PATIENTS LOSS PREDICTION

Group 4

Ruwinda Rowel -s15654
Nipuni Sandunika -s15657
Ganeshi Umayangana -s15669
Bashitha Wijesinghe -s15677
Hiruni Kudagama -s15680
Kavishka Palihena -s15771

Contents

List of Figures

List of Tables

Objectives of the Analysis

The primary objective of building a classification model for the given dataset is to predict whether a patient has survived or died after undergoing surgery. The model aims to analyze relevant features and patterns within the dataset to make accurate predictions regarding the post-surgery mortality status of patients. By utilizing machine learning techniques, the goal is to develop a robust and reliable classification model that can assist healthcare professionals in assessing the risk of mortality for patients while following surgical procedures.

Description of data

28 variables. (1 - Quantitative and others are categorical)
Name of the dataset: **patient_loss.csv** This dataset consists of 10,000 records of

No	Variable	Variable	Description of categories
1	**age**	Age in years	
2	**agecat**	Age category	1 – 45 -54,2 – 55 – 64,3 – 65 – 74,4 – 74+
3	**gender**	Gender	0 – Male, 1 - Female
4	**diabetes**	History of diabetes	0 – No 1 - Yes
5	**bp**	Blood pressure	0 – Hypotension, 1 – Normal, 2 - Hypertension
6	**smoker**	Smoker	0 – No, 1 - Yes
7	**choles**	Cholesterol	0 -Normal, 1 - High
8	**active**	Physically active	0 – No, 1 - Yes
9	**obesity**	Obesity	0 – No, 1 - Yes
10	**angina**	History of angina	0 – No, 1 - Yes
11	**mi**	History of myocardial infarction	0 – No, 1 - Yes
12	**nitro**	Prescribed nitroglycerin	0 – No, 1 - Yes
13	**anticlot**	Taking anti-clotting drugs	0- None, 1 – Aspirin, 2 – Heparin, 3 - Warfarin
14	**site**	Hospital ID	1 – 0001, 2 – 0002, 3 – 0003, 4 – 0004, 5 - 0005
16	**doa**	Dead on arrival	0 – No, 1 - Yes
17	**ekg**	EKG result	-1 – DOA, 0 – No ST elevation, 1 – ST elevation
18	**cpk**	CPK blood result	-1 – DOA, 0 – Normal CPK, 1 – High CPK
19	**tropt**	Troponin T blood result	-1 – DOA, 0 – Normal Troponin T,1 – High Troponin T
20	**clotsolv**	Clot-dissolving drugs	-1 – DOA, 0 – None,1 – Streptokinase,2 - Reteplase ,3 -Alteplase
21	**bleed**	Hemorrhaging	-1 – DOA, 0 – No, 1 - Yes
22	**magnes**	Magnesium	-1 – DOA, 0 – No, 1 - Yes
23	**digi**	Digitalis	-1 – DOA, 0 – No, 1 - Yes
24	**betablk**	Beta blockers	-1 – DOA, 0 – No, 1 - Yes
25	**der**	Died in ER	-1 – DOA, 0 – No, 1 - Yes
26	**proc**	Surgical treatment	-2 – Died before surgery,-1 – DOA,0 – None,1 – PTCA,2 - CABG
27	**comp**	Surgical complications	-3 – No surgery performed,-2 - Died before surgery,-1 – DOA 0 – No,1 - Yes
28	**result**	Surgery result	-3 – No surgery performed,-2 - Died before surgery,-1 – DOA 1 – Well,2 – Stable,3 – Critical,4 - Died

Table 1 - Description of Data

Data pre-processing and Feature Engineering

At the start of the project, our main objective was to find the observations where a surgery had been conducted. We began the feature engineering process by filtering out rows where the result is -3, -2, or -1, as these values indicate scenarios where no surgery was performed, the patient died before surgery, or the patient died on arrival, respectively. Since our objective revolves around predicting post-surgery survival, these observations do not represent our target group and were thus removed. Following this, we eliminated the 'DOA' and 'DER' variables as they all held a value of 0, rendering them redundant. Subsequently, we created a new response variable indicating the outcome after surgery, categorizing 'Well' (1), 'Stable' (2), and 'Critical' (3) as 'Survived'. We recorded 'Survived' and 'Died' as 0 and 1, respectively, for consistency and ease of interpretation. Finally, we assessed the main variables to include in our model, crucial for predicting post-surgery survival effectively.

Descriptive Data Analysis

Univariate analysis

Analysis of Proc (Surgical treatment)

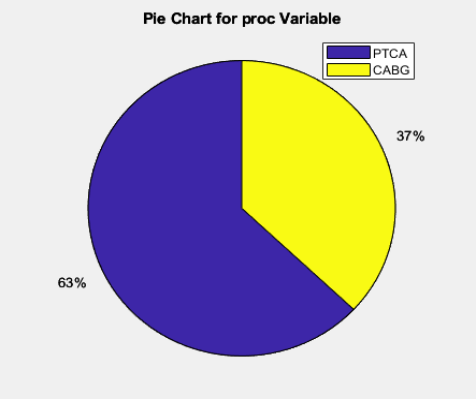


Figure 1- Analysis of Proc

The univariate analysis of the 'proc' variable reveals that it represents surgical treatment, with 63% of patients undergoing Percutaneous Transluminal Coronary Angioplasty (PTCA). PTCA is a minimally invasive procedure aimed at opening clogged coronary arteries to enhance blood flow to the heart muscle. The remaining 37% of patients fall into the Coronary Artery Bypass Graft (CABG) category. CABG is a surgical intervention utilized to treat coronary heart disease by creating grafts to bypass blocked arteries.

Analysis of Comp (Surgical complications)

In here, it was found that 89% of patients did not experience complications, while 11% did encounter surgical complications.

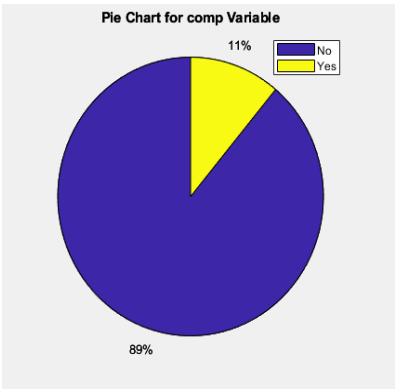


Figure 2 -Analysis of Comp

Analysis of site (Hospital ID)

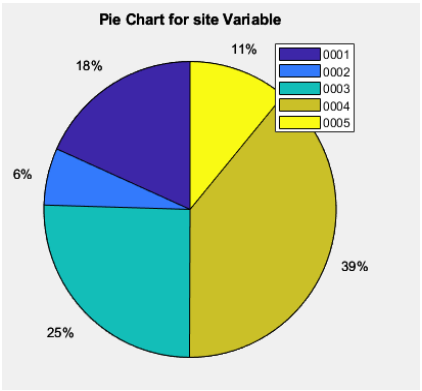
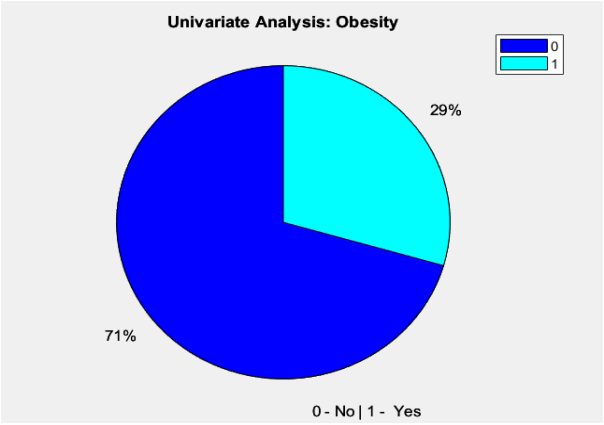


Figure 3 - Analysis of Site

Most patients in the study sought medical care at hospitals identified by the site (hospital ID) 0004.

Analysis of Obesity



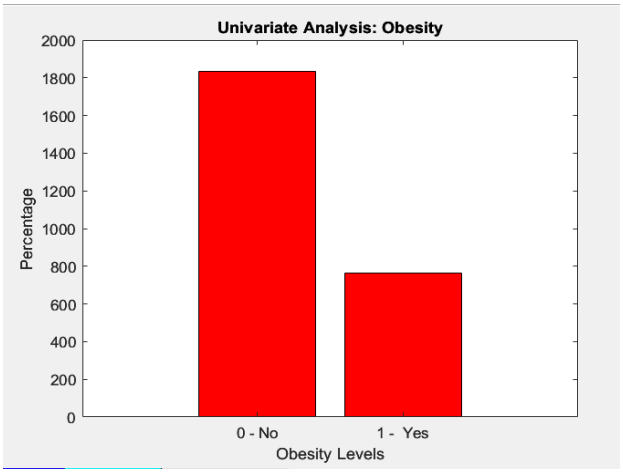


Figure 4 - Analysis of Obesity

A pie chart is used to visually represent the distribution of patients based on their obesity status. The majority of the chart is occupied by the slice representing patients classified as obese. This visualization provides a quick and clear understanding of the prevalence of obesity among the patients.

Bivariate analysis

Analysis of Proc (Surgical treatment) vs Result

The data from the plot suggests that most patients who underwent both PTCA and CABG treatments did not experience fatalities, indicating a relatively favorable outcome in terms of survival for these procedures.

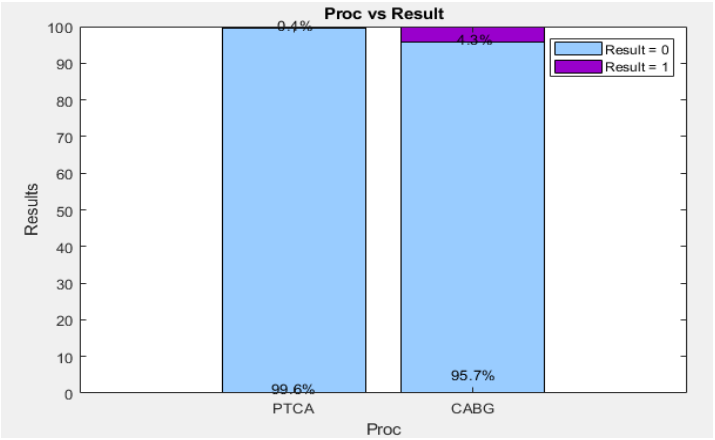


Figure 5 - Analysis of Proc Vs Result

Analysis of Comp (Surgical complications) vs Result

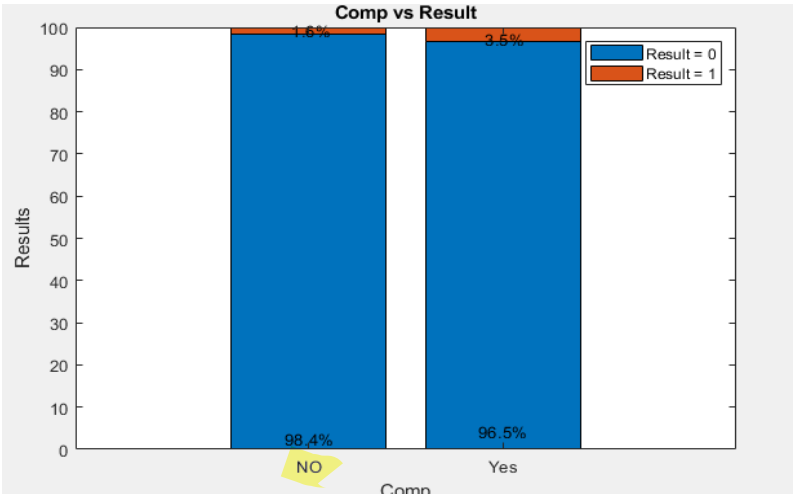


Figure 6 - Analysis of Comp Vs Result

The plot indicates that a significant proportion of patients who experienced surgical complications did not result in fatalities, suggesting a relatively low mortality rate in this context.

The plot indicates that a significant proportion of patients who experienced surgical complications resulted in fatalities

Analysis of site (Hospital ID) vs Result

Based on the provided information, it seems that hospital 0004 has a notably higher survival rate, with 99.1% of patients not experiencing fatalities. In contrast, hospital 0002 has a higher mortality rate, as 4.3% of patients died. This suggests that hospital 0004 may be considered the better option for patient outcomes in this context.

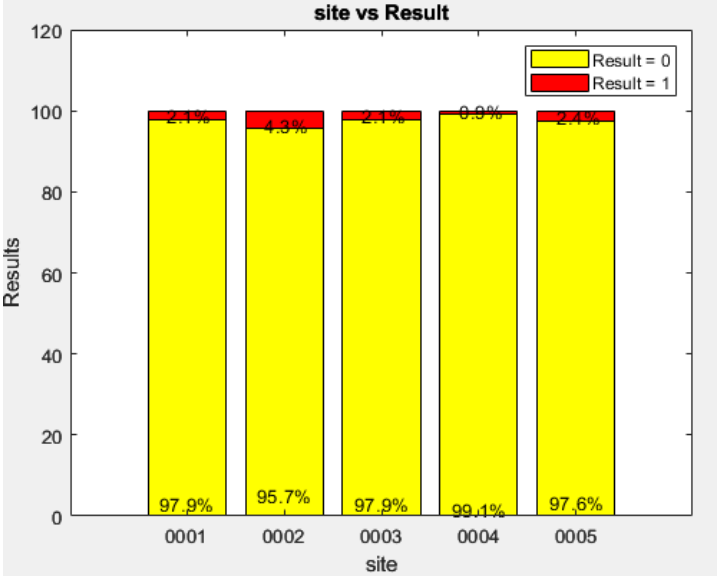


Figure 7 - Analysis of Site Vs Result

Agecat (Age category) Vs Result

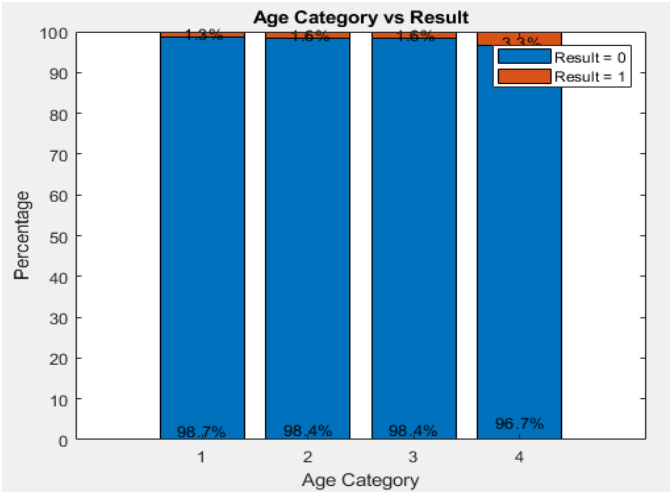


Figure 8 - Agecat Vs Result

In here approximately all the age categories have the same survival percentages. But the old aye category (age 74+) has a somewhat higher percentage of not surviving than other age categories.

Analysis of Obesity Vs Result

The percentage bar chart illustrates the survival rates among patients categorized by their obesity status, with a focus on those classified as high obesity. This chart effectively communicates the relationship between obesity status and survival, the higher survival rate among patients with high obesity.

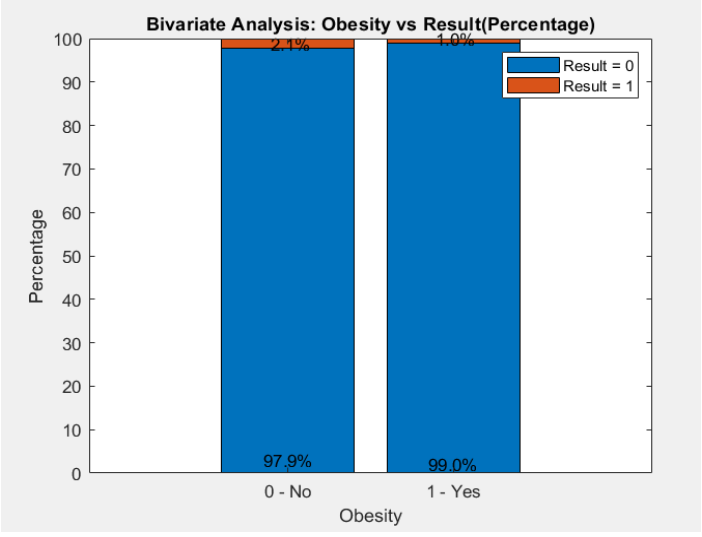


Figure 9 - Analysis of Obesity Vs Result

Advanced Analysis

Addressing imbalanced dataset

One of the key disadvantages of our dataset is that it is severely imbalanced with 4166 observations being patient who survived surgery while 71 are the latter. To address this issue we considered 3 remedies, Under Sampling the Majority Category, Over Sampling the Minority Category and Synthetic Minority Oversampling Technique or SMOTE. Without going into much detail, under sampling involves reducing the number of observations from the majority category, while over sampling and SMOTE involves increasing the number of observations from the minority sample. Out of these 3 methods we found that SMOTE generated the lowest misclassification rate. For reference we've included the misclassification errors and confusion matrix we got from under sampling and SMOTE when we applied the K-Nearest Neighbors.

Sampling Method	**Under sampling**	**SMOTE**
Mis-Classification Rate	0.0775	0.0498

Table 2 - Mis Classification Rate

Table 3 - Confusion Matrix of Under Sampling

	Survived	Did not Survived
Survived	861	5
Did not Survived	48	151

	Survived	Did not Survived
Survived	130	1
Did not Survived	10	1

Table 4 - Confusion matrix of SMOTE

Thus, for our further analysis moved forward with a synthetic dataset generated through SMOTE.

Clustering Technique

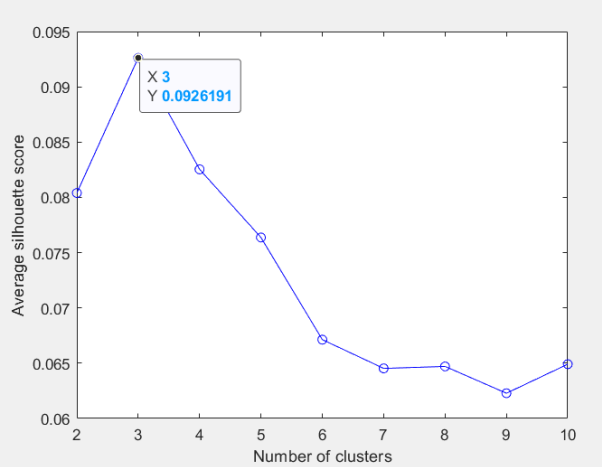


Figure 11 - Average Silhouette Score

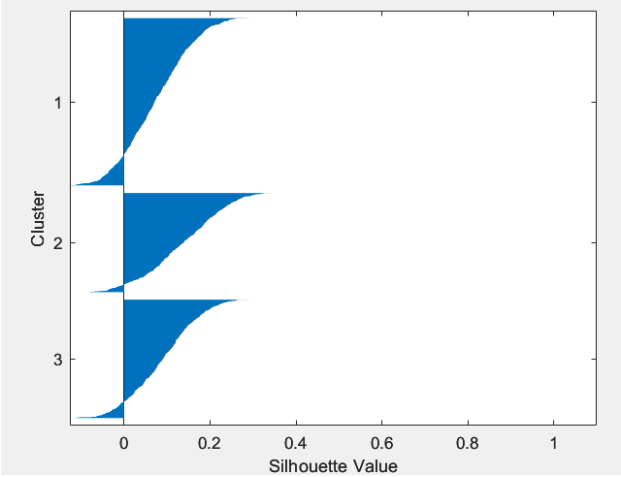


Figure 10 - Clusters.

Dividing patients into homogeneous groups involves clustering patients into groups considering their similarities on the predictor variables. The process of dividing patients into similar groups based on their similarities is done using kmedoids clustering. For the identification of the best number of clusters, the kmedoids clustering algorithm was used with the distance measure, cosine. As shown in figure 11 it ended up with 3 clusters whose average silhouette value is maximum which is 0.0926191. The clusters have been separated as the above figure 10 of silhouette plot which shows how the clusters have been divided. By the silhouette, plot indicates that the clustering is not accurate as some of the observations are on the negative side in all the 3 clusters indicating that they are misclassified to each cluster and the highest value of each cluster is 0.2881,0.31946 and 0.288832 less than 0.5 which indicates that these clusters are not properly clustered.

Probabilistic Neural Network

It appears that a Probabilistic Neural Network was utilized to predict the class of the output variable, distinguishing between survival and non-survival outcomes. This approach allows for a probabilistic assessment, potentially offering more nuanced predictions than a deterministic model. Finally, we get the best spread as 1 and its misclassification error as 0.018497.

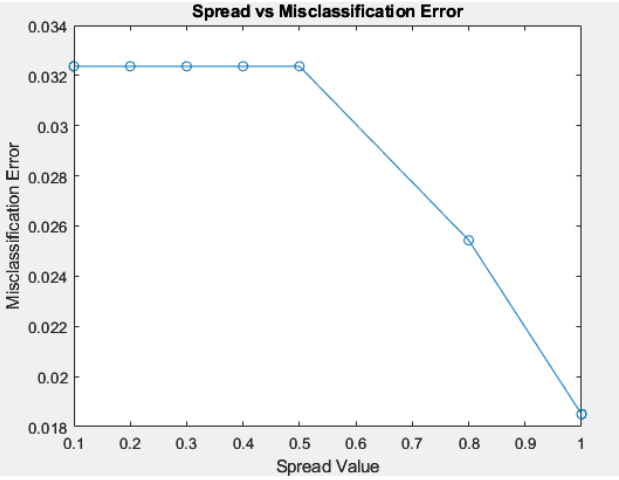


Figure 12 - Spread VS Misclassification Error in PNN

KNN

K Nearest Neighbors (KNN) is a simple yet effective machine learning algorithm that makes predictions based on the majority class (for classification of its nearest neighbors in the feature space. It relies on measuring distances between data points to determine similarities and assign labels, making it intuitive to understand and implement. In this case we considered the distance criterion to be humming distance as nearly all the variables were factorized categorical. The main Tuning Parameter in K Nearest Neighbors is K or the number of Nearest Neighbors we consider using cross validation where we found that the optimal number of K is 3 as evident from the following graph.

Thus, once we set the k value as 3 for the final data, we got the following misclassification rate as 0.0572 and the following confusion matrix.

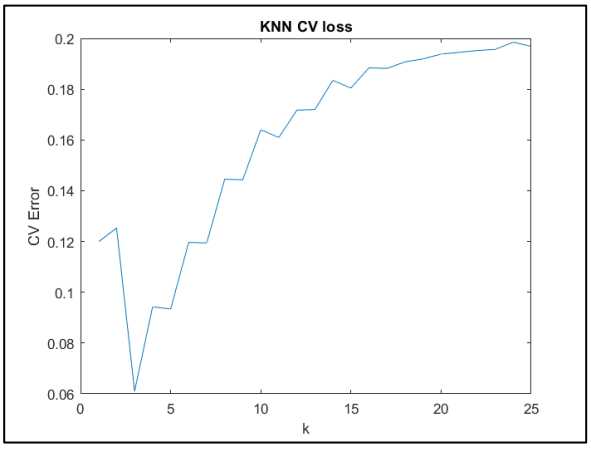


Figure 13 - KNN CV loss.

	Survived	Did not Survived
Survived	862	4
Did not Survived	47	142

Table 5 - Confusion Matrix of KNN

Classification Tree

In this study, we strategically selected key variables using a decision tree approach. The classification tree algorithm, initially applied with default parameters, revealed that at pruning level 6, outperforming 22 other levels, was optimal. By incorporating SMOTE sampling and setting the prune criteria to 'MaxNumSplits',16, we refined the fitted tree in the post-pruning phase. A variable importance plot highlighted critical predictors such as proc, bp, comp, agecat, obesity, and site, offering valuable insights and enhancing our comprehension of the dataset.

The following table shows that resubloss values of the Classification Tree under down-sampling, up-sampling and SMOTE sampling,

	Up-Sampling	Down-Sampling	Smote
resubloss	0. 1730	0.1620	0.0817

Table 6 - resubloss of the classification

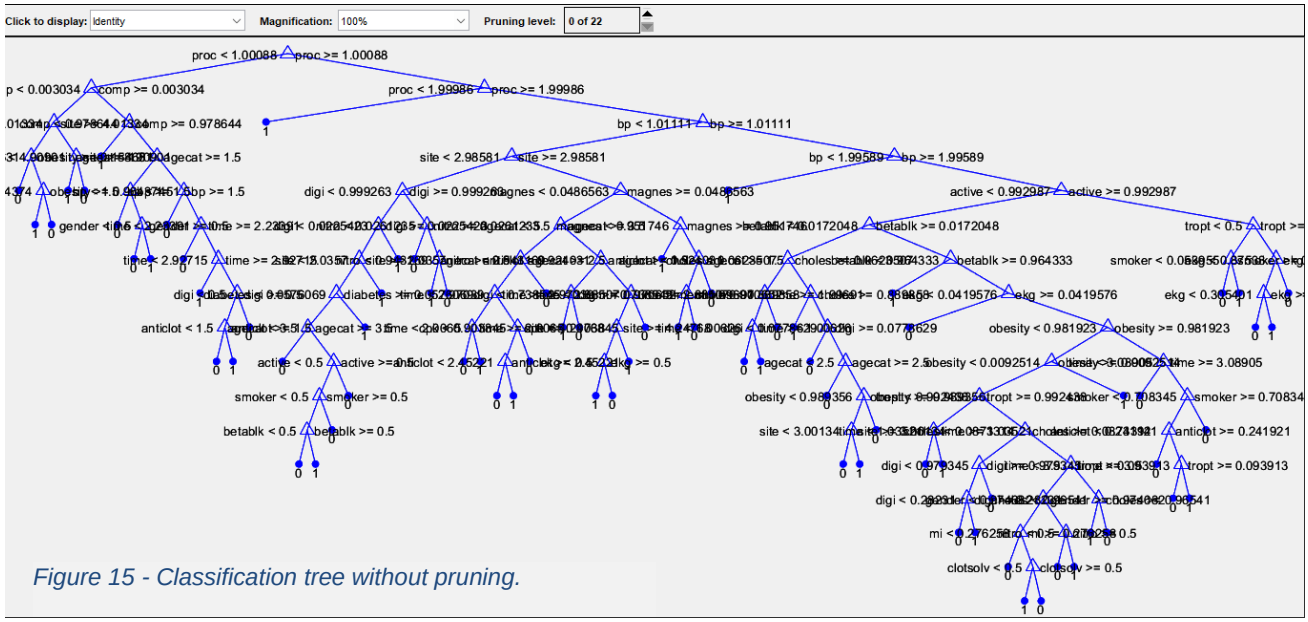


Figure 15 - Classification tree without pruning.

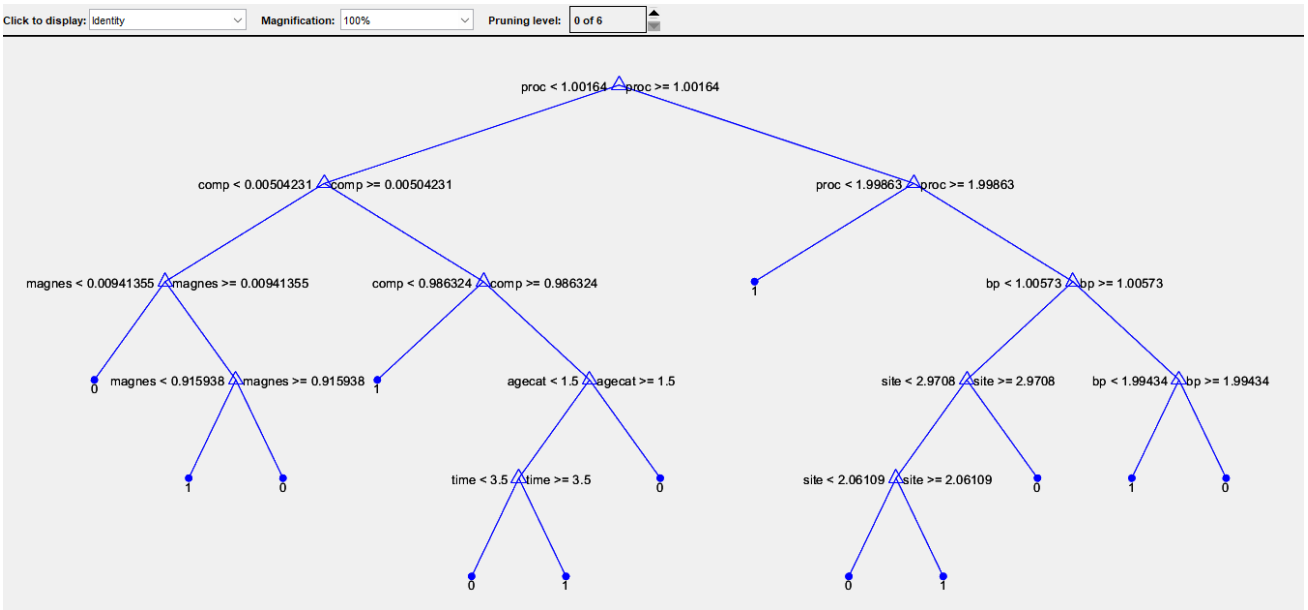


Figure 14 - Classification tree with pruning.

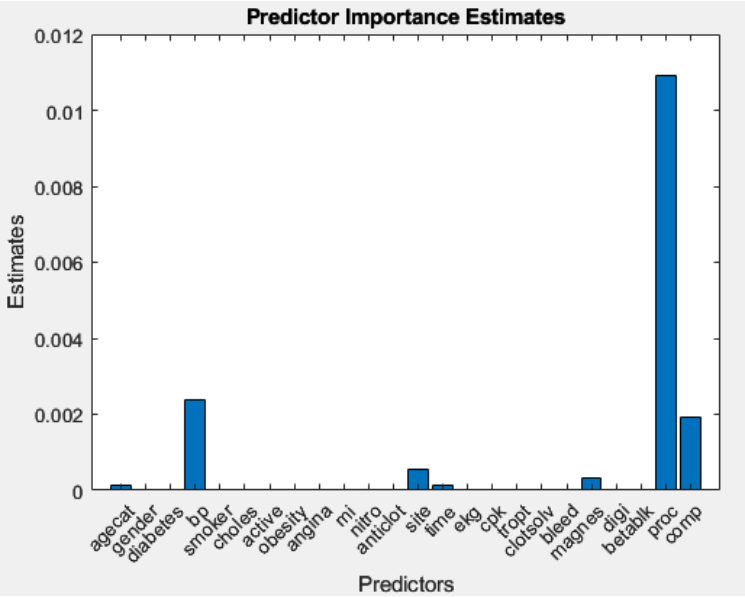


Figure 16 - Predictor Importance Estimates

Random Forrest

Random Forest Classification is an ensemble learning technique that constructs multiple decision trees during training and combines their outputs to make predictions. By using a combination of randomly selected features and bootstrapped samples of the training data, it provides robust and accurate classification results while mitigating overfitting. Given the imbalanced and synthetic nature of our dataset we hoped an ensemble method would yield a much lower classification error compared to other methods. By using cross validation. we found that the optimum number of trees for the Forrest was 60. Further this yielded a classification error of 0.0169.

	Survived	Did not Survived
Survived	865	1
Did not Survived	17	182

Table 7 - Confusion Matrix of Random Forest

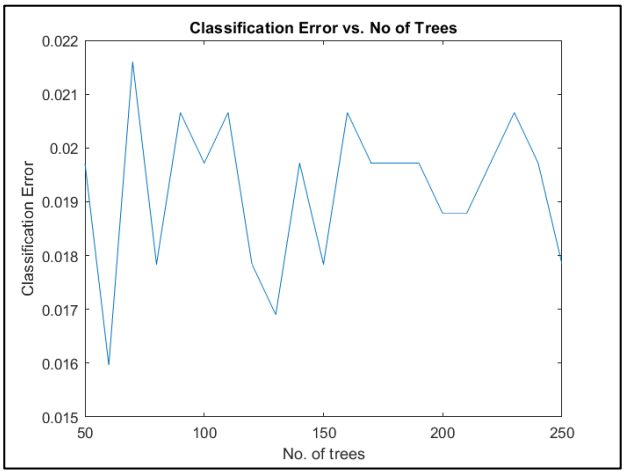


Figure 17 - Classification Error Vs No.of Trees

Concluding Remarks

- ✓ In conclusion, our thorough analysis of the dataset revealed several significant findings regarding the prediction of patient loss. Firstly, we addressed the issue of dataset imbalance by implementing various resampling techniques, including up sampling, down sampling, and SMOTE. Notably, SMOTE emerged as the most effective method, yielding the lowest misclassification error among the three approaches.
- ✓ Furthermore, our exploration into cluster analysis using k-medoids and silhouette plots indicated a lack of significant clusters among the patients, suggesting homogeneity within the dataset.
- ✓ We also employed the Probabilistic Neural Network (PNN) model to predict patient loss status, identifying a spread value of 1 as optimal for minimizing misclassification error.
- ✓ In addition, we utilized K Nearest Neighbors (KNN) with hamming distance for factorized categorical, determining an optimal K value of 3 through cross-validation, resulting in a low misclassification rate.
- ✓ Moreover, our investigation into the impact of various variables on patient mortality revealed significant contributions from factors such as proc, COMP, BP, Sight, Age cat, and Obesity, as identified by the classification tree.
- ✓ Ultimately, our results demonstrate that the Random Forest model, particularly when combined with SMOTE, emerges as the most effective approach for predicting patient loss. This highlights the importance of leveraging advanced machine learning techniques in healthcare analytics to improve patient care outcomes and inform decision-making processes.

Appendix

- Google drive link:
  DM_Codes_Group4
- Github link :
  https://github.com/ruwindarowel/Patient-Loss-Prediction-Group-Project