# Optimize Data for Analysis Needs

Gladys Murage

School of Technology, North Central University

## TIM-8550 v1: DATA PREPARATION METHODS (7057984436)

Dr Mohammad Yavarimanesh

October 15 , 2024

**Managerial Report on Risk Factors for Death**: insights from the Framingham Heart Study.

**To**: the CEO, St. Francis Hospital.

**From**: Gladys Murage, Data Science Director.

**Date**: October 15th 2024

**Executive Summary**

The Framingham Heart Study dataset (frmgham6.csv) offers vital insights into the risk factors associated with mortality. Using machine learning models for analysis: Logistic Regression, Random Forest and XGBoost, identification of several key predictors of death have been unearthed in this report. These insights can inform St Francis Hospital practices and improve patient outcomes. This report presents evidence from data science on why an audit framework at St. Francis Hospital to implement routine patient chart and electronic medical reviews should be sought. The report also highlights strategies to mitigate risks associated with these predictors of death for the patients at St Francis Hospital system in order to improve outcomes.

The Logistic Regression model showshigh accuracy and precision at 96.47% .It is especially efficient at identifying patients who are not at risk of death, which could be useful for resource allocation and prioritizing patients with higher risks. Both the Random Forest and XGBoost models portray an accuracy of 100% in predicting those patients who will die from those that will not die, based on risk factors. Both models are excellent for predicting all patients but especially the high-risk patients for death. This analysis also ranks the risk factors for death in descending order per the logistic coefficients of regression: time to death, age, current smokers, time to hypertension, sex ( men being more predisposed to death as opposed to women), BMI, total cholesterol, systolic blood pressure, glucose, diastolic blood pressure, number of cigarettes smoked in a day and so many others. Refer to Figure 13 for the rest of the factors.

This analysis recommends auditing and monitoring of key risk factors, regular updates made to patient charts with key health metrics (such as age, cholesterol, blood pressure, BMI and smoking status) to ensure the risk predictions are based on the latest data. Tracking these factors at each patient visit could improve early detection of high-risk individuals for death. Secondly, prioritizing monitoring of patients with borderline metrics is recommendedfor patients with metrics close to the decision threshold. More frequent monitoring and follow-up appointments are recommended to reduce the number of false negatives (patients missed as high risk).

The recommendations and insights in this report are therefore firstly to implement chart audits: regular audits of patient charts should be conducted to ensure accurate recording of the most relevant clinical features (age, cholesterol, blood pressure, and smoking).Secondly, it is to develop follow-up programs: patients identified as high-risk based on these factors should be enrolled in follow-up programs, focusing on managing modifiable risks such as smoking cessation and hypertension management. Thirdly, further model validation is required, despite perfect accuracy, it is essential to validate all three models on new data or perform cross-validation to ensure it generalizes well to other patient populations at St Francis Hospital system.

Fourth, clinical decision support systems are recommended; integration of the Random Forest model, XGBoost and Logistic Regression models into St Francis hospital systems electronic medical record, ( MyChart ), will assist clinicians in decision-making, allowing them to flag high-risk patients for closer monitoring by healthcare providers, and thus helping improve patient outcomes. Implementing these changes will also assist the hospital to be compliant with provisions set forth by Centers for Medicare and Medicaid and the Affordable Care Act in achieving healthcare goals for patients based on set standards of care.

1. **Introduction**
   1. **Background**

According to the National Heart Lung Blood Institute NHLBI (2024), in 1948, in the then town of Framingham Massachusetts, an awe-inspiring group of 5,209 Framingham citizens volunteered to address the cardiovascular disease epidemic. They voluntarily enrolled in a research trial to prevent heart disease. The research trial is called the Framingham Heart Study (FHS) and is run by NHLBI. Seventy-five years later, their collective contribution to health research is without equal. The FHS enrolled the first generation of subjects, their children and grandchildren.

To date the FHS has enrolled 15,000 subjects from three generations, including the original cohort, their children, and grandchildren. In those 75 years the FHS has yielded numerous medical breakthroughs, which have informed on how cardiovascular health affects the rest of the body. The FHS, a landmark longitudinal study initiated in 1948, aims to identify common factors contributing to cardiovascular disease. Over decades, this research has led to the identification of numerous risk factors that are now foundational in preventive cardiology.

* 1. **Business Problem**

Indiana Heart Physicians (IHP) is a large thriving cardiovascular clinic located on the south side of Indianapolis Indiana, with over 60 practicing clinical cardiologist, cardiovascular interventionalists, congestive heart failure specialists, nurse practitioners and physician assistants. The practice has recently gone through a merger and been bought by the neighboring hospital St Francis Hospital. St Francis Hospital through its medical practice branch Franciscan alliance owns numerous primary care offices, endocrinology practices and other medical practices. As a result of the merger St Francis has acquired many new family doctors, nurse practitioners and physician assistants. St Francis vision is to harmonize medical guidelines and practices throughout its Franciscan alliance network in order to comply with the Centers for Medicare & Medicaid Services (CMS), and to also improve outcomes for their patients.

One of those visions is to create a cardiovascular computer algorithm that audits patients’ charts to inform the practitioners, and the health system on which patients are at a high risk for death from cardiovascular disease in the next 10 years. Executing the algorithm and elevating standards of care per recommended guidelines will reduce mortality rates from cardiovascular disease. The medical director for Indiana Heart Physicians has been charged with this responsibility and has approached me as the data science director of St Francis to assist with the project. Specifically, the medical director has requested feedback to be given to the chief executive officer of St Francis, and other stakeholders: medical doctors, nurse practitioners, physician assistants, registered nurses and medical assistants.

* 1. **Objectives**

The three main objectives of this report are firstly to identify key risk factors by utilizing statistical analysis to pinpoint significant predictors of mortality from the Framingham Study dataset(frmgnham6.csv). Secondly, is to develop audit strategies by recommending routine audits for patient charts to monitor risk factors for death (mortality). The final objective is to promote preventive care by suggesting actionable strategies to mitigate identified risks and improve patient health outcomes.

1. **Methodology**
   1. **Data Overview**
      1. ***Framingham Data Set (frmgham2.csv)***

The researcher makes an inquiry into the ability of acquiring the teaching data set so as to solve the business problem. The data set is later received from BioLINCC a part of the National Heart Lung Blood Institute (NHLBI) after a personal request is made by the researcher for the Framingham Heart Study. Megan Savage fulfills the request, on July 8th, 2024, when Framingham ZIP-1.7MB is received by the researcher. The Framingham data set is downloaded into a personal Dropbox owned by the researcher, unzipped and the CSV file saved as frmgham2.csv.

* + 1. ***Rationale for Data Selection Process***

The data set is chosen to help answer a business problem question. The particular St Francis problem calls for using clinically evident data with positive outcomes from investigations dealing with the cardiovascular system. It is felt that a great source of such data is to obtain the data directly from the longitudinal Framingham Heart Study (FHS) conducted by the National Heart Lung Blood Institute( NHLBI). Exploratory data research of the version of the study in Kaggle on the internet reveals that a lot of column features have been deleted out and the data set is incomplete.

* 1. **Data Preparation/Cleaning Process**

Data preparation is the process of detecting corrupt or dirty data, removing irrelevant parts of the data and replacing it with the correct data. The main purpose of data cleaning is to remove the error and validate the data. Data can be cross checked to remove the errors and the issue can be resolved by validating the data, Data cleaning asks if the data scientist fixed the potential issues that will cause problems during machine learning and further analysis (Sahoo et al., 2019).

Data cleaning is defined as a technique of abolishing and fixing the significant problems and outliers present in the data set. In the real world, data scientists get data from diverse sources such as the internet, social media, and organization. Data scientists use several kinds of programming languages for data cleansing. Data cleansing is thus an essential step in exploratory data analysis (EDA). Dirty data can make an impact in other departments of a firm; therefore, employing data cleansing techniques to clean data is vital. Data cleaning looks for various kinds of errors present in the data set that make it dirty data: null and incomplete values, duplicate values, inconsistent values, multiple values and inaccurate values (Purohit, 2021).

Data Cleaning of the frmgham2.csv dataset begins after loading and importing pandas library and loading the dataset in Jupyterlab an integrated developing environment owned by Anaconda that runs python. Data preparation begins with a “df head()” command which reveals that the first 5 records and a “df tail ()” command which reveals the last 5 records. Secondly, before starting the cleaning process, it is important to explore the shape and size of the data set with a “df.shape” command which when executed, reveals that the data set is 11627 rows and 39 columns. The data set is thus longer than it is wider. Thirdly, is to get the data types by executing the “df.dftypes” command. This reveals that all data types are either float 64 or integers64, informing the researcher that this data may need less cleaning than was anticipated. The data set, however, has a lot of columns, 39, and the researcher will have to focus on the columns (variables) that bring relevance to the study.

The data cleaning in earnest starts by looking for any rows or variables that have missing (null values) by using and executing the “ df.isna ()” command. If the data set has null values it will return “True” values if it has data it returns “False.” This is however a large data set and it is impossible to visualize all rows. For example, this command returns the top 5 records and bottom 5 records as all having false values, meaning that they have data. To visualize and return the null values for the whole data set, the command “ df. Isna(). Sum()” is executed. It reveals that there are a lot of null values for each variable in the data set. They are visualized as a pie chart with percentages but also as a bar chart due to the overlapping of names in the pie chart. For the bar graph, refer to Figure 1 below:

**Figure 1**

*A Bar Graph of Missing values by variables*

A graph with numbers and lines

Description automatically generated with medium confidence

A decision must be made whether to repair the data or drop the null value. To demonstrate what happens to the data set, if the missing data is dropped by executing the command “df.dropna ()” . The command drops the null and not a number from the data set to clean it up and leaves only 2236 rows and 39 columns. This breaks the fidelity in our responsibility as data scientist to represent as much information in the investigation as possible, especially since this is data that will impact the target variable of death. The best ethical decision is made to repair the null values using the means of each missing variable, otherwise the researcher loses too many variables that are useful for data interpretation. The repair command replaces the null values with the mean and no impact to the shape and size of the data set printed, 11627 rows and 39 columns are left intact.

Secondly, a command is run to check for duplicated variables “df.duplicated” and it returns false values. The drop duplicated command is run to help in removing duplicates as follows “df.drop\_duplicates(inplace=True).” Next the data frame is printed “print(df)” and the data frame remains unchanged.

* 1. **Preparing the Data for Exploratory Data Analysis (EDA)**

Data transformation is the process of formatting data before analysis. For example, data type conversion, and datetime cast. It may also involve translation and mapping and changing text into integer float. The following variables in this data set in the medical field are measured as whole numbers and therefore need to be transformed from floats to integers: SYSBP, DIABP, CIGPDAY, BMI, BPMEDS, HEARTRTE, GLUCOSE, educ, HDLC, LDLC, and TOTCHOL. Once transformation is completed, the command “df.info()” confirms that all the variables in the data set are now represented by integers.

* 1. **Exploratory Data Analysis (EDA)**

Exploratory data analysis (EDA) is an essential step in any research analysis. The word exploratory precisely means exactly that; to explore the data in its native form before the start of doing any serious analysis, model testing or revising hypotheses. This involves pre-processing once data has been visualized and its shape deciphered via python commands. The next step is usually cleaning the data to make sure that we have clean data before conducting further EDA analysis or later performing machine learning algorithms. The primary aim of exploratory data analysis is to examine the data for distribution, outliers and anomalies in order to direct specific testing of the hypothesis. EDA also provides tools for hypothesis generation by visualizing and understanding the data usually through graphical representation. EDA is done early on in initial data analysis and serves many functions (Data et al.,2016).

According to another perspective, exploratory data analysis (EDA) is a popular technique or an operation which is mandatory in any kind of analytical project done by using previously collected data sets. EDA is multipurpose: it finds hidden patterns between two variables present in a data set and conducts hypothesis testing by using the least possible structure. EDA is performed by using several techniques such as descriptive statistics, visualization techniques, extraction of information, interpretation, and this is followed by the decision-making process. Exploratory Data Analysis (EDA) is a technique used for the summarization of data by considering the data set's main characteristics and representing it. EDA considers more narrowly checking the presumptions needed to deal with the null or missing values, to fit the model, hypothesis testing, and making changes in attributes. EDA is carried out in order to address an organization's efficiency and performance and eventually help the business to make more informed decisions using the results of the EDA (Purohit, 2021).

* + 1. ***Non graphical EDA***

Non graphical EDA (NGEDA) provides an idea about the description and distribution of the variables. Univariate NGEDA is a principal form of data analysis that involves only one variable to identify underlying data distribution and the characteristics of population distribution. NGEDA includes finding measures of central tendency (mean, median, mode), Spread (standard deviation, variance and interquartile range (Q3-Q1)), skewness and kurtosis. It also covers outlier detection and for any quantitative variable, univariate EDA helps in making initial assessments on the variable distribution using the data sample. NGEDA are objective and quantitative in nature failing to give complete representation of data (Banu et al., 2022). Univariate means only one variable, exposure or outcome is under study. There are two types of NGEDA: univariate NGEDA and multivariate NGEDA. (Data et al., 2016).

* + 1. ***Graphical Exploratory Data Analysis (GEDA)***

Fundamentally, graphical EDA is the counterpart of traditional NGEDA that analyses the data sets to help summarize their statistical characteristics, focusing on the same for key components namely: measures of central tendency, measures of spread, shape of the distribution and existence of outliers. It is divided into univariate GEDA, Bivariate GEDA and multivariate GEDA (Sahoo et al., 2019). Python is an extremely popular programming language today due to its flexibility and wide collection of inbuilt libraries. These libraries are essential to performing data analytics and complex computations.

Python supports multiple libraries for data analytics: NumPy for mathematical and statistical calculations, and Pandas the Python Data Analysis Library. Data visualization plays a vital role in representing the data. Secondly, complex data relationships can be represented graphically making it easy to understand. Python has many libraries that support displaying data in the form of charts, graphs, plots and animations. Two such popular libraries are Matplotlib and Seaborn (Banu et al.,2022).

* 1. **Data Exploration**

This is the first stage of data analysis which provides knowledge about the content and characteristics of the data set. Data Exploration informs people about the size of the data, can also find the missing values of the data, and find possible relationships among data. Data visualization is done using tabular data and understanding the characteristics (Sahoo et al., 2019)

* 1. **Data Analysis**

Data analysis is the collection of different processes to inspect, clean, transform, and model the data with an objective of discovering potentially useful information, drawing several conclusions, and finally supporting business decision-making. Data analysis has become vital to any business that is data driven. EDA evaluates or comprehends data and is a significant component of any data analysis process in data science or machine learning. EDA helps in exploring the data; understanding the structure and relationships between variables and builds a consistent and valuable output for the end user (Banu et al., 2022).

The FHS data set has presence of outliers (values below the minimum and above the maximum whiskers on box plots), as seen in the summary with the 5th and 95th percentile, in the histograms and boxplots. Features with zero data in the summary table have been removed. Refer to Figure 2 for the boxplots showing outliers below.

In finding outliers and dealing with them, the interquartile range method to calculate data points that fall outside of 1.5 times of the inter quartile range (IQR) (Q3-Q1) is utilized. Therefore + and - 1.5\*IQR means only the data within the constraints is being considered. The describe command is run prior to and after outlier removal to make sure that the changes take place and that data transformation has taken place. Executing the “describe()” function is a very vital part of non-graphical EDA (NGEDA). It gives a brief statistical summary of the features: the mean, standard deviation, minimum, maximum, count, and the 3 quartiles (25%, 50% and 75%). It gives the values of the summary of numerical features. Refer to Figure 3 below for the summary table after removal of outliers.

**Figure 2**

*Box plots for the Framingham Heart Study Data*A close-up of a white sheet

Description automatically generated

**Figure 3**

*The Summary Table of Statistics after Outlier Removal*

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Column1** | **count** | **mean** | **std** | **min** | **25%** | **50%** | **75%** | **max** |
| **SEX** | 4393 | 1.623492 | 0.484565 | 1 | 1 | 2 | 2 | **2** |
| **TOTCHOL** | 4393 | 237.122923 | 39.827759 | 129 | 209 | 238 | 262 | **351** |
| **AGE** | 4393 | 49.881402 | 8.256603 | 32 | 44 | 49 | 56 | **78** |
| **SYSBP** | 4393 | 127.153426 | 17.108971 | 85 | 115 | 125 | 137 | **192** |
| **DIABP** | 4393 | 80.484407 | 9.803525 | 53 | 73 | 80 | 87 | **112** |
| **CURSMOKE** | 4393 | 0.475985 | 0.49948 | 0 | 0 | 0 | 1 | **1** |
| **CIGPDAY** | 4393 | 8.692921 | 11.709025 | 0 | 0 | 0 | 20 | **50** |
| **BMI** | 4393 | 24.656727 | 3.437721 | 16 | 22 | 24 | 27 | **35** |
| **HEARTRTE** | 4393 | 75.531983 | 11.077675 | 45 | 68 | 75 | 82 | **109** |
| **GLUCOSE** | 4393 | 78.664011 | 10.188882 | 52 | 72 | 79 | 84 | **108** |
| **educ** | 4393 | 2.058275 | 1.024193 | 1 | 1 | 2 | 3 | **4** |
| **PREVHYP** | 4393 | 0.258593 | 0.437911 | 0 | 0 | 0 | 1 | **1** |
| **TIME** | 4393 | 1104.58935 | 1199.95526 | 0 | 0 | 0 | 2177 | **4611** |
| **PERIOD** | 4393 | 1.50717 | 0.550299 | 1 | 1 | 1 | 2 | **3** |
| **HDLC** | 4393 | 49 | 0 | 49 | 49 | 49 | 49 | **49** |
| **LDLC** | 4393 | 176 | 0 | 176 | 176 | 176 | 176 | **176** |
| **DEATH** | 4393 | 0.081493 | 0.273622 | 0 | 0 | 0 | 0 | **1** |
| **ANYCHD** | 4393 | 0.005691 | 0.075231 | 0 | 0 | 0 | 0 | **1** |
| **HYPERTEN** | 4393 | 0.661735 | 0.473173 | 0 | 0 | 1 | 1 | **1** |
| **TIMEAP** | 4393 | 8660.36786 | 410.080838 | 5146 | 8766 | 8766 | 8766 | **8766** |
| **TIMEMI** | 4393 | 8660.36786 | 410.080838 | 5146 | 8766 | 8766 | 8766 | **8766** |
| **TIMEMIFC** | 4393 | 8660.36786 | 410.080838 | 5146 | 8766 | 8766 | 8766 | **8766** |
| **TIMECHD** | 4393 | 8642.09356 | 493.072462 | 1675 | 8766 | 8766 | 8766 | **8766** |
| **TIMESTRK** | 4393 | 8660.36786 | 410.080838 | 5146 | 8766 | 8766 | 8766 | **8766** |
| **TIMECVD** | 4393 | 8660.36786 | 410.080838 | 5146 | 8766 | 8766 | 8766 | **8766** |
| **TIMEDTH** | 4393 | 8666.87799 | 390.688409 | 6381 | 8766 | 8766 | 8766 | **8766** |
| **TIMEHYP** | 4393 | 4912.79854 | 3490.78727 | 0 | 1448 | 5171 | 8766 | **8766** |

* 1. **The Rationale for Features Selection**

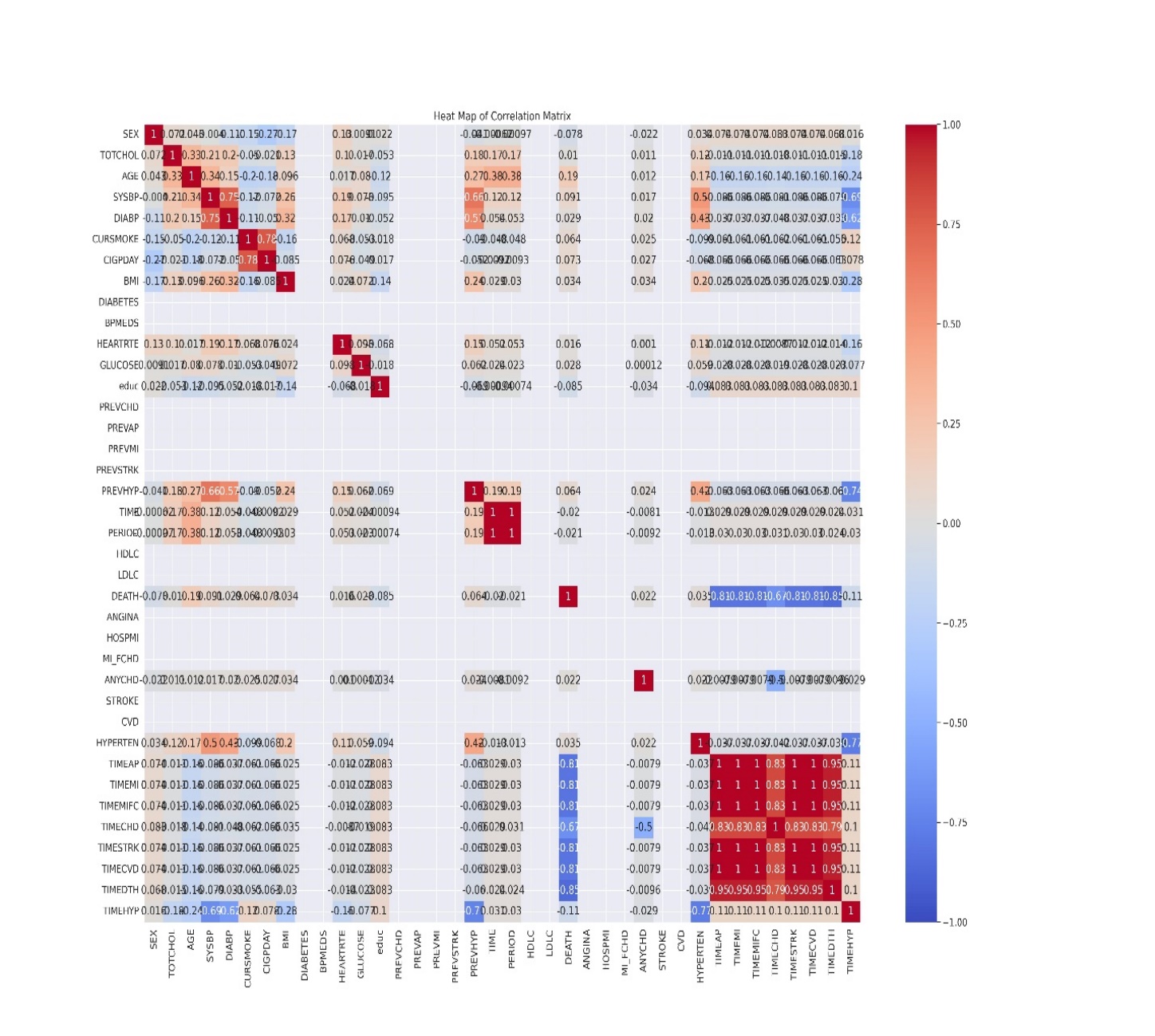
There are 39 features(variables) and this number must be narrowed down to a smaller number for data analysis, and later for model building. To achieve this a heat map of the cleaned and transformed variables is run. A heat map is a data visualization tool that uses color to represent values in a matrix or 2D data. It is particularly useful for showing patterns, correlations, or intensities across different variables or categories. The colors in a heat map help quickly identify high and low values, trends, or clusters within the dataset.

In machine learning and data analysis, heat maps are often used to visualize the correlation between different variables. The correlation matrix is shown with colors representing the strength and direction of relationships. In this analysis, after running a heat map, it is followed by a deletion of all variables that have zero correlations: DIABETES, BPMEDS, PREVCHD, PREVAP, PREVMI, PREVSTRK, HDLC, LDLC, ANGINA, HOSPMI, MI\_FCHD, STROKE, CVD, TIME, PERIOD. Refer to Figure 3 below. Figure 4 shows all features that have correlations in the FHS data set.

A study is then conducted looking for strong and weak correlations. Positive (red) correlation means as an independent variable increases, the target variable decreases. Negative correlation (blue) means as an independent variable decreases, the target variable increases. The heat map also displays correlation coefficients ranging from -1 to 1. These values represent the strength and direction of the relationship between two variables. Based on the correlation heatmap Figure 4, the researcher chooses death as the target variable to study. It is felt that death would end up representing a larger number than ANYCHD and CVD is not an option left on the table since it is deleted for not having any correlation with any variables. Refer to Figure 4 below.

**Figure 3**

*A heat map that includes all features prior to deletion of variables without correlation*



The analysis utilizes a comprehensive cleaned and transformed dataset with the following features: demographics: Age, sex, education level; clinical measurements: total cholesterol (TOTCHOL), systolic blood pressure (SYSBP), diastolic blood pressure (DIABP), body mass index (BMI), heart rate (HEARTRTE), glucose levels (GLUCOSE); lifestyle factors: Smoking status (CURSMOKE), average cigarettes smoked per day (CIGPDAY); health history: previous history of hypertension (PREVHYP), coronary heart disease (ANYCHD), and other relevant medical histories.

**Figure 4**

*Heat Matrix Correlation after dropping variables without correlation*

A screenshot of a graph

Description automatically generated

1. **Findings**
   1. **Death Count as a Comparison to Non-Deaths**

Refer to figure 5 below for interpretation of results with death being (8.1%) count compared to people in the trial who did not die (91.9%). A pie chart and bar chart are provided for side-by-side comparisons.

**Figure 5**

**A blue and orange pie chart

Description automatically generated***Death versus No death Distribution Pie Chart and Bar Graph*

**A graph with a bar

Description automatically generated**

* 1. **Correlation of variables with the target variable Death**

The research involved finding how all the variables compared to death when it comes to correlation. For features that are positive on the graph, it means that as the variable increases death also increases these include: Total cholesterol, age, systolic BP, Diastolic BP, Current smoker, cigarettes per day, BMI, Heart rate, Glucose, prevalent hypertension, ANY Coronary Heart Disease and hypertension. Conversely for features that are negative on the graph, as the independent variable decreases, the target variable, death increases. These include education level, time: to Angina Pectoris, Myocardial Infarction, hospitalized Myocardial Infarction or Fatal Coronary Heart Disease, Coronary Heart disease, Stroke, Cardiovascular Disease, Death and Hypertension. Refer to Figure 6 below:

**Figure 6**

*Correlation of variables with the target variable Death*

A graph with different colored bars

Description automatically generated

* 1. **Death by Gender**

Further research was conducted by Groupby of sex on DEATH. Results showed that more males suffered from death than women. In other words, being a male put you at higher risk for death from any cause including fatal coronary heart disease than being a woman. Refer to figure 7 below:

**Figure 7**

*Death by Gender Category*

**A graph of death and death

Description automatically generated**

* 1. **Multivariate Analysis of SYSBP,TOTCHOL and DEATH**

A multivariate scatterplot analysis of systolic blood pressure and total cholesterol showed that as systolic blood pressure increased and total cholesterol (HDLC and LDLC) increased, the number of deaths greatly increased from left of the graph with the highest increase in the right top quadrant. Refer to Figure 8 below:

**Figure 8**

*Multivariate analysis of Systolic Blood Pressure, Total Cholesterol and Death.*

A graph showing a number of numbers

Description automatically generated with medium confidence

* 1. **Logistic Regression Model**

A logistic regression model is a type of regression model used for binary classification whereby the target variable has two possible outcomes such in this case death or no death. A logistic regression machine model is developed to evaluate the relationship between the selected risk factors and the likelihood of mortality. Target variable: DEATH (binary: 1 = deceased, 0 = survived). Features: 12 predictor variables, including clinical measurements and lifestyle factors are used. The logistic regression model is trained to identify whether a patient is at risk of death based on various health metrics and lifestyle factors. The logistic regression model analyzes various risk factors to predict the likelihood of death based on the Framingham Heart Study dataset.

It is important to note that further data preprocessing occurs before machine learning in logistic regression. That due to the sampling being low and there being an imbalanced dataset, the minority set is oversampled to generate synthetic observations. This technique of data transformation that is used is called the Synthetic Minority Oversampling (SMOTE). The data is split and also standardized via a z score normalization in order to normalize or rescale features so that they have properties of normal distribution.

* + 1. ***Logistic Regression Model Accuracy***

The accuracy of the model was 96.47%. The model's accuracy indicates that 99.47% of the predictions made by the model are correct. This is a strong performance, showing that the model is exceptionally reliable in identifying the outcomes.

* + 1. ***Confusion Matrix***

The confusion matrix provides a detailed breakdown of the model's performance in terms of true positives, true negatives, false positives, and false negatives. Firstly, true negatives (1210): the model overidentifies 800 patients as not at risk of death (i.e., predicted 0 and actual outcome was 0). Secondly, false positives (0): the model correctly predicts that 0 patients were at risk of death when they were not. Thirdly, false negatives (9): the model incorrectly predicts that 9 patients were not at risk of death when they were. Fourth, true positives (99): the model correctly overidentifies 99 patients as *at risk of death* (i.e., predicted 1 and actual outcome was 1). For the logistic regression confusion matrix refer to Figure 9 below.

* + 1. ***Classification Report***

A classification report is a summary of key metrics used to evaluate the performance of a classification model such as the logistic regression model above. It provides detailed insights into how well the model is performing across different classes. The classification section provides additional performance metrics: precision, recall, and F1-score and Support. These performance metrics give more insights into how well the model handles each class (death vs no death).

**Figure 9**

*Logistic Regression Confusion Matrix*

A diagram of a blue and red box

Description automatically generated with medium confidence

Precision measures how many of the predicted positive instances are actually positive. The formula is True Positives divided by (True Positives + False Positives). On the other hand, recall is the sensitivity or true positive rate, which measures how many of the actual positive instances were correctly predicted by the model. The formula is True Positives divided by (True Positives + False negatives). The F1-Score is the harmonic means of precision and recall which balances the two. It ranges from 0 to 1 where 1 indicates perfect precision and recall. The F1 formula is 2 multiplied by the fraction on ( (Precision x Recall)/ (Precision + Recall)),Finally, there is Support which provides the number of actual occurrences of each class in the dataset. It provides context for the Precision, Recall and F1 score values. Refer to figure 10 below:

**Figure 10**

*A Classification Report from the Logistic Regression Model*

| **Metric** | **Class (No Death)** | **Class (Death)** |
| --- | --- | --- |
| **Precision** | 0.99 | 1 |
| **Recall** | 1.00 | 0.92 |
| **F1-Score** | 1 | 0.96 |
| **Support** | 1210 | 108 |

* + 1. ***Explanation of the Classification Report:***

The results of the logistic regression model represented by the classification report from Figure 10 above can be explained as follows:

1. Precision: for patients predicted *not at risk of death (0)*, 99% of those predictions were correct. For patients predicted *at risk of death (1)*, 100% of those predictions were correct.
2. Recall: the model successfully identified all patients in the *not at risk of death* category (100% recall). For the *at risk of death* category, the model identified 92% of those at actual risk (there is room for improvement here in identifying patients at risk).
3. F1-Score: the F1-score balances precision and recall, showing a perfect performance of 1 for patients not at risk of death, and a high score of 0.96 for those at risk once again showing that the model leaves some room for improvement.
4. Support: the model predicts in total 1210 do not die while 108 die. These numbers are higher than the actual numbers.
   * 1. ***Cross validation of the Logistic Regression Model’s ROC AUC***

Cross validation is a technique used in machine learning to assess the performance and generalizability of a model. It helps in assuring that the model performs well not just on the training data but also on unseen data. This is important because it helps prevent overfitting of the model whereby the model performs well on training data but poorly on unseen data. Cross validation of the logistic regression model is carried out using the K-fold cross-validation of 10 folds. The cross-validation results are provided and they show how well the model performs across different folds of the Framingham dataset, specifically using the **ROC AUC** metric. ROC AUC stands for Receiver Operating Characteristic (ROC) and the area under the curve(AUC). Refer to Table 11 below:

**Table 11**

*Cross Validation results of the Logistic Regression Model*

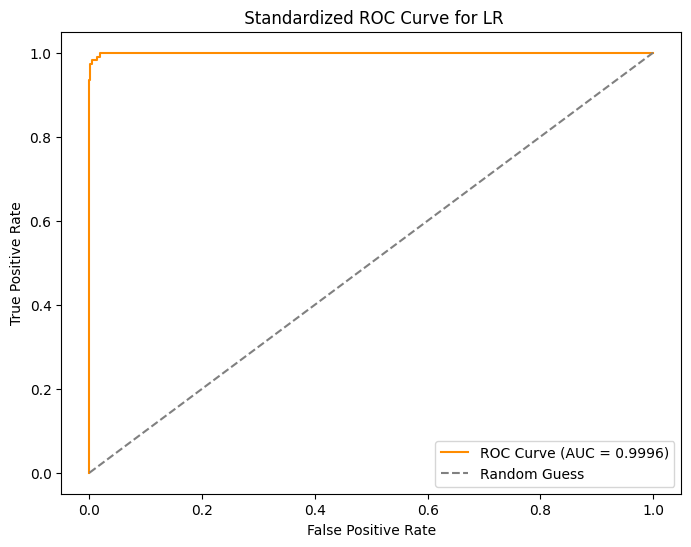
|  |  |  |
| --- | --- | --- |
| **Cross-validation ROC AUC results** | **Mean ROC AUC** | **Standard Deviation** |
| 1.000 | 0.9552 | 0.0335 |
| 0.931 |  |  |
| 0.889 |  |  |
| 0.957 |  |  |
| 0.929 |  |  |
| 0.931 |  |  |
| 0.986 |  |  |
| 0.958 |  |  |
| 0.986 |  |  |
| 0.986 |  |  |

Overall conclusion is that the regression model is performing very well across the different folds, with an average ROC AUC of 0.9552. The low standard deviation means that the performance is consistent, and the model generalizes well across the dataset. While there is a slight dip in performance on some folds (e.g., fold 3 with a score of 0.889), the model's overall ability to classify correctly is strong. ROC AUC measures the ability of the model to discriminate between the positive class( individuals who will die) and the negative class (individuals who do not die).

Plotting a ROC Curve with Python produces a ROC AUC of 0.9996. An area under curve of 99.96% means that the Linear Regression Model ranks a randomly chosen positive instance (death) much higher than a randomly chosen negative instance (no death). In other words, the Linear Regression model has a 99.96% chance of correctly distinguishing between a person who will die and one who will not, based on the available features (variables) such as age, total cholesterol and blood pressure. An AUC of 99.96% in evaluating the model means that the Linear Regression model is doing an excellent job at discriminating and is accurate at distinguishing between classes. This result of 99.96% means that the model has a high predictive power and is suitable for making decisions based on the Framingham data at St Francis hospital. Results are shown in Figure 12 below:

**Figure 12**

*ROC Curve for Logistic Regression Model on Framingham Data*

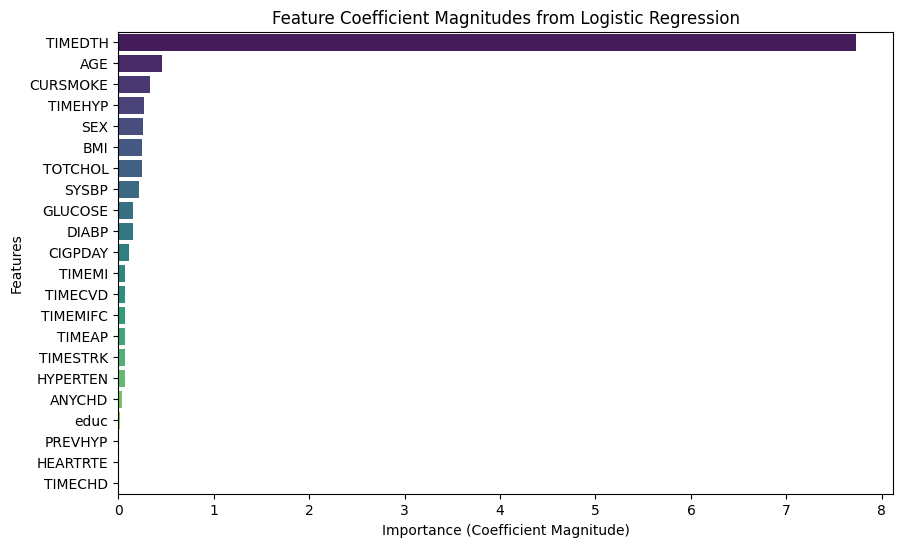


* + 1. ***Standardized Logistic Regression Coefficients***

Further research is conducted to rank the features ( risk factors) by order of importance of contributing to mortality, the dependent variable. Time till death occurs ranks first, followed by age which increases the risk of death, third is current smoking, fourth is time till hypertension, and fifth is sex. From earlier analysis presented here, there is more death in males than females. Being a male predisposes one to death occurring more than by being a female. For the rest of the features ranking, refer to Figure 13 below:

**Figure 13**

*Feature Coefficient Magnitudes from Logistic Regression*



* + 1. ***Actionable Insights for St. Francis Hospital based on the Logistic Regression:***
       1. **High accuracy and precision**: the model shows strong overall performance with a high accuracy of 96.47%. It is especially good at identifying patients who are not at risk of death, which could be useful for resource allocation and prioritizing patients with higher risks.
       2. **Room for improvement in identifying high-risk patients**: the recall for the patients at risk of death category is 92%, which suggests that some patients who are actually at risk may not be flagged by the model. Improvements could include refining the model with additional risk factors or focusing more on improving the recall for thi***s*** class.
       3. **Audit and monitor key risk factors***:* regular updates should be made to patient charts with key health metrics (such as age, cholesterol, blood pressure, BMI, and smoking status) to ensure the risk predictions are based on the latest data. Tracking these factors at each patient visit could improve early detection of high-risk individuals for death.
       4. **Prioritize monitoring of patients with borderline metrics***:* for patients with metrics close to the decision threshold, more frequent monitoring and follow-up appointments could be recommended to reduce the number of false negatives (patients missed as high risk).
       5. ***Conclusions from the Logistic Regression:***

The logistic regression model provides a reliable prediction tool to assess the risk of death in patients. With further refinement, especially improving recall for high-risk patients, it can be integrated into St. Francis Hospital’s patient monitoring system, supporting decision-making and optimizing resource allocation.

1. **Implications for Further Data Research**
   1. **Random Forest Model**

The results of the Logistic Regression Model suggest the need to improve the model based on model accuracy and recall for high-risk patients. A second machine learning model that is more robust is employed, namely the Random Forest. In this analysis, a Random Forest model is applied to predict the risk of death, based on various health and lifestyle factors from the Framingham dataset. It is the hope of the researcher that the model’s performance on predicting death is found to be highly accurate and thus providing potential for clinical applications in monitoring and reducing mortality risk in high-risk patients at St. Francis Hospital.

Random Forest is an ensemble learning method used for classification, regression, and other tasks. It operates by constructing a large number of decision trees during training and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. It is based on the concept of bagging (Bootstrap Aggregating), where multiple models are trained on different subsets of data and combined to improve accuracy and reduce overfitting.

* + 1. ***Random Forest Analysis Objective***

The objective of this analysis is to help St. Francis Hospital identify key risk factors for death in patients and implement a more data-driven approach to patient care. The results from the Random Forest model can be used to:

* Assist in identifying patients with a high risk of death.
* Inform preventive care strategies and follow-up protocols.
* Ensure better documentation and auditing processes during patient visits.
  + 1. ***Advantages of Random Forest Model to Logistic Regression Model***:

1. Handles non-linearity: Random Forest can capture non-linear relationships between features and the target variable because decision trees split data based on complex conditions. Logistic Regression assumes a linear relationship between the features and the log-odds of the outcome, which may not always hold true, especially with complex datasets.
2. Feature importance: Random Forest provides a measure of feature importance based on how much each feature contributes to reducing prediction error in the trees, which helps in feature selection. Logistic Regression can show the relationship between features and the target variable, but it does not have built-in feature importance scores like Random Forest.
3. Better performance on large datasets: Random Forest tends to perform better when dealing with large datasets and high-dimensional data, as it can handle many features and detect complex interactions between them. Logistic Regression may struggle with large datasets if the relationship between variables is highly non-linear.
4. Works well with categorical variables: Random Forest can handle categorical features naturally (via decision trees), without the need to create dummy variables or one-hot encoding. Logistic Regression requires categorical variables to be transformed into dummy variables, which can lead to an increase in dimensionality.
5. Robust to overfitting: Random Forest, due to the averaging of multiple trees, is less prone to overfitting than a single decision tree. The ensemble approach generally prevents it from fitting noise in the training data. Logistic Regression can overfit, especially if there are many correlated features or the model is too complex without regularization.
6. Handles missing data: Random Forest can handle missing data relatively well by splitting data based on the available features and using surrogate splits. Logistic Regression typically requires missing data to be imputed before model training.
7. Works well with large feature sets: Random Forest can manage datasets with many features, as it selects a random subset of features at each split, making it computationally efficient for large feature sets. Logistic Regression may struggle with large feature sets, especially if regularization is not applied, as it does not have any mechanism to handle irrelevant features directly.
   * 1. ***Random Forest Accuracy***

The Random Forest model results for the Framingham Heart Study dataset on death indicate that the model has 100% accuracy. This means that the model predicts all instances (deaths and non-deaths) correctly. This implies perfect classification for this dataset and this level of model accuracy should raise concerns about overfitting. Overfitting means that the model performs too perfectly on this Framingham data, but that it might not generalize well to unseen data for example from St Francis Hospital.

* + 1. ***Confusion Matrix of Random Forest***

There are 802 True Negatives (TN), these are cases where the model correctly predicts that the patient does not die. Secondly, there is 0 False Positives (FP) meaning that there are no cases where the model incorrectly predicts death when there was none. Thirdly, there are 0 False Negatives (FN) meaning that there are no cases where the model fails to predict death when the patient does indeed die. Fourth, there are 77 True Positives (TP) meaning that the model correctly predicts all cases of death. Refer to Figure 14 below.

* + 1. ***Random Forest Classification Report***

Precision (1.00 for both classes). This means that the model predicts all instances accurately. Precision is the proportion of true positives to all predicted positives. This means that there are no false alarms. Recall (1.00 for both classes), meaning that the model detects all actual deaths correctly. F1-Score (1.00 for both classes) which is the harmonic means of precision and recall, indicating a perfect balance. Support, which is the number of actual instances in each class (802 non-deaths, 77 deaths). Refer to Figure 15 below.

**Figure 14**

A blue and white graph

Description automatically generated*Confusion Matrix Results for Random Forest*

**Figure 15**

***Classification Results for Random Forest***

| **Metric** | **Class (No Death)** | **Class (Death)** |
| --- | --- | --- |
| **Precision** | 1 | 1 |
| **Recall** | 1 | 1 |
| **F1-Score** | 1 | 1 |
| **Support** | 802 | 77 |

* + 1. ***Interpretation for St. Francis Hospital from the Random Forest Model:***

The Random Forest model demonstrates a perfect performance in predicting death in the Framingham study dataset. However, in a real-world context, this might suggest overfitting, meaning the model is too finely tuned to the current data and may not generalize well to new patient data at St Francis Hospital. It is important to evaluate the model's performance on unseen data or through cross-validation to ensure robustness.

St Francis hospital can nonetheless use these findings to audit and monitor key risk factors. Secondly, the hospital can ensure that patient charts are updated with relevant risk factors (e.g., age, cholesterol, blood pressure, cigarette smoking) at every visit. Thirdly, the hospital can develop preventive strategies by focusing on high-risk patients that the model identifies so as to reduce mortality rates.

* + 1. ***Key Risk Factors Analyzed***

Here are data on what the Random Forest model reveals about the significant predictors (risk factors) for death. This impacts patient chart audits and documentationand based on the results, St. Francis Hospital should audit patient charts for the following risk factors at every visit and update the patient’s electronic medical record:

1. Total cholesterol (TOTCHOL) a total of LDLC and HDLC: measures the amount of cholesterol in the blood. Higher cholesterol levels are linked to an increased risk of cardiovascular disease and eventual mortality.
2. Age (AGE): older age is a significant predictor of mortality, as risk increases with age.
3. Systolic blood pressure (SYSBP): hypertension characterized by high blood pressure is a major risk factor for heart disease and stroke, both of which increase the risk of death.
4. Diastolic blood pressure (DIABP): diastolic pressure, the bottom number in blood pressure readings also plays a role in assessing cardiovascular risk and mortality.
5. Cigarettes per day (CIGPDAY): smoking is a well-known risk factor for many chronic conditions, including heart disease, lung disease, cancer, and death. Smoking habits should be documented at every patient visit. to indite Encouraging cessation programs for smokers can have a substantial impact on reducing mortality risks.
6. Body Mass Index (BMI): higher BMI, indicating overweight status or obesity, is associated with increased mortality risk.
7. Heart rate (HEARTRTE): elevated resting heart rates can be indicative of cardiovascular stress and are linked to mortality.
8. Glucose (GLUCOSE): high glucose levels are often associated with diabetes, which is a risk factor for cardiovascular disease and death.
9. Education level (educ): lower education levels have been linked to poorer health outcomes due to lifestyle factors and access to healthcare.
10. Previous hypertension diagnosis (PREVHYP): a history of hypertension is a key indicator of future cardiovascular events and death. Ensure that a history of hypertension is flagged, as this strongly correlates with cardiovascular risk.
11. Any coronary heart disease (ANYCHD): coronary heart disease significantly increases the risk of death due to its impact on heart function.
12. Hypertension diagnosis (HYPERTEN): active hypertension is a strong predictor of mortality.

By making sure these key factors are updated and reviewed regularly in patient charts, St Francis can better manage patient health and potentially reduce mortality rates.

* + 1. ***Recommendations for St. Francis Hospital from the Random Forest Model***

1. **Implement chart audits**: regular audits of patient charts should be conducted to ensure accurate recording of the most relevant clinical features (age, cholesterol, blood pressure, and smoking).
2. **Develop follow-up programs**: patients identified as high-risk based on these factors should be enrolled in follow-up programs, focusing on managing modifiable risks such as smoking cessation and hypertension management.
3. **Further model validation**: despite perfect accuracy, it is essential to validate the model on new data or perform cross-validation to ensure it generalizes well to other patient populations.
4. **Clinical decision support systems**: integrate the Random Forest model into hospital systems to assist clinicians in decision-making, allowing them to flag high-risk patients for closer monitoring by healthcare providers.
   * 1. ***Conclusions from the Random Forest Mode:***

The Random Forest model provides valuable insights into the risk factors for death, allowing St. Francis Hospital to focus on high-risk individuals and update their care strategies accordingly. Although the model shows perfect performance, additional validation is necessary to ensure its robustness in a real-world setting. Nonetheless, this analysis forms the foundation for better predictive healthcare and risk management at the hospital.

**4.2 Extreme Gradient Boosting (XGBoost) Model**

This data science analysis also involves a third machine learning model, XGBoost, short for Extreme Gradient Boosting to improve upon the results of the logistic regression. The results obtained were exactly similar to those of Random Forest in terms of 100% accuracy and the confusion matrix. XGBoost is a popular and powerful machine learning algorithm that is particularly well-suited for structured/tabular data. It is an implementation of the gradient boosting framework and has gained widespread use in various data science competitions and practical applications due to its efficiency and accuracy. Refer to Figure 16 below for results. What makes XGBoost so ideal are its key features and concepts:

*4.2.1* **Boosting Algorithm**: XGBoost is an ensemble learning method that builds models in a sequential manner. It combines the predictions of several base learners (often decision trees) to improve the overall performance of the model. Each new model attempts to correct the errors made by the previous models.

*4.2.2* **Gradient Boosting**: the algorithm uses gradient descent optimization to minimize the loss function. It calculates the gradients (i.e., direction and rate of change of the loss function) to update the model iteratively. This approach helps in improving performance and reducing errors.

*4.2.3* **Regularization**: XGBoost includes L1 (Lasso) and L2 (Ridge) regularization techniques to prevent overfitting. This is a significant advantage over traditional gradient boosting methods.

*4.2.4* **Tree Pruning**: XGBoost employs a more sophisticated method of tree pruning, which allows XGBoost to grow trees in a more efficient manner. It uses a depth-first approach to optimize the construction of trees and can stop splitting nodes when certain conditions are met.

*4.2.5* **Handling Missing Values**: XGBoost can automatically handle missing values in the dataset, making it easier to work with real-world data where missing entries are common.

*4.2.6* **Parallel and Distributed Computing**: the algorithm is designed to make use of multiple CPU cores, enabling faster computation. It can also be scaled across distributed computing environments, which is beneficial for large datasets.

*4.2.7* **Flexibility**: XGBoost supports various objective functions, including regression, classification, and ranking problems, making it a versatile tool for a range of applications.

*4.2.8* **Feature Importance**: XGBoost provides several methods to assess the importance of different features in the model, which can help in feature selection and understanding the data better. XGBoost had 100% accuracy, a precision of 1 in the positive and negative instances for death, and likewise the recall was 1 for both positive and negative instances, while the F1 score was also 1 for both groups. Given that this model avoids overfitting, this should give St Francis confidence to adopt this algorithm for all its patients both high and low risk.

**Figure 16**

*XGBoost Standardized Confusion Matrix*

A blue and white graph

Description automatically generated

**5. Recommendations for St Francis Hospital from Data Science Analysis**

1. **Implement chart audits**: regular audits of patient charts and electronic medical records should be conducted to ensure accurate recording of the most relevant clinical features (age, cholesterol, blood pressure, and smoking).
   1. **Develop follow-up programs**: patients identified as high-risk based on factors mentioned should be enrolled in follow-up programs, focusing on managing modifiable risks such as smoking cessation and hypertension management.
   2. **Further model validation**: Despite the great accuracy of the Logistic Regression model and the perfect accuracy depicted by Random Forest and XGBoost, it is essential to validate the model on new data or perform cross-validation to ensure it generalizes well to other patient populations at St Francis.
   3. **Clinical decision support systems**: integrate the Logistic Regression Model, Random Forest and XGBoost models into the St Francis hospital, and Franciscan Alliance network of systems to assist clinicians in decision-making, allowing them to flag high-risk patients for closer monitoring.
   4. **Public awareness campaigns**: Increase awareness of risk factors associated with coronary heart disease through public health initiatives.
2. **Targeted Interventions**: Develop targeted lifestyle modification programs aimed at high-risk groups

**6 Conclusion**

The longitudinal Framingham Heart Study data remains a golden standard and critical in cardiovascular risk assessment. These findings underscore its utility in predicting DEATH and the importance of addressing modifiable risk factors to improve patient outcomes. According to O’Donnell and Elosua (2008), epidemiology involves the study of disease frequency and its determinants within the population. Cardiovascular epidemiology began in the 1930s as a result of changes observed in the causes of death and in the 1950s, several epidemiological studies were set in motion with the aim of clarifying the cause of cardiovascular disease. Four years after the Framingham Heart Study started, researchers had identified high cholesterol and high blood pressure levels as important factors in the development of cardiovascular disease. In subsequent years, the Framingham study and other epidemiological studies have helped to identify other risk factors, which are now considered classical risk factors.

By coining the expression "risk factor", the Framingham Heart Study helped to bring about a change in the way medicine is practiced. Today, a risk factor is defined as a measurable characteristic (feature). that is causally associated with increased disease frequency and that is a significant independent predictor of an increased risk of presenting with the disease. This wide-ranging overview describes some of the most important insights into the causes of cardiovascular disease to have come from the Framingham Heart Study. The emphasis is on the identification of risk factors, and the assessment of their predictive ability and their implications for disease prevention( O’Donnell & Elosua, 2008). This data science analysis has helped identify those risk factors that cause death based on Logistic Regression, Random Forest and XGBoost. St Francis hospital should adopt the recommendations and be on its way to reducing cardiovascular disease and death.

References

Banu, A., Reddy, S., & Rama, M.(2022). Graphical exploratory data analysis (GEDA): A case study on employee attrition. *Journal of Science and Technology,7*(09), 01-11. <https://doi.org/10.46243/jst.2022.v7.109.pp01-11>

Data, M. C., Komorowski, M., Marshall, D. C., Salciccioli, J. D., & Crutain, Y. (2016). Exploratory data analysis. Secondary analysis of electronic health records. Springer, Cham. <https://doi.org/10.1007/978-3-319-43742-2_15>

O'Donnell, C. J., & Elosua, R. (2008). Cardiovascular risk factors. Insights from framingham heart study. *Revista Española de Cardiología (English Edition)*, *61*(3), 299-310. <https://doi.org/10.1016/S1885-5857(08)60118-8>

Purohit, K. (2021). Separation of Data Cleansing Concept from EDA. *International Journal of Data Science and Analysis* *7*(3): 89-97. [https://doi.org/10 .11648/j.ijdsa.20210703.16](https://doi.org/10%20.11648/j.ijdsa.20210703.16)

Sahoo, K., Samal, A. K., Pramanik, J., & Pani, S. K. (2019). Exploratory data analysis using Python. *International Journal of Innovative Technology and Exploring Engineering, 8*(12), 4727-4735. [https://doi.org/10.35940/ijitee.L3591 .1081219](https://doi.org/10.35940/ijitee.L3591%20.1081219)