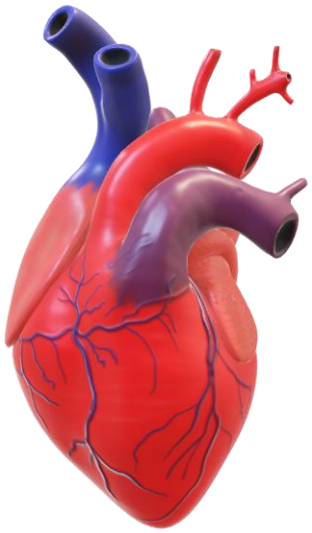
HEART FAILURE CLINICAL PREDICTION



NAME: THINES KUMAR NADARAJA
STUDENT ID: TP063097
ADVANCED BUSINESS ANALYTICS AND VISUALIZATION
LECTURER: DR PREETHI SUBRAMANIAN

Contents

Heart Failure Clinical Prediction

Introduction

Data published by the World Health Organization (WHO) in 2018 shows that Cardiovascular Diseases (CVD) are a contributing factor in causes of death, on a global scale (Lanzer et al., 2020). Big Data Analytics in Healthcare is pivotal in ensuring that CVD analysis and failure prediction can aid cardiologists and doctors alike (Sammani et al., 2019), with the appropriate treatment for patients (Kim, 2021). Big Data tools such as Apache Spark, is used for Predictive and Prescriptive Analytics in Healthcare (Thammasudjarit et al., 2018). Apache Spark is useful in healthcare organizations as it provides an open-source framework (Rammal & Z., 2018), which complements the Hadoop Distributed File System (HDFS) (Lanzer et al., 2020). Running HDFS along with Machine Learning (ML) algorithms, will certainly enable Big Data Analytics in Healthcare in prediction of CVD (Alexander & Wang, 2017).

Aims and Project Methodology

The aim of this project is to apply Big Data Analytics in Heart Failure Prediction, which will in turn, provide prevention of future diseases, based on factors that cause CVD. Procuring the dataset from Kaggle ("Heart Failure Prediction", 2021) containing 300 patients which have been previously affected with heart failure. Factors contributing to heart failure such as Anaemia, Level of CPK enzyme in blood, Diabetes, Blood Ejection Fraction, Hypertension, Blood Platelets, Creatinine Serum in Blood, Sodium Serum in Blood, Gender, Cigarette Smoking, Follow-Up Period and Death Events are analysed for this dataset. This dataset is then cleaned, using pre-processing techniques. Once dataset has been pre-processed, data analytics is then applied. Finally, using various data visualization techniques, a report that contains the analysis of the dataset is produced. Applying various Machine Learning modelling techniques on the dataset would provide a solution for Heart Failure Prediction, that would be deemed useful for the healthcare industry. The software used for this project are Tableau and SAS-Miner Studio.

Project Scope

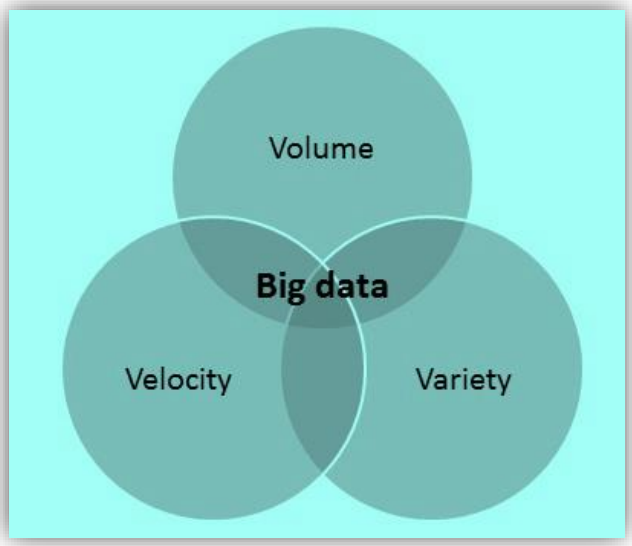
The scope of this project is to provide a solution for the healthcare industry using Big Data Analytics. Big Data Analytics ensure that knowledge procured from data in the healthcare industry, can be used to affect change and disrupt said industry (Roy et al., 2020). Big Data Analytics consist of 3 components, known as the 3 Vs of Big Data (Seward, 2021):

1) Volume – Volume of data is high, which would require analytics to be performed, to process and provide analysis of data.

2) Velocity – Data moves at a quick rate, and with real-time, processing such data is important for analytics.

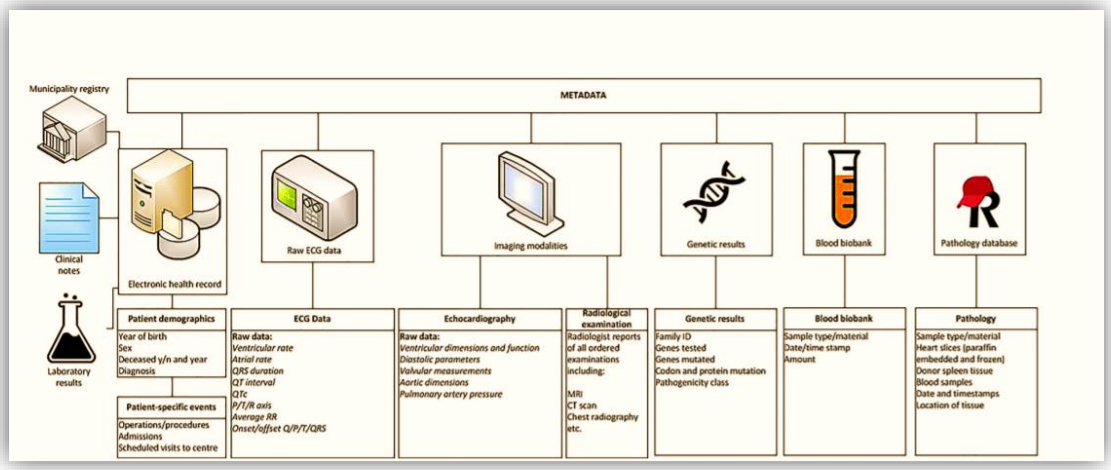
3) Variety – Data comes in different forms; structured, unstructured, and semi-structured data. This would require pre-processing of data, which is a key component of Big Data Analytics.

The 3 Vs of Big Data is shown in the figure below:

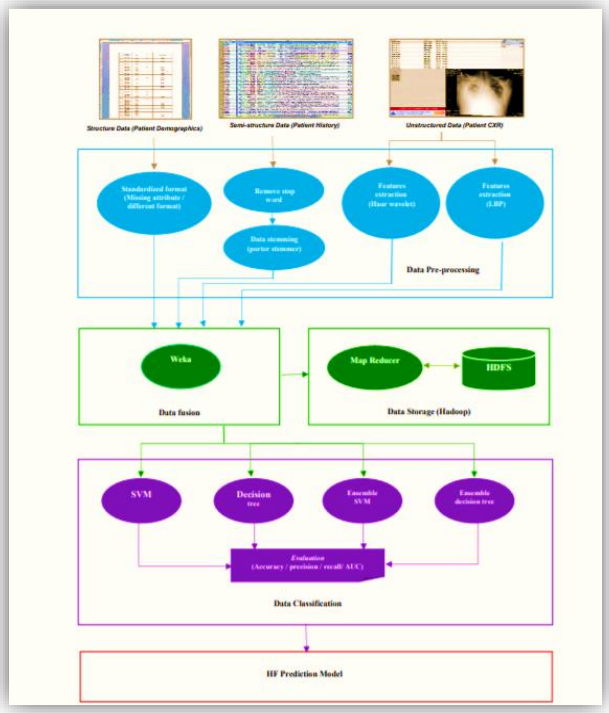


Big Data Analytics in the healthcare industry, which is the domain for this project, can be applied with Predictive and Prescriptive Analytics (Roy et al., 2020). Predictive Analytics provide an indication of the future of the data, whilst Prescriptive Analytics provide a solution that considers historical data (Khan & Alotaibi, 2020). Machine Learning is an example of statistical techniques used for Predictive and Prescriptive Analytics (Deka, 2014).

Big Data Analytics can be applied in the healthcare industry, with the schematic diagram shown below:



Big Data Analytics can be applied in the healthcare industry, by processing all forms of data, storing data in Hadoop, and finally, classifying data using Predictive Analytics. Next, Prescriptive Analytics is then applied to the data, as shown in below:



Analysis and Discussion

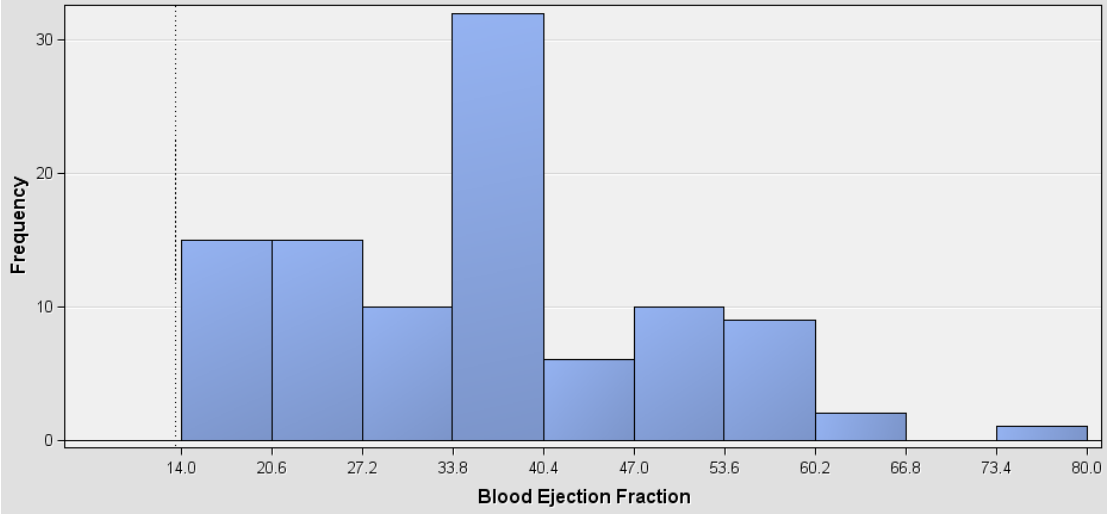
This assignment details the analysis performed using SAS Enterprise Miner, to predict heart failure based on a set of variables. The dataset used is shown in the table below:

Metadata	Type	Description
Age	Interval	Age of patients (Continuous variable)
Anemia	Nominal	Anemia in patients (Categorical variable)
Creatinine Level	Interval	Level of creatinine in patients (Continuous variable)
Diabetes	Nominal	Diabetes in patients (Categorical variable)
Blood Ejection Fraction	Interval	Blood Ejection Fraction in patients (Continuous variable)
High Blood Pressure	Nominal	High Blood pressure in patients (Categorical variable)
Platelets	Interval	Platelets count in patients (Continuous variable)
Serum Creatinine Level	Interval	Serum Creatinine Level in patients (Continuous variable)
Serum Sodium Level	Interval	Serum Sodium Level in patients (Continuous variable)
Gender	Nominal	Gender of patients (Categorical variable)
Smoking Status	Nominal	Smoking status of patients (Categorical variable)
Follow-up Period	Interval	Days of follow-up period for patients after first visit (Continuous variable)
Death Event	Nominal	Death event that occurs in patients (Categorical variable)

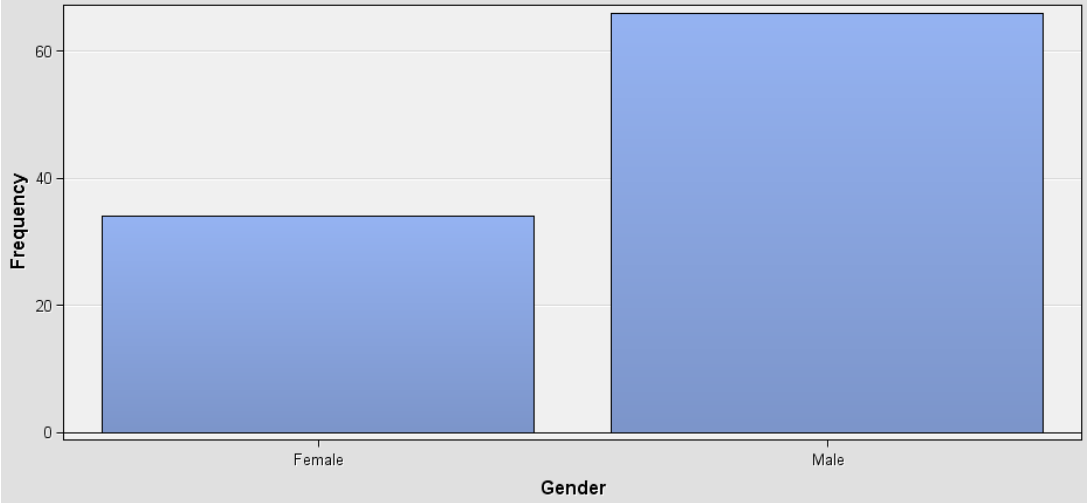
The target variable is Death Event (YES indicates Death occurring), based on the input variables. The aim of this assignment is to investigate the variables that affect the predicted outcome, which is Death Event.

The dataset was loaded onto SAS Enterprise Miner (SAS EM) and the dataset was explored.

Histograms of the dataset was created by using SAS EM, as shown in the figures below:



Blood ejection fraction shows a normal distribution (continuous variable)



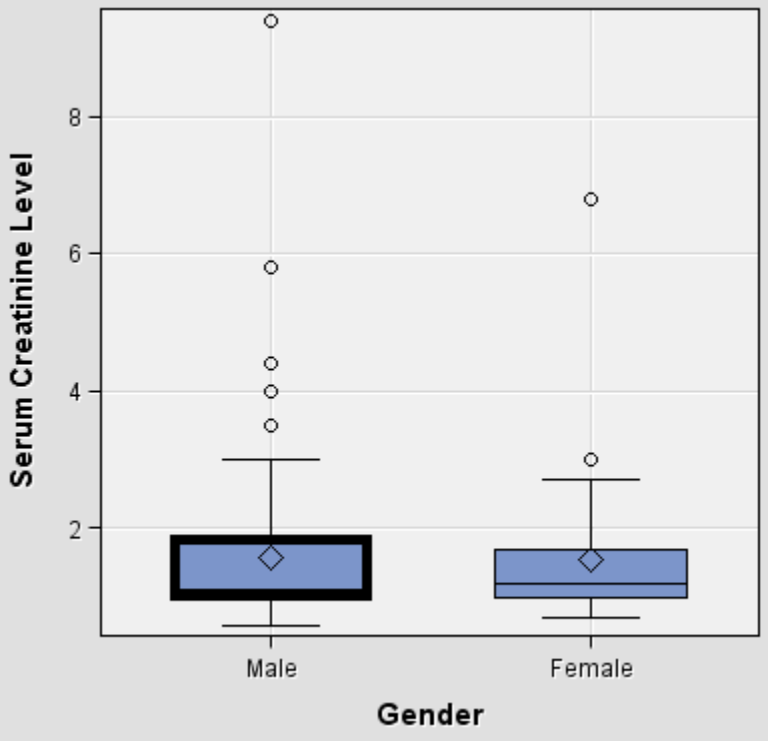
Gender shows a higher frequency of Male (65%) compared to Female (35%) for the dataset

Next, missing values for the data was explored and imputed, if necessary. Using the StatExplore function on SAS EM,

Ordered Inputs	Data Role	Variable	Median	Missing	Non Missing	Minimum	Maximum	Mean	Standard Deviation	Skewness	Kurtosis	Role	Label	Abs C.V.	Coefficient of Variation	Sign
1	TRAIN	Creatinine_...	250	0	299	23	7861	581.8395	970.2879	4.46311	25.14905	INPUT	Creatinine ...	1.667621	1.667621	+
2	TRAIN	Serum_Cre...	1.1	0	299	0.5	9.4	1.39388	1.03451	4.455996	25.82824	INPUT	Serum Cre...	0.74218	0.74218	+
3	TRAIN	Follow_Up...	115	0	299	4	285	130.2609	77.61421	0.127803	-1.21205	INPUT	Follow-Up ...	0.595837	0.595837	+
4	TRAIN	Platelets	262000	0	299	25100	850000	263358	97804.24	1.462321	6.209255	INPUT	Platelets	0.371374	0.371374	+
5	TRAIN	Blood_Ejec...	38	0	299	14	80	38.08361	11.83484	0.555383	0.041409	INPUT	Blood Ejecti...	0.310759	0.310759	+
6	TRAIN	Age	60	0	299	40	95	60.83389	11.89481	0.423062	-0.18487	INPUT	Age	0.195529	0.195529	+
7	TRAIN	Serum_So...	137	0	299	113	148	136.6254	4.412477	-1.04814	4.119712	INPUT	Serum Sodi...	0.032296	0.032296	+

There were no missing values for all the continuous variables in this dataset. Hence, imputation is not necessary for this dataset.

Next, using the GraphExplore function on SAS EM, plots were created for Exploratory Data Analysis of this assignment:



Boxplot shows that the Serum Creatinine level for both genders to be somewhat similar, although Males have more outliers than females, which indicates a significantly high Serum Creatinine Level. This would have to be investigated further, to study which variables affect the outcome (Death Event) of the patient

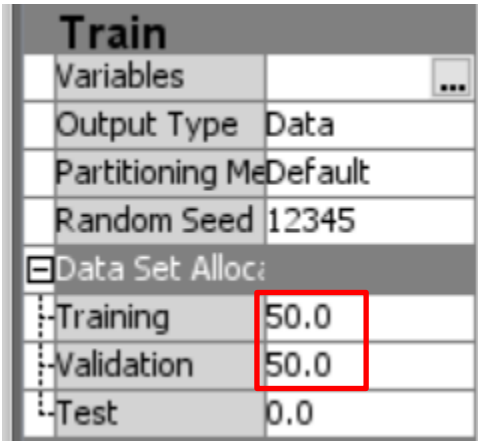
Next, predictive modelling techniques were applied to the dataset.

Model 1: Decision Tree

Experiment 1

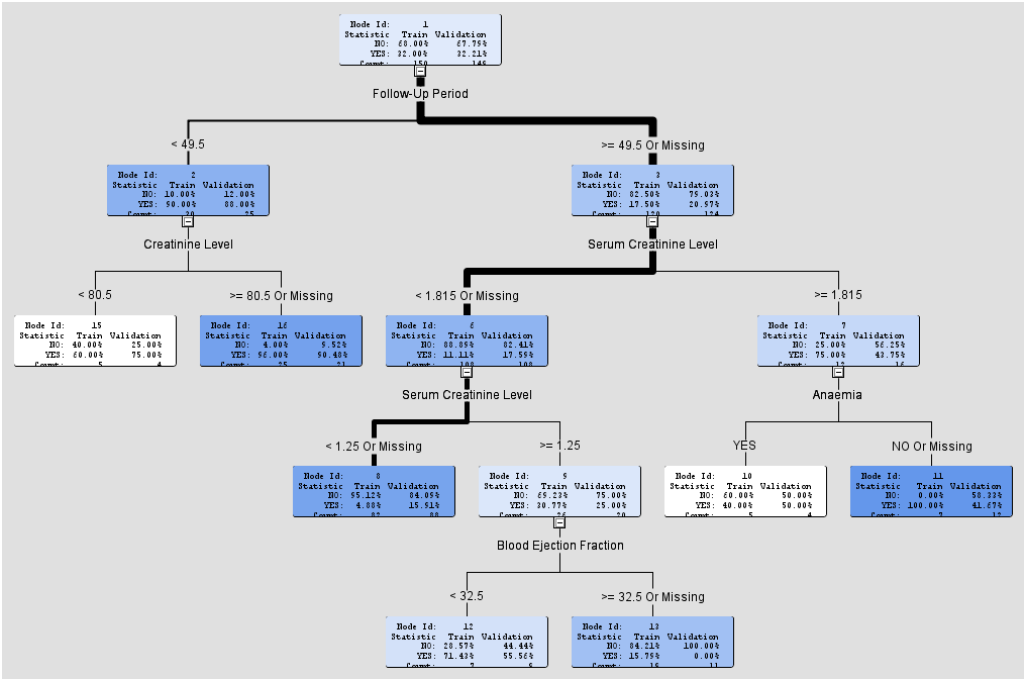
Using decision trees and setting the target variable of "Death Event", decision trees predictive modelling was performed.

Firstly, data partition was performed on the dataset.



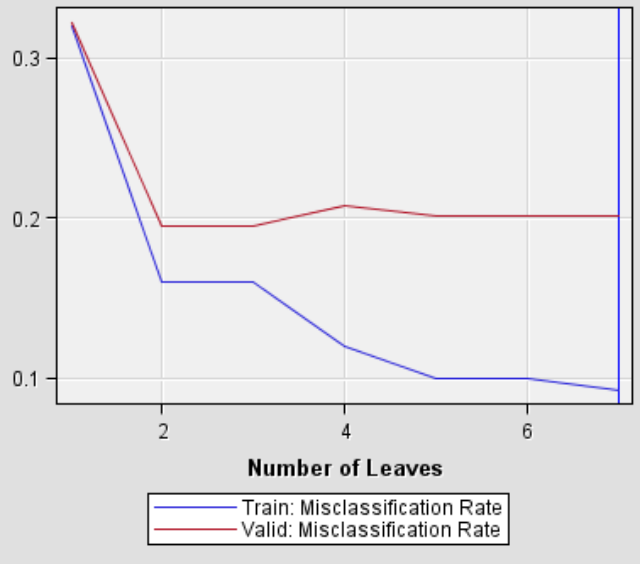
The allocation for data partition was 50-50, for Training and Validation. Test was left at 0.

Next, using the Decision Tree model in SAS EM, and with Death Event being the Target Variable, the Interactive Decision Tree is shown below:



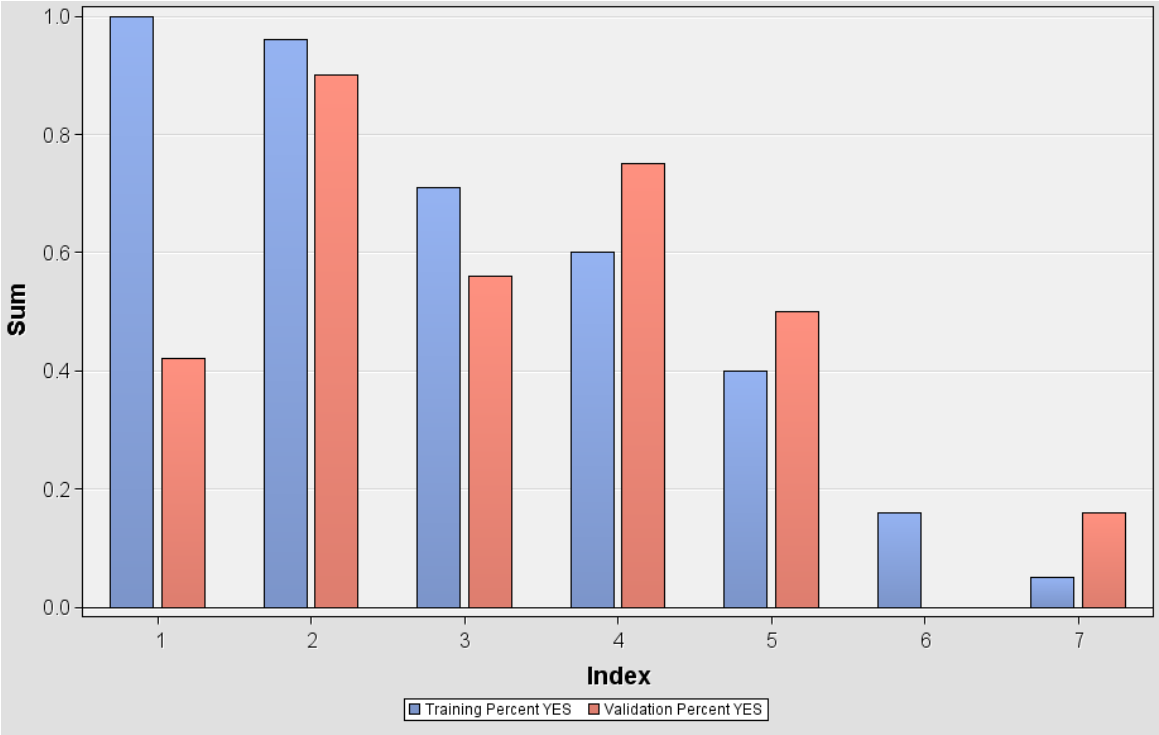
Follow-Up Period (the number of days for follow-up for each patient after their first visit) is the first node. Serum Creatinine Level is the second node, Anemia as the third node and Blood Ejection Fraction as the final node. These nodes were selected based on the high logworth (-Log(p)) value of these variables in the decision tree.

The misclassification rate of the decision tree is depicted below:



Based on the misclassification rate above, the target value that is predicted above 0.3 has a primary outcome classification, whilst target value below 0.3 has a secondary outcome classification. However, with only 6 leaves, this classification is incorrect for all instances that are unassigned inside this class. The maximum leaves of this tree is only 6, with misclassification occurring after 2 leaves. 7 is shown as the optimal under the misclassification rate.

Next, the leaf statistic for the decision tree is shown below:



There are differences in bar heights between Training and Validation data. It shows that the predicted outcomes (training data) have variations with the observed outcomes (validation data) which could be the consequence of lesser case amounts in the subsequent leaf.

Finally, using the Node Rules section of the Results, the following was produced:

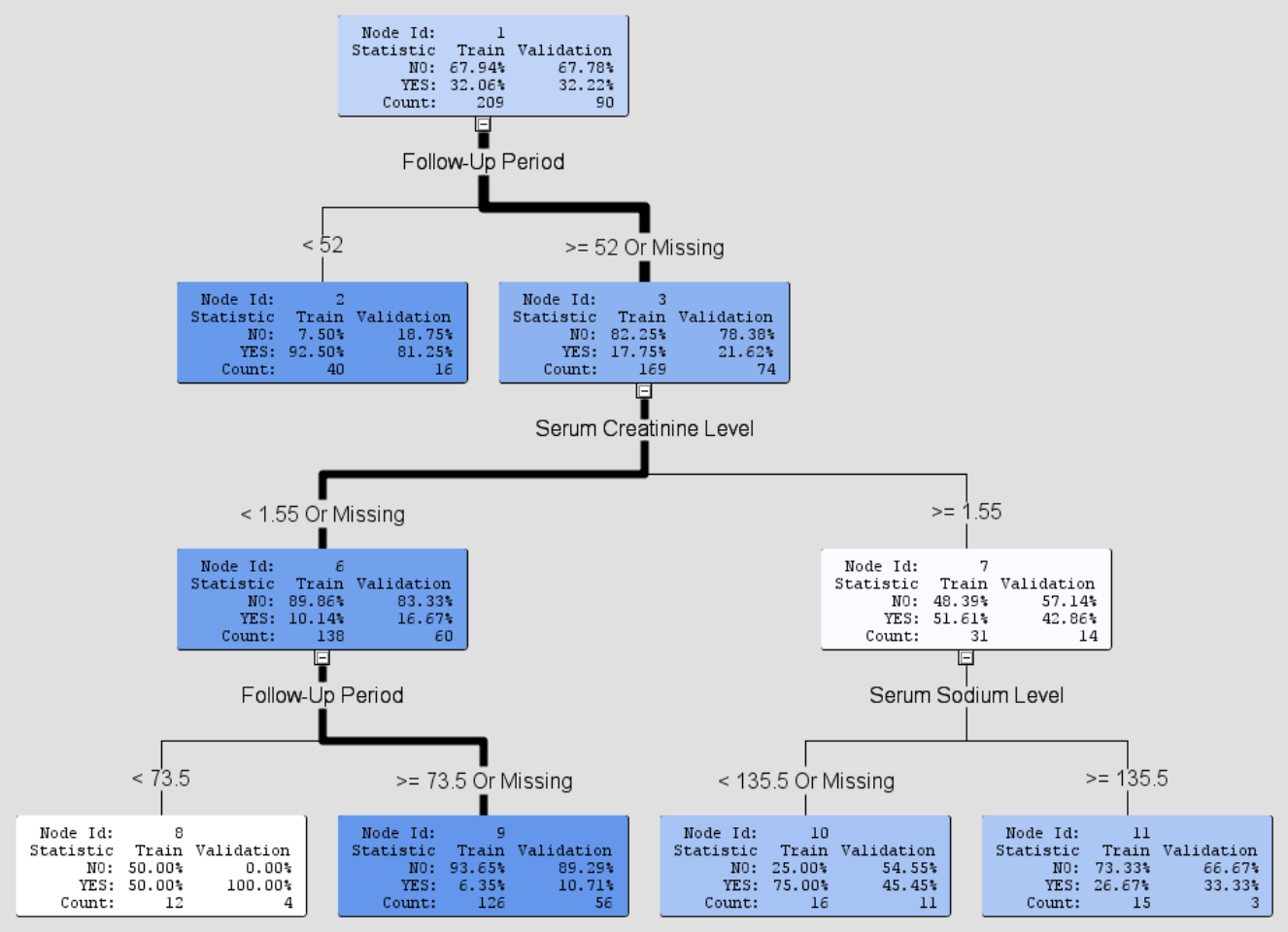
```
*-------------------------------------------------------------*
 Node = 16
*-------------------------------------------------------------*
if Follow-Up Period < 49.5
AND Creatinine Level >= 80.5 or MISSING
then
 Tree Node Identifier   = 16
 Number of Observations = 25
 Predicted: Death_Event=YES = 0.96
 Predicted: Death_Event=NO = 0.04
```

```
*-------------------------------------------------------------*
 Node = 11
*-------------------------------------------------------------*
if Serum Creatinine Level >= 1.815
AND Follow-Up Period >= 49.5 or MISSING
AND Anaemia IS ONE OF: NO or MISSING
then
 Tree Node Identifier   = 11
 Number of Observations = 7
 Predicted: Death_Event=YES = 1.00
 Predicted: Death_Event=NO = 0.00
```

These nodes have the highest predicted Death Event as YES, with Node 11 having a predicted Death Event of YES occurring at a chance of 1, IF Serum Creatinine Level is equal or greater than 1.815 AND Follow-up period is equal or greater than 49.5 days AND Anemia is NOT PRESENT. This seems to be the optimum level that a patient can achieve, for a result of Death Event. Moreover, Node 16 also predicts a high Death Event of YES occurring is 0.96, if Creatinine Level is equal or greater than 80.5, for follow-up period lesser than 49.5 days. Hence, variables of Creatinine Level and Follow-Up days majorly affect the outcome of the patient, which is Death Event.

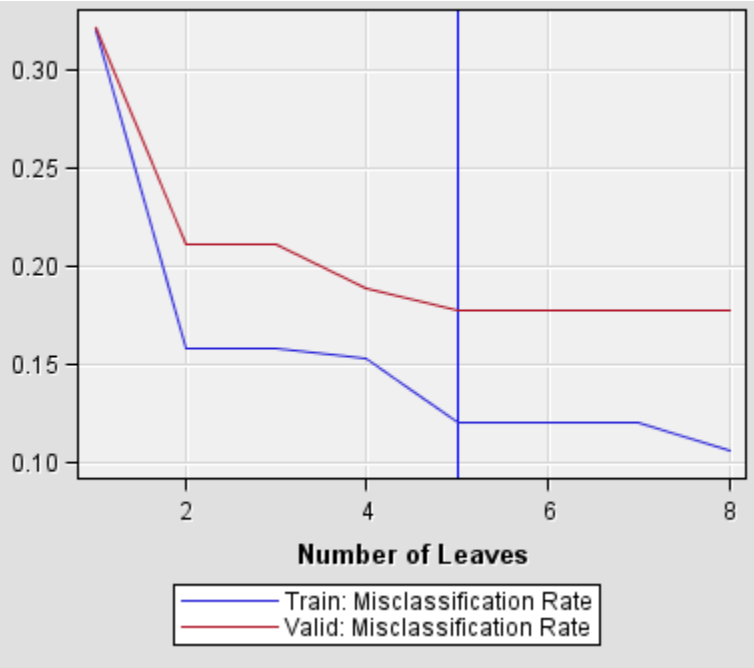
Now, by changing the data partitioning, using Training Data of 70% and Validation data of 30%, the following outputs were produced:

Data Set Alloc	
Training	70.0
Validation	30.0
Test	0.0



The decision tree is taking Serum Creatinine Level, Follow-up period and Serum Sodium Level as variables of interest.

The misclassification rate for this experiment is shown below:



This time, 5 leaves is the optimal leaf, as compared to 7 from Experiment 1. Target value for predicted is above 0.3 for primary outcome classification, whilst target value below 0.3 has a secondary outcome classification, which is the same as experiment 1. Experiment 2 has 8 leaves, as compared to 6 from experiment 1.

The leaf statistic for Experiment 2 is shown below:

Index 3 is the only instance where the heights for both bars are not similar, whilst the other indexes show that training and validation data are somewhat similar.

Next, the Node Rules graph is produced, as depicted below:

```
*-------------------------------------------------------------*
 Node = 2
*-------------------------------------------------------------*
if Follow-Up Period < 52
then
 Tree Node Identifier   = 2
 Number of Observations = 40
 Predicted: Death_Event=YES = 0.93
 Predicted: Death_Event=NO = 0.08
```

```
*-------------------------------------------------------------*
 Node = 10
*-------------------------------------------------------------*
if Serum Sodium Level < 135.5 or MISSING
AND Serum Creatinine Level >= 1.55
AND Follow-Up Period >= 52 or MISSING
then
 Tree Node Identifier   = 10
 Number of Observations = 16
 Predicted: Death_Event=YES = 0.75
 Predicted: Death_Event=NO = 0.25
```

Based on the Node Rules, Node 2 shows that If Follow Up Period is lesser than 52 days, the Predicted Death Event of Yes is 0.93 chance of occurring, whilst Node 10 shows that If Serum Sodium Level is 135.5 AND Serum Creatinine Level is greater or equal to 1.55 AND Follow-up Period is greater or equal to 52 days, then Predicted Death Event of Yes is 0.75 chance of occurring.

Based on the 2 experiments, the summary of Predicted Death Event is:

Experiment 1:

Follow-Up Period of greater than 49.5 days for Serum Creatinine Level greater or equal to 1.815 (P = 1)

Experiment 2:

Follow-Up Period of greater than 52 days for Serum Creatinine Level greater or equal to 1.55 (P = 0.75)

The 2 experiments produce somewhat similar results for predicted Death Event in a patient.

Model 2: Logistic Regression

Experiment 1

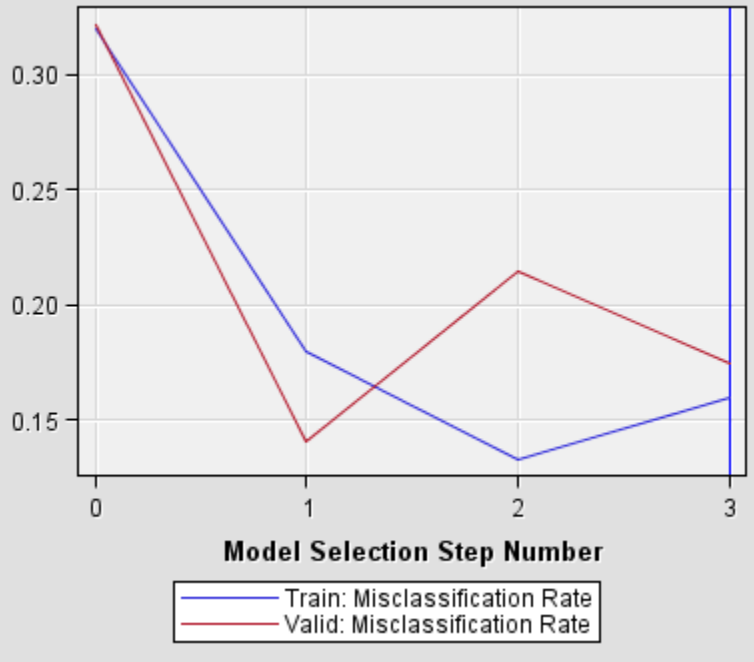
For experiment 1, the target and validation data are partitioned into 50-50.

Logistic regression is the second model to be tested for the predictive modelling section of this assignment.

Class Targets	
Regression Ty	Logistic Regres
Link Function	Logit
Model Options	
Suppress Inter	No
Input Coding	Deviation
Model Selectio	
Selection Mode	Stepwise
Selection Crite	Default
Use Selection D	Yes

Since the target variable (Death Event) is a categorical variable, Logistic Regression is used. Stepwise regression is chosen to analyse which variables were entered into the regression model (based on significance of < 0.005) to enable understanding of which variables affect the outcome (Death Event = YES).

Based on this model, the misclassification rate graph is produced, as shown below:



From the misclassification graph above, the iteration for both training and validation data stops after 3 selections, using stepwise regression. After step 1 (Step 0 is entering the intercept) which is adding the Follow-Up period variable, the training data misclassification rate drops to 0.160. Step 2 is the addition of the Serum Creatinine Level variable, while Step 3 is the addition of the Blood Ejection Fraction variable. However, the significance of the Blood Ejection Fraction is 0.0205, which is greater than 0.005. Hence, other variables were not added into the stepwise regression.

Next, the Maximum Likelihood Estimates graph is analysed, which is shown below:

Analysis of Maximum Likelihood Estimates

Parameter	Death_ Event	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	YES	1	0.8609	1.0779	0.64	0.4245
Blood_Ejection_Fraction	YES	1	-0.0475	0.0211	5.06	0.0245
Follow_Up_Period	YES	1	-0.0176	0.00362	23.66	<.0001
Serum_Creatinine_Level	YES	1	1.6249	0.5046	10.37	0.0013

The equation for the logistic regression is:

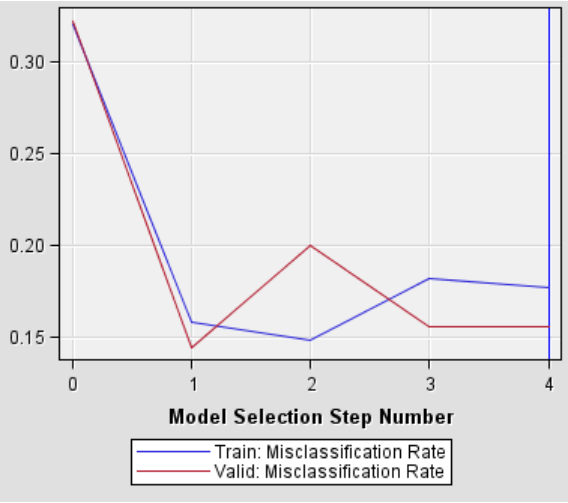
Death Event = $0.8609 - 0.0475\ X_1 - 0.0176X_2 + 1.6429\ X_3$

Whereby X_1 is the Blood Ejection Fraction, X_2 is the Follow-up period and X_3 is the Serum Creatinine Level. Logistic Regression also proves that variables Follow-up period and Serum Creatinine Level affect the outcome, which is Death Event, in a patient.

Experiment 2

For experiment 2, the target data is 70%, while validation data is 30%.

The misclassification rate for experiment 2 is shown below:



Based on the misclassification rate, 4 steps were used in the stepwise regression, as compared to 3 in Experiment 1.

At Step 1, Follow-up period was entered into the regression equation. At Step 2, Serum Creatinine Level is entered. At Step 3, Blood Ejection Fraction is entered. Finally, at step 4, Age is entered.

The maximum likelihood estimate is shown below:

Analysis of Maximum Likelihood Estimates

Parameter	Death_ Event	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	YES	1	-0.1455	1.2838	0.01	0.9098
Age	YES	1	0.0398	0.0176	5.12	0.0237
Blood_Ejection_Fraction	YES	1	-0.0658	0.0191	11.94	0.0005
Follow_Up_Period	YES	1	-0.0184	0.00324	32.09	<.0001
Serum_Creatinine_Level	YES	1	1.0431	0.3113	11.23	0.0008

The equation for the logistic regression is:

Death Event = -0.1455- 0.0658 X_1 – 0.0184X_2 + 1.0431 X_3 + 0.0398 X_4

Whereby X_1 is the Blood Ejection Fraction, X_2 is the Follow-up period, X_3 is the Serum Creatinine Level and X_4 is the age. Logistic Regression also proves that variables Follow-up period and Serum Creatinine Level affect the outcome, which is Death Event, in a patient. Age, however, has a significance of 0.0237, which is greater than 0.005.

Experiment 1 and Experiment 2 show somewhat similar variables affecting the predicted outcome, which is Follow-up period and Serum Creatinine Level. Hence, based on both the predictive models and experiments, it can be stated that the variables that affect Death Event are Follow-up period and Serum Creatinine level. Blood ejection fraction and Age would be the next variables that would affect the predicted outcome.

Conclusion

Big Data Analytics in the Healthcare industry is a pivotal form of analytics using Data Science, and for the purpose of this project, predicting the rate of heart failure based on various dependent and independent variables, is sufficient in providing an accurate solution for prediction. This project outlines the methodology and purpose of carrying out Big Data Analytics for Heart Failure Prediction, using various techniques available in data science.

Based on the predictive modelling performed using SAS EM, the main variables that would cause heart failure in a patient would be Follow up period, Serum Creatinine Level, Blood Ejection fraction and Age.

Patients who have a higher follow up period (more than 50 days) and a high serum creatinine level (more than 1.55) have a higher likelihood of death event occurring.

Based on this assignment, heart failure can be predicted with the usage of predictive modelling. This would be beneficial to the healthcare industry, as with the usage of data science, predictive analytics can be performed to aid medical care of patients in a hospital. Machine Learning techniques such as Decision Tree, Regression, Cluster Analysis and Neural Networks can certainly benefit the healthcare industry.

References

1. Heart Failure Prediction. Kaggle.com. (2021). Retrieved 10 July 2021, from https://www.kaggle.com/andrewmvd/heart-failure-clinical-data.

2. Alexander, C., & Wang, L. (2017). Big Data Analytics in Heart Attack Prediction. Journal Of Nursing & Care, 06(02). https://doi.org/10.4172/2167-1168.1000393

3. Deka, G. (2014). Big Data Predictive and Prescriptive Analytics. Advances In Data Mining and Database Management, 370-391. https://doi.org/10.4018/978-1-4666-5864-6.ch015

4. Khan, Z., & Alotaibi, S. (2020). Applications of Artificial Intelligence and Big Data Analytics in m-Health: A Healthcare System Perspective. Journal Of Healthcare Engineering, 2020, 1-15. https://doi.org/10.1155/2020/8894694

5. Kim, J. (2021). Big Data, Health Informatics, and the Future of Cardiovascular Medicine.

6. Lanzer, J., Leuschner, F., Kramann, R., Levinson, R., & Saez-Rodriguez, J. (2020). Big Data Approaches in Heart Failure Research. Current Heart Failure Reports, 17(5), 213-224. https://doi.org/10.1007/s11897-020-00469-9

7. Rammal, H., & Z., A. (2018). Heart Failure Prediction Models using Big Data Techniques. International Journal Of Advanced Computer Science And Applications, 9(5). https://doi.org/10.14569/ijacsa.2018.090547

8. Roy, A., Bruce, C., Schulte, P., Olson, L., & Pola, M. (2020). Failure prediction using personalized models and an application to heart failure prediction. Big Data Analytics, 5(1). https://doi.org/10.1186/s41044-020-00044-2

9. Sammani, A., Jansen, M., Linschoten, M., Bagheri, A., de Jonge, N., & Kirkels, H. et al. (2019). UNRAVEL: big data analytics research data platform to improve care of patients with cardiomyopathies using routine electronic health records and standardised biobanking. Netherlands Heart Journal, 27(9), 426-434. https://doi.org/10.1007/s12471-019-1288-4

10. Seward, J. (2021). Paradigm Shift in Medical Data Management.

11. Thammasudjarit, R., Pattanateepon, A., & Pattanaprateep, O. (2018). Big Data Analytics in Healthcare. Ramathibodi Medical Journal, 41(2), 116-123. Retrieved 13 July 2021.