**Data Mining**

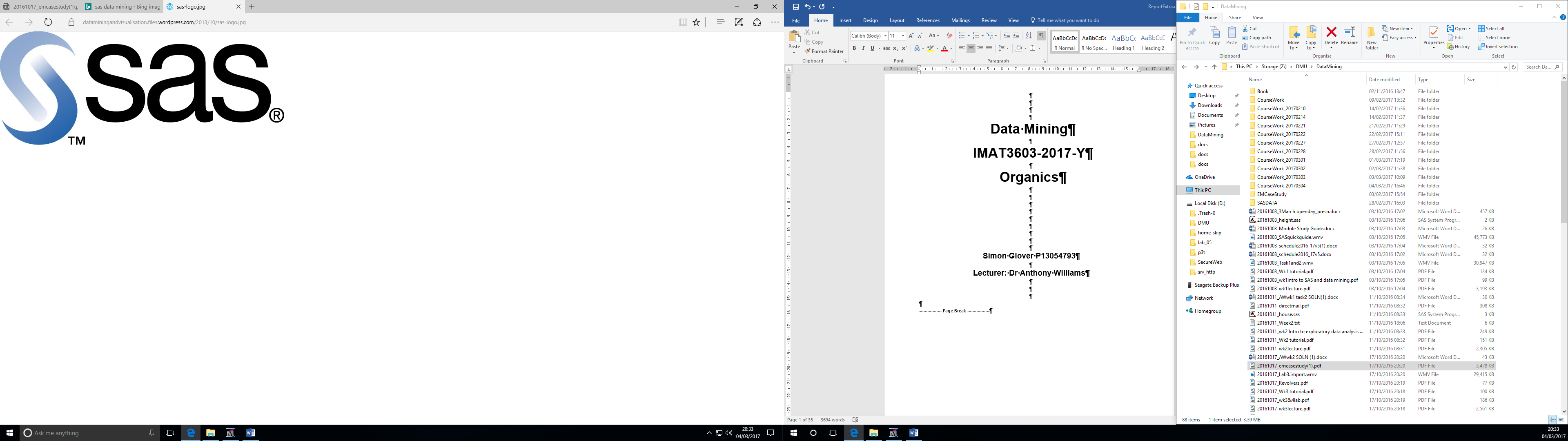
**IMAT5238**

**P2586104**

**20th May 2020**

**Framingham Heart Study**

(Total word Count: 2070)



**Lecturer: Dr Anthony Williams**

# ABSTRACT

An analysis to identify the risk factors that contribute to cardiovascular disease CVD in the Framingham Heart Study Data using predictive models. The analysis uses directed data mining techniques such as decision trees, neural networks, logistic regression and gradient boosting to predict the risk factors. The predictability, stability and accuracy of the model are evaluated using lift charts, %response -scope and diagnostic charts, respectively.

The best model for the analysis was the imputed backward selection logistic regression model with 43% true positives, 76% and predictability of 2.14 times better than random at 10% depth. The model predicts Systolic blood pressure, sex (female), smoking, weight, cholesterol, height and weight status (overweight) as the contributing risk factors.

Contents

[Abstract 2](#_Toc40835656)

[Introduction 4](#_Toc40835657)

[Business Problem 4](#_Toc40835658)

[Data Mining Representation 4](#_Toc40835659)

[Methodology – Data Mining Techniques 4](#_Toc40835660)

[Exploratory Data Analysis 4](#_Toc40835661)

[Data Partition 5](#_Toc40835662)

[Data Modelling 5](#_Toc40835663)

[Logistic Regression 5](#_Toc40835664)

[Modification Nodes 6](#_Toc40835665)

[Selection Methods 6](#_Toc40835666)

[Model Assessment 6](#_Toc40835667)

[Regression Equation 7](#_Toc40835668)

[Neural Networks 7](#_Toc40835669)

[Model Development 7](#_Toc40835670)

[Model Assessment 7](#_Toc40835671)

[Neural Network Architecture 7](#_Toc40835672)

[Decision Trees 8](#_Toc40835673)

[Model Development 8](#_Toc40835674)

[Model Assessment 8](#_Toc40835675)

[Critical Path 8](#_Toc40835676)

[Target Path Of Interest 8](#_Toc40835677)

[Overfitting And Limitations 9](#_Toc40835678)

[Research Node – Gradient Boosting 9](#_Toc40835679)

[Theory 9](#_Toc40835680)

[Settings 9](#_Toc40835681)

[Analysis 10](#_Toc40835682)

[Model Performance 10](#_Toc40835683)

[Analysing The Best Model 10](#_Toc40835684)

[Conclusion 11](#_Toc40835685)

[Recommendations 12](#_Toc40835686)

[References And Bibliography 13](#_Toc40835687)

[Appendix 14](#_Toc40835688)

[My Reflections On The Process – What Did I Learn From This Exercise? 16](#_Toc40835689)

# Introduction

The Framingham Heart Study was conducted in 1948 at the National Heart Institute to understand the cause of the steady increase in heart disease-related deaths and severe illnesses by identifying the risk factors that contribute to Cardiovascular disease (CVD). The researchers recruited participants (from Framingham, Massachusetts) comprised of 5,209 men and women between the ages of 30 and 62 over a long duration.

## Business Problem

Build a predictive model that helps The National Health Serve predict risk factors that contribute to Cardiovascular Disease (CVD) using the Framingham Data.

## Data Mining Representation

The business problem is translated into a data mining process. Directed data mining style is applied to explore and analyse the data set using predictive modelling to obtain the best fit. CVD leads to death and illness; hence; death is used as an indicator to identify the effects of the parameters on CVD. Therefore, the target (response) variable for this analysis is status. This analysis is a Binary classification, Estimation, Prediction and Classification problem (Layoff and Berry, 2013).

## Methodology – data mining Techniques

During the data mining process, a comparative methodological study is conducted using different Machine Learning algorithms – Logistic regression, neural networks, decision tree and gradient boosting. Using SEMMA (Sample, Explore, Modify, Model and Assess) steps in the directed data mining methodology. The universal criterium for selecting the appropriate set of hyperparameters was the evaluation of 10% depth of cumulative lift charts, % response scope and classification charts (see *Appendix H*). SAS ® Enterprise Miner Workstation™14.1. software is used for this analysis.

## Exploratory Data Analysis

The data mining roles are input for all variables except Status and Death age – SAS rejects variables with >50% missing values. The ratio of alive to dead for the target variable is 61.78% to 38.22% -approximately 2:1(see *Appendix A*)*.*

|  |  |  |
| --- | --- | --- |
| Variable Name | Role | Level |
| Status | Target | Binary |
| Bpressure\_Status | Input | Ordinal |
| Chol\_Status | Input | Ordinal |
| Cholesterol | Input | Interval |
| Death\_age | Rejected (missing values >50%) | Interval |
| Diastolic | Input | Interval |
| Height | Input | Interval |
| Sex | Input | Binary |
| Smoke\_Status | Input | Ordinal |
| Smoking | Input | Interval |
| Systolic | Input | Interval |
| Weight | Input | Interval |
| Weight\_Status | Input | Ordinal |

Table 1. Summary of observations, measurement levels and data mining roles

|  |  |
| --- | --- |
| Variable | The comments on the shape of the distribution in *italics* and descriptive statistics in standard text |
| Cholesterol | *Approximately 42% of participants have cholesterol 190 and 230.* There is little variability in the data with standard deviation as 44.99% with a mean value of 227.4. 2.9% of the data is missing. |
| Diastolic | *Approximately 63% had diastolic blood pressure between 72 and 94. The normal level is 80, so most have a normal diastolic level with outliers from 127 to 160).* The mean value of 85 also suggests that most participants had a normal diastolic level. |
| Height | *Approximately 70% are between 61 and 69 inches tall. The mean height is 64.8.* |
| Systolic | *Approximately 62% had systolic blood pressure between 103.8 and 147.4. The normal level is 120, so most have a normal diastolic level with outliers from 191 even 300.* *The distribution is positively skewed.* The mean value of 136.90 suggests variability in the data. |
| Weight | *Approximately 78% weigh between 113kg and 183kg. There are outlier values below 254kg till 300kg.* The mean weight is 153. |
| BPressure | *Has three categories, 41% had high bpressure status* |
| Chol\_Status | *Has three categories, 39% had borderline cholesterol status.2.9% of data is missing* |
| Smoke\_Status | *49% are non-smokers. Has five categories* |
| Weight\_Status | *68% of participants are overweight. Has three categories* |

Table 2. Descriptive Analysis

STat-Explore

From the chi-squared plot, Systolic, Bpressure\_Status, Diastolic and Sex have the strongest effects on the response variable.

Graph Explore

Systolic and Diastolic are colinear, i.e. as systolic increases, so does diastolic increase for both target categories. (see *Appendix C*).

## Data PArtition

Data is partitioned using simple random sampling method as training, validation and test set at 40%, 30% and 30% respectively using DMU student id number - 86104 in the random seed generator (see *Appendix B*). The initial model (seen data) is built with the training set; the validation set evaluates the training set while the test set is to be applied on unseen data to assess the effectiveness of the model (see *Appendix B).*

# Data Modelling

## Logistic Regression

The Regression Model estimates the relationship between the input variables and the target variable. A logistic regression model is used in this binary classification task.

### Modification Nodes

To improve the performance of the data model, the following techniques are applied to modify the distribution of the input variables—from the descriptive statistics recall; the skewed distributions, outliers, and missing values.

Filter

To remove extreme values and outliers from the distribution.

Replacement Cap

To cap the missing values with of 3 Standard Deviation from the mean of the data without losing observations.

Transformation

A mathematical function Log10(Variable + 1) is used to eliminate skewness in an attempt to make it normal.

Impute

This method replaces the missing values in the class variables with the highest count and the Median value for interval variables. The original variables are replaced with the imputed variables.

### Selection Methods

None

All candidate effect on target model is included in the final model.

Forward

This method starts with a null model and continually adds input variables to the model according to the significance level of the most correlation to the target variable.

Backward

This method starts with the full model and continually eliminates the input variables from the model according to the significance level of the least correlation to the target variable.

Stepwise

Like the Forward Selection method, this method begins with a null model and adds the input variable according to the most significant. However, it eliminated the previously added variables and continues until no further input variables can be added or eliminated.

### Model Assessment

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Model | Non-Cumulative % Response Scope | Cumulative Lift 10% Depth | True Negative | False  Negative | True Positive | False Positive | Missclasification Rate |
| Full Regression | 70.51 | 2.10 | 818 | 335 | 261 | 147 | 0.31715 |
| Forward Selection | 70.51 | 2.09 | 827 | 343 | 253 | 138 | 0.30814 |
| Backward Selection | **76.92** | **2.14** | **826** | **340** | **256** | **139** | **0.30685** |
| Stepwise Selection | 70.51 | 2.09 | 827 | 343 | 253 | 138 | 0.30814 |

Table 3. Regression Performance

### Regression equation

**Logit P** = -1.6578+0.00365(IMP\_Chloesterol) + 0.0313(Systolic)+).000665(IMP\_Weight)+0.0205(IMP\_Smoking)-0.0755(IMP\_Height)-0.3323(IMP\_Weight\_Status = Overweight)-0.3880(Sex=Female).

## Neural Networks

Neural network behaves like the regression model being it takes non-linear functions of linear combinations of input variables by creating connections according to the weight of the variables.

### Model Development

The model uses 2,3 and 4 hidden nodes with one hidden layer for each node. Iteration value for misclassification rate is very crucial in selecting the best model.

### Model ASSessment

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Model  (No. of Hidden Nodes) | Non-Cum % Resp Scope | Cumulative Lift 10% depth | True Negative | False  Negative | True Positive | False Positive | Misclassification Rate |
| 5 | 74.36 | 2.05 | 804 | 353 | 243 | 161 | 0.32928 |
| 4 | 65.38 | 1.97 | 819 | 351 | 245 | 146 | 0.31839 |
| 3 | **69.23** | **2.05** | **824** | **355** | **241** | **141** | **0.31775** |
| Direct 3 | 70.51 | 2.00 | 814 | 339 | 263 | 181 | 0.32928 |

Table 4. Model Assessment for Neural Networks

MODEL WEIGHTS

BP\_status = Normal in hidden node 1, BP\_status = Normal in Hidden 2, Sex= Female have a positive effect on the probability of risk factor associated with CVD.

Systolic, Sex= Female, Diastolic have a negative effect on the probability of risk factors associated with CVD. (see *Appendix G*).

### Neural network architecture

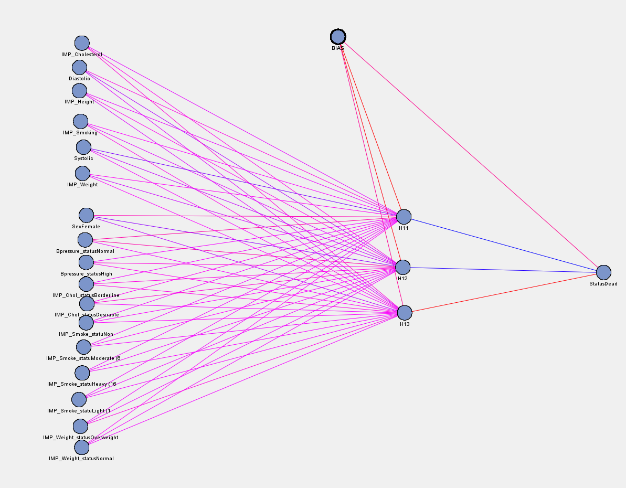


Figure . NN Architecture

## Decision Trees

The decision tree creates segments according to the hierarchy of the data. The first segment is the root node of the tree, which is partitioned into two or more segments by applying rules based on the value of the target variable. The partitioning continues until there is no further segment in the hierarchy. The hierarchy is the tree, and each segment is a node.

### model Development

The decision tree is built on the best splitting criterion for binary target variables which is the Chi-square test at 0.02 significance level. Four models were implemented; Default, three-way, interactive and pruned (don’t split node less than 50 cases and don’t create node with less than 25 cases) (see *Appendix E*).

### model Assessment

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Model | Non-Cum % Resp Scope | Cumulative Lift 10% depth | True Negative | False Negative | True Positive | False Positive | Misclassification Rate |
| Default | 73.24 | 1.93 | 892 | 411 | 185 | 73 | 0.31006 |
| Pruned | **72.99** | **2.01** | **902** | **418** | **178** | **63** | **0.30814** |
| Interactive | 73.24 | 1.93 | 892 | 411 | 185 | 73 | 0.31006 |

Table . DT Performance

### Critical path

When Systolic is <147 for 1076 participants of 1542 in the training data and 799 participants of 1146, then the participants are Alive.

### Target path of interest

When Systolic is >=147 and Systolic >=169

If Systolic is >147 for 329 of 540 for training data and 212 of 415 for validation data and Systolic is >169 for 107 participants out of 146 and the participant Sex = Male then the participant is dead.

If Systolic is >169 and sex = Male, then 107 out of 146 participants are dead.

If Systolic is >=169 and Sex = Female or Missing and Cholestrol >=2115 then dead.

### Overfitting and limitations

After pruning the leaves, 6 nodes are identified in the model. Misclassification rate increases with an increasing number of nodes. When this happens, the model becomes unstable and begins to fit noise in the data, which leads to overfitting. Tree Stability is determined by the lowest misclassification rate on the validation data. The limitation is this model may be unreliable on unseen data for a scope of 72%.

## Research node – Gradient Boosting

### Theory

In the gradient boosting procedure, partitioning algorithm uses best split for the data. The model resamples the input variables several times to produce output which sums to create a weighted average for each resampling of the data (Patel A., 2020). The model creates multiple decision trees that produce a singles predictive model. Like decision trees, this model is not affected by missing values. The node is expected to improve the poor data fits from the decision tree model. (SAS, 2018).

### Settings

The maximum tree split is set to 2, and maximum depth 5 - number of successive hierarchical splits.

A screenshot of a cell phone

Description automatically generated

Figure . Settings for Gradient Boosting

### Analysis

Iteration rapidly decays for both training and validation. Systolic, Diastolic, BP\_Status and Weight have the highest validation importance.

### Model Performance

The decision tree had superior performance to the gradient boosting, being that the misclassification rate is 0.31006 (see *Appendix F*).

## ANALYSING the best model

The imputed backward regression model is the best model; the model predicts the highest lift value at a 10% depth as 2.14 times better than random, it has the highest number of true positives and % response has the most extended scope.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Model | Non- Cumulative % Res. Scope | Cumulative Lift @10% depth | True -% | False - % | True + % | False+ % |
| Decision Trees | 72.99 | 2.01 | 93.5% | 6.5% | 29% | 70% |
| Neural Networks | 69.23 | 2.05 | 85.4% | 14.6% | 40.4% | 59.6% |
| Backwards Regression | **76.92** | **2.14** | 85.6% | 14.4% | **43%** | 57% |

A screenshot of a social media post

Description automatically generated

Figure . Cumulative lift of the best performing model

A screenshot of a map

Description automatically generated

Figure Non-cumulative lift of the best performing models

A picture containing screenshot

Description automatically generated

Figure . Diagnostic Charts

# Conclusion

In all the directed data mining models, systolic, sex, and cholesterol had significant effects. Diastolic and BP status had weights in the neural network model. The scope of the decision tree rapidly decreases in the training set. The imputed backward regression model is the champion model for this task; the validation set the highest accuracy and good stability.

From Logit P-probability of risk factor contributing to CVD, the significance of effects can be evaluated according to the p-value of parameters estimates. In conclusion, the most critical risk factors predicted that contribute to CVD are Systolic Blood Pressure, Sex = Female, Smoking, Cholesterol, Height, Weight and Weight Status = Overweight.

# Recommendations

1. Apply the prediction model to actual cardiovascular participants to measure the exact degree of a risk factor to the predicted factors and improve actions towards the prevention of death and severe illness.
2. Conduct further analysis using other machine algorithms to tackle colinear variables and the poor performance of the regression model.

# References and Bibliography

Linoff, G. and Berry, M., 2013. Data Mining Techniques. Hoboken, N.J.: Wiley.

SAS Institute Inc. 2018. *Data Mining Using SAS® Enterprise MinerTM: A Case*

*Study Approach*, *Fourth Edition.* Cary, NC: SAS Institute Inc.

Sarma, Katta-muri S., PhD, 2017. *Predictive Modeling with SAS® Enterprise Miner™: Practical Solutions for Business Applications, Third Edition.* Cary, NC: SAS Institute Inc.

SAS Institute Inc. 2018. *Getting Started with SAS® Enterprise Miner™ 15.1.* Cary, NC: SAS Institute Inc.

other Sources

Patel, Ahmed A. *Assessment 3: Boosting,* PowerPoint slides. *IMAT5238: Data Mining, Techniques and applications.* Available at <https://vle.dmu.ac.uk/> (Accessed 10th May 2020).

## Appendix

APPENDIX A: Data Mining Roles

A screenshot of a social media post

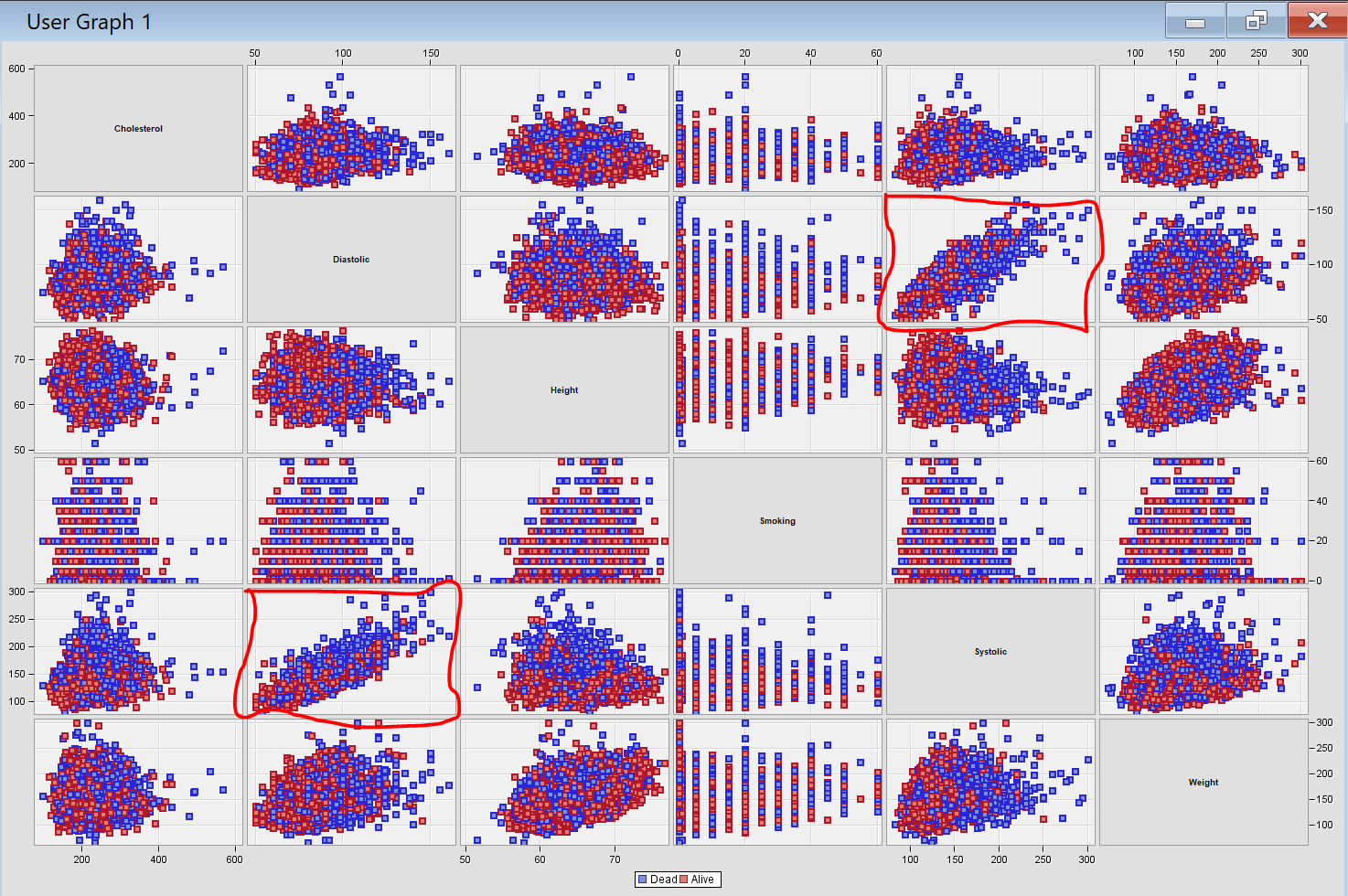
Description automatically generated

APPENDIX B: Data PArtition

A screenshot of a cell phone

Description automatically generated

Appendix C: Correlation of variables grouped by Status



Detail 1. Systolic and Diastolic are colinear

appendix D: lift and Classification chart – Regression

A picture containing screenshot, cabinet

Description automatically generated

Appendix E: Best Decision Tree

A screenshot of a social media post

Description automatically generated

Appendix F: Model ASSESSMENT Decision Trees and Gradient BoostingA picture containing screenshot

Description automatically generated

Appendix G: Best nEural Network.

A screenshot of a computer

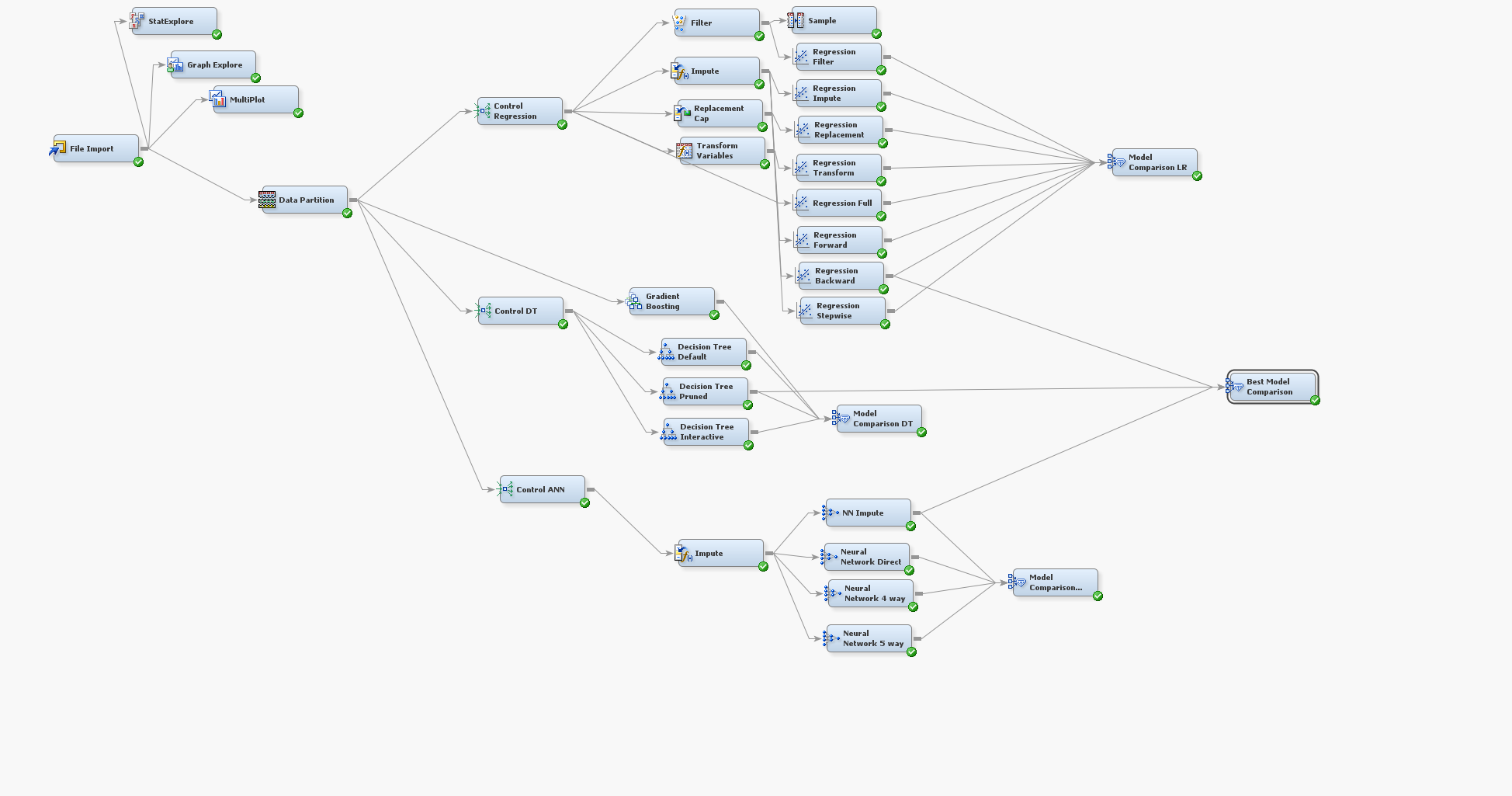
Description automatically generated

APPENDIX H: Neural network performance

A screenshot of a computer

Description automatically generated

appendix I: workFLOw Diagram



## My Reflections on the process – What did I learn from this exercise?

Data mining was by far the most challenging module for me this semester, from my first day in class, I was certain I would require more learning hours to under the concepts and statistical terminologies. The best part of this module is the wealth of resources available to the student and the pathed learning Anthony established, making it easy to follow at my preferred pace.

My journey with data mining and machine learning had only just begun, I can firmly say that at this point I have received a proper beginners level training and I am ready to advance with more materials and courses.

After completing this module, one thing is certain – I am a novice, and I need to get acquainted with statistics and machine learning.

Overall, the three most exciting things about this course are

* Time will never be enough; Anthony was right about the minimum number of hours required to understand the module properly.
* The availability of materials means you can ask relevant questions early.
* The discussion boards provided answers to multiple problems and created familiarity amongst student.