance and variable importance by number segment number

A final maximum of 8 represents best clustering, with the second lowest RMSSTD (avg.) of 0.6267. An initial assessment of segments 1, 3, 7 and 8 indicates their data has a similarly low variance, with 8 the tightest and cluster 2 with largest variance (Fig. 19). Clusters 1, 3, 7 and 8 are the smallest and similar in size, with cluster 4 the largest (Fig. 20). The clusters show good separation, with higher DNCS (avg.) than clusters with greater final max. (number of segments), and with no overlap (Fig. 21). Clusters 1 and 3 have the least separation, while clusters 2 and 4 appear to represent a significant collection of outliers given their radius and location within the two-dimensional space. Cluster order of variable importance is acceptable, with the first two related to blood pressure and aligning with Cramer's V.

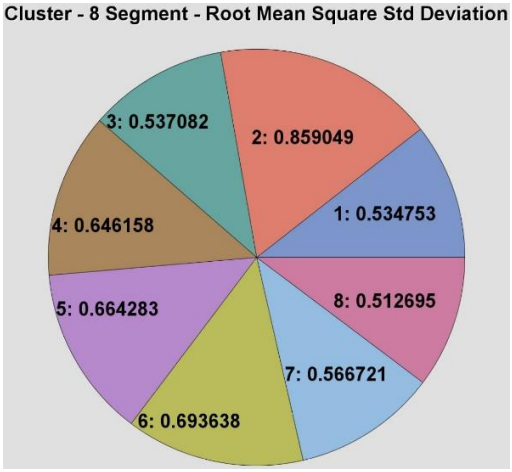

Fig. 19 RMSSTD for 8 segment cluster

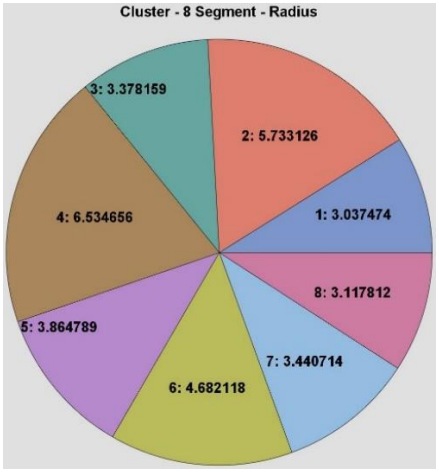

Fig. 20 Radius for 8 segment cluster

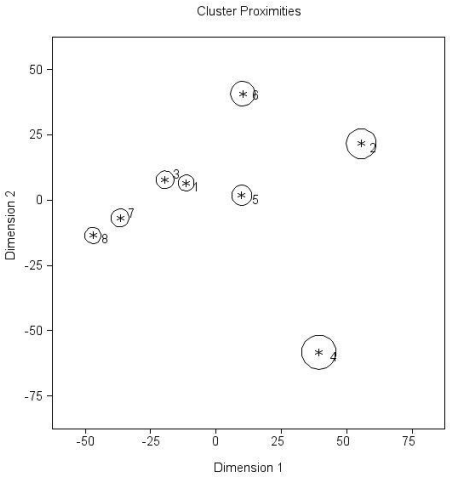

Fig. 21 Proximity for 8 segment cluster

The input means plot visually illustrates the features of each cluster (Fig. 22). All variables have been normalized with min. value of 0 and max. of 1. Beginning interpretation with extremes, cluster 8 (freq. 1234) groups the lowest "*Weight*", "*Cholesterol*", "*Diastolic*", "*Systolic*" and "*Smoking*" levels for a female majority (99%). Cluster 2 (freq. 237) consists of group with 82% females recording below average "*Smoking*", above average "*Weight*" and "*Cholesterol*", but with maximum "*Diastolic*" and "*Systolic*" which is evidence for its greater separation on the proximity plot. Cluster 6 (freq. 343) groups the heaviest with below average "*Cholesterol*" but well above average "*Diastolic*" and "*Systolic*", and below

average "*Smoking*", for 84% males. Cluster 4 (freq. 517) is a segment with 93% females with around average "*Systolic*" and "*Diastolic*", and below average "*Smoking*", but with the highest "*Cholesterol*" (mean 302.01) distinctly separating this cluster furthest from others (Fig. 21). Cluster 5 (536) groups the heaviest smokers (mean 27.98) with well above average "*Weight*", "*Cholesterol*", "*Diastolic*" and "*Systolic*" for 95% males.

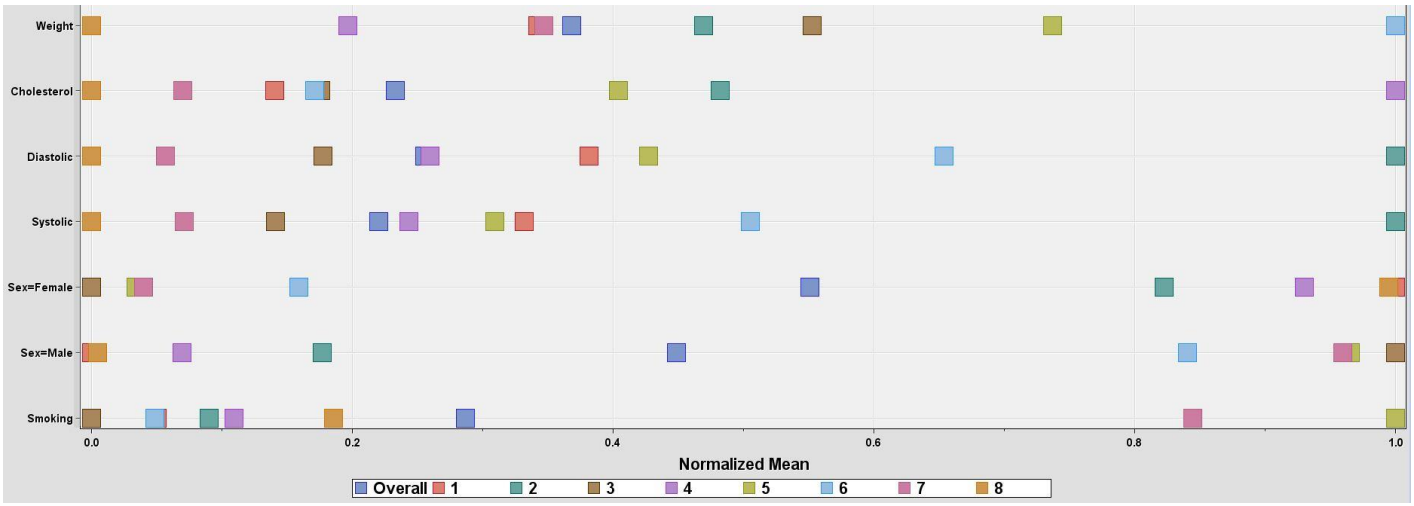

Fig. 22 Input means plot for 8 segment cluster

Cluster 3 (777) groups males with below average "*Systolic*", "*Diastolic*" and "*Cholesterol*", but above average "*Weight*", for non-smokers, and the only cluster with variable purity ("*Sex*", "*Smoking*"). Cluster 1 (865) is below average for "*Weight*", "*Cholesterol*" and "*Smoking*", but above average for "*Diastolic*" and "*Systolic*" for almost only females (>99%). Clusters 1 and 3 likely appear closer on the proximity plot due to their closeness to mean for several variables. Cluster 7 (700) groups below average for "*Cholesterol*", and average for "*Weight*", "*Diastolic*" and "*Systolic*", and second highest mean for "*Smoking*" (23.93), for 95% males.

In simplified terms clustering works by starting with one large cluster of all observations then subdividing it into 2 smaller clusters according to a question using one of the input variables, as determined by a weighted variable value correlation. The splitting continues until no cluster content variation threshold is exceeded or the *Preliminary Max*. property (150) is reached. These smaller clusters are then merged by minimum variance until the specified *Final Max*. (8) number of clusters is assembled. The questions asked during cluster splitting are fully transparent and easily interpretable by viewing the decision tree representation (snapshot Fig. 23, full tree *Appendix F*). 5209 observations in the root node are first split by "*Sex*", with female observations (2873) to the left branch and males (2336) to the right. Females are then further subdivided by "*Cholesterol*" < 266.5 or >= 266.5 and then by "*Systolic*" and "*Weight*", and so subdivision continues.

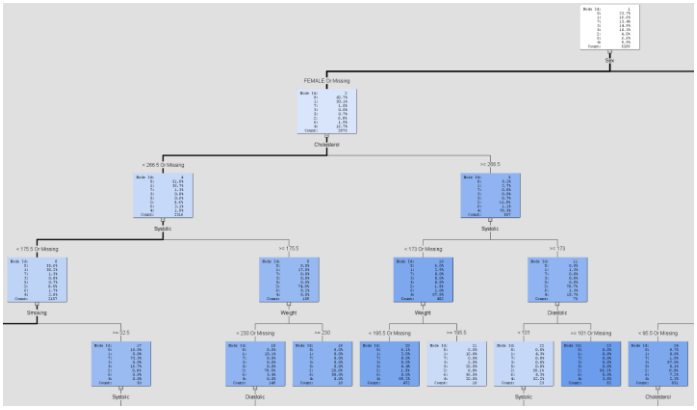

Fig. 23 Snapshot of 8 segment clsuter decision tree

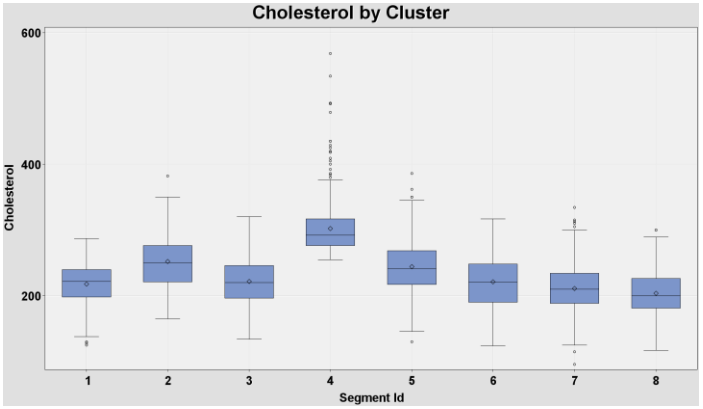

Fig. 24 Cholesterol value by cluster boxplot

Connecting a *Graph Explore* node to the *Cluster* node allows visualisation grouped by cluster (segment) id. To use all observations property *Sample Properties* > *Size* is changed to *Max*. The boxplot in Fig. 24 illustrates cluster 4 grouping high to extreme "*Cholesterol*", having a higher inter-quartile, and positively skewed with a short min. whisker and many outliers beyond max. whisker.

As stated earlier, clustering does not classify, but it very much delivers complimentary analysis in researching a classification problem by providing an alternative perspective on the truth without "bias" from the target "*Status*" label. The clustering outcome makes sense if we consider risk due to incorrect classification is mortality and the business case is preventative action to reduce the risk factors associated with C.V.D. Cramer's V and Chi-Square testing showed moderate association at best between the independent variables and target dependent variable "*Status*". This supports the assumption of a complex interaction between multiple risk factors, as proposed by [1] with "*Coronary prediction estimates tend to be most reliable when the data are most concentrated and can be particularly useful when subjects have multiple mild abnormalities that act synergistically to increase CHD risk*".

The clustering exercise tells us there is a sub-group (cluster 2) with high "*Diastolic*" (mean 115.33) and "*Systolic*" (mean 202.06) rates, mostly females, that suffer a mortality rate of 75% (Fig. 25, 26). Given the authors of [3] advise "*our results provide strong support for lowering blood pressure to systolic blood pressures less than 130 mm Hg*" the mean systolic of 202.06 in this group is way above recommended. Conversely, those females in cluster 8, grouped by common characteristics of lowest "*Weight*", "*Cholesterol*", "*Diastolic*", "*Systolic*" and "*Smoking*", show a survival rate of 82% (Fig. 26), a significant 31% of the total "*Alive*" population. Cluster 8 has mean "*Diastolic*" of 75.04 and mean "*Systolic*" of 118.46, within the recommended guideline. According to [3], "*Overall, a 10 mm Hg reduction in systolic blood pressure reduced the risk of major cardiovascular disease events by 20%*", broadly supported by comparing cluster 4 with mean "*Systolic*" of 138.81 against cluster 5 with mean "Systolic" of 144.33, where "*Alive*" frequency is 302 for cluster 4 and 235 for cluster 5. Such cluster analysis can help focus efforts to reduce C.V.D, for example by specific promotion of healthy lifestyle choices through channels reaching these populations. Full mean statistics for clusters are given in *Appendix F*.

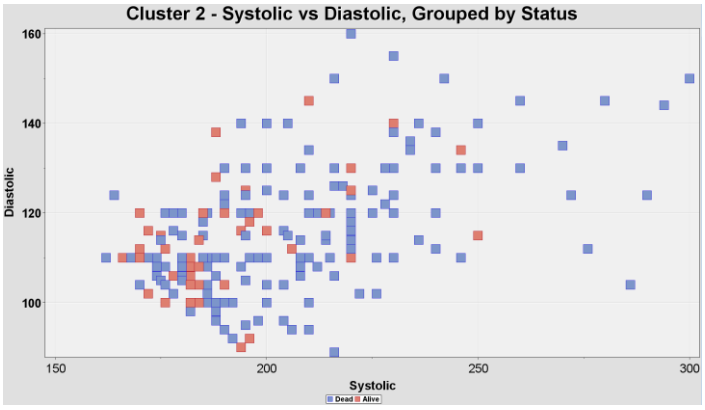

Fig. 25 Cluster 2 Systolic vs Diastolic, Grouped by Status (Blue=Dead)

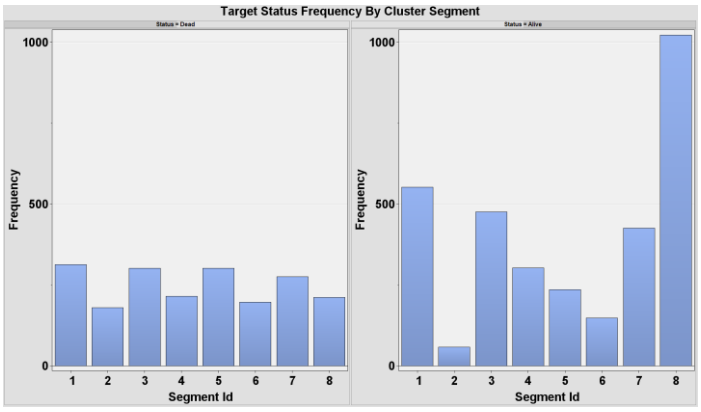

Fig. 26 Status frequency count by cluster segment, dead left, alive right

4. Data Preparation for Classification

Earlier analysis has uncovered interesting dataset characteristics. 8 of 13 variables have missing data, with "*Death_age*" only relevant for minority class "*Dead*" and therefore missing in 61.78% of observations. Variables "*Systolic*" and "*Diastolic*" have a moderate linear correlation and the strongest association with the target event, whilst variable "*Height*" shows negligible association. Variables "*Diastolic*", "*Systolic*" and "*Cholesterol*" register extreme outliers beyond +3SD. Categorical variables such as "*Smoking_status*" and "*Bpressure_status*" give poorer Chi-Square and Cramer's V values compared with the interval variable counterpart(s) they represent. Clustering provided evidence linking higher blood pressure rates for certain cohorts with greater mortality rate by C.V.D. The quality and relationship of information content is now understood. The objective of assembling the best

classifiers for predicting mortality due to C.V.D is clear – misclassifying future high-risk C.V.D. observations as low risk comes at far greater personal cost to the individual than classifying low-risk as high. To extract the best predictive signals from the dataset several data wrangling strategies need to be tested, as summarized in Table 5. "*Death_age*" is excluded by rejection in the *File Import* node as it is only present in observations with target "*Status*" level "*Dead*", and therefore adversely weight any classification and therefore considered redundant for further analysis. "*Height*" is also dropped due to Cramer's V negligible association.

TABLE 5 Data Strategy Summary

Action	**Data Strategy**						
	DS1	*DS2*	*DS3*	*DS4*	*DS5*	*DS6*	*DS7*
Drop "Death_age" & "Height"	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Drop Categorical	No	Yes	No	Yes	Yes	Yes	Yes
Impute Missing	No	No	Yes	Yes	Yes	Yes	Yes
Replacement	No	No	No	No	Yes	No	No
Filter	No	No	No	No	No	Yes	No
Transform	No	No	No	No	No	No	Yes

Dataset strategy variation 1 (DS1), with no further data manipulation and using default classifier configuration, will provide a baseline for benchmarking other strategies to gauge the effectiveness of data wrangling and transformation techniques. Cramer's V analysis ranked the associative strength of categorical status variables with target event "*Status*" lower than their corresponding interval variables, so DS 2, 4, 5, 6 and 7 will use interval variables only, with the *Drop* node removing categorical variables from the dataset. Comparing performance of these strategies with DS1 should show if the categorical variables contribute noise or predictive value and indicate if classifier like neural networks respond better to a reduced variable subset.

Noise introduced by null values causing additional segmentation was observed during clustering tests, so DS 3–7 include imputation with missing "*Smoking*" values is assumed to represent non-smoker and therefore imputed with 0. "*Cholesterol*" and "*Weight*" use the "*Tree*" impute method as evidence of a linear relation, albeit weak, with other input variables like "*Systolic*", suggests values can be more sensibly determined than by using mean (Fig. 27). Imputation is also critical as classifiers such as neural networks, unlike decision trees, are unable to handle missing values and therefore exclude observations. Imputation will ensure the same dataset without loss is being used across the board.

In addition to imputation DS5 will employ the *Replacement* node to replace any interval variable outlier values beyond +/- 3SD with the computed 3SD lower/upper cut-off limit (Fig. 28). Like the *Impute* node, the *Replacement* node replaces original input variables by newly calculated variables. The assumption is valuable signals represented by extreme values in this problem domain will be lost, and testing will support or reject this concern with comparison of classifier metrics.

Variable Name	Impute Method	Impute Value
Chol_status	COUNT	Borderline
Cholesterol	TREE	
Smoke_status	COUNT	Non-smoker
Smoking	MIN	0
Weight	TREE	
Weight_status	COUNT	Overweight

Fig. 27 Impute Node with applied methods

Variable	Limits Method	Lower Replacement Value	Upper Replacement Value
Diastolic	STDDEV	47.00387	124.6626
IMP_Cholesterol	STDDEV	94.98486	358.1555
IMP_Weight	STDDEV	65.04597	241.5787
Systolic	STDDEV	65.06378	210.719

Fig. 28 Replacement Node with applied method/cut-off limits for interval variables

DS6 follows imputation with the addition of a *Filter* node. Unlike the *Replacement* node which retains all observations, the *Filter* node removes observations matching criteria from the training partition on which the classifier models are fitted. The interval variable default filtering method is set to +/- 3SD, the same method used in *Replacement* and thus relevant comparison under same conditions. After execution of the *Filter* node the training set is reduced, and benchmarking will highlight whether any overfitting is managed with exclusion of outliers or valuable signals are lost.

DS7 adds two variations of the *Transform Variables* node to imputation. The first node configured with property *Default Methods* > *Interval Inputs* > *Multiple* generates a series of transformations for each interval using methods including log, log10 and optimal binning, and is exclusively used by the regression classifier in "*Stepwise selection model*" mode to automatically select the best transformations. The second node variation with property "*Best"* selects a single transformation per input interval variable that gives the best Chi-Square value relationship with the target variable.

5. Classifier Overview

Predicting a binary outcome of survival or death due to C.V.D. requires the training of classifier models using a supervised learning method where guidance in the form of the known target outcome for each dataset observation is given. In predicting the likelihood of mortality due to C.V.D. medical care can be focused on those most at risk who can also be encouraged to make lifestyle changes to improve cardio health. The data discovery process highlighted the presence of normal and skewed distributions, some linear variable correlations, weak and moderate associations with target outcomes and clustering of population sub-groups most at risk. A variety of models will be tested to find the best fit to these data interactions and the transformations trialled, that includes logistic regression, the standard decision tree, gradient boosting and neural networks.

The classifier first tested is the logistic regression model, developed by David Cox in 1958, who considered analysis of trials with "*the observation on any one trial taking one of two forms, such as "success" or "failure", "defective" or "not-defective", and so on. Denote the possible observations by 0 and 1, each series of trials therefore giving a sequence of O's and l's"* with *"we suspect that the probability that a particular trial gives say the outcome 1 depends on the corresponding values of the independent variables"* [16]. The author goes on to introduce the logistic form as a method of fitting a linear model to a binary response by proposing "*A linear relation is unsuitable, except over a narrow range, because of the restriction that θ_i must lie in [0, 1] and, in the absence of special considerations for a particular problem, the best form seems to be the logistic law*". Implementation is with the S.E.M. *Regression* node with "*Regression Type*" configured as "*Logistic Regression*". Calculating probability of target variable level "*Dead*" takes form $p(Dead) = \frac{Number\ of\ Events}{Total\ Number\ of\ Events}$ with the equation taking observation 80 as an example, using a full model ("*Cholesterol*" 305, "*Diastolic*" 78, "*Height*" 60.5, "*Sex*" Female, "*Smoking*" 0, "*Systolic*" 154 , "*Weight*" 117), becomes $p(Dead) = \frac{e^{\beta_0+\beta_1+\beta_2+\beta_3+\beta_4+\beta_5+\beta_6+\beta_7}}{1+e^{\beta_0+\beta_1+\beta_2+\beta_3+\beta_4+\beta_5+\beta_6+\beta_7}}$ where e is 2.71828 (Euler's number), β the coefficient (Fig. 29), 0 the intercept and 1-7 the input variables. Writing the combination gives:

2.71828 ^ (-1.5774 + (0.00474*305) + (0.000242*78) + (-0.0624*60.5) + (-0.4737*1) + (0.0148*0) + (0.0287*154) + (-0.00036*117) = 1.016084

Using the logistic regression method to calculate probability of death becomes P(Dead) = 1.016084/ 2.016084 = 0.5039 or 50.39%. With probability 0.5039 exceeding the default threshold of 0.5 the observation is classified as "*Dead*" despite target label of "*Alive*" (Fig. 30), likely due to the more highly weighted "*Systolic*" and "*Cholesterol*" positive values (estimates) considered risk indicators.

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-1.5774	1.1305	1.95	0.1629
Cholesterol		1	0.00474	0.000928	26.09	<.0001
Diastolic		1	0.000242	0.00519	0.00	0.9629
Height		1	-0.0624	0.0170	13.40	0.0003
Sex	Female	1	-0.4737	0.0593	63.76	<.0001
Smoking		1	0.0148	0.00363	16.61	<.0001
Systolic		1	0.0287	0.00299	92.18	<.0001
Weight		1	-0.00036	0.00178	0.04	0.8398

Fig. 29 Regression intercept / variable coefficients

Observation Number	Status	Sex	Height	Weight	Diastolic	Systolic	Smoking	Cholesterol	Residual: Status=Dead	Into: Status
80	Alive	Female	60.5	117	78	154	0	305	-0.5039	DEAD

Fig. 30 Regression node Status=Dead probability for observation 80

The decision tree is the next binary classifier model tested, "*which takes a divide and conquer approach to a given classification problem, starting at the root node of a tree with the whole dataset and splitting*

it left and right into lower nodes according to questions asked" [4]. The S.E.M. *Decision Tree* node is configured to use the ProbChisq (Probability Ch-Square) method for determining the variable value questions asked based on the significance relationship between input variables and the target variable. Segmentation continues until the predicted target variable level is of the same value within a node subset, or a configured threshold like "*Maximum Depth*" is exceeded. An inverted tree-like structure with a root node at top, (all observations, with distribution to each binary target "*Status*" level given), decision nodes (with question based on input variable value) and terminal nodes (final observation classification distributions) is visualized in the "Tree" view (Fig. 31), which illustrates the ease of interpretability the decision tree model offers. Another strength of the model is the ability to accommodate missing values.

As observed earlier during Chi-Square significance analysis in section 3.1, "*Systolic*" showed the strongest relationship with target "*Status*" level "*Dead*", and so with splitting method "*ProbChiSquare*" variable "*Systolic*" has the highest variable importance and determines formation of the single biggest subset by asking the question "*Is Systolic < 141 Or Missing?*", with 2087 of 3124 observations going left (critical path highlighted black). Chi-Square analysis also provided evidence for variables including "*Height*" having very weak significance value and hence a reasonable basis for concluding why the tree remains at depth of 5 with variable "*Height*" unused when "*Maximum Depth*" of 6 is configured.

The third classifier is the Gradient Boosting model, an ensemble of regression trees each based on an independent sample of data without replacement, with the final outcome probability calculated as the average of predictive classification probabilities from all trees. The term "*boosting*" originates from "*a method that re-weights at each step in order to resolve deficiencies from prediction errors in previous tree steps, and hence the ensemble is grown in an adaptive fashion*" [4]. Friedman described gradient boosting as "*competitive, highly robust, interpretable procedures for both regression and classification, especially appropriate for mining less than clean data*" [17]. The S.E.M. *Gradient Boosting* node can calculate deficiencies, or how badly its doing, with loss functions including average square error and misclassification rate, looking to minimize loss with each tree grown. Boosting is less prone to overfit the data than a single decision tree [18], and selected to evaluate if it performs better than the standard decision tree as stated by Leo Breiman with "*Boosting > Bagging > Single Tree*" [19].

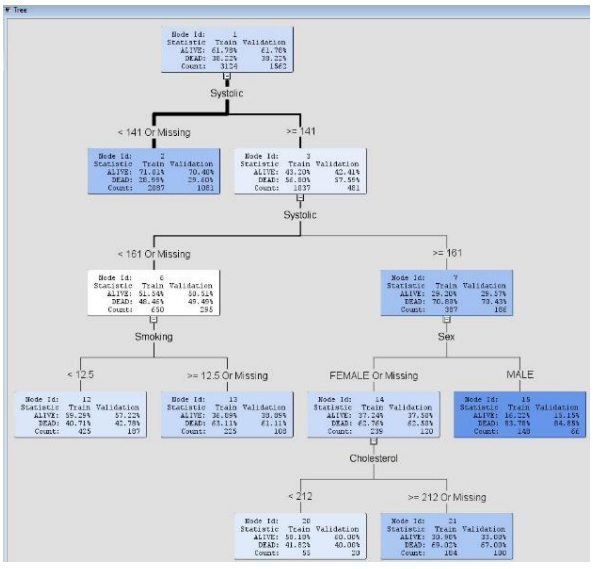

Fig. 31 Decision Tree with strategy DS1 (no imputation)

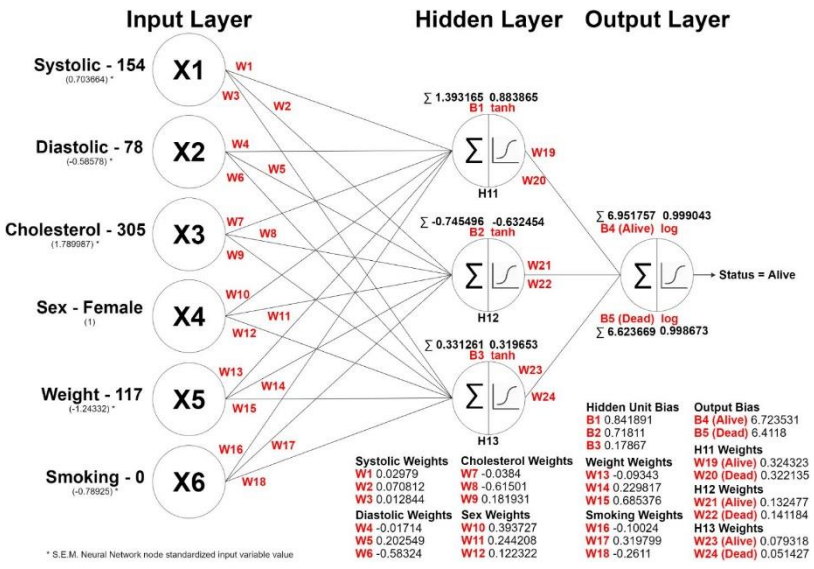

Fig. 32 Multilayer perceptron Neural Network, 1 hidden layer with 3 units

The final classifier tested is the neural network, a model inspired by the human brain with neurons that combine multiple input signals and produce output signals. Neural networks are "*are relatively complex algorithms taken as a whole, although there is great simplicity in their individual elements and elegance in their construction*" [4]. The hierarchical architecture of a multilayer perceptron neural network is presented in Fig. 32 with data from observation 80. It consists of an input layer, 1 hidden

layer and an output layer, with a feed-forward flow from input through hidden to output. The variable final weight and unit bias values shown are calculated during training of the model, starting with estimates which are refined through minimisation of loss error functions.

The input layer consists of units that correspond to each of the input variables and receive their value (standardized), which are then passed forward with estimated variable weights to each hidden unit in the hidden layer, the example containing 3 units each imitating a neuron.

A combination function within each hidden unit, shown as ∑, combines the variable "signals" multiplied by associated weight then adds estimated bias in a linear summation, with the following calculation for unit H11 using input variable standardized values for observation 80:

Bias 0.841891 + (Systolic 0.703664 * 0.02979) + (Diastolic -0.58578 * -0.01714) + (Cholesterol 1.789987 * -0.0384) + (Sex – Female 1 * 0.393727) + (Weight -1.24332 * -0.09343) + (Smoking -0.78925 * -0.10024) = **1.39316**

Each hidden unit also contains a transfer function, represented by the sigmoid S curve in Fig. 32. This function receives the result of the combination function, which is a real number, and transforms it with non-linear functions such as tanh and log into a value range scaled relevant to the target and problem domain. It is these functions that give neural networks their powerful non-linear behaviour. The appropriate choice of function depends on the data, how it's been normalized, and when the hidden unit "neuron" should light up or activate. Tanh centres around 0, with range -1 to 1 and a steeper gradient in comparison to log (0 to 1). The example shows transformation of combination function value with tanh configured in the transfer function. The combination function and transfer function together form the activation function. Model testing with the range of sigmoid functions will confirm the one best matched with the data strategies used and objective of maximising true positive classifications and activating those neurons to predict mortality by C.V.D.

The output or target layer also has an activation function with weighted combination and transfer, using a log transfer function with range 0 to 1 chosen to scale output suitable for binary target variable "*Status*" (1 "*Dead*", 0 "*Alive*"). The illustrated neural network correctly classified observation 80 with target level "*Dead*". Analysing variable importance with a comparison of order between Cramer's V and network hidden unit weighting, hidden unit H11 gives closest match with gender and blood-pressure related values a higher order ("*Sex*", "*Systolic*", "*Diastolic*", "*Cholesterol*", "*Weight*", "*Smoking*"), with "*Sex*" assigned a high positive weight. Unit H12 shows a weak match ("*Smoking*", "*Sex*", "*Weight*", "*Diastolic*", "*Systolic*", "*Cholesterol*") with "*Cholesterol*" a large negative weight. H13 shows a moderate match ("*Weight*", "*Cholesterol*", "*Sex*", "*Systolic*", "*Smoking*", "*Diastolic*"), with "*Weight*" assigned a stronger positive weight but "*Smoking*" and "*Cholesterol*" assigned negative weighting. Hidden units H11 and H13 with larger "*Alive*" weights (W19, W23) support correct classification.

6. Evaluation Design

A quantitative after-only study design will be used to evaluate classifier predictive performance, with empirical measurements collected from execution of the models within S.E.M. The *Data Partition* node typically splits the dataset into training, validation and test sets. The training set is used for developing and fitting the model to the data. The validation set, unseen during model assembly, is used to evaluate the models. Learnings from this initial analysis highlighted the moderate and weak influence of several independent variables on target "*Status*", with an imbalanced dataset in which minority target variable level "*Dead*" is represented by only 38% of observations and females 55% of the population. A variety of configurations to increase the range of variable values and data relationships available to model training and evaluation were tested, including 50|50, 60|40, 60|30, 62|38 and 66|34. These were all found to overfit the models where they showed good performance on the training set but much poorer on validation. Changing partition allocation back to default of 40|30 for training/validation showed comparable gain and lift performance, evidence the overfitting issue was minimized and models better able to generalize with the validation set and any new data with this default allocation. The default

"*Stratified*" partitioning method for binary targets is kept, meaning each observation with target label "*Dead*" has an equal probability of being assigned to the partition sets, same for "*Alive*".

All observations will be quantitatively measured by ratio scale in the form of misclassification rate for the validation set, the lower the rate the better, with the lowest misclassification rate determining the winning model from a predictive performance view. This metric is appropriate for quantifying classifiers where the target is binary and calculated as the proportion of incorrectly classified observations divided by the total number of observations. Let's say the validation set contains 1773 observations, and a model outcome consists of 423 false negatives and 144 false positives. The misclassification rate is calculated as (423 + 144)/1773 = 0.3197, or an error rate of 31.97%.

Each of the 7 data strategies will be connected to 4 models, meaning there will be 7 logistic regression, 7 decision tree, 7 neural network and 7 gradient boosting models. The 7 models within each model type group will be connected to a *Model Comparison* node with "*Report Selection Criteria*" set to "*Valid: Misclassification Rate*" to identify each winning group model, which then compete with each other using the same evaluation criteria to determine the final best predictive model and data strategy.

Earlier data distribution analysis showed a 55% female majority in which "*Dead*" was the imbalanced minority class with 31.19% that recorded higher mean for "*Systolic*", "*Diastolic*" and "*Cholesterol*" than for the same male class. Strong clustering of gender purity was also observed, with cluster 8 99% female, cluster 4 93% female and cluster 5 95% male. These observations justified additional model evaluation. Separation of data into two gender groups will be tested. For each group both a single blood pressure (BP) component ("*Systolic*") and a dual B.P. component ("*Systolic*" and "*Diastolic*") will be tested. These tests also provide the opportunity to confirm if "*In men, a model including both SBP and DBP was significantly better than a model including either measure alone. In women, a model with both SBP and DBP was not significantly better than a model with only SBP*" [20] holds true.

7. Classification Build & Tune

The final S.E.M. modelled solution is shown in Fig. 33, with the flow from left to right. The Framingham dataset is first imported. Initial exploratory and secondary data analysis was supported by the blue box labelled "*Distr. Analysis*" (in-depth distribution review, characteristics including Chi-Square significance and Cramer's V association strength) and yellow box labelled "*Clustering Analysis*". The red box labelled "*Variable Drop/Filter*" dropped "*Death_age*" and "*Height*" as standard and "*Diastolic*" as required for single B.P. component analysis ("*Systolic*" only), with the *Filter* node ("*Tables to Filter*" = "*All Data Sets*") filtering on "*Sex*" for gender specific classification. Note dropping and filtering of variables is done directly after file import and before data wrangling including imputation to prevent any data leakage over into fitted datasets.

The green box labelled "*Data Strategy Preparation*" prepares the strategies described in Section 4, including dropping of categorical status variables like "*Weight_status*" and imputation of missing values. There are two variations of the *Transform Variables* node. The upper variant labelled "*TraVar (Multiple)*" performs several transformations per input variable and connected only to the logistic regression models configured with a "*Stepwise*" selection to choose only a group of variables that are statically significant. The lower variant labelled "*Best*" calculates several transformations per input variable and only outputs the transformation with best Chi-Square significance with target "*Status*". Both variants had "*Optimal Binning > Number of Bins*" set to 10 to explore if more granular segmentation of intervals such as "*Systolic*" is effective. A *Control Point* node for each strategy distributes the prepared dataset to each of the 4 model types.

The purple box labelled "*Models*" shows the 28 models named with model type and used data strategy, such as "*RG (DS1)*" a logistic regression classifier consuming DS1. Each model with a "family" group is connected to a group *Control Point*, which in turn is connected to a *Model Comparison* node, to rank for example the regression classifiers in order of validation misclassification rate (descending). The

best model from each group is then connected to a second control point as shown in the yellow box labelled "*Model Compare (Best)*" to determine the best overall model. The light grey box labelled "*Classification Threshold Analysis*" facilitated exploratory adjustment of threshold value from standard of 0.5 and briefly outlined in Section 9.

Developing the refined model with hyperparameter tuning per classifier group will use input from DS4. This strategy imputes missing values and drops categorical status variables, a reasonable "*generic*" choice to negate possible observation exclusion due to nulls and possible conflict noise between interval and associated categorical representation. The purpose of model tuning is to extract best possible performance over default, however the challenge lies in finding the best combination of configuration parameters bound by time and computing constraints. Ideally each model would be specifically tuned to the characteristics of its own input, however it's reasonable to assume tuning with DS4 is the best compromise. Once a reasonable number of important parameter settings for each model are tested with objective of lowest misclassification rate for validation set, the model configuration will be deployed to other models of same type except for the DS1 model variant acting as baseline with default hyperparameter settings. Note tuning is done manually without access to any automated grid search or random search mechanisms.

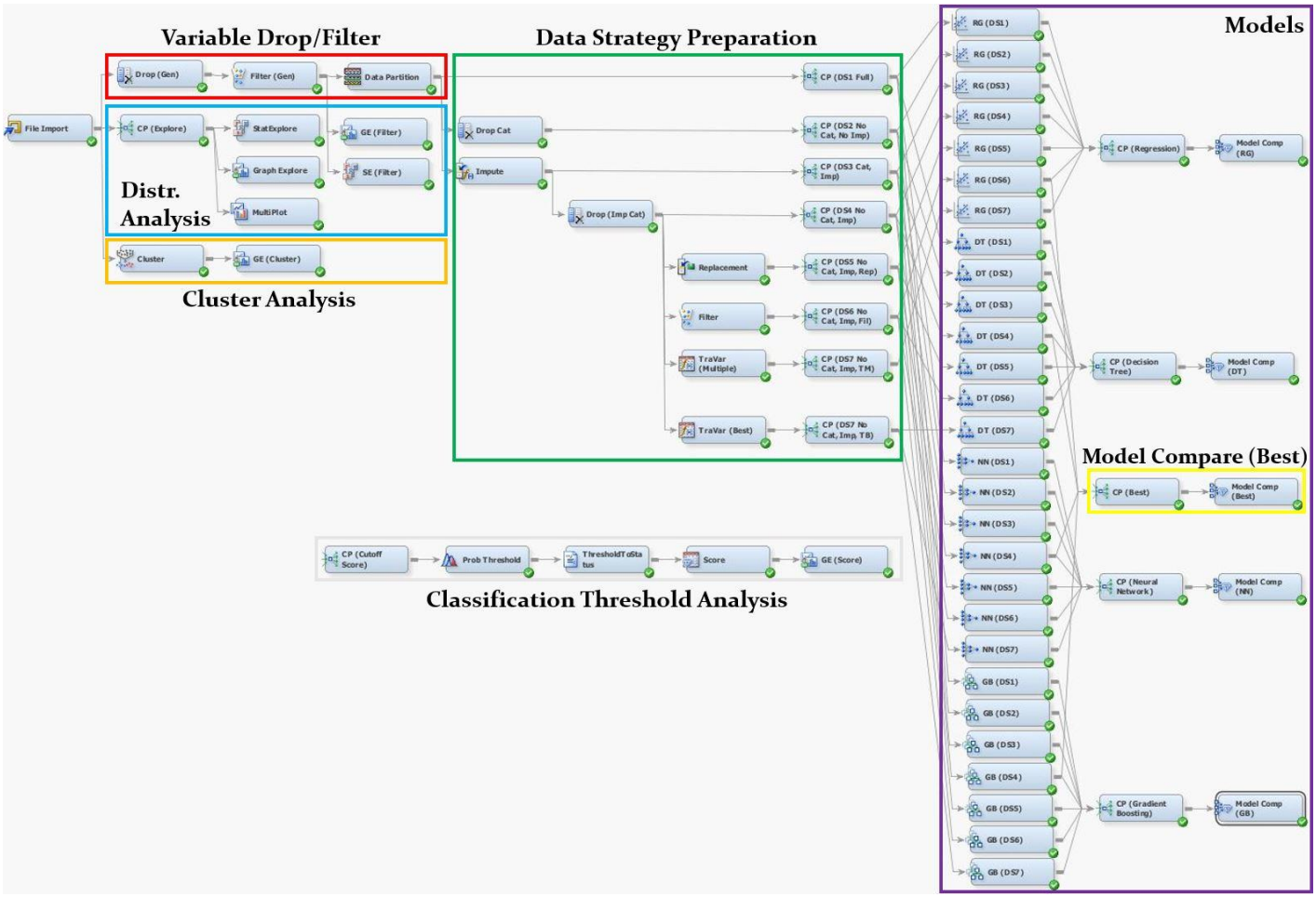

Fig. 33 Final S.E.M. modelling solution

For all tuning "*Selection Criterion*" was set to evaluation metric "*Valid Misclassification Rate*". Tuning the logistic regression model included testing of properties *Link Function* (Logit, Cloglog), *Selection Model* (None, Stepwise, Forward, Backward), *Optimization Options* (default, Congra, Dbldog, Newrap, Nrridg, Quanew, Trureg), *Max Iterations* (default, 600, 800) and *Max Function Calls* (default, 1000, 1200). Best configuration over standard modified *Link Function* to *Cloglog* and *Selection Model* to Stepwise showed minimal improvement, decreasing misclassification rate from 0.325909 to 0.325322.

Despite extensive Decision Tree tuning efforts, the default configuration was found the optimal with *Nominal Target Criterion* = *ProbChisq*, *Max. Branch* = 2, *Max. Depth* = 6, *Leaf Size* = 5, *Number of Rules* =5 and *Significance Level* = 0.2. Any adjustment of these properties either showed no change or degraded misclassification rate. The best positive "tune" for the Decision Tree was not the model properties themselves but reverting of the *Partitioning* node back to standard allocation of 40|30, as the tree was heavily overfitting the test set.

The final configuration for the neural network modified the default "*Number Hidden Units"* from 3 to 7, then setting "*Architecture"* to "*User"* which opened the "*Hidden Layer Activation Function"* for change to *SoftMax*. "*Target Layer Activation"* stays with default of *Logistic* for a binary target. Other properties including "*Direct Connection*" and "*Max Iterations*" were tested with negative results. Final tuning decreased misclassification rate from 0.320905 to 0.300266. The improved performance with SoftMax was unexpected given the intended use is with multi-class classification (not binary) [21].

Attempts to tune the Gradient Boosting model (*N iterations*, *Shrinkage*, *Max. Branch*, *Max. Depth*, *Reuse Variable*, *Interval Bins*, *Leaf Fraction*) resulted in same or worse performance as the default configuration. Finding Gradient Boosting performant "out-of-box" and less receptive to tuning than other classifiers is no surprise as the author discovered in previous research [4, p. 10].

8. Final Classifier Evaluation

Table 6 presents final misclassification rates per scenario and data strategy using the validation set. Baseline strategy includes categorical variables. "*No Cat*" indicates no categorical variables, "*Cat*" includes categorical variables, "*Imp*" applies imputation (Fig. 27), "*Repl*" applies replacement (Fig. 28), "*Filter*" filtering and "*Tra*" transformation. Under category, "*F/M*" represents use of both female and male observations, with "*F*" female only and "*M*" male only. "*Dual BP*" indicates dual blood pressure component using both "*Systolic*" and "*Diastolic*", whilst "*Single BP*" only "Systolic" (and dropping of categorical variable "*Bpressure_status*"). Misclassification rates in bold highlight best (lowest) rate for model (RG Logistic Regression, DT Decision Tree, NN Neural Network, GB Gradient Boosting). Rates highlighted red indicate lowest rate for given category test. Score rank is given beside model type per category. "*Wins*" totals number of highlighted rates per strategy, highest in red. "*Improve Baseline*" gives misclassification rate improvement as best minus baseline.

All scenarios result in misclassification rate less than 0.5 so classify the binary target "*Status*" better than random chance. In 22 of 24 test scenarios there is an improvement over baseline configuration, albeit insignificant except for neural network in category 3 (0.041812), and regression in category 4 (0.032857) and 6 (0.031429). Across all scenarios DS6 is the most effective single data preparation strategy with a win count of 14. This is somewhat unexpected as earlier assumptions suspected filtering out extremes would lose valuable signals. A reasonable assumption is the filtering, only applied to the training set, allows most model scenarios to better generalise when classifying the validation set.

It's clear that no single classifier and data strategy combination works for all scenarios. Neural networks score lowest misclassification in 4 of 6 categories (1, 2, 3, 4). Regression wins category 5. Gradient Boosting wins category 6, and beats decision trees in every category, justifying its inclusion and indicating boosting brings improved performance over standard trees. DS3 with N.N. is the most effective combination without filter on gender with either single or dual component B.P. DS6 is best for female-only observations. As illustrated on the boxplot in Fig. 40, *Appendix C*, females have a higher inter-quartile, higher mean, longer max. whisker and more extreme outliers beyond max. whisker for "*Systolic*". It's reasonable to suggest DS6 with filtering using limiting condition +/-3SD is effective in outlier management and preventing overfitting. DS4 and DS2, both without filtering or replacement, work best for male-only observations. These findings also justify trials in splitting by gender and how cluster analysis showing segmentation by "*Sex*" helped shape subsequent steps in the data mining flow.

With both genders together the dual B.P. component is more effective (NN, DS3) than single component (GB, DS3), evidence also that the categorical variables do contribute predictive value alongside their interval variable counterparts (beating NN, DS4 without categorical). When looking at male-only scenarios, the misclassification rate difference (0.004285) between dual B.P. (0.351429) and single B.P. (0.355714) is marginally more pronounced than in female scenarios (dual 0.268293 vs single 0.269454, diff. 0.001161), so the earlier hypothesis outlined at end of Section 6 "In *women, a model with both SBP and DBP was not significantly better than a model with only SBP*" [20]" can be accepted. "*In men, a model including both SBP and DBP was significantly better than a model including either measure alone"* [20] can also be accepted.

TABLE 6 Final Classifier Validation Set Misclassification Rate

Scenario		**Data Strategy Score by Validation Set Misclassification Rate**							
Category	***Model***	***DS1 Baseline***	***DS2 No Cat***	***DS3 Cat/Imp***	***DS4 No Cat/Imp***	***DS5 No Cat/Imp/Repl***	***DS6 No Cat/Imp/Filter***	***DS7 No Cat/Imp/Tra***	***Improve Baseline***
1 F/M Dual BP	RG 3rd	**0.322663**	0.327145	**0.322663**	**0.322663**	**0.322663**	0.325224	0.330986	0
	DT 4th	0.328425	0.327785	0.327145	0.327145	0.327145	**0.325224**	0.330346	0.003201
	NN 1st	0.328425	0.336748	**0.316261**	0.334827	0.382202	0.338028	0.329065	0.012164
	GB 2nd	0.321383	**0.320743**	0.325224	0.324584	0.324584	**0.320743**	0.331626	0.00064
2 F/M Single BP	RG 3rd	0.324584	0.327145	**0.322663**	**0.322663**	**0.322663**	0.323944	0.330986	0.001921
	DT 4th	0.328425	0.327785	0.327145	0.327145	0.327145	**0.325224**	0.330346	0.003201
	NN 1st	0.327785	0.336108	**0.316901**	0.327785	0.382202	0.334187	0.325864	0.010884
	GB 2nd	0.322023	0.322023	0.323944	0.324584	0.324584	**0.320743**	0.330986	0.00128
3 F Dual BP	RG 2nd	0.279907	0.270616	0.270616	0.270616	**0.269454**	**0.269454**	0.28223	0.010453
	DT 4th	0.29036	0.296167	0.29036	0.296167	0.296167	**0.277584**	0.296167	0.012776
	NN 1st	0.310105	0.289199	0.2741	0.288037	0.311266	**0.268293**	0.297329	0.041812
	GB 3rd	0.28223	0.281069	**0.276423**	0.283391	0.283391	0.277584	0.279907	0.005807
4 M Dual BP	RG 4th	0.402857	0.374286	0.374286	0.374286	0.374286	0.374286	**0.37**	0.032857
	DT 3rd	0.367143	0.367143	0.365714	0.365714	0.365714	**0.364286**	0.374286	0.002857
	NN 1st	0.37	0.368571	0.37	**0.351429**	0.468571	0.352857	0.368571	0.018571
	GB 2nd	0.358571	0.36	0.362857	0.361429	0.361429	**0.355714**	0.372857	0.002857
5 F Single BP	RG 1st	0.281069	0.270616	0.270616	0.270616	**0.269454**	**0.269454**	0.28223	0.011615
	DT 3rd	0.29036	0.296167	0.29036	0.296167	0.296167	**0.276423**	0.296167	0.013937
	NN 4th	0.288037	0.289199	0.29849	**0.278746**	0.311266	**0.278746**	0.283391	0.009291
	GB 2nd	**0.270616**	**0.270616**	0.271777	0.271777	0.271777	0.271777	0.279907	0
6 M Single BP	RG 4th	0.401429	0.374286	0.374286	0.374286	0.374286	0.374286	**0.37**	0.031429
	DT 3rd	0.367143	0.367143	0.365714	0.365714	0.365714	**0.364286**	0.374286	0.002857
	NN 2nd	0.377143	**0.361429**	0.398571	0.365714	0.465714	**0.361429**	0.365714	0.015714
	GB 1st	0.36	**0.355714**	0.361429	0.362857	0.362857	0.362857	0.374286	0.004286
Wins		2	4	5	4	4	**14**	2	

Of all scenarios category 3, female-only with dual B.P. component when using DS6 achieves lowest overall misclassification of 0.268293, in stark contrast to the misclassification rate of 0.351429 for male-only category 4. Table 7 presents the classification table with metrics to describe the predicted classifier outcome and actual target "*Status*" levels. True positives (TP) is number of correctly predicted "*Dead*". True negatives (TN) is number of correctly predicted "*Alive*". False positives (FP) is number of "Alive" predicted as "Dead" by the classifier. False negatives (FN) is number of "*Dead*" predicted as "*Alive*". Under scores, "*Precision*" indicates the proportion of observations predicted as "*Dead*" did suffer mortality due to C.V.D. "*Sensitivity*" is the proportion of observations "*Dead*" the classifier correctly predicted. Finally, "*Specificity*" gives the proportion of observations that did not die from C.V.D. that were correctly predicted, and opposite of sensitivity. Category 3 female-only with NN DS6 scores lowest misclassification, recalling the formula = (FP+FN)/Obs., therefore (46+185)/861 = 0.268293. Category 4 male only misclassification = (181+65)/700 = 0.351429. This suggests the lower number of FN plus high number of TN contributing to observation total contributes to low misclassification rate for category 3, whereas higher number of FN and lower TN results in higher rate for category 4. Splitting of the dataset by gender is the right approach when considering the risk due

to incorrect classification of "*Dead*" as "*Alive*" is higher than "*Alive*" as "*Dead*", if the purpose of the classifier is to drive medical guidelines based on mortality by C.V.D. This is supported by the highest sensitivity rates achieved by category 3 and 4. See *Appendix G* for winning model classification chart.

TABLE 7 Category Best Model Classification Table

Scenario			**Classification Table**				**Scores**		
Category	***Model***	***Strategy***	***FN***	***TN***	***FP***	***TP***	***Precision*** $\frac{TP}{TP+FP}$	***Sensitivity*** $\frac{TP}{TP+FN}$	***Specificity*** $\frac{TN}{TN+FP}$
1 F/M Dual BP	NN	DS3	366	837	128	231	0.6434	0.3869	0.8673
2 F/M Single BP	NN	DS3	365	835	130	232	0.6408	0.3886	0.8652
3 F Dual BP	NN	DS6	46	547	185	83	0.3097	0.6434	0.7472
4 M Dual BP	NN	DS4	65	307	181	147	0.4481	0.6933	0.6290
5 F Single BP	RG	DS6	194	555	38	74	0.6607	0.2761	0.9359
6 M Single BP	GB	DS2	192	315	57	136	0.7046	0.4146	0.8467

The cumulative lift chart (Fig. 34, train top, validate bottom) measures model effectiveness in predicting binary classification compared with random. The lift value indicates how many times better than random the model is. Observing both sets we see the gradient boosting and regression models track each other, suggesting they are robust, not overfitting the training set and able to generalize. The decision tree performs better at 5% depth in "*train*" than "*validate*", indicating an overfitting issue. Regression, neural network and gradient boosting show comparable performance on "*train*" while the decision tree starts to degrade after around 23% of data. Reviewing the validation set, the decision tree looks unreliable, predicitability decreasing with a low lift of 1.63 at 5% depth, picking up to 1.74 at 15% and then decaying markedly compared with other models which are more consistent. At 20% depth the neural network is best at 1.71 times better than random, then gradient boosting 1.68 and regression 1.67. The ratio of "*Alive*" (61.78%) to "*Dead*" (38.22%) is approximately 3 to 2. Our validation set contains 1562 observations. For the first 10%, or 156 observations, random suggests finding 60 "*Dead*". The performance of the neural network model is 1.93 times better than random at 10% depth, so the model finds 115 "*Dead*" (115 / random 60 = ~1.9, the neural network lift at 10%).

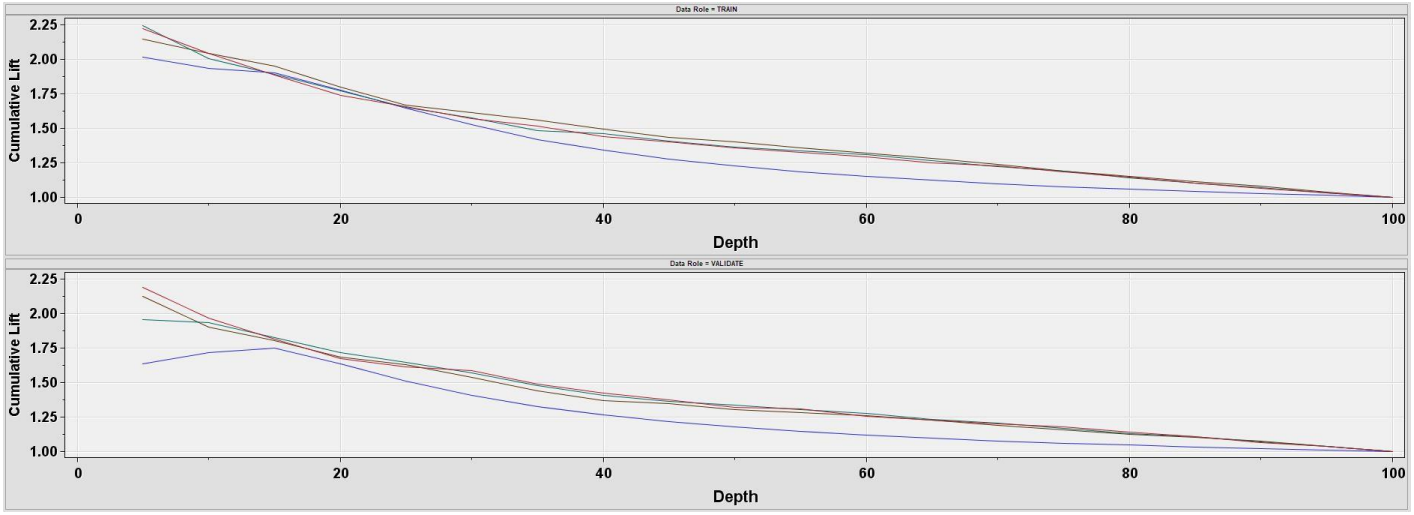

Fig. 34 Category 1 F/M dual B.P. cumulative lift winning model comparison. DT (DS6) blue, RG (DS3) red, NN (DS3) green, GB (DS2) brown

The worst performer in category 1, the decision tree, exhibits lift that drops to worse than random (1.0) after only 23% depth (Fig. 35). The neural network ranking 1st shows more stable lift decay.

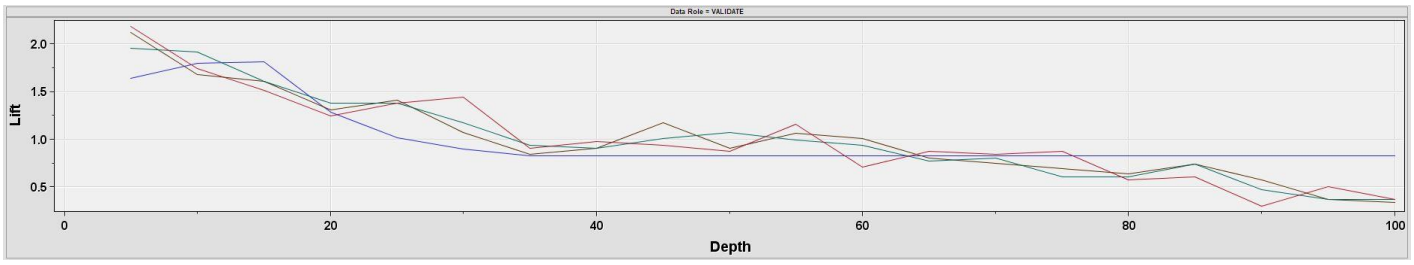

Fig. 35 Category 1 F/M dual B.P. lift (non-cumulative) winning model comparison. DT (DS6) blue, RG (DS3) red, NN (DS3) green, GB (DS2) brown

9. Classification Thresholds & Probability

In S.E.M. the probability threshold for positive binary classification is 0.5. When predicting "*Dead*" by C.V.D. an observation class probability less than 0.5 has an outcome of "*Alive*", otherwise "*Dead*". The number of true positives detected can be improved by adjusting the threshold cutoff point using the *Cutoff* node, say to 0.4, with property "*Cutoff User Input*". Probabilities above 0.4 become a positive classification with "*Status*" set to "*Dead*". Adjusting classification often comes at cost to other metrics like misclassification due to increased false positives for example, but in this real-world scenario it is the true positives we're most interested in. Ordering of observations by calculated probability in descending order can also help focus on top N risks. See *Appendix H* for more on threshold adjustment.

10. Final Conclusion & Recommendations

All analysis indicates higher blood pressure levels are the biggest contributing factors to mortality by C.V.D., this is well known. A blood pressure rate below 130/85 (systolic/diastolic) is recommended to reduce C.V.D. risk, informed by the distribution and cluster analysis, variable importance and classification rules used. This is further supported by "*evidence that keeping SBP between 120 and 139 mm Hg is optimal for both middle-aged and elderly adults with or without prior CVD*" [5]. Cluster analysis highlighted particularly at-risk groups, for example cluster 2 consisting 82% females recording above average "*Weight*" and "*Cholesterol*" and very high "*Diastolic*" and "*Systolic*". Efforts should be focused on specifically targeting these groups with health guidance including promotion of exercise and healthy diet through typical channels like their G.P., and dedicated channels such as lifestyle magazines and dedicated online resources. The author of [22] agrees with "*the potential benefits of a healthy diet, weight control, and regular exercise cannot be overemphasized. These lifestyle treatments have the potential to improve BP control and even reduce medication needs*" [22].

It is recommended to implement lifetime patient observation, improving data collection to include age, alcohol consumption, demographic, family medical history and other conditions like diabetes. This will have several benefits. First, a richer model will provide greater insight into patient wellbeing over time and help develop a better understanding of the relationship between C.V.D and contributing factors that might also be race or geographically influenced. For example, it is known that in elderly patients systolic rates actually decrease, putting them at risk, as [23] highlights with "*At very old age, however, systolic blood pressure eventually decreases again and it is this decline in systolic blood pressure that is associated with a worse prognosis*" [23]. Second, changes in patient health can be monitored, such as any decline in blood pressure, which can help gauge the effect of these health promotion campaigns. Third, constant monitoring with a richer dataset will highlight any change in current identified and newly developing risks and allow models to be continuously adapted and improved.

Best sensitivity to death by C.V.D. was realised on datasets split by gender and this approach is recommended going forward, although there is further work to be done in lowering misclassification rate with the male cohort. With regard to model selection a two-step approach is recommended. Classification should continue with neural networks using a dual B.P. component as they deliver best misclassification rate with lowest risk of classifying "*Death*" as "*Alive*", but harder to interpret. Further work should refine decision tree performance and rules used, including on SAS and trialling other platforms such as Python, incidentally where the author observed significant performance improvement with decision tree tuning [4, p. 10]. The decision tree produces results far easier to understand and share with health professionals like G.P.s and bodies including heart organisations, and easier to operationalise, including any porting even to Excel.

Further work to explore observations labelled "*Alive*" but classified "*Dead*" (false positives) is recommended to assess if participants exhibit risk-level high blood pressure and cholesterol rates who should receive urgent advice and preventative treatment.

Appendix A – Framingham Heart Dataset Structure

The structure table below provided to author of this study.

Variable name	Variable description
Status	Person's status (*Alive, Dead*)
Sex	Sex (*Female, Male*)
Height	Person's height in inches
Weight	Person's weight in kg
Diastolic	Person's blood pressure in the arteries when the heart rests between beats. A normal diastolic blood pressure measurement is 80 or below.
Systolic	Person's blood pressure in the arteries: when the heart beats, it contracts and pushes blood through the arteries to the rest of the body. A normal systolic blood pressure measurement is 120 or below.
Smoking	Number of cigarettes smoked per week
Death_age	Age at death in years
Cholesterol	Person's cholesterol measurement
Chol_status	Cholesterol status (*borderline, desirable, high*)
Bpressure_status	Blood pressure (*High, Normal, Optimal*)
Weight_status	Weight status (*Underweight, Normal, Overweight)*
Smoke_status	Smoking status (*Nonsmoker, Light, Moderate, Heavy, Very heavy)*

Appendix B – Blood Pressure and Weight Categories

BLOOD PRESSURE CATEGORY	SYSTOLIC mm Hg (upper number)		DIASTOLIC mm Hg (lower number)
NORMAL	LESS THAN 120	and	LESS THAN 80
ELEVATED	120 – 129	and	LESS THAN 80
HIGH BLOOD PRESSURE (HYPERTENSION) STAGE 1	130 – 139	or	80 – 89
HIGH BLOOD PRESSURE (HYPERTENSION) STAGE 2	140 OR HIGHER	or	90 OR HIGHER
HYPERTENSIVE CRISIS (consult your doctor immediately)	HIGHER THAN 180	and/or	HIGHER THAN 120

Fig. 36 American Heart Association blood pressure cateogries [8]

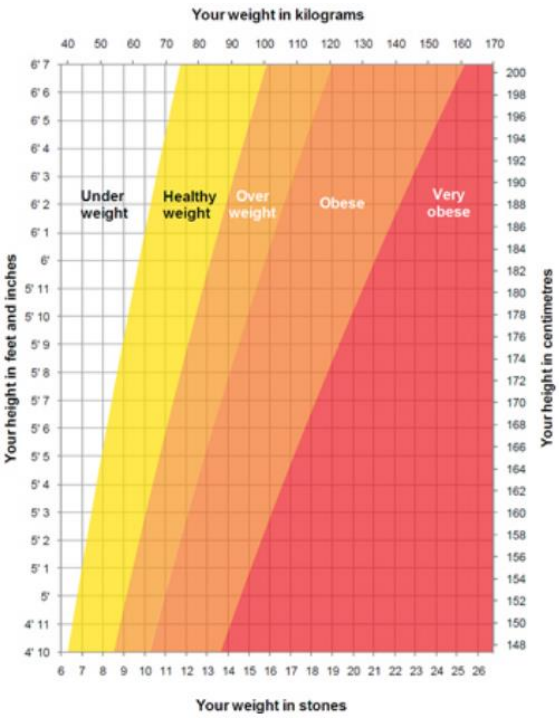

Fig. 37 NHS weight categories [9]. 1 stone = 6.35kg

Appendix C – Initial Data Analysis – Further Graphs

The plots from the Multiplot node (Fig. 38, 39) below show input levels on the horizontal axis and the frequency of the target class "*Status*" on the vertical axis. Observe "*Systolic*" by "*Status*" is positively skewed with greater distribution of "*Dead*" as systolic rates rise, while "*Height*" by "*Status*" is more normally distributed.

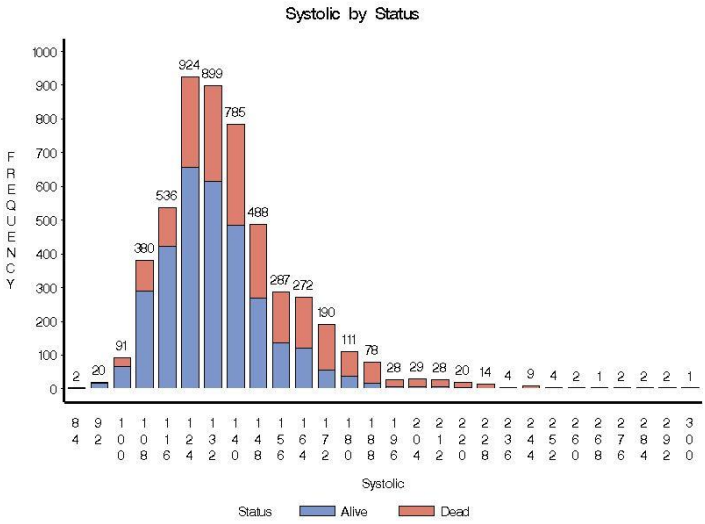

Fig. 38 Frequency count Systolic by Status

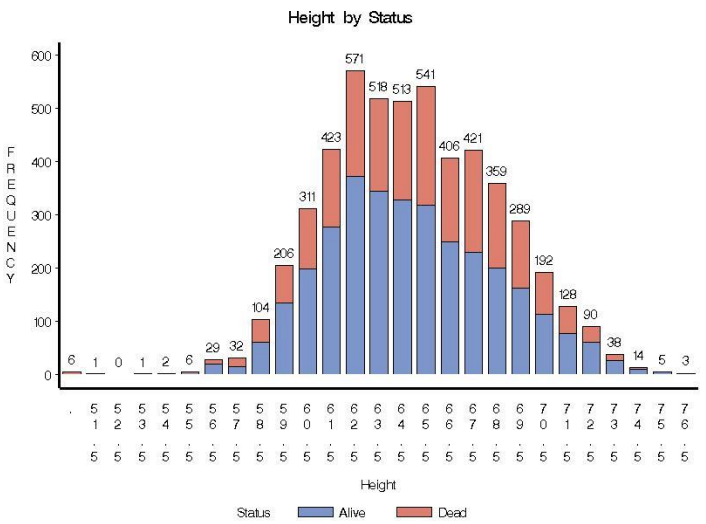

Fig. 39 Frequency count Height By Status

Fig. 40 shows broader inter-quartile distribution of "*Systolic*" for females, with a longer maximum whisker suggesting skewed distribution, with higher number of extreme outliers.

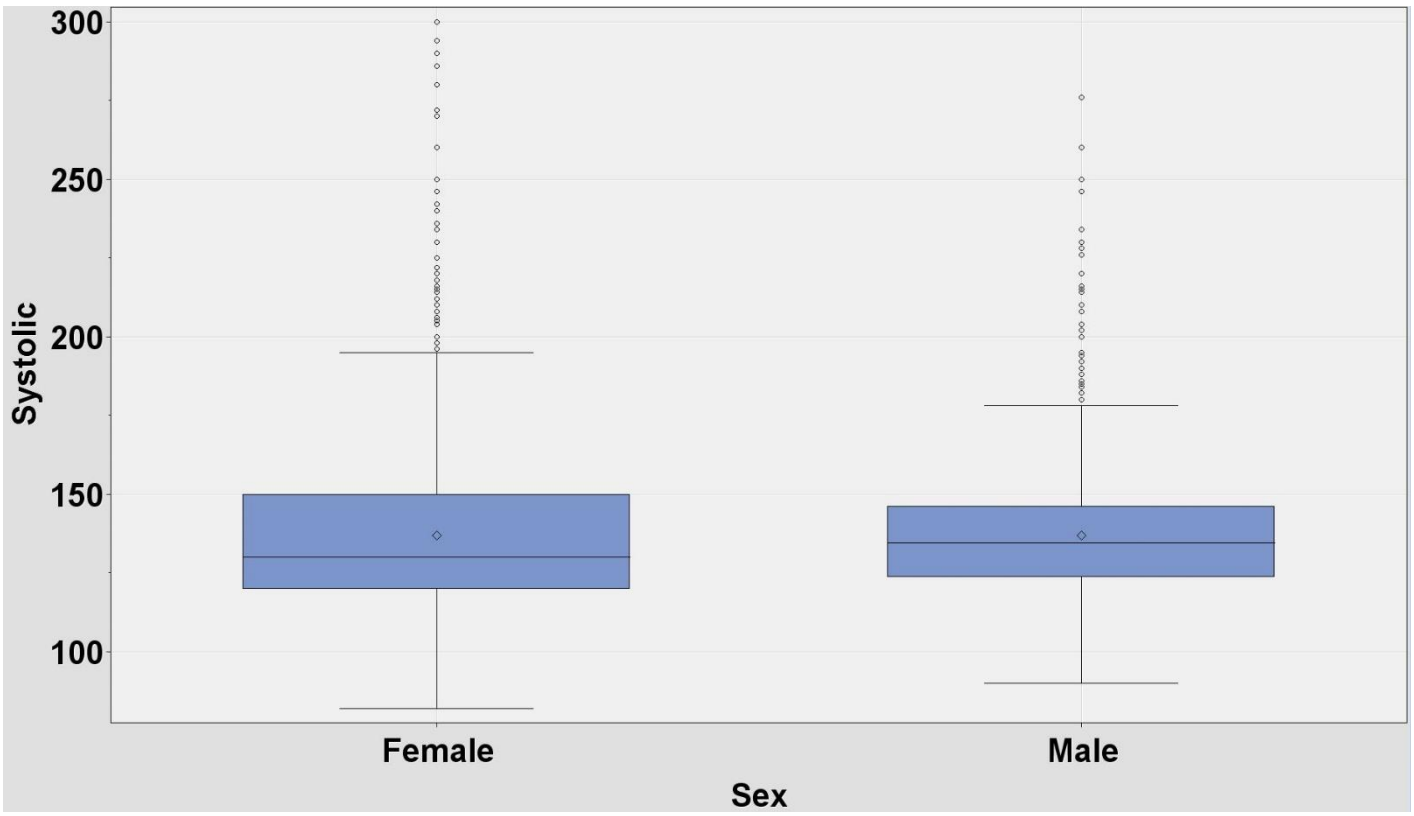

Fig. 40 Boxplot "*Systolic*" distribution by "*Sex*"

Appendix D – Chi-Squared Test – Input Variable Bins Below and Above Critical Value 3.841

Sorted by Chi-Square ascending, Fig. 41 shows input variable bins below critical value of 3.841 meaning null hypothesis "no statistical significance" can be accepted with 95% confidence. Fig. 42 shows input variables above critical value.

Target	Input	Target: Formatted Value	Input: Formatted Value	Frequency Count	Chi-Square ▲
Status	Smoking	Alive	10	158	0.001384
Status	Smoking	Dead	10	97	0.002236
Status	Smoking	Alive	35	30	0.002427
Status	Smoking	Dead	35	19	0.003923
Status	Cholesterol	Alive	379.2 -473.6	9	0.007673
Status	Cholesterol	Dead	379.2 -473.6	6	0.012402
Status	Weight_status	Alive	Underweight	113	0.012503
Status	Weight_status	Dead	Underweight	68	0.020208
Status	Smoking	Alive	55	1	0.044907
Status	Height	Alive	LOW-56.5	17	0.046786
Status	Smoking	Dead	55	1	0.072583
Status	Height	Dead	LOW-56.5	12	0.075619
Status	Weight	Alive	.	3	0.134722
Status	Weight_status	Alive	.	3	0.134722
Status	Smoke_status	Alive	Moderate (6-15)	363	0.144089
Status	Smoking	Alive	60	8	0.173138
Status	Weight	Dead	.	3	0.217748
Status	Weight_status	Dead	.	3	0.217748
Status	Smoking	Alive	15	205	0.225935
Status	Smoke_status	Dead	Moderate (6-15)	213	0.232888
Status	Smoking	Dead	60	7	0.279839
Status	Height	Alive	56.5 -61.5	631	0.339072
Status	Smoking	Dead	15	116	0.365173
Status	Cholesterol	Alive	190.4 -284.8	2145	0.471712
Status	Height	Dead	56.5 -61.5	367	0.548033
Status	Cholesterol	Dead	190.4 -284.8	1379	0.762415
Status	Smoking	Alive	45	5	1.143999
Status	Chol_status	Alive	Borderline	1186	1.147216
Status	Chol_status	Alive	.	83	1.26574
Status	Cholesterol	Alive	.	83	1.26574
Status	Height	Alive	71.5 - HIGH	113	1.347569
Status	Height	Alive	61.5 -66.5	1635	1.37283
Status	Systolic	Alive	125.6 -169.2	1752	1.424418
Status	Smoking	Alive	20	538	1.686019
Status	Smoke_status	Alive	.	16	1.750777
Status	Smoking	Alive	.	16	1.750777
Status	Smoking	Alive	1	81	1.794088
Status	Smoking	Dead	45	8	1.849014
Status	Chol_status	Dead	Borderline	675	1.854215
Status	Smoking	Alive	5	311	1.85612
Status	Smoking	Alive	25	65	1.934421
Status	Height	Alive	.	1	1.976446
Status	Weight	Alive	253.4 - HIGH	4	2.023355
Status	Chol_status	Dead	.	69	2.045781
Status	Cholesterol	Dead	.	69	2.045781
Status	Height	Dead	71.5 - HIGH	51	2.17804
Status	Height	Dead	61.5 -66.5	936	2.218869
Status	Systolic	Dead	125.6 -169.2	1166	2.302248
Status	Smoking	Dead	20	383	2.725068
Status	Smoke_status	Alive	Non-smoker	1610	2.729468
Status	Smoking	Alive	0	1610	2.729468
Status	Smoke_status	Dead	.	20	2.829733
Status	Smoking	Dead	.	20	2.829733
Status	Smoke_status	Alive	Heavy (16-25)	603	2.887333
Status	Smoking	Dead	1	32	2.899737
Status	Smoking	Alive	30	113	2.958193
Status	Smoking	Dead	5	155	2.999997
Status	Smoking	Alive	50	9	3.105096
Status	Smoking	Dead	25	60	3.126553
Status	Weight	Alive	206.8 -253.4	110	3.177272
Status	Height	Dead	.	5	3.194477
Status	Weight	Dead	253.4 - HIGH	9	3.270295
Status	Smoke_status	Alive	Light (1-5)	392	3.29048

Fig. 41 Chi-Squared below critical value 3.841

Status	Weight	Alive	LOW-113.6	234	3.88769
Status	Smoke_status	Dead	Non-smoker	891	4.411566
Status	Smoking	Dead	0	891	4.411566
Status	Smoke_status	Dead	Heavy (16-25)	443	4.666718
Status	Smoking	Dead	30	102	4.781248
Status	Diastolic	Alive	138 - HIGH	3	4.794954
Status	Weight_status	Alive	Overweight	2090	4.847584
Status	Cholesterol	Dead	473.6 - HIGH	5	4.992482
Status	Smoking	Dead	50	17	5.018683
Status	Weight	Dead	206.8 -253.4	101	5.13534
Status	Diastolic	Alive	72 -94	2127	5.194228
Status	Smoke_status	Dead	Light (1-5)	187	5.318315
Status	Height	Alive	66.5 -71.5	821	5.381759
Status	Weight	Dead	LOW-113.6	99	6.283569
Status	Smoking	Alive	40	68	6.853206
Status	Diastolic	Dead	138 - HIGH	13	7.749955
Status	Weight	Alive	113.6 -160.2	1943	7.807227
Status	Weight_status	Dead	Overweight	1460	7.835021
Status	Diastolic	Dead	72 -94	1150	8.395292
Status	Height	Dead	66.5 -71.5	620	8.698392
Status	Systolic	Dead	256.4 - HIGH	10	9.984963
Status	Smoking	Dead	40	83	11.07665
Status	Smoke_status	Alive	Very Heavy (> 25)	234	11.15538
Status	Weight_status	Alive	Normal	1012	11.58322
Status	Cholesterol	Alive	284.8 -379.2	246	12.3551
Status	Weight	Dead	113.6 -160.2	1009	12.61861
Status	Weight	Alive	160.2 -206.8	924	14.34258
Status	Smoke_status	Dead	Very Heavy (> 25)	237	18.03015
Status	Cholesterol	Alive	LOW-190.4	735	18.58657
Status	Weight_status	Dead	Normal	460	18.72164
Status	Chol_status	Alive	Desirable	998	19.47759
Status	Cholesterol	Dead	284.8 -379.2	252	19.96921
Status	Systolic	Alive	212.8 -256.4	8	20.44546
Status	Diastolic	Alive	116 -138	27	21.0877
Status	Chol_status	Alive	High	951	21.83684
Status	Diastolic	Alive	LOW-72	640	22.52428
Status	Bpressure_status	Alive	Normal	1497	22.63396
Status	Sex	Alive	Female	1977	23.01871
Status	Weight	Dead	160.2 -206.8	770	23.18152
Status	Sex	Alive	Male	1241	28.31025
Status	Cholesterol	Dead	LOW-190.4	280	30.04097
Status	Chol_status	Dead	Desirable	407	31.4811
Status	Systolic	Dead	212.8 -256.4	48	33.04545
Status	Diastolic	Dead	116 -138	76	34.08348
Status	Chol_status	Dead	High	840	35.2943
Status	Bpressure_status	Alive	Optimal	626	35.51181
Status	Diastolic	Dead	LOW-72	219	36.40538
Status	Bpressure_status	Dead	Normal	646	36.58267
Status	Sex	Dead	Female	896	37.20452
Status	Sex	Dead	Male	1095	45.7571
Status	Systolic	Alive	LOW-125.6	1342	47.57264
Status	Diastolic	Alive	94 -116	421	48.0943
Status	Bpressure_status	Dead	Optimal	173	57.39679
Status	Bpressure_status	Alive	High	1095	66.64078
Status	Systolic	Dead	LOW-125.6	458	76.89039
Status	Diastolic	Dead	94 -116	533	77.73352
Status	Systolic	Alive	169.2 -212.8	116	81.80537
Status	Bpressure_status	Dead	High	1172	107.7097
Status	Systolic	Dead	169.2 -212.8	309	132.2198

Fig. 42 Chi-Squared above critical value 3.841

Appendix E – Variable Relationship with Scatter – Further Graphs

Fig. 43 shows scatter generated from Graph Explore node of "*Systolic*" vs "*Cholesterol*", grouped by "*Status*". There is no linear relationship. Where there are outliers for both the outcome is mostly "*Dead*". Fig. 44 is "*Systolic*" vs "*Weight*", grouped by "*Status*", with an extremely weak linear relationship. "*Systolic*" again is the biggest influencer in outcome "*Dead*".

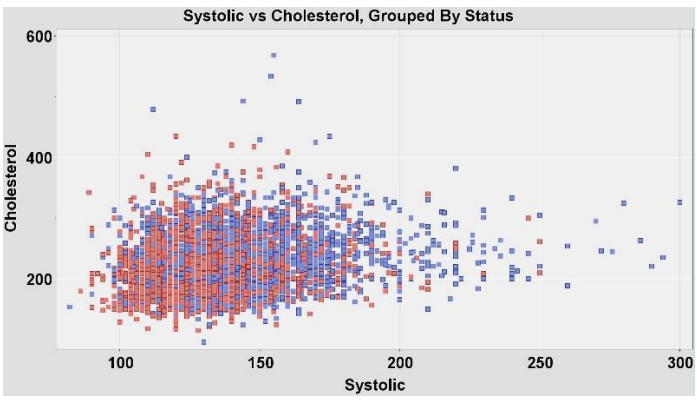

Fig. 43 Systolic vs Cholesterol, Grouped by Status (Blue=Dead, Red=Alive)

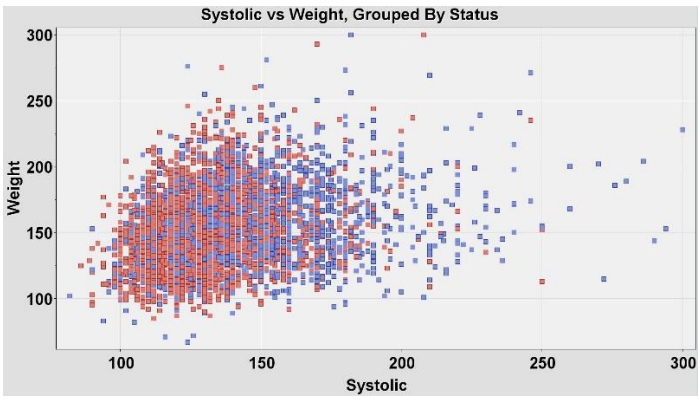

Fig. 44 Systolic vs Weight, Grouped by Status (Blue=Dead, Red=Alive)

Appendix F – Clustering – Further Graphs

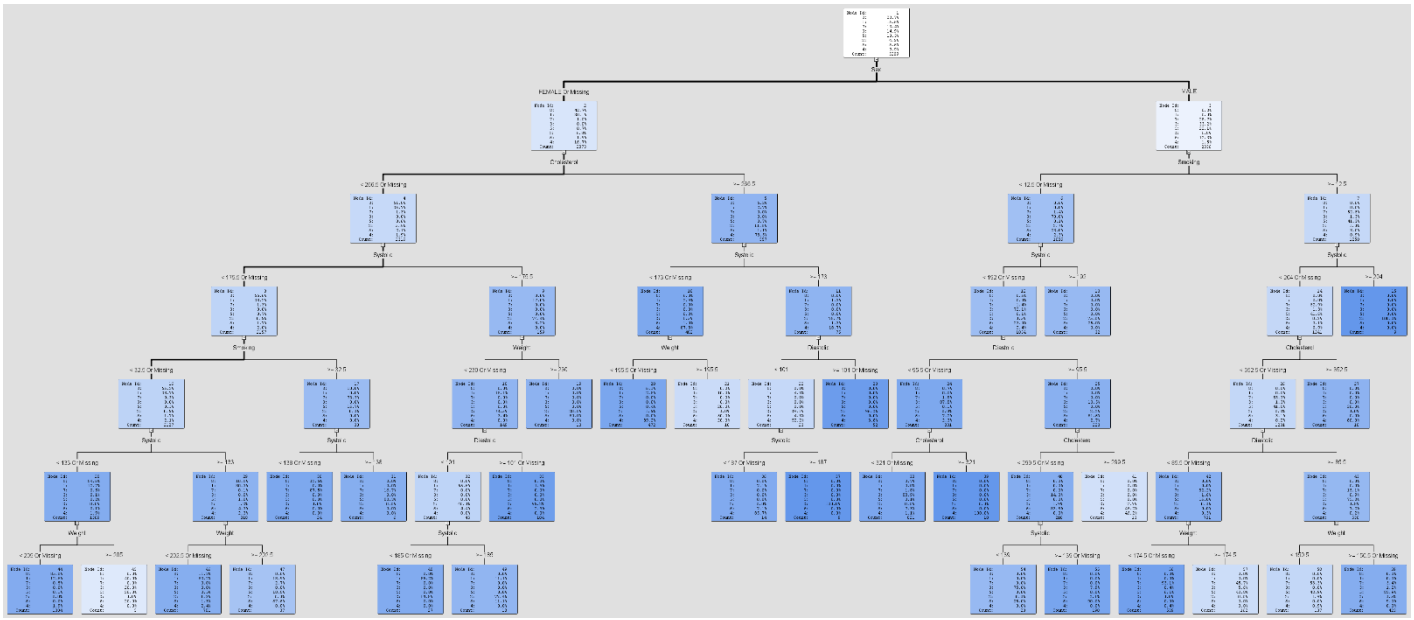

Fig. 45 Cluster 8 segment full tree

.. Property	Value
General	
Node ID	Clus
Imported Data	
Exported Data	
Notes	
Train	
Variables	
Cluster Variable Role	Segment
Internal Standardization	Standardization
Number of Clusters	
Specification Method	Automatic
Maximum Number of Clusters	10
Selection Criterion	
Clustering Method	Ward
Preliminary Maximum	150
Minimum	2
Final Maximum	8
CCC Cutoff	3
Encoding of Class Variables	
Ordinal Encoding	Rank
Nominal Encoding	GLM
Initial Cluster Seeds	
Seed Initialization Method	Default
Minimum Radius	0.0
Drift During Training	No
Training Options	
Use Defaults	Yes
Settings	
Missing Values	
Interval Variables	Mean
Nominal Variables	Default
Ordinal Variables	Default
Scoring Imputation Method	None
Score	
Cluster Variable Role	Segment
Hide Original Variables	Yes
Cluster Label Editor	

Fig. 46 Cluster 8 segment property settings

Name	Use
Bpressure_statu	No
Chol_status	No
Cholesterol	Yes
Death_age	No
Diastolic	Yes
Height	No
Sex	Yes
Smoke_status	No
Smoking	Yes
Systolic	Yes
Weight	Yes
Weight_status	No

Fig. 47 Cluster 8 segment variables used

The scatter plot in Fig. 48 shows "*Systolic*" vs "*Diastolic*" grouped by "*Status*", for cluster segment 8 with females only, grouped by common characteristics of lowest "*Weight*", "*Cholesterol*", "*Diastolic*", "*Systolic*" and "*Smoking*", show a survival rate of 82%, a significant 31%. Note low systolic and diastolic rates.

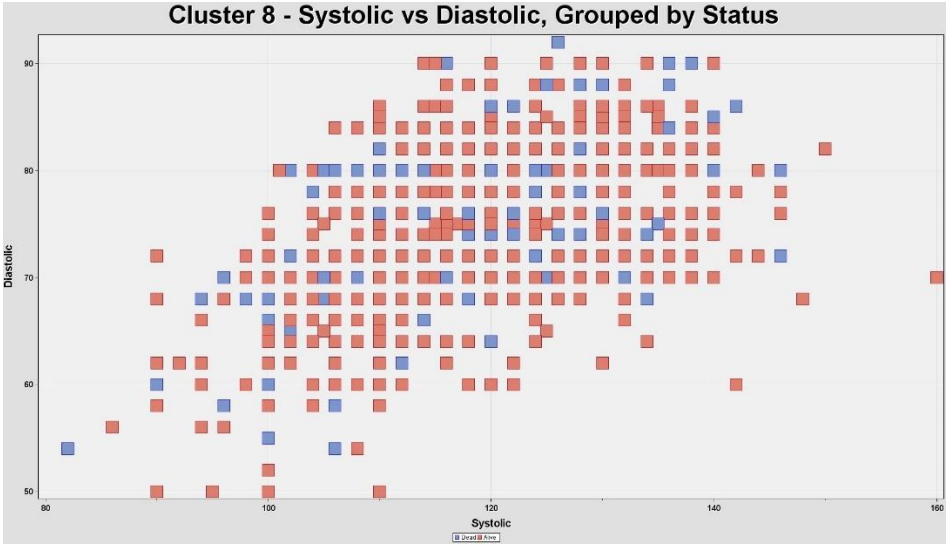

Fig. 48 Cluster segment 8 – Females only – Scatter plot Systolic vs Diastolic, grouped by Status

Mean Statistics

Clustering Criterion	Maximum Relative Change in Cluster Seeds	Improvement in Clustering Criterion	Segment Id	Frequency of Cluster	Root-Mean-Square Standard Deviation	Maximum Distance from Cluster Seed	Nearest Cluster	Distance to Nearest Cluster	Cholesterol	Diastolic	Smoking	Systolic	Weight	Sex=Female	Sex=Male
0.589672	0.025673	.	1	865	0.534753	3.037474	8	1.899481	218.4128	90.43699	3.185201	146.2301	151.2914	1	6.77E-15
0.589672	0.025673	.	2	237	0.859049	5.733126	6	2.827655	251.6561	115.3376	4.232488	202.0675	159.9831	0.822785	0.177215
0.589672	0.025673	.	3	777	0.537082	3.378159	7	1.971787	221.8694	82.2175	1.86688	130.2896	165.6526	0.001287	0.998713
0.589672	0.025673	.	4	517	0.646158	6.534656	1	1.990263	302.0174	85.50677	4.724113	138.8143	141.3909	0.930368	0.069632
0.589672	0.025673	.	5	536	0.664283	3.864789	7	1.877656	244.0547	92.25933	27.98507	144.334	178.2035	0.035448	0.964552
0.589672	0.025673	.	6	343	0.693638	4.682118	3	2.248667	221.4206	101.3994	3.147422	160.7434	196.1108	0.16035	0.83965
0.589672	0.025673	.	7	700	0.566721	3.440714	5	1.877656	211.5599	77.34	23.93833	124.4171	151.6158	0.041429	0.958571
0.589672	0.025673	.	8	1234	0.512695	3.117812	1	1.899481	204.7486	75.04943	6.728173	118.4692	127.9741	0.995138	0.004862

Fig. 49 Cluster 8 segment mean statistics

Appendix G – Classification Chart Winning Models

Fig. 50 shows classification chart for winning models, NN (DS3) best.

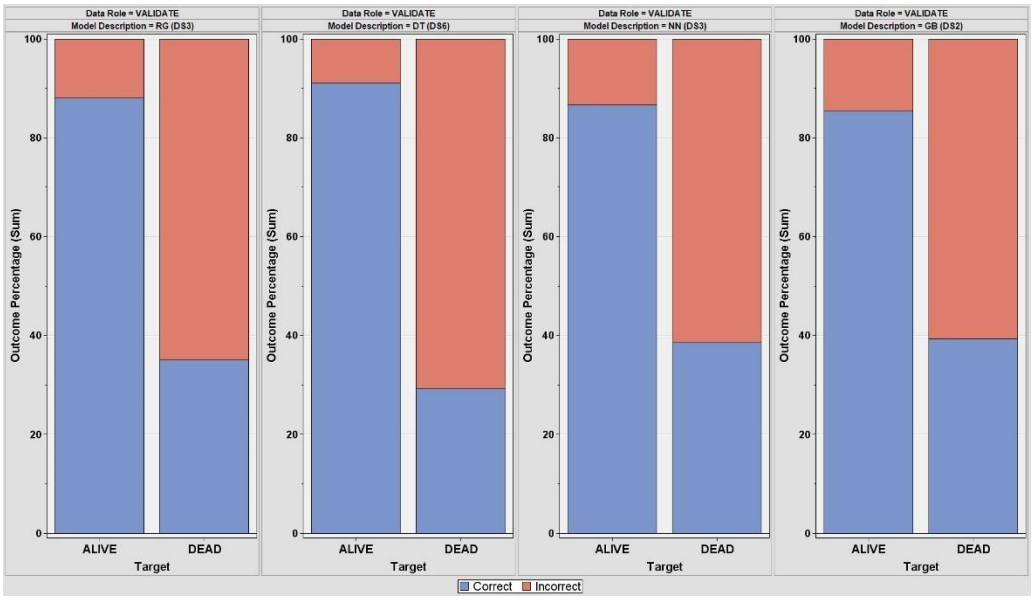

Fig. 50 Classification chart winning models

Fig. 51 Threshold adjustment and probability scoring solution in S.E.M.

Property	Value
General	
Node ID	CUT
Imported Data	
Exported Data	
Notes	
Train	
Variables	
Depth Scale %	1
Score	
Cutoff Method	User Input
Cutoff User Input	0.4
Status	
Create Time	08/05/18 13:27
Run ID	f0cb3963-f5e2-4a78-84ef-b58dde
Last Error	
Last Status	Complete
Last Run Time	27/05/18 21:19
Run Duration	0 Hr. 0 Min. 8.15 Sec.
Grid Host	
User-Added Node	No

Fig. 52 Cutoff Node threshold adjustment to 0.4

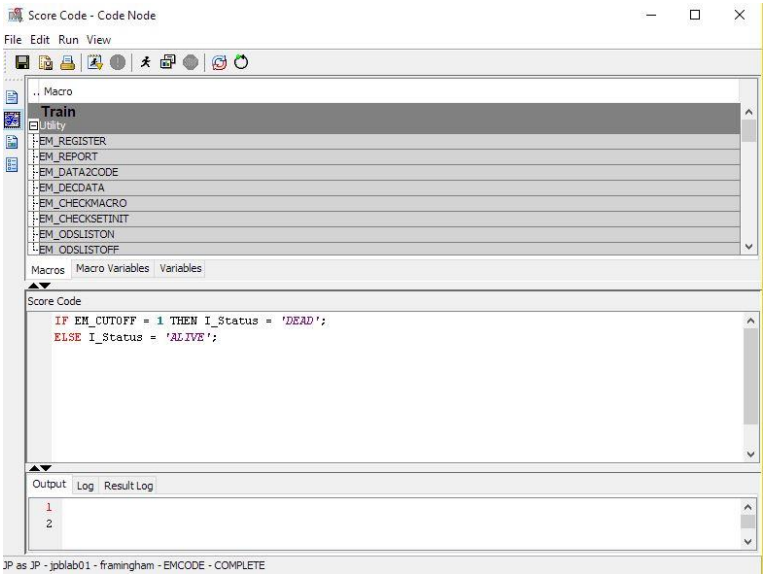

Fig. 53 SAS Code node score code to adjust classification

Note observation 51 in Fig. 54 with original target "*Status*" level "*Alive*" has probability calculated as 0.441557, and as the cutoff threshold has been adjusted to 0.4, this observation is classified as "*Dead*". In contrast observation 64 has calculated probability of "*Dead*" as 0.396818, therefore classified as "*Alive*" as under threshold. This output can be used to explore observations labelled "*Alive*" but classified "*Dead*" (false positives) to assess if participants exhibit risk-level indicators and need advice/treatment.

Sample Table

Observation Number	Status	Sex	Probability for level DEAD of Status	Target Variable: Status	Prediction for Status
51	Alive	Female	0.441557	ALIVE	DEAD
52	Alive	Female	0.064555	ALIVE	ALIVE
53	Dead	Female	0.263539	DEAD	ALIVE
55	Alive	Female	0.269308	ALIVE	ALIVE
57	Alive	Female	0.214575	ALIVE	ALIVE
58	Alive	Female	0.042744	ALIVE	ALIVE
60	Alive	Female	0.145665	ALIVE	ALIVE
64	Alive	Female	0.396818	ALIVE	ALIVE

Fig. 54 Output from Score node with actual/predicted classification and probability

References

[1] P. W. Wilson, R. B. D'Agostino, D. Levy, A. M. Belanger, H. Silbershatz and W. B. Kannel, "Prediction of Coronary Heart Disease Using Risk Factor Categories," *Science Volunteer,* vol. 97, no. 18, pp. 1837-1847, 1998.

[2] S. S. Franklin, M. G. Larson, S. A. Khan, N. D. Wong, M. Eric P. Leip, W. B. Kannel and D. Levy, "Does the Relation of Blood Pressure to Coronary Heart Disease Risk Change With Aging?," *US National Library of Medicine National Institutes of Health,* vol. 103, no. 9, pp. 1245-1249, 2001.

[3] D. Ettehad, C. A. Emdin, A. Kiran, S. G. Anderson, T. Callender, J. Emberson, J. Chalmers, A. Rodgers and K. Rahimi, "Blood pressure lowering for prevention of cardiovascular disease and death - a systematic review and meta-analysis," *The Lancet,* vol. 387, no. 10022, pp. 957-967, 2016.

[4] J.-P. Boyd and H. Lidgley, "Predicting Customer Behaviour With A Variety Of Classifiers," 2018.

[5] A. S. Koh, M. Talaei, A. Pand, R. Wang, J.-M. Yuan and W.-P. Koh, "Systolic blood pressure and cardiovascular mortality in middle-aged and elderly adults - The Singapore Chinese Health Study," *International Journal of Cardiology,* vol. 219, pp. 404-409, 2016.

[6] N. I. O. H. US National Library Of Medicine, "Health Effects of Light and Intermittent Smoking: A Review," [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2865193/.

[7] D. Wilson, J. Parsons and M. Wakefield, "The Health-Related Quality-of-Life of Never Smokers, Ex-smokers, and Light, Moderate, and Heavy Smokers," *Preventive Medicine,* vol. 29, no. 3, pp. 139-144, 1999.

[8] A. H. Association, "Understanding Blood Pressure Readings," [Online]. Available: http://www.heart.org/HEARTORG/Conditions/HighBloodPressure/KnowYourNumbers/Understanding-Blood-Pressure-Readings_UCM_301764_Article.jsp#.WvMD_oiFPmE.

[9] N. U. Kingdom, "Height/Weight Chart," [Online]. Available: https://www.nhs.uk/livewell/loseweight/pages/height-weight-chart.aspx.

[10] T. H. Institute, "Cholesterol," [Online]. Available: https://www.texasheart.org/heart-health/heart-information-center/topics/cholesterol/.

[11] K. Pearson, "On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling," *Philosophical Magazine,* vol. 5, no. 50, pp. 157-175, 1900.

[12] M. L. McHugh, "The Chi-square test of independence," *Biochemia Medica,* vol. 23, no. 2, pp. 143-149, 2013.

[13] MEDCALC, "Statistical Tables - Values of the Chi-squared distribution," [Online]. Available: https://www.medcalc.org/manual/chi-square-table.php.

[14] H. Cramer, Mathematical Methods of Statistics, Princeton University Press, 1946.

[15] L. M. Rea and R. A. Parker, Designing and Conducting Survey Research - A Comprehensive Guide, Jossey-Bass, 2014.

[16] D. R. Cox, "The Regression Analysis of Binary Sequences," *Journal of the Royal Statistical Society,* vol. 20, no. 2, pp. 215-242, 1958.

[17] J. H. Freidman, "Greedy Function Approximation: A Gradient Boosting Machine," in *IMS 1999 Reitz Lecture*, 1999.

[18] H. Bayat, "Boosting Algorithm," School of Computer Science and Informatics, De Montfort University, Leicester, UK, 2018.

[19] T. Hastie, "Boosting," 2003. [Online]. Available: http://web.stanford.edu/~hastie/TALKS/boost.

[20] R. J. Glynn, G. J. L'Italien, H. D. Sesso, E. A. Jackson and J. E. Buring, "Development of Predictive Models for Long-Term Cardiovascular Risk Associated With Systolic and Diastolic Blood Pressure," *Science Volunteer,* vol. 39, pp. 105-110, 2002.

[21] S. Raschka, "What is the intuition behind SoftMax function?," 2018. [Online]. Available: https://www.quora.com/What-is-the-intuition-behind-SoftMax-function.

[22] J. PA, O. S, C. BL and e. al, "2014 Evidence-Based Guideline for the Management of High Blood Pressure in Adults Report From the Panel Members Appointed to the Eighth Joint National Committee (JNC 8)," *JAMA,* vol. 311, no. 5, pp. 507-520, 2014.

[23] T. v. Bemmel, E. Holman, J. Gussekloo, G. Blauw, J. Bax and R. Westendorp, "Low blood pressure in the very old, a consequence of imminent heart failure - the Leiden 85-plus Study," *Journal of Human Hypertension,* vol. 23, pp. 27-32, 2009.

[24] I. Myrtveit, E. Stensrud and U. H. Olsson, "Analyzing Data Sets with Missing Data: An Empirical Evaluation of Imputation Methods and Likelihood-Based Methods," *IEEE TRANSACTIONS ON SOFTWARE ENGINEERING,* vol. 27, no. 11, pp. 999-1013, 2001.

[25] W. Akinfaderin, "Missing Data Conundrum: Exploration and Imputation Techniques," [Online]. Available: https://medium.com/ibm-data-science-experience/missing-data-conundrum-exploration-and-imputation-techniques-9f40abe0fd87.

[26] S. H. Walker and D. B. Duncan, "Estimation of the probability of an event as a function of several independent variables," *Biometrika,* vol. 54, no. 1, pp. 167-179, 1967.

[27] J. V. Hulse, T. M. Khoshgoftaar and A. Napolitano, "Experimental Perspectives on Learning from Imbalanced Data," in *Proceedings of the 24th international conference on Machine learning*, Corvalis, Oregon, USA, 2007.