# CSCI 6409 – Process of Data Science

# Project Report

## DiabAnalytica: Predictive Modeling for Type2 Diabetes


## Group Members

Pallavi Singh – B00955705 (pl414338@dal.ca)
Priyadharshini Sridharan – B00955679 (pr423614@dal.ca)
Syanthan Vullingala – B00952624 (sy454227@dal.ca)

Professor Evangelos Milios

## INDEX


| S No | Title | Page No |
|---|---|---|
| **1** | INTRODUCTION | 3 |
| **2** | BACKGROUND INFORMATION | 3 |
| **3** | PROBLEM STATEMENT | 3 |
| **4** | PROBLEM SIGNIFICANCE | 4 |
| **5** | PROJECT OBJECTIVES | 4 |
| **6** | PROJECT SCOPE | 4 |
| **7** | DATA COLLECTION & PREPARATION | 5 |
| **8** | DATA ANALYSIS | 7 |
| **9** | MODELING | 8 |
| **10** | RESULTS | 9 |
| **11** | INTERPRETATION OF RESULTS | 11 |
| **12** | IMPLICATIONS | 12 |
| **13** | LIMITATIONS | 12 |
| **14** | CONCLUSIONS | 12 |
| **15** | REFERENCES | 14 |
| **16** | APPENDIX | 17 |

## INTRODUCTION

Diabetes remains an important health challenge with its widespread presence steadily increasing around the world. It can lead to other serious issues such as cardiovascular diseases, renal failure, and neuropathy and contribute to rising mortality rates in diabetic patients. Considering the situation, accurately predicting mortality risks in diabetic patients is crucial for several reasons like identifying high-risk individuals, forming proper diabetic treatment options and mitigating mortality among older patients.

## BACKGROUND INFORMATION

Our project mainly focuses on Type 2 Diabetes Mellitus which generally occurs when insulin is not effectively used by the body due to excess body weight and physical inactivity. According to the reports of WHO (World Health Organisation) a total of around 1.5 million people died due to diabetes in the year 2019 and around 48% people are less than 70 years of age. T2DM constitutes more than 95% in the diabetic people and people effected with T2DM usually live 6 years lesser compared to healthy 50-year-old individuals as per the Centres for Disease Control and Prevention (CDC).[1]

Studies conducted by WHO and CDC indicate that risk of mortality is very high in diabetic patients and it's hard to predict the insulin amount required for each patient and it gets progressively harder when the patient has additional complications and comorbidities like HIV/Depression/Alcohol Abuse etc. Many influencing factors that could affect the diabetic complications need to be considered and hence there is a need to develop an all-cause mortality prediction model that could be utilised by health-practitioners for devising better diabetic treatment plans, identifying sensitive individuals and controlling the mortality rate.

## PROBLEM STATEMENT

Diabetes is a serious health condition which can lead to complications like vision loss, heart strokes, kidney failures and limb amputations. According to previous studies, this illness increases the risk of mortality and the onset of both microvascular and macrovascular complications. However, these risks can be reduced by following proper diabetic treatment plans and meeting treatment objectives. This leads us to our problem statement where accurately predicting remaining life expectancy remains a significant challenge.

Mortality prediction models that are developed in the past are failed to apply for wide range of populations, offering limited accuracy, especially in the context of multiclass classification. This limitation affects our expectations of forming optimized treatment plans and identifying high risk individuals, which shows the need for a more refined predictive model.

## PROBLEM SIGNIFICANCE

The significance of addressing this problem lies in the potential health benefits and improved quality of life for older adults with T2DM. By accurately predicting mortality risk, healthcare providers can implement more effective, personalized treatment plans, potentially reducing premature death rates and improving overall patient outcomes. Furthermore, this research could offer valuable insights into the factors most indicative of mortality risk, thereby guiding future healthcare strategies and policies.

## PROJECT OBJECTIVES

The main objective of this project is to develop a multiclass classification methodology to predict all-cause mortality in elderly individuals diagnosed with Type 2 Diabetes Mellitus, considering various individual-level risk factors. The aim is to develop a single mortality prediction model for patients aged 65 years or older with Type 2 Diabetes Mellitus, considering their associated comorbidities and complications. By comparing the performance of various machine learning algorithms, the project aims to identify the most effective method for accurately classifying patients into distinct risk categories, thereby facilitating more precise and effective treatment planning.

## PROJECT SCOPE

This project is specifically focused on older adults aged 65 and above with T2DM, leveraging a structured dataset of U.S. military Veterans. The study tracks the mortality status of patients at both the 5-year and 10-year intervals. The project however will not cover the patients with Type 1 Diabetes or gestational diabetes and usage of external datasets for the validation of model is also not within the scope of the project.

## METHODOLOGY

## DATASET COLLECTION AND PREPARATION

The data utilized in this study was meticulously selected from a publicly accessible dataset via the **Mendeley Data Repository**, comprising records of U.S. Military Veterans aged 65 years and above diagnosed with Type 2 Diabetes Mellitus (T2DM) over the years 2004 to 2016. It contains 275,190 rows and 70 columns consisting of demographics (Age, Gender, Race etc.), clinical biomarkers (BMI, Blood Pressure, cholesterol etc.) and comorbidities (AIDS/HIV, Blood Loss, Alcohol Abuse etc). [2]

### Creating the Target Variable:

A new column was created as ‘mortality’ based on DEATH_5’ and ‘DEATH_10’ columns. The variable ‘DEATH_5’ indicates if the remaining life of the patient is up to 5 years, ‘DEATH_10’ indicates if the remaining life of the patient is up to 10 years. (See Appendix Fig. 2 for distribution)

| Death_5 | Death_10 | Mortality Class | Description |
|---|---|---|---|
| 1 | 1 | Class 1 | Remaining life upto 10 years |
| 0 | 1 | Class 1 | Remaining life upto 10 years |
| 0 | 0 | Class 2 | Remaining life more than 10 years |

**Data Understanding:**

The dataset contains 68 continuous variables including the target variable (Float64(9), Int64(59)) and 2 categorical variables (String) originally (See Appendix Fig 3 for description). The Analytics Base Table report gives deeper insights about the continuous variables and categorical variables (See Appendix Fig.4 and Fig.5).

**Continuous Variables:**

ABT report of continuous variables show different attributes like mean, median, 1st quartile, 3rd quartile, cardinality, missing values, etc. It is observed that many of them have cardinality value '2'. They are considered binary variables, and the remaining are considered numerical variables. Analysis, visualizations and data preparation are performed separately for both these types.

Box plots are created for numerical variables to visualize their distribution across different classes of the target variable (See Appendix Fig.6). Bar plots are used to understand the distribution of target classes over the binary variable values (See Appendix Fig.7-59).

**Categorical Features:**

For the categorical features, first string columns are converted to categorical data type to facilitate their analysis. ABT report is created for categorical features to find out the mode value, mode frequency and proportion for each categorical feature. Bar plots are created to understand the distribution of the various classes of each categorical feature over the target classes (See Appendix Fig.60-61). The variable 'RACE {1,2,3}' is converted from numerical to categorical {'Black', 'White', 'Other'} since it has an extremely low cardinality value and can be logically classified as mentioned (See Appendix Fig. 62).

**Data Issues:**

1. **Missing Values:** It is identified from ABT report that columns 'MICROALB', 'SERUMALB' have a high proportion of missing values - 74.15% and 30.47% respectively. The variables 'HDL', 'LDL', 'SERUMCRE', 'TRI' also had a few missing values.

2. **Duplicates:** Some duplicate instances were observed in the dataset.

3. **Outliers:** The box plots of numerical variables clearly indicate the presence of Outliers in few predictor variables. These columns include “AGE”, “BMI” [7], “A1C” [8], “SERUMALB” [9], “SERUMCRE” [10], “N_IP”, “N_OP”, “SYSTOLIC” [6], “DIASTOLIC” [11], “TRI” [12], “LDL” [13], “HDL”, “FRAILTY”.

4. **Imbalance of Data:** Many binary variables show high imbalance in the True / False instance. Some have too many zeros and some have too many 1s that are sometimes distributed across target classes with very slight variations causing insignificant impact on the target variable (See Appendix Fig. 63).

**Data Quality Plan and Preparation:**

To mitigate the above-mentioned data issues, a Data Quality Plan is schemed (See Appendix Fig. 64-66). Major actions taken were:

- Removing Missing instances / features based on the proportion
- Binning variables with outliers
- Removing variables with too much imbalance in 0/1 values
- Removing the duplicate instances

**Data Analysis**

**Chi Square Test:**

The chi-square test was used to determine the strength of the relationship between the predictor variables and the target variable.[3]

**H0 (Null Hypothesis) -** There is no significant relationship between the predictor variables and target variable.
**H1 (Alternate Hypothesis) -** There is a significant relationship between the predictor variables and target variable.

Based on the hypothesis testing result, we have identified that "Sex" column's p value (0.03) is greater than the significance level. It shows that we failed to reject the null hypothesis for the "Sex" column. However, for the remaining columns, the p-values were found to be less than the significance level. (See Appendix Fig. 67)

**Cramer’s Test:**

The Cramer's V test was used 9us to gauge the strength of association between pairs of categorical predictors. In Cramer's V test, values range from 0 (indicating no association) to 1 (indicating a perfect association).[4]

Following the application of Cramer's V test, we found no significant correlations among the predictor variables. Additionally, we observed no significant associations between the predictor variables and the mortality target variable. (See Appendix Fig. 68)

**Phi coefficient:**

The Phi coefficient measures the strength and direction of association between two binary variables. In our case, the target variable is mortality, and each predictor is a binary variable indicating the presence or absence of a specific condition or characteristic. (See Appendix Fig. 69). For results inference, refer Appendix Fig. 84.[14]

**Feature Selection – IV & WOE:**

To enhance the predictive performance of the models regarding mortality, statistical measures Information Value (IV) and Weight of Evidence (WOE) were used for feature selection. Higher IV values indicate stronger predictive power. WOE, on the other hand, transforms raw features into a more meaningful form by capturing the relationship between each feature's categories and the likelihood of the target variable. After calculating IV the predictor variables were split int 5 categories. (See Appendix Fig. 85). All the "not useful" predictors are removed (See Appendix Fig. 70).[5]

Subsequently, selected predictors underwent one-hot encoding to prepare categorical data for modeling algorithms. Mutual Information (MI) was then used to identify best 20 features by using the SelectKBest method (See Appendix Fig. 71). This approach focussed on predictors with the greatest relevance to mortality, enhancing the effectiveness and interpretability of our predictive models.

**MODELING**

The dataset is divided into train and test with 25% used for testing and 75% for training. Baseline model (Majority Classifier), Linear models (Logistic Regression) and Tree-based models (Random Forest, XGBoost, Decision Tree) were employed to perform classification.

**Baseline Model - Majority Classifier:**

A Majority Class Classifier is used as a baseline model as it can be used as straightforward but insightful benchmark against which the performance of more intricate models can be measured. This model is used when there is imbalance in the target variable, and it is computationally inexpensive. It is implemented using the 'DummyClassifier'. [15]

**Candidate Models**

**Logistic Regression:**
Logistic Regression model was employed since the target variable is binary variable and it is more interpretable for binary classification tasks. 'L2-regularization' is used in the model to reduce overfitting.[16]

**Random Forest Classifier:**
Random Forest Classifier being an ensemble technique gives high predictive accuracy and is robust against noise and outliers. It also reduces overfitting and hence is employed for this classification task. Here, 'max_depth' was restricted to 15 to reduce complexity of individual trees and hence overfitting.[17]

**Random Forest Classifier with Hyperparameter Tuning:**
Going a step ahead, the hyperparameters of Random Forest are tuned using Randomized Search Cross Validation to get the best parameters for the model to obtain higher accuracy. The model is then built using the best set of parameters returned by Randomized Search (See Appendix Table. 1).

**XGBoost Classifier:**
XGBoosting, another ensembling technique that works by improvising the weak learners is used to capture the complex interactions in the data and is hence employed in this binary classification. The data is first sanitized to remove >, < symbols from feature names to avoid parsing errors while building XGB Classifier. XGB Classifier is built by tuning parameters using Randomized Search CV (See Appendix Table. 1).[18]

**Decision Tree Classifier:**
Decision Trees are known for their interpretability and ability to accommodate non-linear relationship between target and predictor variables. They use Gini index or entropy to split the data based on the most importance features. Decision tree model is built with Grid Search hyper parameter tuning technique (See Appendix Table 1).[19]

**Model Evaluation:**
These models are the compared and evaluated using the following metrics:

- Training and test accuracy
- Confusion Matrix (TPR, FPR, Recall, Precision, F1- Score)
- AIC Score
- AUC (Area Under ROC Curve)
- Cohen Kappa

**RESULTS**:

**MODEL PERFORMANCE ANALYSIS**:

The Baseline Model 'Majority Classifier' has an accuracy of 57% on test data which is much less compared to all the other candidate models. For evaluation the results, Training and Test Accuracy Curve, Confusion Matrix [20], AIC, AUC score and Kappa Score are considered.

1. **Training and Test Accuracy Curves:**

The training and testing accuracy scores differ marginally for all the models especially Logistic regression, XGBoost and Hyper parameter tuned Random Forest. The testing accuracy is consistently around 67%-69% for all the models. Also, the difference between the training and

testing accuracy is also negligible, merely 1% to 2%. Hence, it can be inferred that the model is performing consistently well on both training and test data. There is no overfitting in the plot for any of the models. The best training and testing accuracy curves are seen for logistic regression where they almost overlap with each other. But this measure is not sufficient to decide the best model. Other metrics are also considered to evaluate the model. (See Appendix Fig. 72-75).

2. **Inference based on confusion matrix:**

**Logistic Regression (Accuracy – 68%):**

- Moderate TPR and FPR.
- Balanced precision and recall.
- Overall F1-score is reasonable but could be improved (See Appendix Fig. 76, Fig. 81).

**Random Forest (Accuracy - 67%):**

- Similar performance to logistic regression, with slightly lower TPR and slightly higher FPR.
- Balanced precision and recall.
- F1-score is comparable to logistic regression (See Appendix Fig. 77, Fig.81).

**XGBoost (Accuracy – 68%):**

- Similar performance to logistic regression and random forest, with slightly lower TPR and FPR.
- Balanced precision and recall.
- F1-score is comparable to logistic regression and random forest (See Appendix Fig. 78, Fig. 81).

**Random Forest Hyperparameter Tuned (Accuracy – 68%):**

- Similar performance to the regular random forest model.
- Slightly higher TPR and slightly lower FPR compared to other models.
- Balanced precision and recall (See Appendix Fig.79, Fig.81).

**Decision Tree (Accuracy – 67%):**

- Slightly higher FPR and similar TPR compared to other models
- Balanced precision and recall
- F1-score is comparable to other models (See Appendix Fig. 80, Fig. 81)

From these results, it can be inferred that all models have similar recall rates, but Logistic Regression and XGBoost slightly outperform the others in terms of overall F1 score. This could suggest they have a better balance of false positives and false negatives in this specific scenario. However, the differences are quite minimal, indicating that all models perform similarly on this dataset.

3. **Inference based on AKAIKE INFORMATION CRITERION** (AIC)**:**

The Akaike Information Criterion (AIC) is a widely used measure for model selection that is based on the concept of information entropy. AIC assesses the goodness of fit of the model while penalizing for the number of parameters and it is used for comparing different models on the same dataset. A lower AIC value indicates a better model. It was observed that XGBoost model has the lowest AIC Score (See Appendix Fig. 82). [21]

4. **Inference based on AUC (Area Under the Receiver Operating Characteristic) Curve:**

The AUC score, or Area Under the Receiver Operating Characteristic (ROC) Curve, is a performance metric for binary classification problems. The AUC score measures the overall ability of the model to discriminate between positive and negative classes. Logistic Regression model has the highest AUC score (0.674), which means it performs best at distinguishing between the two classes out of the models listed. (See Appendix Fig.83.) [22]

5. **Inference based on Cohen's Kappa:**

Cohen's Kappa score [23] is used to measure the level of agreement between two raters (or models) who each classify the categories. To the overall consistency of the models, we determined Cohen kappa values between the models.

Inferences:

- The models Logistic Regression and XGBoost have the highest agreement with 96% similarity in predictions.
- Logistic regression - Random Forest hyperparameter tuned and XGBoost - Random Forest also have a good agreement of 92% for both.
- Overall, all the models have showcased a high consistence in making predictions with above 79% between every combination.

**INTERPRETATION OF RESULTS:**

The purpose of the analysis was to create a predictive model for mortality in older people with Type 2 Diabetes Mellitus (T2DM) to identify important clinical biomarkers and high-risk individuals. Some interesting clinical biomarkers like peripheral vascular diseases, congestive heart failure, cardiac arrhythmias, age, and serum creatinine levels were found to be significant predictors of death. This emphasizes how important heart-related illnesses are to the mortality associated with T2DM.

The prediction results indicated that the 80–84 age group was particularly vulnerable, indicating a higher likelihood of death in this group. All things considered, these results highlight the significance of thorough cardiac evaluation and vigilant serum creatinine level monitoring to lower the risk of death in older type 2 diabetes patients.

**IMPLICATIONS:**

It's important to recognize that, despite the fact that our investigation has found important factors linked to mortality in senior people with Type 2 Diabetes Mellitus (T2DM), the machine learning models did not attain adequate accuracy. As a result, further investigation and refinement of the predictive models are necessary to fulfill the objective of our problem statement.

Involving domain experts in the analysis process can provide in-depth insights on business logic (Clinical reasoning) and enhance the feature selection process. By collaborating with experts in the field, we can refine our model and incorporate additional factors that may influence mortality risk in T2DM patients.

**LIMITATIONS:**

The fact that our study relies on a dataset that is restricted to T2DM-afflicted U.S. military veterans 65 years of age and older limits the applicability of our findings to larger populations. The lack of representation from other diverse demographics or healthcare systems may limit the applicability of our results to non-veteran elderly individuals or those from different countries.

Despite efforts to address data issues such as missing values and outliers, residual biases, errors incurred during the study or other inaccuracies may persist in the dataset. These flaws might affect the predictive models' accuracy and the validity of the results, which could result in erroneous interpretations.

Alternative algorithms and methods may yield different results, and exploring additional approaches could enhance the robustness of our study. Additionally, various statistical tests were used for identifying relationship between predictors and target variables. Most of the used statistical tests are for finding linear relationship, we could have explored more statistical tests which can capture non-linear relationship.

## CONCLUSION

**Short Summary of Key Findings:**

The study aimed to develop a predictive model for mortality among elderly individuals with Type 2 Diabetes Mellitus (T2DM), utilizing machine learning techniques. Through comprehensive analysis, significant predictors such as age, cardiovascular diseases, and serum creatinine levels were identified, providing valuable insights into factors influencing mortality risk in this population. Notably, peripheral vascular diseases, congestive heart failure, and cardiac arrhythmias emerged as critical predictors, highlighting the importance of cardiac evaluation in managing mortality risk in T2DM patients.

Furthermore, the models exhibited moderate performance. Evaluation of model performance using metrics such as confusion matrices, AIC, AUC scores, and Cohen’s Kappa provided valuable insights into the effectiveness of different modelling approaches. While all

models exhibited similar recall rates, Logistic Regression and XGBoost showed slightly superior performance in terms of overall F1 score. Additionally, the AIC criterion favoured the Logistic Regression model, indicating its potential as the preferred model for mortality prediction. However, further refinement and exploration of alternative algorithms could enhance model accuracy and reliability.

**Achievement of Objectives:**

Our project effectively addressed the problem statement by developing a predictive model for mortality in elderly T2DM patients and identifying key predictors associated with mortality risk. Through rigorous analysis and modeling, we contributed valuable insights into the clinical management of T2DM patients and provided a foundation for future research in this area. However, our study encountered some limitations, including reliance on a specific dataset (limited to U.S. military veterans), which may restrict the generalizability of our findings to broader populations.

Despite these limitations, our project lays the groundwork for future studies aimed at improving mortality prediction and clinical decision-making in T2DM management. Moving forward, incorporating diverse datasets and exploring alternative modeling techniques could enhance the robustness and applicability of our findings. One example could be adding and removing features based on mutual information values and checking model performance. Overall, our project represents a significant step towards addressing the challenges associated with mortality prediction in elderly T2DM patients and underscores the importance of interdisciplinary collaboration in healthcare research.

## REFERENCES

1. World Health Organization. (2023, April 5). *Diabetes*. https://www.who.int/news-room/fact-sheets/detail/diabetes
2. Kevin Griffith. (2021, December 21). A *Novel Dataset of Predictors of Mortality for Older Veterans Living with Type II Diabetes*. Mendeley Data. https://data.mendeley.com/datasets/kn8v3678n9/3
3. Sage Pub. *PEARSON'S R, CHI-SQUARE, T-TEST, AND ANOVA*. https://www.sagepub.com/sites/default/files/upm-binaries/33663_Chapter4.pdf
4. Stats Test. *Cramer's V.* https://www.statstest.com/cramers-v-2/
5. Anik Chakraborty. (2021, September 5). *Weight of Evidence (WoE) and Information Value (IV) — how to use it in EDA and Model Building?* Medium. https://anikch.medium.com/weight-of-evidence-woe-and-information-value-iv-how-to-use-it-in-eda-and-model-building
6. Centers for Disease Control and Prevention. (2021, May 18). *High Blood Pressure Symptoms and Causes*. https://www.cdc.gov/bloodpressure/about.htm
7. Centers for Disease Control and Prevention. (2022, June 3). *Assessing Your Weight.* https://www.cdc.gov/healthyweight/assessing/index.html
8. Centers for Disease Control and Prevention. (2022, September 30). *All About Your A1C.* https://www.cdc.gov/diabetes/managing/managing-blood-sugar/a1c.html
9. University of Rochester Medical Center. *Albumin (Blood). Health Encyclopedia.* https://www.urmc.rochester.edu/encyclopedia/content.aspx?contenttypeid=167&contentid=albumin_blood
10. Mount Sinai. *Creatinine blood test.* https://www.mountsinai.org/health-library/tests/creatinine-blood-test

11. Centers for Disease Control and Prevention. (2021, May 18). *High Blood Pressure Symptoms and Causes.* https://www.cdc.gov/bloodpressure/about.htm
12. Mayo Clinic. (2022, September 3). *Triglycerides: Why do they matter?* https://www.mayoclinic.org/diseases-conditions/high-blood-cholesterol/in-depth/triglycerides/art-20048186
13. Cleveland Clinic. (2022, July 28). *Cholesterol Numbers and What They Mean.* https://my.clevelandclinic.org/health/articles/11920-cholesterol-numbers-what-do-they-mean
14. Statistics How To. *Phi Coefficient (Mean Square Contingency Coefficient).* https://www.statisticshowto.com/phi-coefficient-mean-square-contingency-coefficient/
15. Scikit Learn. *Dummy Classifier.* https://scikit-learn.org/stable/modules/generated/sklearn.dummy.DummyClassifier.html
16. Anshul Saini. (2024, April 4). *A Beginner's Guide to Logistic Regression.* Analytics Vidhya. https://www.analyticsvidhya.com/blog/2021/08/conceptual-understanding-of-logistic-regression-for-data-science-beginners/
17. Sruthi E R. (2024, January 3). *Understand Random Forest Algorithms with Examples.* Analytics Vidhya. https://www.analyticsvidhya.com/blog/2021/06/understanding-random-forest/
18. Guest Blog. (2024, April 4). *Introduction to XGBoost Algorithm in Machine Learning.* Analytics Vidhya. https://www.analyticsvidhya.com/blog/2018/09/an-end-to-end-guide-to-understand-the-math-behind-xgboost/
19. Anshul Saini. (2024, January 5). *Decision Tree – A Step-by-Step Guide*. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2021/08/decision-tree-algorithm/

*20.* Sarang Narkhede. (2018, May 9). *Understanding Confusion Matrix.* Medium. https://towardsdatascience.com/understanding-confusion-matrix-a9ad42dcfd62

21. Bonakdari. H, Zeynoddin. M (2022). *Goodness of Fit & Precision Criteria*. Science Direct. https://www.sciencedirect.com/science/article/abs/pii/B9780323917483000033

22. Sarang Narkhede. (2018, June 26). *Understanding AUC - ROC Curve.* Medium https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5

23. Data Tab. (2024). *Cohen's Kappa.* https://datatab.net/tutorial/cohens-kappa

## APPENDIX

### Tasks execution per team member(s):

| Tasks | Members |
|---|---|
| Brainstorming | Pallavi, Priyadharshini, Syanthan |
| Dataset Collection | Pallavi, Priyadharshini, Syanthan |
| Data Preparation | Pallavi, Priyadharshini, Syanthan |
| Feature Engineering | Pallavi, Priyadharshini |
| Modelling | Pallavi, Priyadharshini |
| Evaluation | Pallavi, Priyadharshini |
| Report Writing | Pallavi, Priyadharshini, Syanthan |
| Presentation | Pallavi, Priyadharshini, Syanthan |

### Figures and Tables:

### Fig 1 – PREDICTORS DESCRIPTION

| # | Category | Attributes | Variable Name |
|---|---|---|---|
| 1 | **Demographics** | Age (in years) | AGE |
| 2 | | Gender | SEX |
| 3 | | Marital Status | MARRIED |
| 4 | | Race | RACE |
| 5 | | Enrolment priority | PRIORITY |
| 6 | **Clinical Biomarkers** | Body Mass Index (BMI) | BMI |
| 7 | | Diastolic blood pressure | DIASTOLIC |
| 8 | | Systolic blood pressure | SYSTOLIC |
| 9 | | HDL Cholesterol | HDL |
| 10 | | LDL Cholesterol | LDL |
| 11 | | Hemoglobin A1c | A1C |
| 12 | | Triglycerides | TRI |
| 13 | | Urine microalbumin | MICROALB |
| 14 | | Serum albumin | SERUMALB |

| | | | |
|---|---|---|---|
| 15 | | Serum creatinine | SERUMCRE |
| 16 | **Prescriptions of Diabetes Medication** | Alpha-glucosidase inhibitor prescription | ALPHA |
| 17 | | Biguanide prescription | BIGUAN |
| 18 | | Insulin prescription | INSULIN |
| 19 | | Sulfonylurea prescription | SULF |
| 20 | | Thiazolidinedione prescription | TZD |
| 21 | | Blood Pressure prescription | BP_RX |
| 22 | | Other diabetes medication | OTHER_MED |
| 23 | **Comorbidities** | Acute myocardial infarction | AMI |
| 24 | | AIDS/HIV | HIV |
| 25 | | Alcohol abuse | ALCOHOL |
| 26 | | Ankle-brachial index | ABI |
| 27 | | Blood loss anemia | BLOODLOSS |
| 28 | | Cardiac arrhythmias | ARRHYTHMIA |
| 29 | | Chronic pulmonary disease | PULMONARY |

| # | Category | Attributes | Variable Name |
|---|---|---|---|
| 30 | **Comorbidities** | Coagulopathy | COAG |
| 31 | | Congestive heart failure | CHF |
| 32 | | Coronary arterial disease | CAD |
| 33 | | Coronary Artery Bypass Grafting | CABG |
| 34 | | Deficiency anemia | ANEMIA |
| 35 | | Depression | DEPRESSION |
| 36 | | Diabetes, complicated | DMCX |
| 37 | | Diabetic foot infections | FEET |
| 38 | | Drug abuse | DRUGS |
| 39 | | End-stage liver disease for cirrhosis or alcoholic fatty liver | ESLD |
| 40 | | Fluid & electrolyte disorders | FLUIDSLYTES |
| 41 | | Hyperglycemia | HYPERG |
| 42 | | Hypertension, complicated | HTNCX |

| | | | |
|---|---|---|---|
| 43 | | Hypertension, uncomplicated | HTN |
| 44 | | Hypothyroidism | HYPOTHYROID |
| 45 | | Liver disease | LIVER |
| 46 | | Lower limb amputation | AMPUTATION |
| 47 | | Lymphoma | LYMPHOMA |
| 48 | | Metastatic cancer | METS |
| 49 | | Obesity | OBESITY |
| 50 | | Other neurological disorders | NEUROOTHER |
| 51 | | Paralysis | PARALYSIS |
| 52 | | Peptic ulcer disease excluding bleeding | PUD |
| 53 | | Percutaneous coronary interventions | PCI |
| 54 | | Peripheral vascular disorders | PVD |
| 55 | | Psychoses | PSYCHOSES |
| 56 | | Pulmonary circulation disorders | PHTN |
| 57 | | Renal failure | RENAL |

| # | Category | Attributes | Variable Name |
|---|---|---|---|
| 58 | **Comorbidities** | Retinopathy | RETINOPATHY |
| 59 | | Retinopathy screening | RETSCREEN |
| 60 | | Rheumatoid arthritis/collagen vascular diseases | RHEUMATIC |
| 61 | | Severe depression | SEVERE_DEP |
| 62 | | Smoking | SMOKER |
| 63 | | Solid tumor without metastasis | TUMOR |
| 64 | | Valvular disease | VALVULAR |
| 65 | | Weight loss | WEIGHTLOSS |
| 66 | **Others** | Frailty Index | FRAILTY |
| 67 | | Inpatient days | N_IP |
| 68 | | Outpatient visits | N_OP |

**Fig 2**

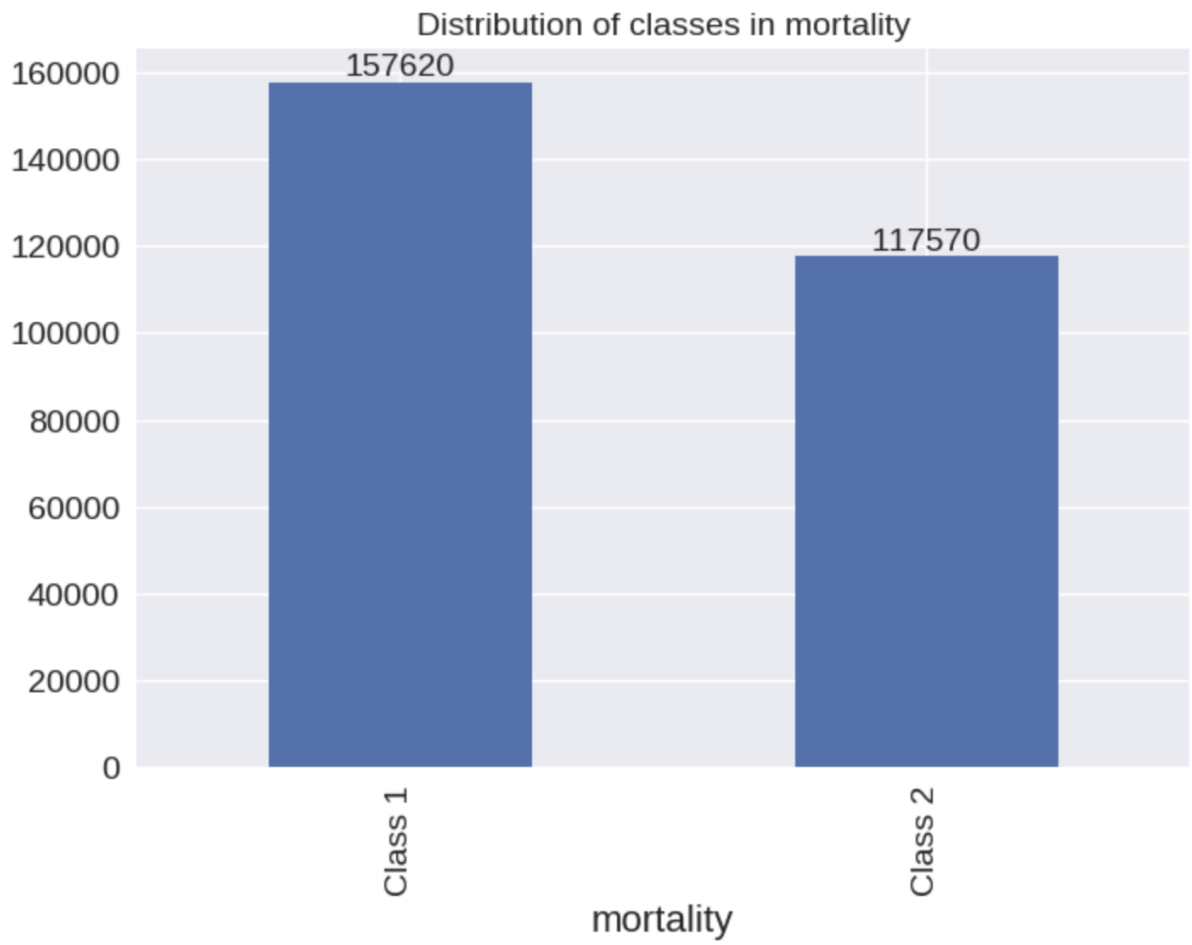



**Fig 3 – DATA INFORMATION**

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 275190 entries, 0 to 275189
Data columns (total 70 columns):
 #   Column       Non-Null Count   Dtype
---  ------       --------------   -----
 0   SYSTOLIC     275190 non-null  Int64
 1   A1C          275190 non-null  Float64
 2   PRIORITY     275190 non-null  string
 3   AGE          275190 non-null  Int64
 4   BMI          275190 non-null  Float64
 5   MARRIED      275190 non-null  string
 6   HDL          267456 non-null  Float64
 7   LDL          257953 non-null  Float64
 8   DIASTOLIC    275190 non-null  Int64
 9   TRI          267886 non-null  Float64
 10  SERUMCRE     263433 non-null  Float64
 11  MICROALB     71130 non-null   Float64
 12  SERUMALB     191327 non-null  Float64
 13  N_IP         275190 non-null  Int64
 14  N_OP         275190 non-null  Int64
 15  INSULIN      275190 non-null  Int64
 16  SULF         275190 non-null  Int64
 17  OTHER_MED    275190 non-null  Int64
 18  BIGUAN       275190 non-null  Int64
 19  TZD          275190 non-null  Int64
 20  ALPHA        275190 non-null  Int64
 21  SEX          275190 non-null  Int64
 22  RACE         275190 non-null  Int64
 23  DEATH_5      275190 non-null  Int64
 24  DEATH_10     275190 non-null  Int64
 25  CHF          275190 non-null  Int64
 26  ARRHYTHMIA   275190 non-null  Int64
 27  VALVULAR     275190 non-null  Int64
 28  PHTN         275190 non-null  Int64
 29  PVD          275190 non-null  Int64
 30  HTN          275190 non-null  Int64
 31  HTNCX        275190 non-null  Int64
 32  PARALYSIS    275190 non-null  Int64
 33  NEUROOTHER   275190 non-null  Int64
 34  PULMONARY    275190 non-null  Int64
 35  DMCX         275190 non-null  Int64
 36  HYPOTHYROID  275190 non-null  Int64
 37  RENAL        275190 non-null  Int64
 38  LIVER        275190 non-null  Int64
 39  PUD          275190 non-null  Int64
 40  HIV          275190 non-null  Int64
 41  LYMPHOMA     275190 non-null  Int64
 42  METS         275190 non-null  Int64
 43  TUMOR        275190 non-null  Int64
 44  RHEUMATIC    275190 non-null  Int64
 45  COAG         275190 non-null  Int64
 46  OBESITY      275190 non-null  Int64
 47  WEIGHTLOSS   275190 non-null  Int64
 48  FLUIDSLYTES  275190 non-null  Int64
 49  BLOODLOSS    275190 non-null  Int64
 50  ANEMIA       275190 non-null  Int64
 51  ALCOHOL      275190 non-null  Int64
 52  DRUGS        275190 non-null  Int64
 53  PSYCHOSES    275190 non-null  Int64
 54  DEPRESSION   275190 non-null  Int64
 55  SEVERE_DEP   275190 non-null  Int64
 56  CAD          275190 non-null  Int64
 57  AMI          275190 non-null  Int64
 58  RETINOPATHY  275190 non-null  Int64
 59  HYPERG       275190 non-null  Int64
 60  ABI          275190 non-null  Int64
 61  RETSCREEN    275190 non-null  Int64
 62  CABG         275190 non-null  Int64
 63  PCI          275190 non-null  Int64
 64  SMOKER       275190 non-null  Int64
 65  ESLD         275190 non-null  Int64
 66  FEET         275190 non-null  Int64
 67  AMPUTATION   275190 non-null  Int64
 68  FRAILTY      275190 non-null  Float64
 69  BP_RX        275190 non-null  Int64
dtypes: Float64(9), Int64(59), string(2)
memory usage: 164.8 MB
```

**Fig 4 - ABT REPORT CONTINUOUS FEATURES**

| | Count | Miss % | Card. | Min | 1st Qrt. | Mean | Median | 3rd Qrt | Max | Std. Dev. |
|---|---|---|---|---|---|---|---|---|---|---|
| **SYSTOLIC** | 275190 | 0.00 | 144 | 79.00 | 128.00 | 136.84 | 136.00 | 145.00 | 240.00 | 13.27 |
| **A1C** | 275190 | 0.00 | 5128 | 1.20 | 6.30 | 7.08 | 6.85 | 7.60 | 19.73 | 1.17 |
| **AGE** | 275190 | 0.00 | 36 | 65.00 | 71.00 | 75.21 | 75.00 | 79.00 | 101.00 | 5.58 |
| **BMI** | 275190 | 0.00 | 514 | 10.50 | 26.50 | 30.00 | 29.40 | 32.80 | 92.30 | 5.09 |
| **HDL** | 275190 | 2.81 | 8099 | 10.00 | 33.60 | 40.86 | 39.33 | 46.50 | 90.00 | 10.45 |
| **LDL** | 275190 | 6.26 | 13435 | 20.00 | 75.75 | 92.56 | 90.67 | 107.00 | 220.00 | 25.11 |
| **DIASTOLIC** | 275190 | 0.00 | 84 | 36.00 | 65.00 | 70.38 | 70.00 | 75.00 | 144.00 | 7.71 |
| **TRI** | 275190 | 2.65 | 11685 | 10.50 | 100.67 | 164.00 | 141.67 | 200.25 | 2000.00 | 96.35 |
| **SERUMCRE** | 275190 | 4.27 | 7263 | 0.10 | 1.00 | 1.26 | 1.15 | 1.39 | 14.57 | 0.51 |
| **MICROALB** | 275190 | 74.15 | 12192 | 0.00 | 7.50 | 65.06 | 16.60 | 44.60 | 4957.16 | 205.64 |
| **SERUMALB** | 275190 | 30.47 | 2836 | 1.44 | 3.80 | 4.01 | 4.02 | 4.25 | 5.00 | 0.36 |
| **N_IP** | 275190 | 0.00 | 391 | 0.00 | 0.00 | 2.36 | 0.00 | 0.00 | 730.00 | 18.83 |
| **N_OP** | 275190 | 0.00 | 407 | 1.00 | 15.00 | 30.88 | 23.00 | 38.00 | 630.00 | 26.85 |
| **INSULIN** | 275190 | 0.00 | 2 | 0.00 | 0.00 | 0.16 | 0.00 | 0.00 | 1.00 | 0.36 |
| **SULF** | 275190 | 0.00 | 2 | 0.00 | 0.00 | 0.45 | 0.00 | 1.00 | 1.00 | 0.50 |
| **OTHER_MED** | 275190 | 0.00 | 2 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 1.00 | 0.04 |
| **BIGUAN** | 275190 | 0.00 | 2 | 0.00 | 0.00 | 0.42 | 0.00 | 1.00 | 1.00 | 0.49 |
| **TZD** | 275190 | 0.00 | 2 | 0.00 | 0.00 | 0.10 | 0.00 | 0.00 | 1.00 | 0.30 |
| **ALPHA** | 275190 | 0.00 | 2 | 0.00 | 0.00 | 0.01 | 0.00 | 0.00 | 1.00 | 0.11 |
| **SEX** | 275190 | 0.00 | 2 | 0.00 | 1.00 | 0.99 | 1.00 | 1.00 | 1.00 | 0.11 |
| **RACE** | 275190 | 0.00 | 3 | 1.00 | 1.00 | 1.17 | 1.00 | 1.00 | 3.00 | 0.45 |
| **CHF** | 275190 | 0.00 | 2 | 0.00 | 0.00 | 0.19 | 0.00 | 0.00 | 1.00 | 0.39 |
| **ARRHYTHMIA** | 275190 | 0.00 | 2 | 0.00 | 0.00 | 0.26 | 0.00 | 1.00 | 1.00 | 0.44 |
| **VALVULAR** | 275190 | 0.00 | 2 | 0.00 | 0.00 | 0.12 | 0.00 | 0.00 | 1.00 | 0.33 |
| **PHTN** | 275190 | 0.00 | 2 | 0.00 | 0.00 | 0.02 | 0.00 | 0.00 | 1.00 | 0.15 |
| **PVD** | 275190 | 0.00 | 2 | 0.00 | 0.00 | 0.23 | 0.00 | 0.00 | 1.00 | 0.42 |
| **HTN** | 275190 | 0.00 | 2 | 0.00 | 1.00 | 0.92 | 1.00 | 1.00 | 1.00 | 0.28 |
| **HTNCX** | 275190 | 0.00 | 2 | 0.00 | 0.00 | 0.11 | 0.00 | 0.00 | 1.00 | 0.31 |
| **PARALYSIS** | 275190 | 0.00 | 2 | 0.00 | 0.00 | 0.01 | 0.00 | 0.00 | 1.00 | 0.11 |
| **NEUROOTHER** | 275190 | 0.00 | 2 | 0.00 | 0.00 | 0.05 | 0.00 | 0.00 | 1.00 | 0.22 |
| **PULMONARY** | 275190 | 0.00 | 2 | 0.00 | 0.00 | 0.24 | 0.00 | 0.00 | 1.00 | 0.43 |
| **DMCX** | 275190 | 0.00 | 2 | 0.00 | 0.00 | 0.40 | 0.00 | 1.00 | 1.00 | 0.49 |
| **HYPOTHYROID** | 275190 | 0.00 | 2 | 0.00 | 0.00 | 0.12 | 0.00 | 0.00 | 1.00 | 0.32 |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **RENAL** | 275190 | 0.00 | 2 | 0.00 | 0.00 | 0.12 | 0.00 | 0.00 | 1.00 | 0.33 |
| **LIVER** | 275190 | 0.00 | 2 | 0.00 | 0.00 | 0.03 | 0.00 | 0.00 | 1.00 | 0.16 |
| **PUD** | 275190 | 0.00 | 2 | 0.00 | 0.00 | 0.03 | 0.00 | 0.00 | 1.00 | 0.16 |
| **HIV** | 275190 | 0.00 | 2 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 1.00 | 0.03 |
| **LYMPHOMA** | 275190 | 0.00 | 2 | 0.00 | 0.00 | 0.01 | 0.00 | 0.00 | 1.00 | 0.09 |
| **METS** | 275190 | 0.00 | 2 | 0.00 | 0.00 | 0.01 | 0.00 | 0.00 | 1.00 | 0.09 |
| **TUMOR** | 275190 | 0.00 | 2 | 0.00 | 0.00 | 0.17 | 0.00 | 0.00 | 1.00 | 0.38 |
| **RHEUMATIC** | 275190 | 0.00 | 2 | 0.00 | 0.00 | 0.03 | 0.00 | 0.00 | 1.00 | 0.18 |
| **COAG** | 275190 | 0.00 | 2 | 0.00 | 0.00 | 0.05 | 0.00 | 0.00 | 1.00 | 0.21 |
| **OBESITY** | 275190 | 0.00 | 2 | 0.00 | 0.00 | 0.21 | 0.00 | 0.00 | 1.00 | 0.40 |
| **WEIGHTLOSS** | 275190 | 0.00 | 2 | 0.00 | 0.00 | 0.03 | 0.00 | 0.00 | 1.00 | 0.18 |
| **FLUIDSLYTES** | 275190 | 0.00 | 2 | 0.00 | 0.00 | 0.12 | 0.00 | 0.00 | 1.00 | 0.33 |
| **BLOODLOSS** | 275190 | 0.00 | 2 | 0.00 | 0.00 | 0.01 | 0.00 | 0.00 | 1.00 | 0.12 |
| **ANEMIA** | 275190 | 0.00 | 2 | 0.00 | 0.00 | 0.08 | 0.00 | 0.00 | 1.00 | 0.27 |
| **ALCOHOL** | 275190 | 0.00 | 2 | 0.00 | 0.00 | 0.02 | 0.00 | 0.00 | 1.00 | 0.14 |
| **DRUGS** | 275190 | 0.00 | 2 | 0.00 | 0.00 | 0.01 | 0.00 | 0.00 | 1.00 | 0.09 |
| **PSYCHOSES** | 275190 | 0.00 | 2 | 0.00 | 0.00 | 0.04 | 0.00 | 0.00 | 1.00 | 0.19 |
| **DEPRESSION** | 275190 | 0.00 | 2 | 0.00 | 0.00 | 0.14 | 0.00 | 0.00 | 1.00 | 0.35 |
| **SEVERE_DEP** | 275190 | 0.00 | 2 | 0.00 | 0.00 | 0.02 | 0.00 | 0.00 | 1.00 | 0.15 |
| **CAD** | 275190 | 0.00 | 2 | 0.00 | 0.00 | 0.41 | 0.00 | 1.00 | 1.00 | 0.49 |
| **AMI** | 275190 | 0.00 | 2 | 0.00 | 0.00 | 0.09 | 0.00 | 0.00 | 1.00 | 0.29 |
| **RETINOPATHY** | 275190 | 0.00 | 2 | 0.00 | 0.00 | 0.22 | 0.00 | 0.00 | 1.00 | 0.41 |
| **HYPERG** | 275190 | 0.00 | 2 | 0.00 | 0.00 | 0.03 | 0.00 | 0.00 | 1.00 | 0.17 |
| **ABI** | 275190 | 0.00 | 2 | 0.00 | 0.00 | 0.05 | 0.00 | 0.00 | 1.00 | 0.23 |
| **RETSCREEN** | 275190 | 0.00 | 2 | 0.00 | 0.00 | 0.20 | 0.00 | 0.00 | 1.00 | 0.40 |
| **CABG** | 275190 | 0.00 | 2 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 1.00 | 0.04 |
| **PCI** | 275190 | 0.00 | 2 | 0.00 | 0.00 | 0.01 | 0.00 | 0.00 | 1.00 | 0.07 |
| **SMOKER** | 275190 | 0.00 | 2 | 0.00 | 0.00 | 0.12 | 0.00 | 0.00 | 1.00 | 0.32 |
| **ESLD** | 275190 | 0.00 | 2 | 0.00 | 0.00 | 0.01 | 0.00 | 0.00 | 1.00 | 0.09 |
| **FEET** | 275190 | 0.00 | 2 | 0.00 | 0.00 | 0.07 | 0.00 | 0.00 | 1.00 | 0.26 |
| **AMPUTATION** | 275190 | 0.00 | 2 | 0.00 | 0.00 | 0.01 | 0.00 | 0.00 | 1.00 | 0.12 |
| **FRAILTY** | 275190 | 0.00 | 24 | 0.00 | 0.00 | 0.90 | 0.62 | 1.56 | 7.19 | 1.01 |
| **BP_RX** | 275190 | 0.00 | 2 | 0.00 | 1.00 | 0.93 | 1.00 | 1.00 | 1.00 | 0.25 |

**Fig 5 – ABT REPORT FOR CATEGORICAL VARIABLES**

| | Count | Miss % | Card. | Mode | Mode Freq | Mode % | 2nd Mode | 2nd Mode Freq | 2nd Mode % |
|---|---|---|---|---|---|---|---|---|---|
| **PRIORITY** | 275190 | 0.00 | 9 | GROUP 5 | 105228 | 38.24 | GROUP 8 | 95068 | 34.55 |
| **MARRIED** | 275190 | 0.00 | 3 | MARRIED | 178645 | 64.92 | WIDOWED | 53290 | 19.36 |
| **RACE** | 275190 | 0.00 | 3 | White | 238067 | 86.51 | Black | 28442 | 10.34 |

**Fig 6 – BOX PLOT FOR NUMERICAL VARIABLES**

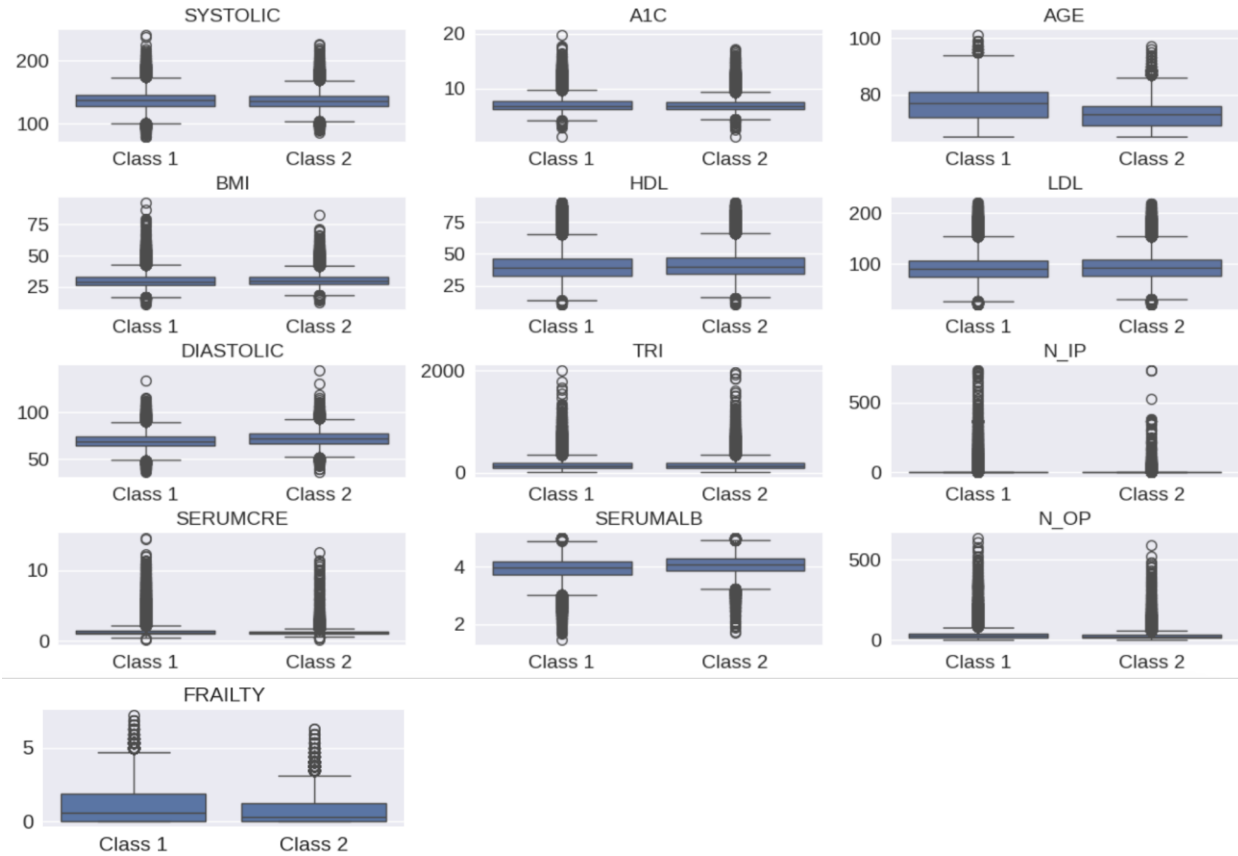



**Fig 7**

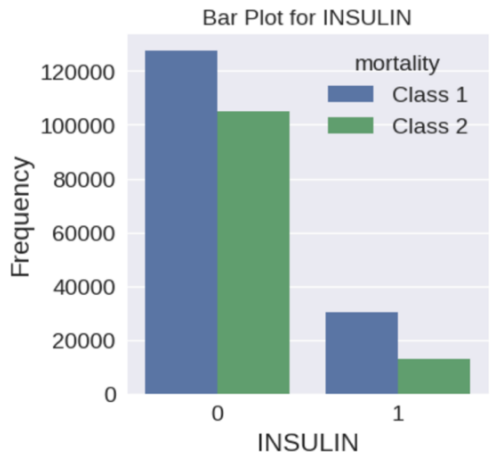



**Fig 8**

**Fig 9**

**Fig 10**

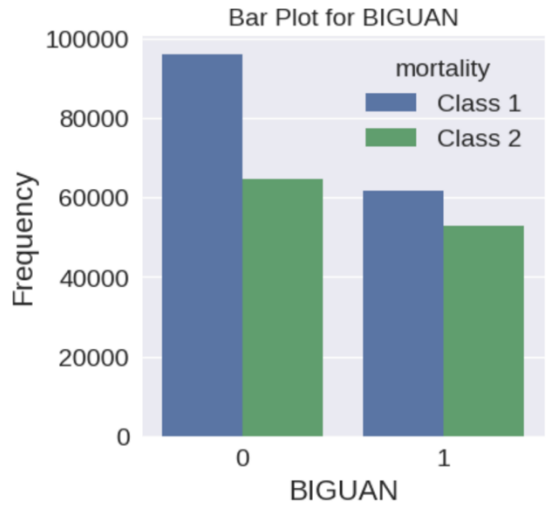



**Fig 11**

**Fig 12**

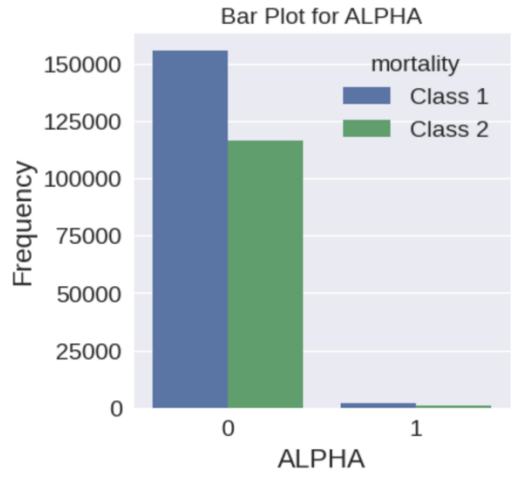



**Fig 13**

**Fig 14**

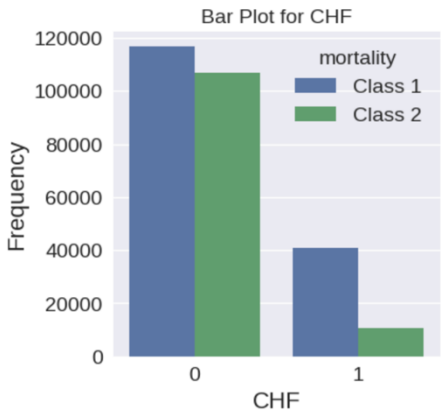



**Fig 15**

**Fig 16**

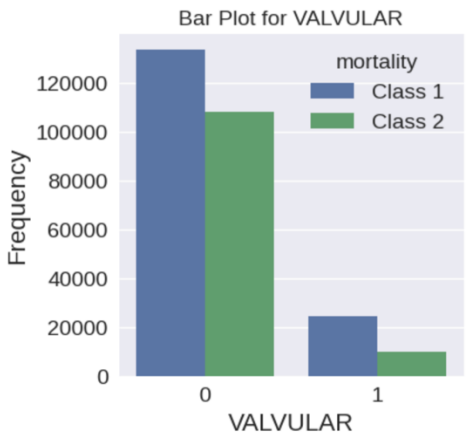



**Fig 17**

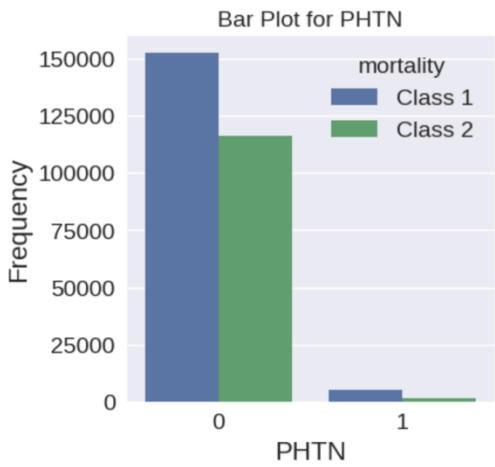



**Fig 18**

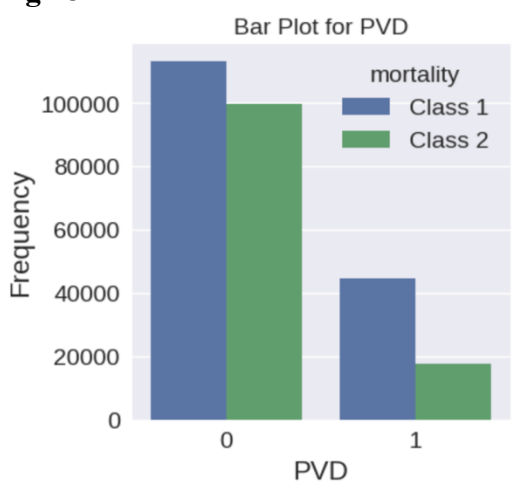



**Fig 19**

**Fig 20**

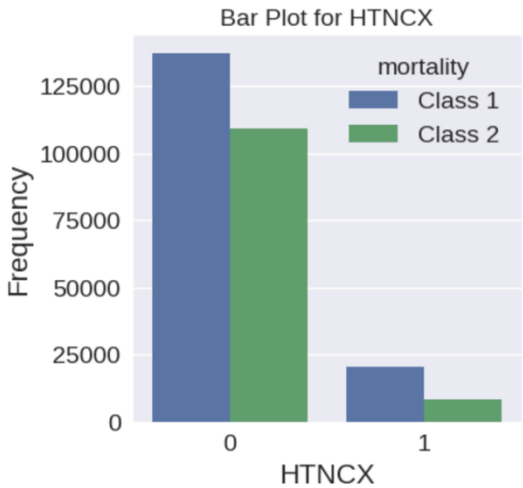



**Fig 21**

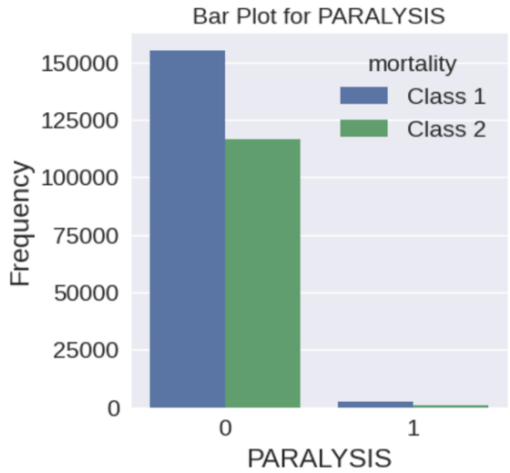



**Fig 22**

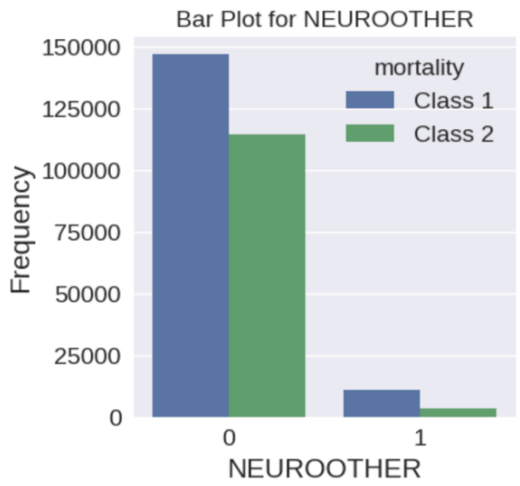



**Fig 23**

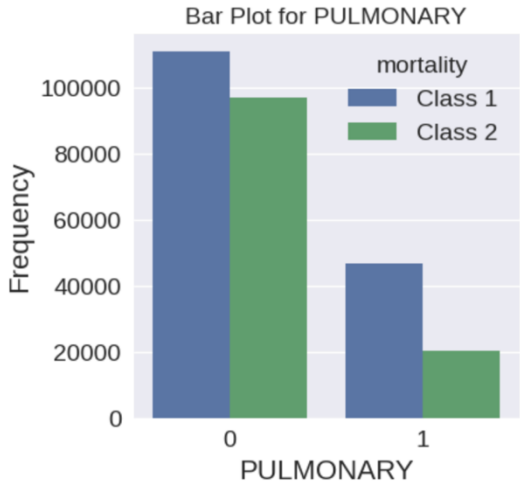



**Fig 24**

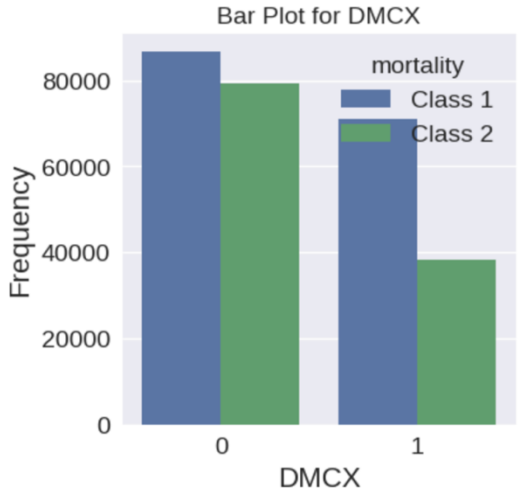



**Fig 25**

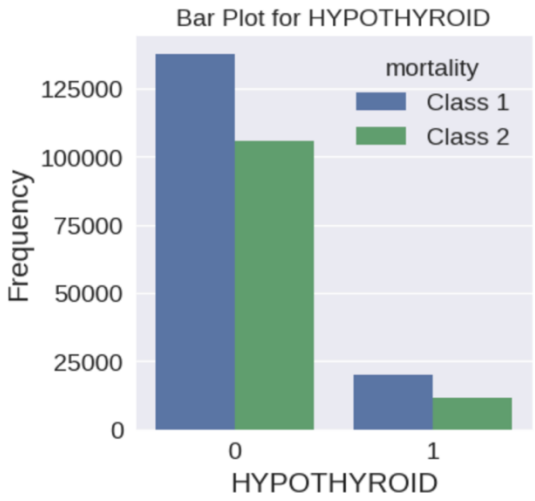



**Fig 26**

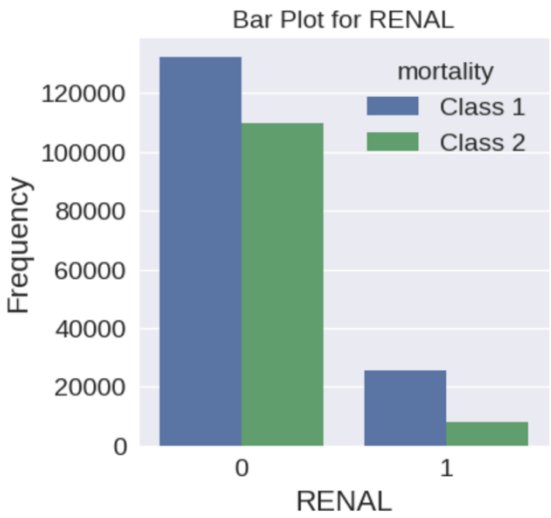



**Fig 27**

**Fig 28**

**Fig 29**

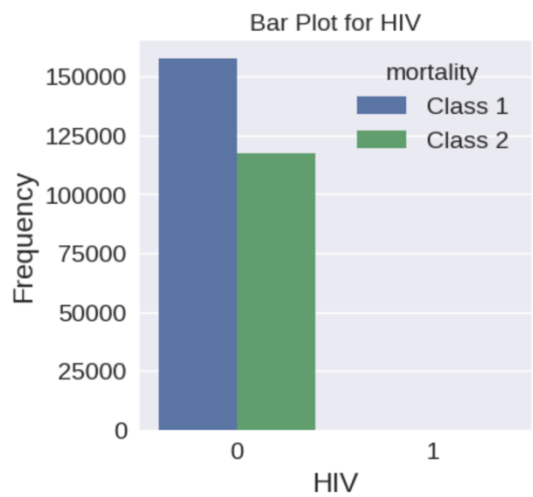



**Fig 30**

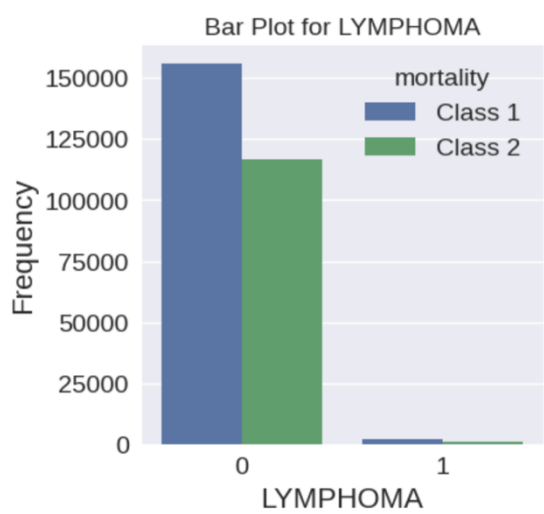



**Fig 31**

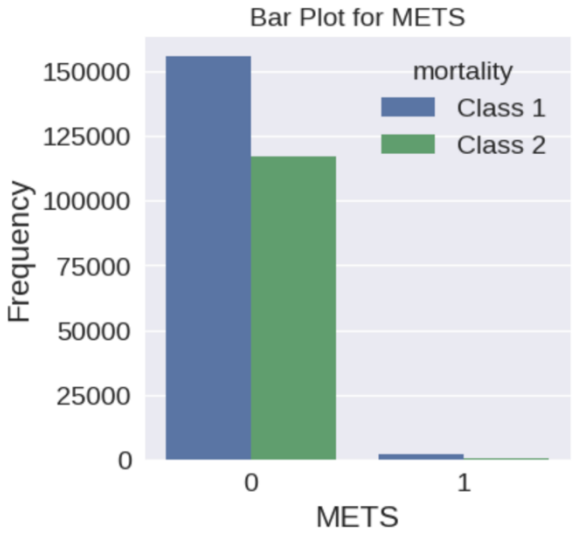



**Fig 32**

**Fig 33**

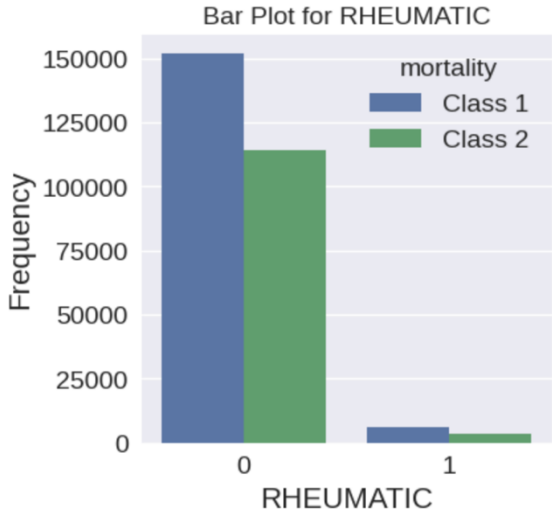



**Fig 34**

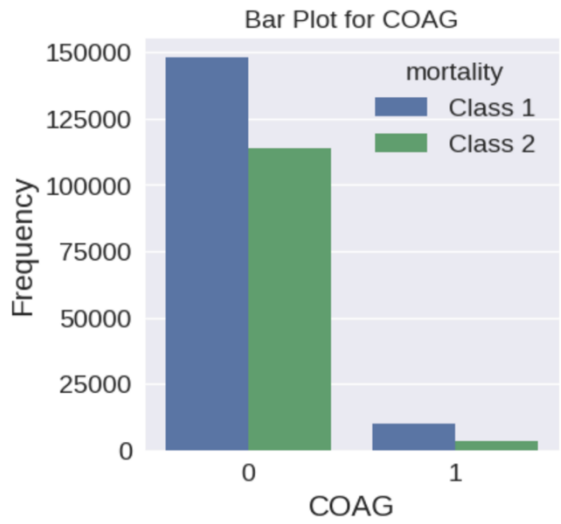



**Fig 35**

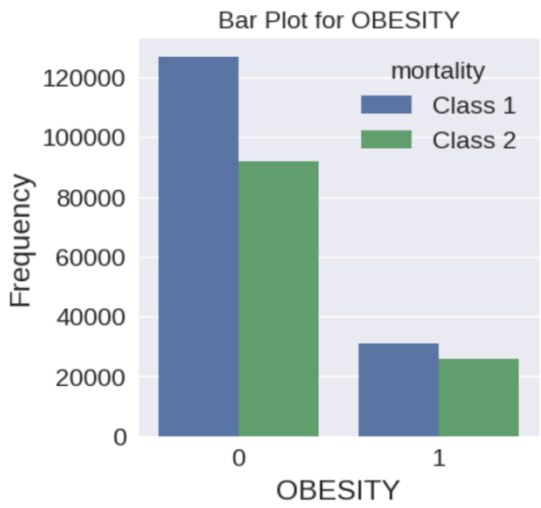



**Fig 36**

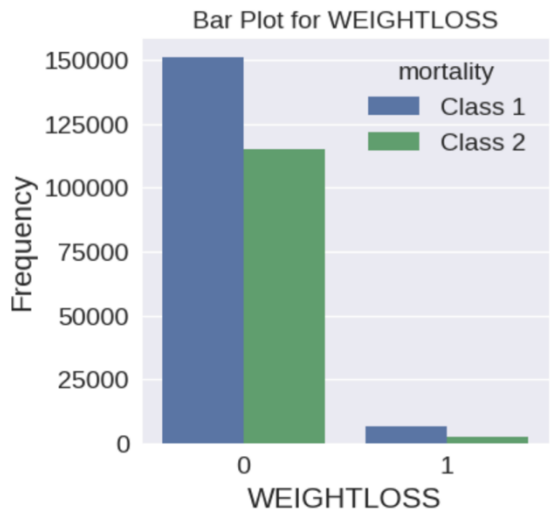



**Fig 37**

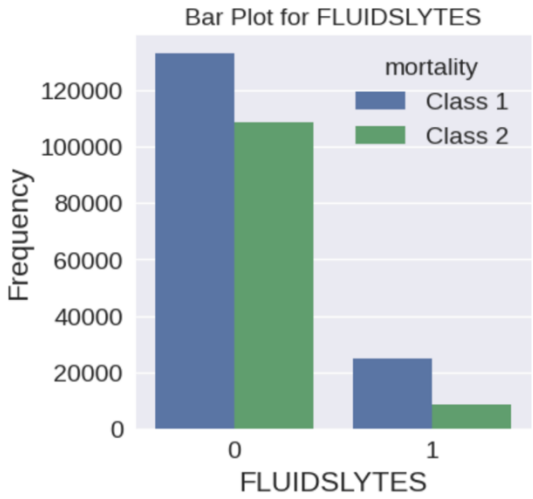



**Fig 38**

**Fig 39**

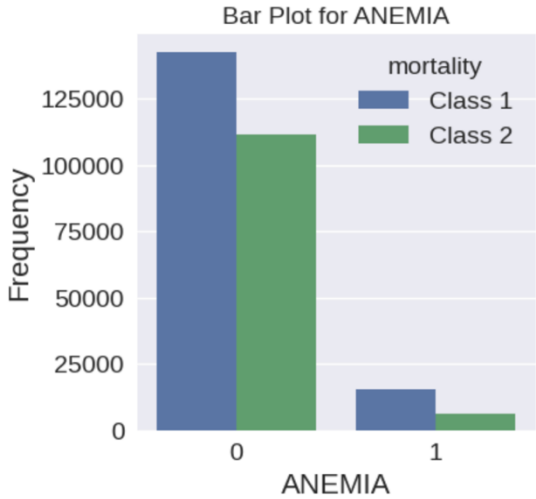



**Fig 40**

**Fig 41**

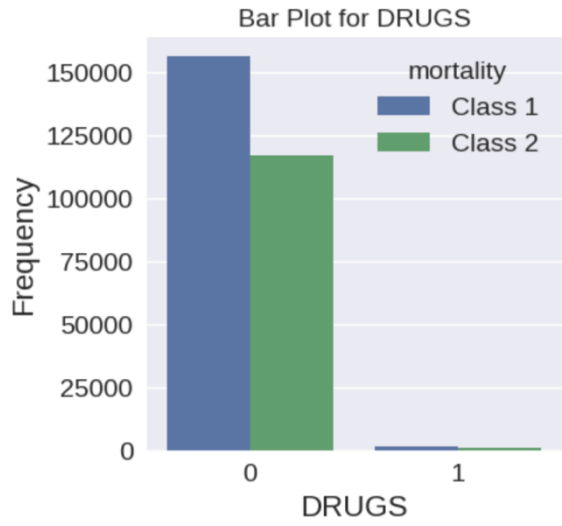



**Fig 42**

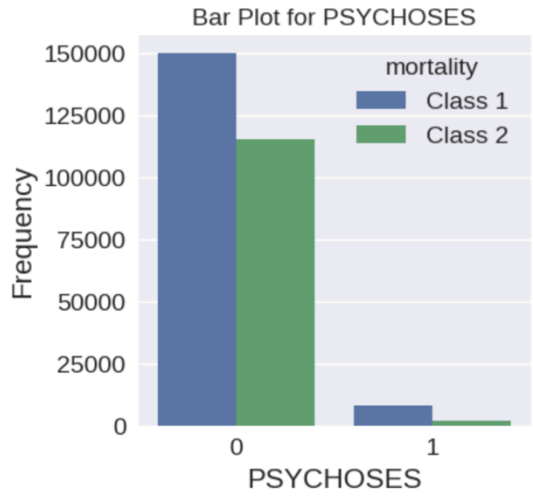



**Fig 43**

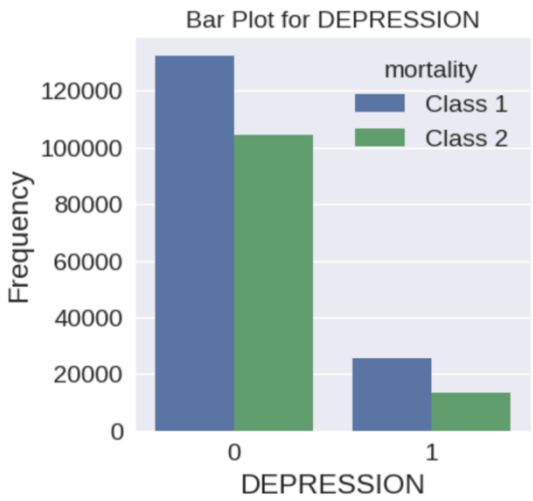



**Fig 44**

**Fig 45**

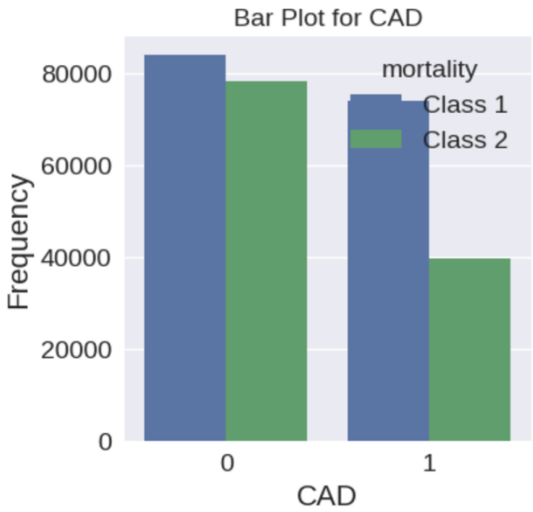



**Fig 46**

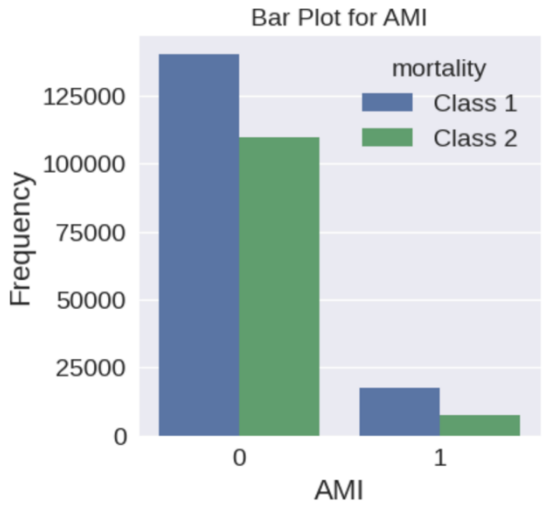



**Fig 47**

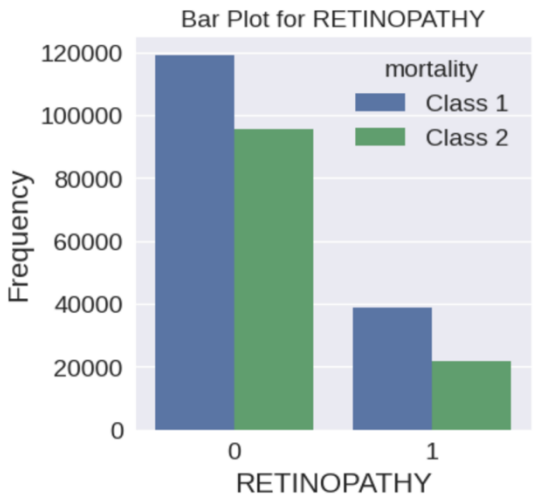



**Fig 48**

**Fig 49**

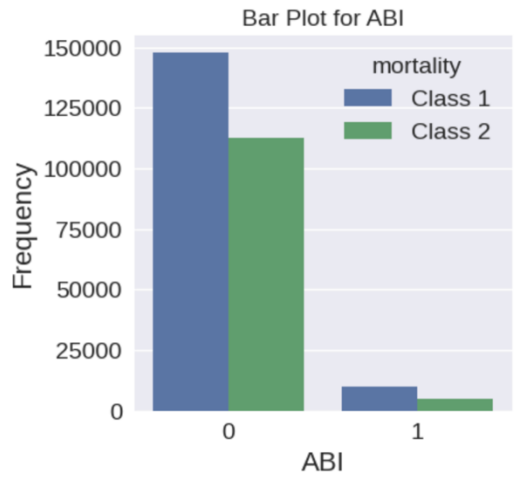



**Fig 50**

**Fig 51**

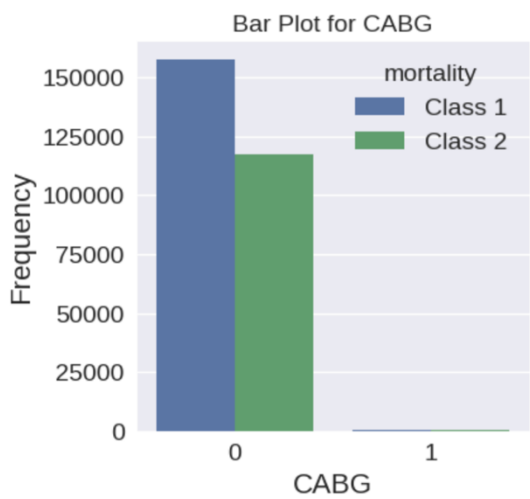



**Fig 52**

**Fig 53**

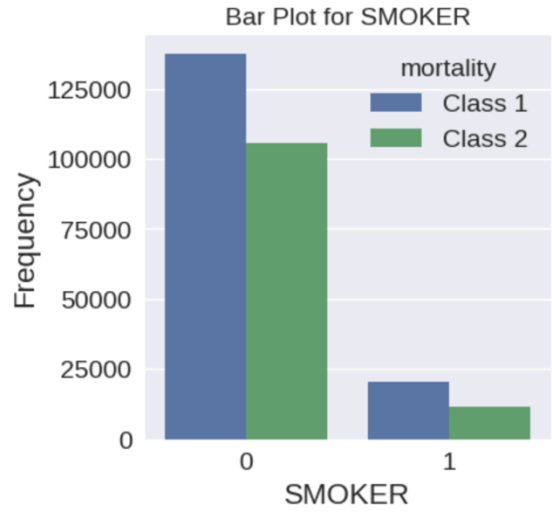



**Fig 54**

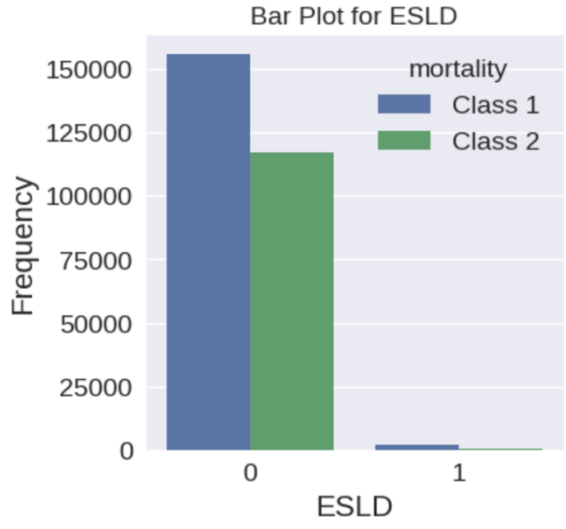



**Fig 55**

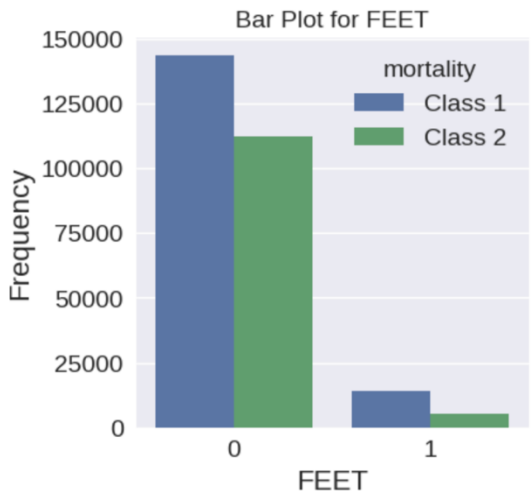



**Fig 56**

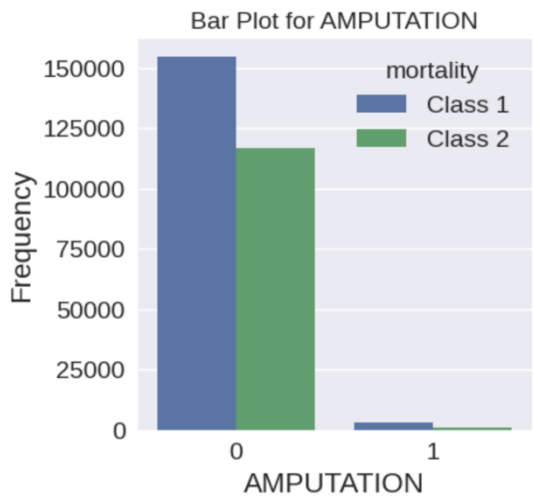



**Fig 57**

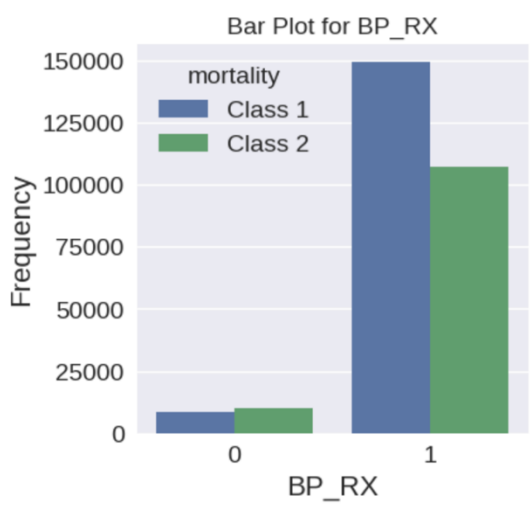



**Fig 58**

**Fig 60**

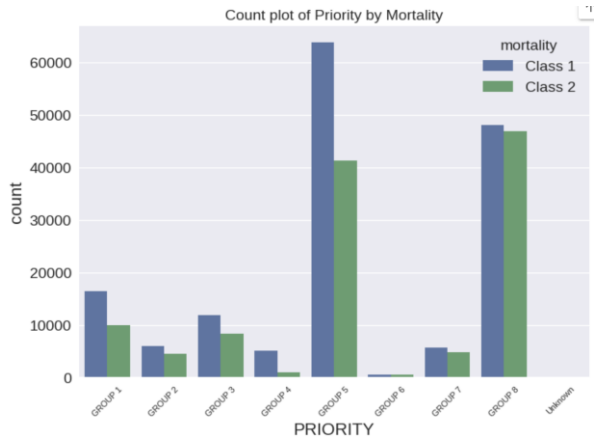



**Fig 61**

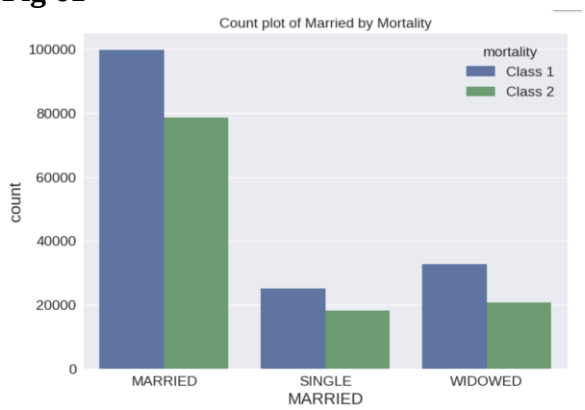
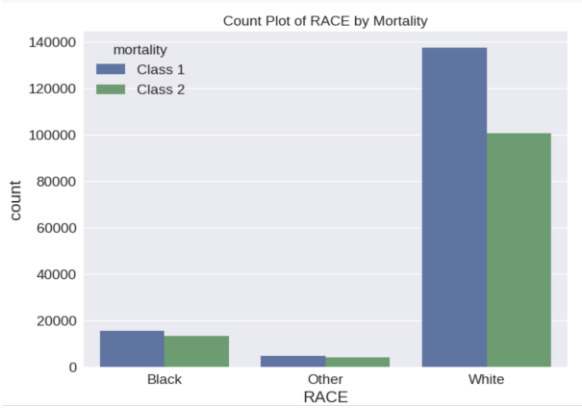



**Fig 62**




## Fig 63 – IMBALANCE OF DATA IN BINARY VARIABLES

| Features | Observations | Action |
|---|---|---|
| **INSULIN** | High Variability | No action |
| **SULF** | Moderate Variability | No action |
| **OTHER_MED** | Mostly 0 | Remove Variable |
| **BIGUAN** | Moderate Variability | No action |
| **TZD** | Mostly 0 | Remove Variable |
| **ALPHA** | Mostly 0 | Remove Variable |
| **SEX** | Mostly 1 | Remove Variable |
| **CHF** | Moderate Variability | No action |
| **ARRHYTHMIA** | Moderate Variability | No action |
| **VALVULAR** | Mostly 0 | Remove Variable |
| **PHTN** | Mostly 0 | Remove Variable |
| **PVD** | High Variability | No action |
| **HTN** | Mostly 1 | Remove Variable |
| **HTNCX** | Mostly 0 | Remove Variable |
| **PARALYSIS** | Mostly 0 | Remove Variable |
| **NEUROOTHER** | Mostly 0 | Remove Variable |
| **PULMONARY** | Moderate Variability | No action |
| **DMCX** | Moderate Variability | No action |
| **HYPOTHYROID** | Mostly 1 | Remove Variable |
| **RENAL** | Moderate Variability | No action |
| **LIVER** | Mostly 0 | Remove Variable |
| **PUD** | Mostly 0 | Remove Variable |
| **PCI** | Mostly 0 | Remove Variable |
| **SMOKER** | Data Imbalanced | Need more Analysis |
| **ESLD** | Mostly 0 | Remove Variable |
| **BP_RX** | Mostly 1 | Remove Variable |

| Features | Observations | Action |
|---|---|---|
| **HIV** | Mostly 0 | Remove Variable |
| **LYMPHOMA** | Mostly 0 | Remove Variable |
| **METS** | Mostly 0 | Remove Variable |
| **TUMOR** | Moderate Variability ; But has Imbalance | Need more Analysis |
| **RHEUMATIC** | Mostly 0 | Remove Variable |
| **COAG** | Mostly 0 | Remove Variable |
| **OBESITY** | Data Imbalanced | Need more Analysis |
| **WEIGHTLOSS** | Mostly 0 | Remove Variable |
| **FLUIDSLYTES** | Data Imbalanced | Need more Analysis |
| **BLOODLOSS** | Mostly 0 | Remove Variable |
| **ANEMIA** | Data Imbalanced | Need more Analysis |
| **ALCOHOL** | Mostly 0 | Remove Variable |
| **DRUGS** | Mostly 0 | Remove Variable |
| **PSYCHOSES** | Mostly 0 | Remove Variable |
| **DEPRESSION** | Data Imbalanced | Need more Analysis |
| **SEVERE_DEP** | Mostly 0 | Remove Variable |
| **CAD** | Good Variation | No action |
| **AMI** | Data Imbalanced | Need more Analysis |
| **RETINOPATHY** | Data Imbalanced | Need more Analysis |
| **HYPERG** | Mostly 0 | Remove Variable |
| **ABI** | Mostly 0 | Remove Variable |
| **RETSCREEN** | Data Imbalanced | Need more Analysis |
| **CABG** | Mostly 0 | Remove Variable |
| **FEET** | Data Imbalanced | Need more Analysis |
| **AMPUTATION** | Mostly 1 | Remove Variable |

## Fig 64 – DATA QUALITY PLAN

| FEATURES | OBSERVATIONS | ACTIONS |
|---|---|---|
| | | |
| SYSTOLIC | Outlier | Binning into ranges '<120', '120-129', '130-139', '140-179', '>=180' based on clinically fixed ranges |
| A1C | Outlier | Binning into ranges '<5.7', '5.7-6.4', '>6.5' based on clinically defined ranges |
| AGE | Outliers and Few Missing Values | Binning into ranges '65-69', '70-74', '75-79', '80-84', '85-89', '90-94', '>=95'. Removing missing instances |
| BMI | Outliers | Binning into ranges '<18.5', '18.5-24.9', '25- 29.9', '>30' based on clinically defined ranges |
| HDL | Outliers and Few Missing Values | Binning into ranges '<40', '40-59.99', '>=60' based on clinically defined ranges. Removing missing instances |
| LDL | Outliers and Few Missing Values | Binning into ranges '<100', '100-159.99', '>160' based on clinically defined ranges. Removing missing instances. |
| DIASTOLIC | Outliers | Binning into ranges '<80', '80-89', '>=90' based on clinically defined ranges |
| TRI | Outliers and Few Missing Values | Binning into ranges '<150', '150-199.99', '>=200' based on clinically defined ranges. Removing missing instances. |
| SERUMCRE | Outliers (High) and Few Missing Values | Binning into ranges '<1.5', '1.5-3.0', '>3.0' based on clinically defined ranges. Removing missing instances |
| MICROALB | 74% Missing data | Removing the feature |
| SERUMALB | 30 % Missing data | Removing the feature |
| N_IP | Outliers (High) | Binning into ranges '0-5', '>5' |
| N_OP | Outliers (High) | Binning into ranges '0-5', '6-30', '>30' |
| OTHER_MED | High number of instances for 1 | Removing the feature |
| TZD | High number of instances for 0 | Removing the feature |
| ALPHA | High number of instances for 0 | Removing the feature |
| VALVULAR | High number of instances for 0 | Removing the feature |
| PHTN | High number of instances for 0 | Removing the feature |
| HTN | High number of instances for 1 | Removing the feature |
| HTNCX | High number of instances for 0 | Removing the feature |

**Fig 65 – DATA QUALITY PLAN**

| | | |
|---|---|---|
| PUD | High number of instances for 0 | Removing the feature |
| HIV | High number of instances for 0 | Removing the feature |
| LYMPHOMA | High number of instances for 0 | Removing the feature |
| METS | High number of instances for 0 | Removing the feature |
| TUMOR | Imbalance in data | Need more Analysis |
| RHEUMATIC | High number of instances for 0 | Removing the feature |
| COAG | Imbalance in data | Need more Analysis |
| OBESITY | Imbalance in data | Need more Analysis |
| WEIGHTLOSS | High number of instances for 0 | Removing the feature |
| FLUIDSLYTES | Imbalance in data | Need more Analysis |
| BLOODLOSS | High number of instances for 0 | Removing the feature |
| ANEMIA | Imbalance in data | Need more Analysis |
| ALCOHOL | High number of instances for 0 | Removing the feature |
| DRUGS | High number of instances for 0 | Removing the feature |
| PSYCHOSES | High number of instances for 0 | Removing the feature |
| DEPRESSION | Imbalance in data | Need more Analysis |
| SEVERE_DEP | High number of instances for 0 | Removing the feature |
| AMI | Imbalance in data | need more Analysis |
| RETINOPATHY | Imbalance in data | Need more Analysis |

**Fig 66 – DATA QUALITY PLAN**

| | | |
|---|---|---|
| PARALYSIS | High number of instances for 0 | Removing the feature |
| PRIORITY | Imbalance in data over different categories. 'Unknown' category had no instances | Removing 'Unknown' category with least data instances |
| MARRIED | Imbalance in data over different categories | Need more Analysis |
| RACE | Imbalance in data over different categories | Need more Analysis |
| SEX | High number of instances for 1 | Removing the feature |
| ESLD | High number of instances for 0 | Removing the feature |
| FEET | Imbalance in data | Need more Analysis |
| AMPUTATION | High number of instances for 0 | Removing the feature |
| FRAILTY | Outliers (High) | Binning the data into 'Non-frail', 'Pre-frail', 'Mild', 'Moderate', 'Severe' |
| NEUROOTHER | High number of instances for 0 | Removing the feature |
| HYPOTHYROID | High number of instances for 1 | Removing the feature |
| LIVER | High number of instances for 0 | Removing the feature |
| BP_RX | High number of instances for 1 | Removing the feature |
| RETSCREEN | Imbalance in data | Need more Analysis |
| CABG | High number of instances for 0 | Removing the feature |
| PCI | High number of instances for 0 | Removing the feature |
| SMOKER | Imbalance in data | Need more Analysis |
| HYPERG | High number of instances for 0 | Removing the feature |
| ABI | High number of instances for 0 | Removing the feature |

## Fig 67- CHI-SQUARED TEST RESULTS

```
Sorted by p-value:
       Column  Chi-Square  p-value  Degrees of Freedom
13        SEX        5.96     0.01                   1
9         TRI       34.13     0.00                   2
14       RACE      175.83     0.00                   2
0    SYSTOLIC      221.30     0.00                   2
7         LDL      231.07     0.00                   2
1         A1C      269.07     0.00                   2
5     MARRIED      411.45     0.00                   2
6         HDL      509.52     0.00                   2
8   DIASTOLIC     1267.38     0.00                   2
2    PRIORITY     3796.12     0.00                   8
3         AGE    23380.27     0.00                   6
4         BMI     1536.45     0.00                   3
10   SERUMCRE     7234.05     0.00                   2
11       N_IP     3095.71     0.00                   1
12       N_OP     4165.81     0.00                   2
15    FRAILTY     1503.18     0.00                   2
```

**Fig 68 – CRAMER'SV TEST**

```
Matrix of Cramer's V results:
           SYSTOLIC  A1C  PRIORITY  AGE  BMI  MARRIED  HDL  LDL  DIASTOLIC  \
SYSTOLIC       1.00 0.02      0.02 0.01 0.04     0.01 0.02 0.05       0.14
A1C            0.02 1.00      0.03 0.03 0.03     0.01 0.03 0.02       0.01
PRIORITY       0.02 0.03      1.00 0.06 0.03     0.17 0.02 0.03       0.02
AGE            0.01 0.03      0.06 1.00 0.12     0.15 0.03 0.04       0.09
BMI            0.04 0.03      0.03 0.12 1.00     0.02 0.11 0.01       0.04
MARRIED        0.01 0.01      0.17 0.15 0.02     1.00 0.01 0.03       0.03
HDL            0.02 0.03      0.02 0.03 0.11     0.01 1.00 0.05       0.01
LDL            0.05 0.02      0.03 0.04 0.01     0.03 0.05 1.00       0.07
DIASTOLIC      0.14 0.01      0.02 0.09 0.04     0.03 0.01 0.07       1.00
TRI            0.02 0.06      0.02 0.07 0.12     0.01 0.21 0.03       0.02
SERUMCRE       0.02 0.03      0.04 0.08 0.03     0.01 0.06 0.02       0.04
N_IP           0.05 0.02      0.18 0.03 0.04     0.08 0.03 0.02       0.06
N_OP           0.02 0.02      0.21 0.02 0.03     0.07 0.03 0.02       0.06
RACE           0.04 0.03      0.12 0.04 0.04     0.09 0.09 0.06       0.05
FRAILTY        0.03 0.01      0.13 0.10 0.01     0.10 0.01 0.05       0.04

            TRI  SERUMCRE  N_IP  N_OP  RACE  FRAILTY
SYSTOLIC  0.02      0.02  0.05  0.02  0.04     0.03
A1C       0.06      0.03  0.02  0.02  0.03     0.01
PRIORITY  0.02      0.04  0.18  0.21  0.12     0.13
AGE       0.07      0.08  0.03  0.02  0.04     0.10
BMI       0.12      0.03  0.04  0.03  0.04     0.01
MARRIED   0.01      0.01  0.08  0.07  0.09     0.10
HDL       0.21      0.06  0.03  0.03  0.09     0.01
LDL       0.03      0.02  0.02  0.02  0.06     0.05
DIASTOLIC 0.02      0.04  0.06  0.06  0.05     0.04
TRI       1.00      0.05  0.00  0.02  0.11     0.01
SERUMCRE  0.05      1.00  0.11  0.07  0.07     0.04
N_IP      0.00      0.11  1.00  0.32  0.08     0.08
N_OP      0.02      0.07  0.32  1.00  0.08     0.10
RACE      0.11      0.07  0.08  0.08  1.00     0.09
FRAILTY   0.01      0.04  0.08  0.10  0.09     1.00
```

**Fig 69 – PHI COEFF**

```
Phi coefficients:
INSULIN: -0.7890064895525275
SULF: nan
OTHER_MED: -0.004391207878871904
BIGUAN: 0.35543972621502845
TZD: -0.11113925050526789
ALPHA: nan
CHF: -1.7685094410905036
ARRHYTHMIA: -4.059525452498272
VALVULAR: -0.7669873096491108
PHTN: -0.25140079258819575
PVD: -1.0447168212048488
HTN: nan
HTNCX: -0.3122113696512452
PARALYSIS: nan
NEUROOTHER: -0.25886128920113777
PULMONARY: nan
DMCX: -1.2483217870412708
HYPOTHYROID: nan
RENAL: nan
LIVER: -0.07214175623571355
PUD: -0.06060914157455442
HIV: -0.0012395426284444874
LYMPHOMA: -0.03114880400492898
METS: -0.04031790144025192
TUMOR: nan
RHEUMATIC: -0.04737493655194464
COAG: -0.2696644486161873
OBESITY: nan
WEIGHTLOSS: -0.11197590879583796
FLUIDSLYTES: nan
BLOODLOSS: nan
ANEMIA: -0.26064511238153215
ALCOHOL: nan
DRUGS: -0.014466880391143777
PSYCHOSES: nan
DEPRESSION: nan
SEVERE_DEP: -0.07918924969702033
CAD: -0.8532841844540198
AMI: nan
RETINOPATHY: nan
HYPERG: -0.12562554590120878
ABI: -0.12076793265781881
RETSCREEN: -0.11102660211926035
CABG: -0.00031493885563733996
PCI: -0.01092013720391748
SMOKER: nan
ESLD: -0.03891542085451911
FEET: -0.5171605164945307
AMPUTATION: nan
BP_RX: nan
```

**Fig 70 – IV & WOE OUTPUT**

| | Feature | Score | Category |
|---|---|---|---|
| 0 | SYSTOLIC | 0.00 | not useful |
| 1 | A1C | 0.00 | not useful |
| 2 | PRIORITY | 0.07 | weak |
| 3 | AGE | 0.45 | strong |
| 4 | BMI | 0.03 | weak |
| 5 | MARRIED | 0.01 | not useful |
| 6 | HDL | 0.01 | not useful |
| 7 | LDL | 0.00 | not useful |
| 8 | DIASTOLIC | 0.02 | weak |
| 9 | TRI | 0.00 | not useful |
| 10 | SERUMCRE | 0.14 | moderate |
| 11 | N_IP | 0.06 | weak |
| 12 | N_OP | 0.07 | weak |
| 13 | INSULIN | 0.06 | weak |
| 14 | SULF | 0.02 | weak |
| 15 | OTHER_MED | 0.00 | not useful |
| 16 | BIGUAN | 0.01 | not useful |
| 17 | TZD | 0.00 | not useful |
| 18 | ALPHA | 0.00 | not useful |
| 19 | RACE | 0.00 | not useful |
| 20 | CHF | 0.21 | moderate |
| 21 | ARRHYTHMIA | 0.10 | moderate |
| 22 | VALVULAR | 0.05 | weak |
| 23 | PHTN | 0.02 | weak |
| 24 | PVD | 0.11 | moderate |
| 25 | HTN | 0.01 | not useful |
| 26 | HTNCX | 0.04 | weak |
| 27 | PARALYSIS | 0.01 | not useful |
| 28 | NEUROOTHER | 0.04 | weak |
| 29 | PULMONARY | 0.09 | weak |
| 30 | DMCX | 0.07 | weak |
| 31 | HYPOTHYROID | 0.01 | not useful |
| 32 | RENAL | 0.10 | weak |
| 33 | LIVER | 0.00 | not useful |
| 34 | PUD | 0.00 | not useful |
| 35 | HIV | 0.00 | not useful |
| 36 | LYMPHOMA | 0.00 | not useful |
| 37 | METS | 0.01 | not useful |
| 38 | TUMOR | 0.02 | not useful |
| 39 | RHEUMATIC | 0.00 | not useful |
| 40 | COAG | 0.02 | weak |
| 41 | OBESITY | 0.00 | not useful |
| 42 | WEIGHTLOSS | 0.02 | not useful |
| 43 | FLUIDSLYTES | 0.07 | weak |
| 44 | BLOODLOSS | 0.01 | not useful |
| 45 | ANEMIA | 0.03 | weak |
| 46 | ALCOHOL | 0.00 | not useful |
| 47 | DRUGS | 0.00 | not useful |
| 48 | PSYCHOSES | 0.03 | weak |
| 49 | DEPRESSION | 0.02 | not useful |
| 50 | SEVERE_DEP | 0.00 | not useful |
| 51 | CAD | 0.08 | weak |
| 52 | AMI | 0.03 | weak |
| 53 | RETINOPATHY | 0.02 | weak |
| 54 | HYPERG | 0.02 | not useful |
| 55 | ABI | 0.01 | not useful |
| 56 | RETSCREEN | 0.00 | not useful |
| 57 | CABG | 0.00 | not useful |
| 58 | PCI | 0.00 | not useful |
| 59 | SMOKER | 0.01 | not useful |
| 60 | ESLD | 0.01 | not useful |
| 61 | FEET | 0.03 | weak |
| 62 | AMPUTATION | 0.02 | not useful |
| 63 | FRAILTY | 0.03 | weak |
| 64 | BP_RX | 0.02 | not useful |

**Fig 71 – MUTUAL INFORMATION ( BEST 20 VARIABLES)**

```
SERUMCRE_<1.5      40.31
N_OP_6-30          31.09
CHF                27.93
DIASTOLIC_<80      26.12
N_IP_0-5           26.04
CAD                24.14
N_OP_>30           22.36
AGE_80-84          22.20
DMCX               21.35
AGE_65-69          20.05
SULF               20.00
ARRHYTHMIA         18.22
PVD                18.07
BMI_>30            17.20
PULMONARY          16.99
SERUMCRE_1.5-3.0   16.26
AGE_70-74          15.20
PRIORITY_GROUP 5   14.69
PRIORITY_GROUP 8   14.36
AGE_85-89          13.76
dtype: float64
```

**Fig 72 - LEARNING CURVE**

**Fig 73 - LEARNING CURVE**

**Fig 74 - LEARNING CURVE**

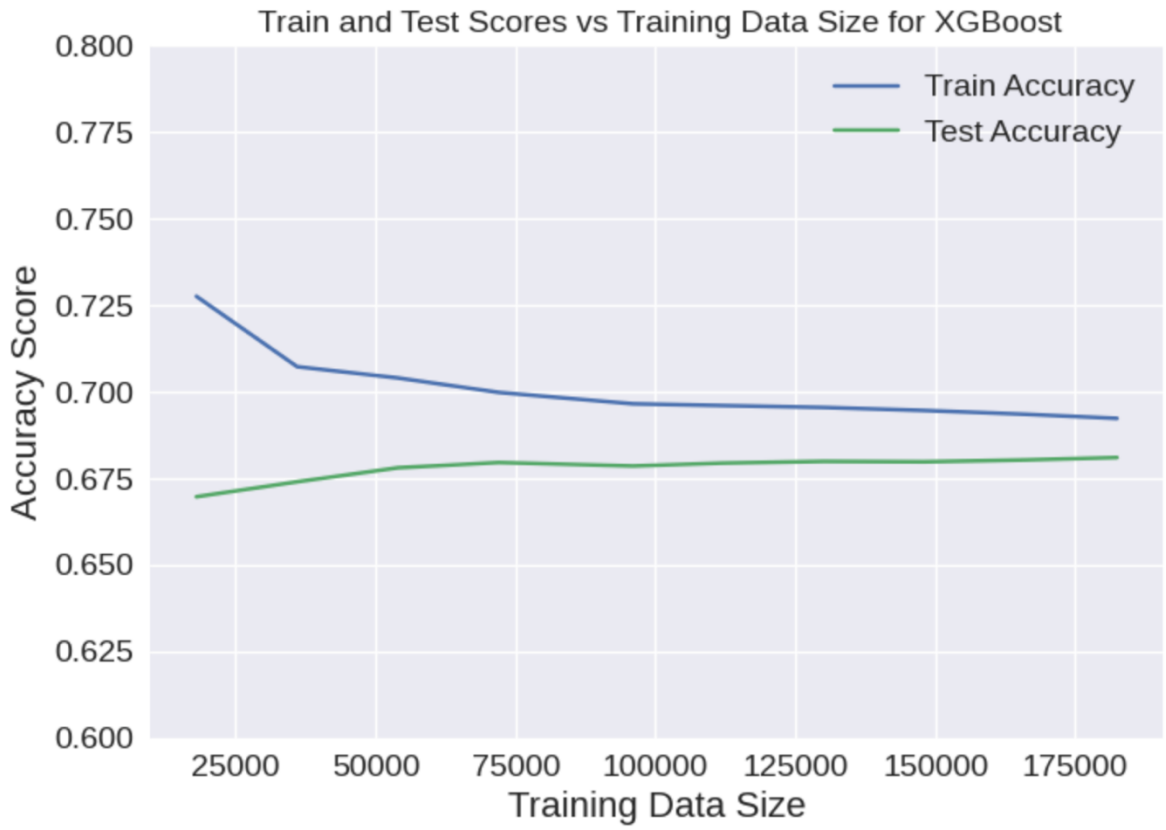



**Fig 75 - LEARNING CURVE**

**Fig 76 - CONFUSION MATRIX FOR LOGICTIC REGRESSION**

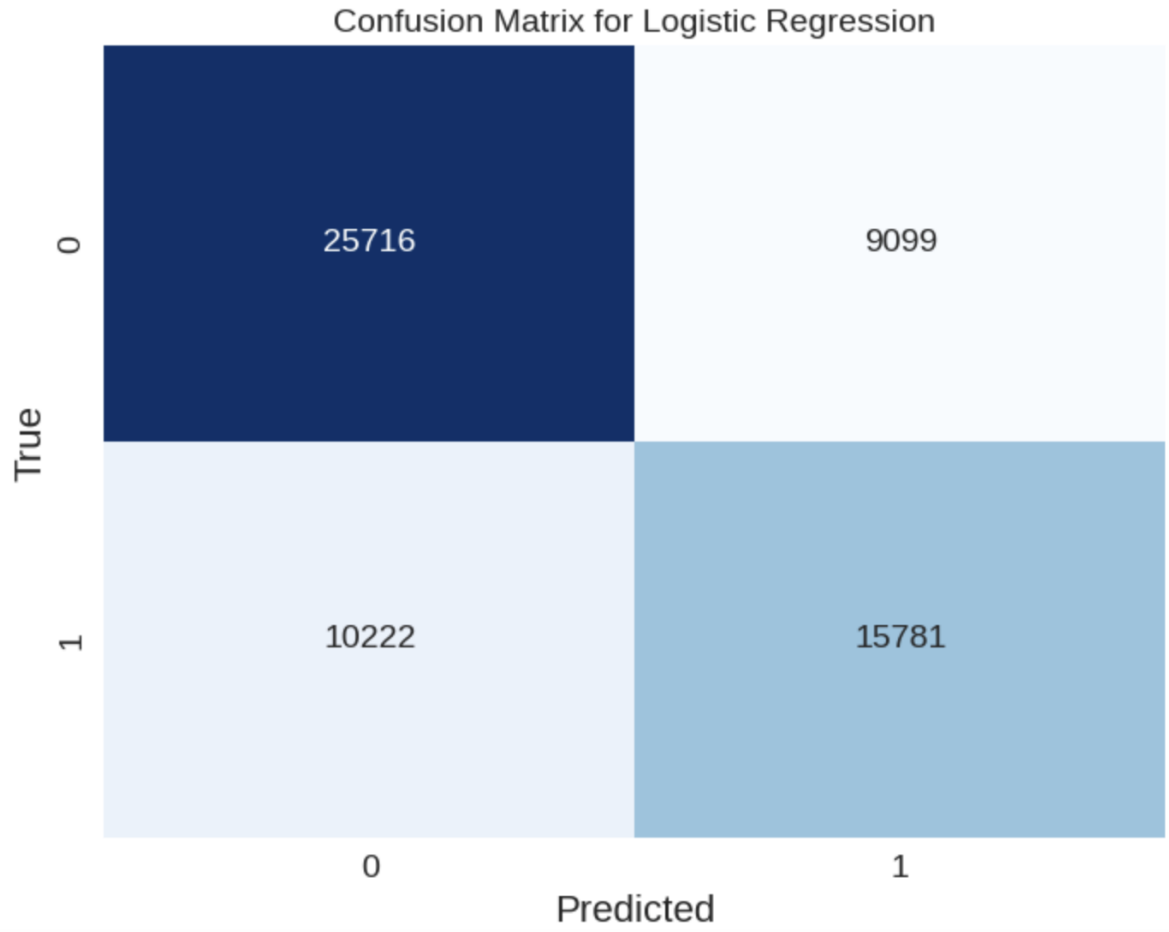



**Fig 77**

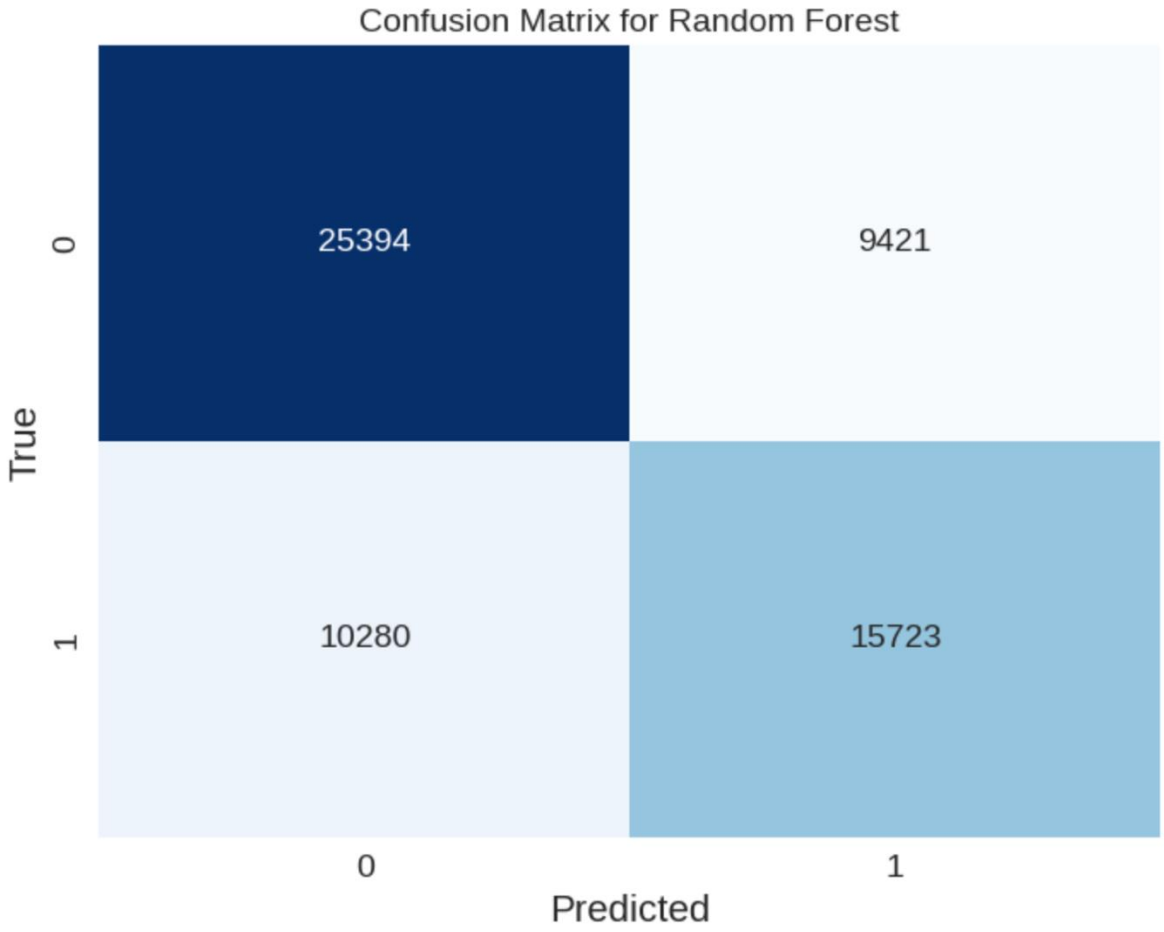

Confusion Matrix for Random Forest
25394
9421
10280
15723
True
0
1
0
1
Predicted

**Fig 78**

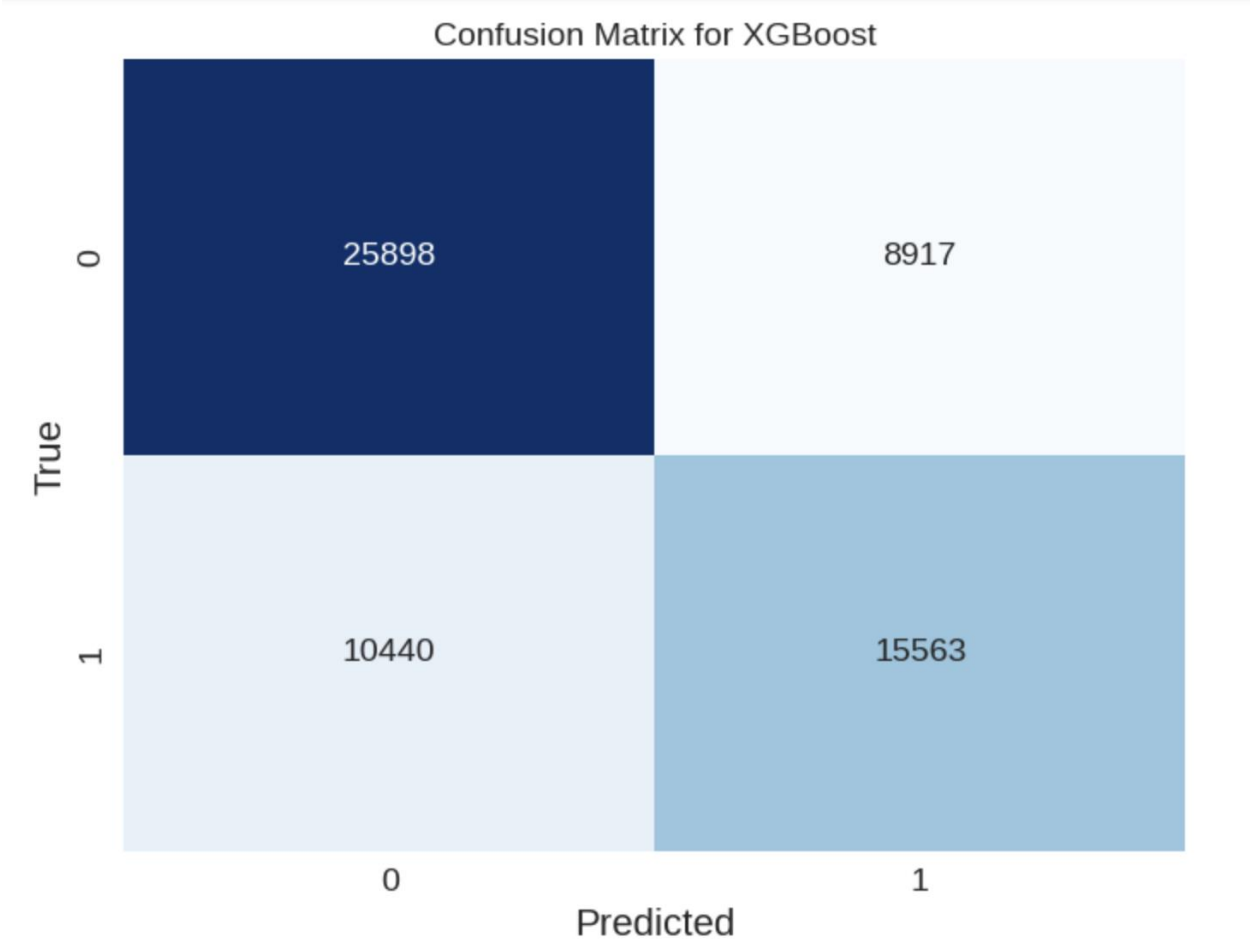

Confusion Matrix for XGBoost
25898
8917
10440
15563
True
0
1
0
1
Predicted

**Fig 79**

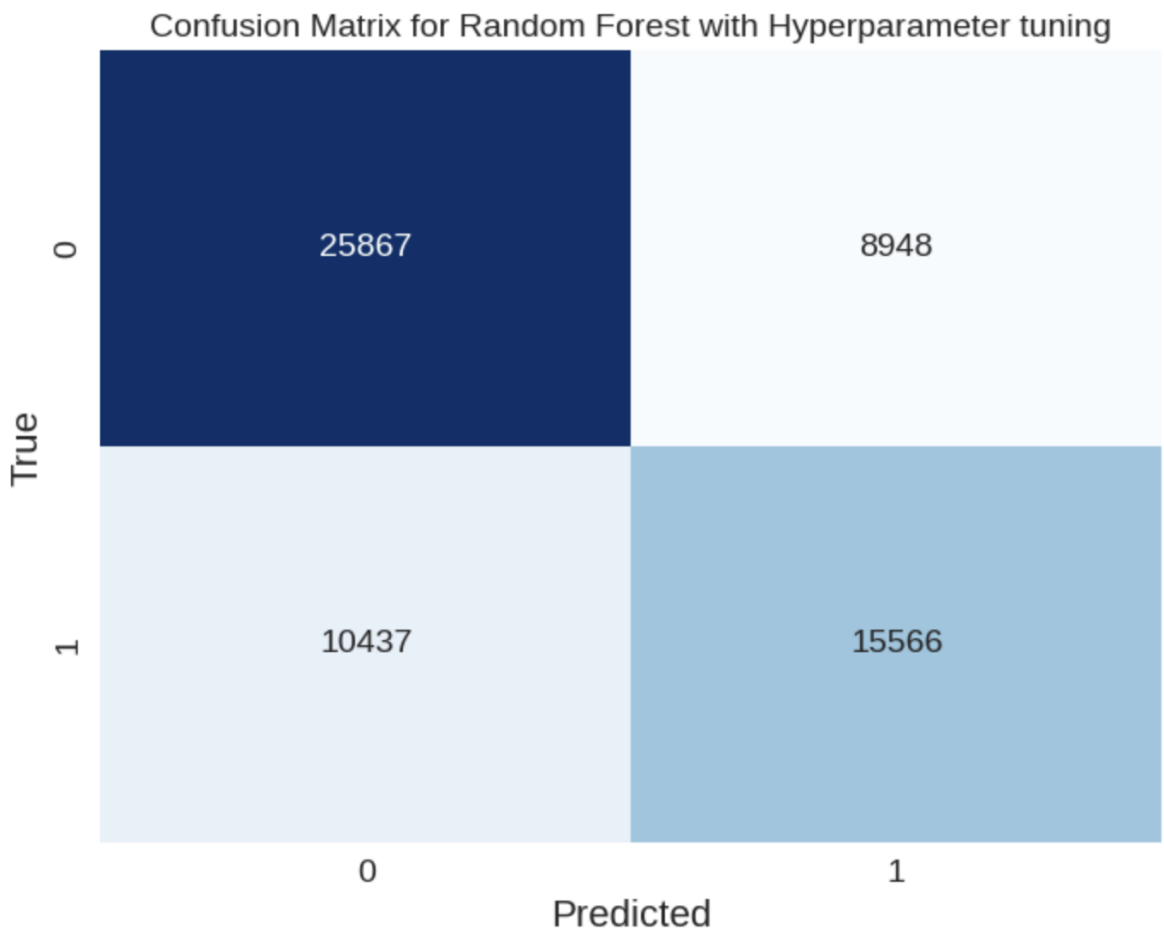

Confusion Matrix for Random Forest with Hyperparameter tuning
25867
8948
10437
15566
True
0
1
0
1
Predicted

**Fig 80**

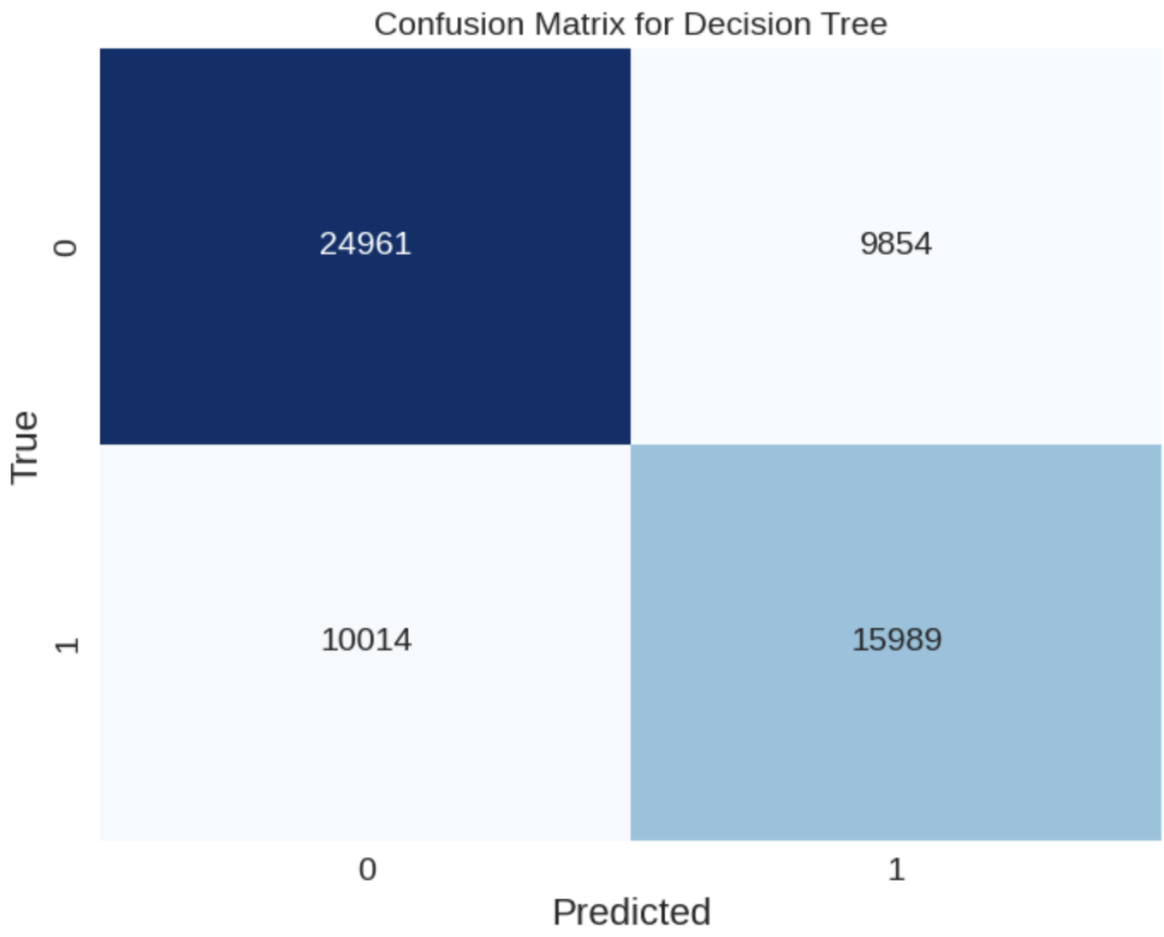



**Fig 81**

| | Model | TPR | FPR | Recall | Precision | F1-Score |
|---|---|---|---|---|---|---|
| **0** | Logistic | 0.61 | 0.26 | 0.63 | 0.61 | 0.62 |
| **1** | Random Forest | 0.60 | 0.27 | 0.63 | 0.60 | 0.61 |
| **2** | XG Boost | 0.60 | 0.26 | 0.64 | 0.60 | 0.62 |
| **3** | XG Boost | 0.60 | 0.26 | 0.64 | 0.60 | 0.62 |
| **4** | Random Forest hyperparameter tuned | 0.60 | 0.26 | 0.63 | 0.60 | 0.62 |
| **5** | Decision Tree | 0.61 | 0.28 | 0.62 | 0.61 | 0.62 |

**Fig 82 – AIC SCORE FOR ALL USED MODELS**

| | Model | Score |
|---|---|---|
| 0 | Logistic Regression | 71433.00 |
| 1 | Random Forest | 73262.94 |
| 2 | XGBoost | 71347.01 |
| 3 | Random Forest with hyper parameter tuning | 71690.99 |
| 4 | Decision Tree | 75073.76 |

**Fig 83 AUC SCORE FOR ALL USED MODELS**

| | Model | Score |
|---|---|---|
| 0 | Logistic Regression | 0.673 |
| 1 | Random Forest | 0.667 |
| 2 | XGB Boost | 0.671 |
| 3 | Decision Tree | 0.666 |
| 4 | Random Forest with hyper parameter tuning | 0.671 |

**Fig 84 - PHI COEFF INFERENCE**

1. INSULIN: **Strong Association**

2. SULF, ALPHA, HTN, PARALYSIS, PULMONARY, HYPOTHYROID, RENAL, TUMOR, OBESITY, FLUIDSLYTES, BLOODLOSS, LYMPHOMA, SMOKER, ALCOHOL, PSYCHOSES, DEPRESSION, AMI, RETINOPATHY, AMPUTATION, CAD, PCI: **No association**

3. OTHER_MED, BIGUAN, TZD, VALVULAR, PHTN, HTNCX, NEUROOTHER, DMCX, LIVER, PUD, HIV, METS, RHEUMATIC, COAG, WEIGHTLOSS, ANEMIA, DRUGS, SEVERE_DEP, RETSCREEN, ABI, CABG, ESLD, FEET, BP_RX: **Weak association**

**Fig 85 IV & WOE CATEGORIES**

| Information Value (IV) | Predictive Power of the Variable |
|---|---|
| < .02 | Useless for prediction |
| .02 to .10 | Weak predictor |
| .10 to 0.30 | Medium Predictor |
| .30 to .50 | Strong predictor |
| > .50 | Suspicious behaviour |

**TABLES**

**Table 1**

| **Random Forest Hyper Parameter Tuned** | XGBoost | Decision tree |
|---|---|---|
| **n_estimators: randint(100, 200),**<br><br>**max_depth: [None, 5, 10],**<br><br>**min_samples_split: randint(2, 6)**<br><br>**min_samples_leaf:randint(2,6)** | **n_estimators: [100, 200, 300]**<br><br>**max_depth: [3, 5, 7]**<br>**learning_rate: [0.01, 0.05, 0.1]**<br><br>**subsample: [0.8, 0.9, 1.0]**<br><br>**colsample_bytree: [0.8, 0.9, 1.0]**<br><br>**gamma: [0, 0.1, 0.2]** | **max_depth : [3, 5, 7, 10]**<br><br>**min_samples_split: [2, 5, 10]**<br><br>**min_samples_leaf: [1, 2, 4],**<br><br>**max_features: ['auto', 'sqrt', 'log2']**<br><br>**criterion: ['gini', 'entropy']** |