HUMBER INSTITUTE OF TECHNOLOGY

AND ADVANCED LEARNING

BIA-5401-0GA

# FINAL GROUP PROJECT

# Predicting Death Events for Heart Failure

Submitted by: Group 4 Grade/Comments:


| First Name | Last Name | Student Number |
|---|---|---|
| Amarachi | Ezeji | N01510367 |
| Aparajita | Roy | N01511087 |
| Ayesha | Karadia | N01514568 |
| Gbenga | Adebowale | N01513109 |
| Ivin | Alexander | N01474172 |
| Navdeep | Chauhan | N01514720 |


Submitted to: Professor Haytham Qushtom

Submission Date: 08/16/2023

# TABLE OF CONTENTS

# INTRODUCTION

Cardiovascular diseases (CVDs) stand as the foremost cause of global mortality, claiming an estimated 17.9 million lives annually, representing approximately 31% of total worldwide deaths. Heart failure is a serious condition that affects millions of people worldwide, including many of our loved ones. It occurs when the heart is unable to pump blood effectively enough to meet the body's needs. This can lead to a variety of symptoms, including shortness of breath, fatigue, and swelling in the legs. Heart failure can also be fatal.

At the Westway Medical Clinic, we are committed to providing the best possible care for our patients with heart failure. You know that early diagnosis and treatment are essential for improving patient outcomes. However, you also know that it can be difficult to identify patients who are at risk of developing heart failure or who are not responding to treatment.

With the high number of patients and mortality in recent times, the organization's executives and the medical team decided to reach out to the analysis/IT department to use the historical clinical records of patients to predict the death events of new patients as well as old patients. In this report, we will explore the potential of using BI to improve the diagnosis and treatment of heart failure in our patients. We will first provide an overview of the heart failure dataset that we will be using. We will then discuss the benefits of using BI for heart failure research and care. Finally, we will propose a specific BI solution that could be used to improve the diagnosis and treatment of heart failure in your patients.
Through this process, the predicted death events (yes or No) can be used to offer the right treatment, approach, diagnosis, care, and information to the medical team, the patient, and the patient's families about their condition. The Westway Medical Clinic is planning to save at least 50-100 patients per year who suffer from this health condition.
The Westway Medical Clinic will capture the below attributes to predict Heart Failure death events.

1. Age

2. Anaemia

3. Diabetes

4. Creatinine Phosphokinase

5. Ejection fraction

6. High blood pressure

7. Platelets

8. Serum creatinine

9. Serum sodium

10. Sex

11. Smoking
12.Time

In the report, our analytics team has used a machine learning model wherein the patient information mentioned above is entered and the model predicts a death event (Yes or No), which can then be used to administer the right treatment, care, and information to the patient and their families by the medical team/ specialists for each new patient.
If this model is successful and gives a good prediction which can save a life, Westway Medical Clinic will be able to save 50-100 patients per year, which would help save families from incurring debt.

## EXECUTIVE SUMMARY

Cardiovascular diseases (CVDs) are a major global health concern, accounting for a substantial amount of global mortality. This concerning pattern is influenced by heart failure, a disease in which the heart struggles to adequately pump blood. The Westway Medical Clinic understands the need for early detection and treatment for heart failure patients. To solve this issue, we went on a data-driven journey, employing Business Intelligence (BI) approaches to forecast death occurrences and improve patient care.

Our analysis describes the entire strategy used to attain these goals. We begin by presenting a summary of the heart failure dataset, which includes important patient characteristics such as age, anaemia, diabetes, and more. Using this information, we want to develop a prediction algorithm that can reliably forecast the risk of fatal events in heart failure patients.

The benefits of this BI solution are numerous. Key benefits include early identification and intervention, effective resource allocation, educated clinical judgments, lower death rates, cost savings, and improved patient outcomes. The Westway Medical Clinic hopes to save 50-100 people each year by utilizing predictive analytics, ultimately having a substantial influence on families and communities.

Despite its intriguing promise, putting this BI solution into action poses certain challenges. Ensuring data quality and privacy, identifying important characteristics, dealing with unbalanced data, and overcoming algorithmic bias all require careful thought. Integrating predictive models into clinical procedures, dealing with ethical problems, and adhering to legal and regulatory requirements are all critical components of effective implementation.

The analysis explores into the solution's technological implementation, beginning with data pretreatment and exploratory data analysis. We demonstrate our findings using interesting visuals that shed light on important relationships and distributions within the dataset. The predictive modelling approach is then described in depth, using Random Forest, Neural Network, and Logistic Regression models. These models have varied degrees of accuracy and have the capacity to forecast mortality occurrences depending on patient characteristics.

In summary, our BI-driven approach presents a significant opportunity to revolutionize heart failure patient care, reduce mortality rates, and allocate resources efficiently. By combining medical expertise with cutting-edge analytics, the Westway Medical Clinic is poised to make a substantial positive impact on patient outcomes and healthcare practices.

## PROBLEM STATEMENT

In the field of healthcare, dealing with heart problems like heart failure is tough. Hospitals want to give patients the best care and use their resources wisely. But to do that, they need to understand what factors affect how bad heart failure gets.

Now, the company wants to reduce the mortality rate for new and old patients. But the company doesn't know how to go about this with the data it has. So, the company wants to build a system that would predict if there will be a death event or no death event based on factors like Age, anaemia (yes or no), diabetes, creatinine phosphokinase, ejection fraction, high blood pressure, platelets, serum creatinine, serum sodium, sex, smoking, and time. Based on the existing clinical records of old patients, this projection is anticipated to be useful in helping patients prepare themselves for any event (death) and also for the organisation to know what actions to take to reduce the mortality rate.

The challenge here is to make a system that can predict if someone with heart failure might die. We have a bunch of information about patients, and we want to find out which attributes are linked to bad outcomes. The main aim is to build a model that can predict if someone with heart failure is likely to die. If we can do this, doctors can use it to choose treatments, use resources better, and help patients more.

Because catching problems early and having the right plan is super important in healthcare, solving this challenge would really help. It could mean fewer people dying from heart issues, smarter use of medical resources, and better results for patients.

## SOLUTION IMPLEMENTATION

We prepared our dataset for analysis by cleaning it. Using the "data.isnull().sum()" code to count the amount of missing values in each column of the Data Frame data, we first attempted to identify missing values in the dataset. Since all the counts are 0, it appears that

there are no missing values in any of the columns in this instance. This is a strong signal that there aren't many missing values in the collection.
Our solution utilizes a dataset containing various patient features and outcomes. The dataset, sourced from Kaggle, comprises a total of 300 entries with 13 essential columns, as described below:

Age: Age of the patient.

Anaemia: Indicates whether the patient has anaemia (0: No, 1: Yes).

creatinine phosphokinase: Level of creatinine phosphokinase in the blood.

Diabetes: Indicates whether the patient has diabetes (0: No, 1: Yes).

Ejection fraction: Ejection fraction of the heart.

High blood pressure: Indicates whether the patient has high blood pressure (0: No, 1: Yes).

Platelets: Platelet count in the blood.

Serum creatinine: Level of serum creatinine in the blood.

Serum sodium: Level of serum sodium in the blood.

Sex: Gender of the patient (0: Female, 1: Male).

Smoking: Indicates whether the patient smokes (0: No, 1: Yes).

Time: Follow-up period in days.

DEATH_EVENT: Outcome variable, indicates whether the patient experienced a death event during the follow-up (0: No, 1: Yes).

We started by performing data preprocessing, which includes handling missing values (none observed), data type conversions (ensuring appropriate data types for each feature), and data scaling if necessary.

We created several visualisations to help you comprehend the dataset. To illustrate the distribution of specific health issues (such as anaemia, diabetes, high blood pressure, and

smoking) in relation to the occurrence of mortality events, this visualisation creates a 2x2 grid of countplots.

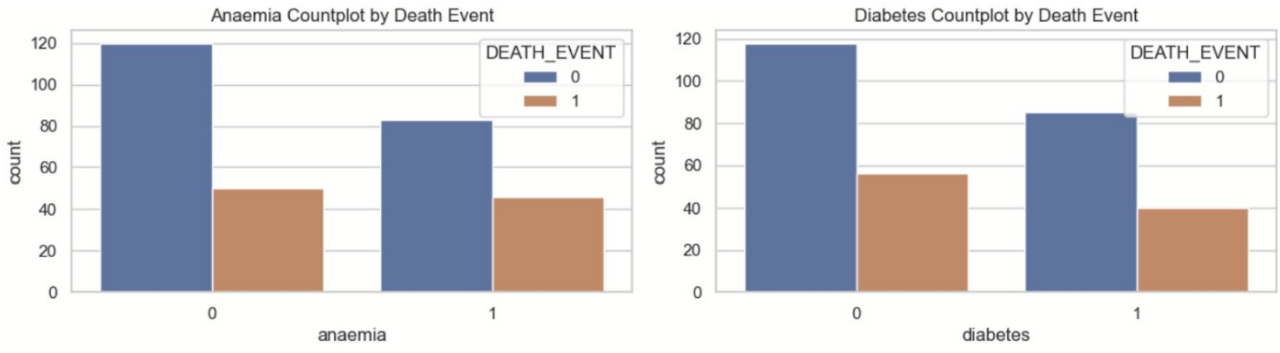


*Fig 1: Count Plot for Anaemia, Diabetes, High blood pressure & Smoking*

We deduced from the above diagram that deaths caused by diabetes and anaemia were nearly identical.

The following visualisation uses a histogram to display the dataset's age distribution. Using sns.histplot and a KDE (Kernel Density Estimate) overlay, the "age" column from the dataset is plotted to create a smooth curve. This makes the dataset's age distribution easier to interpret. According to the below graph, adults between the ages of 60 and 70 had the highest prevalence of heart disease. adults between 65 and 70 years old had the second-highest prevalence.

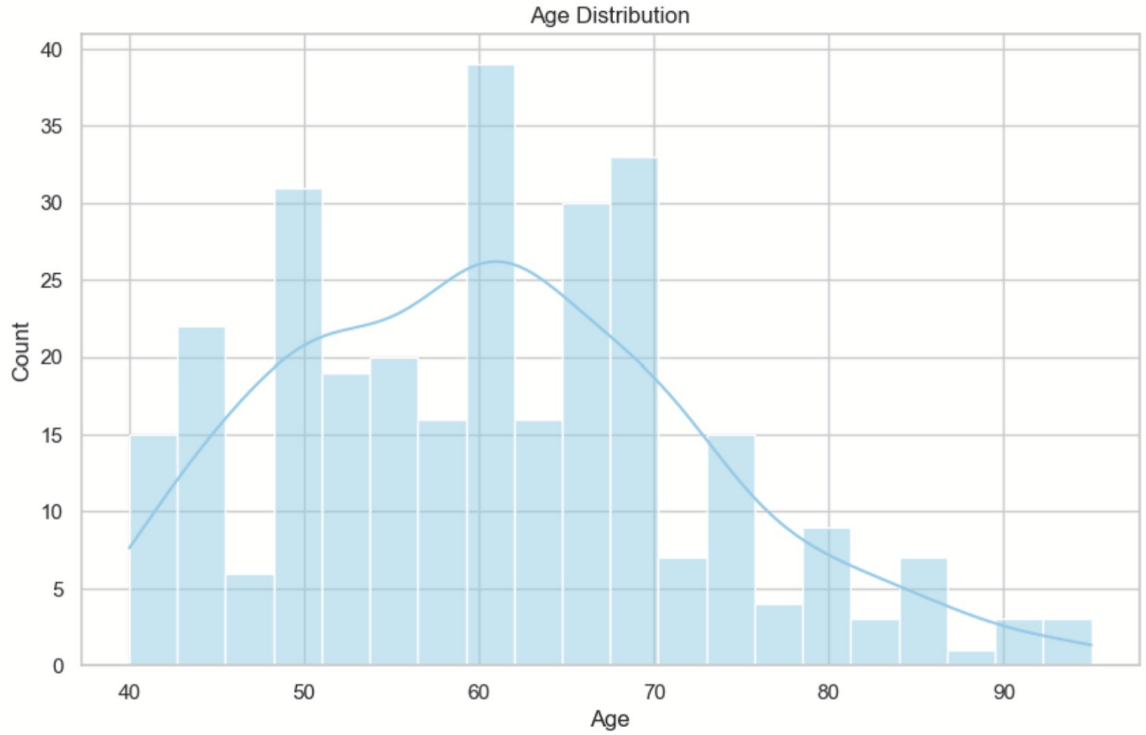


*Fig 2: Age Distribution*

The correlation between the dataset's numerical features is depicted in this heatmap. The DataFrame's.corr() method is used to calculate the correlation matrix. Then, a heatmap with annotations displaying the correlation coefficients is produced using the sns.heatmap function. Both positive and negative correlations are represented using the "coolwarm" colormap.

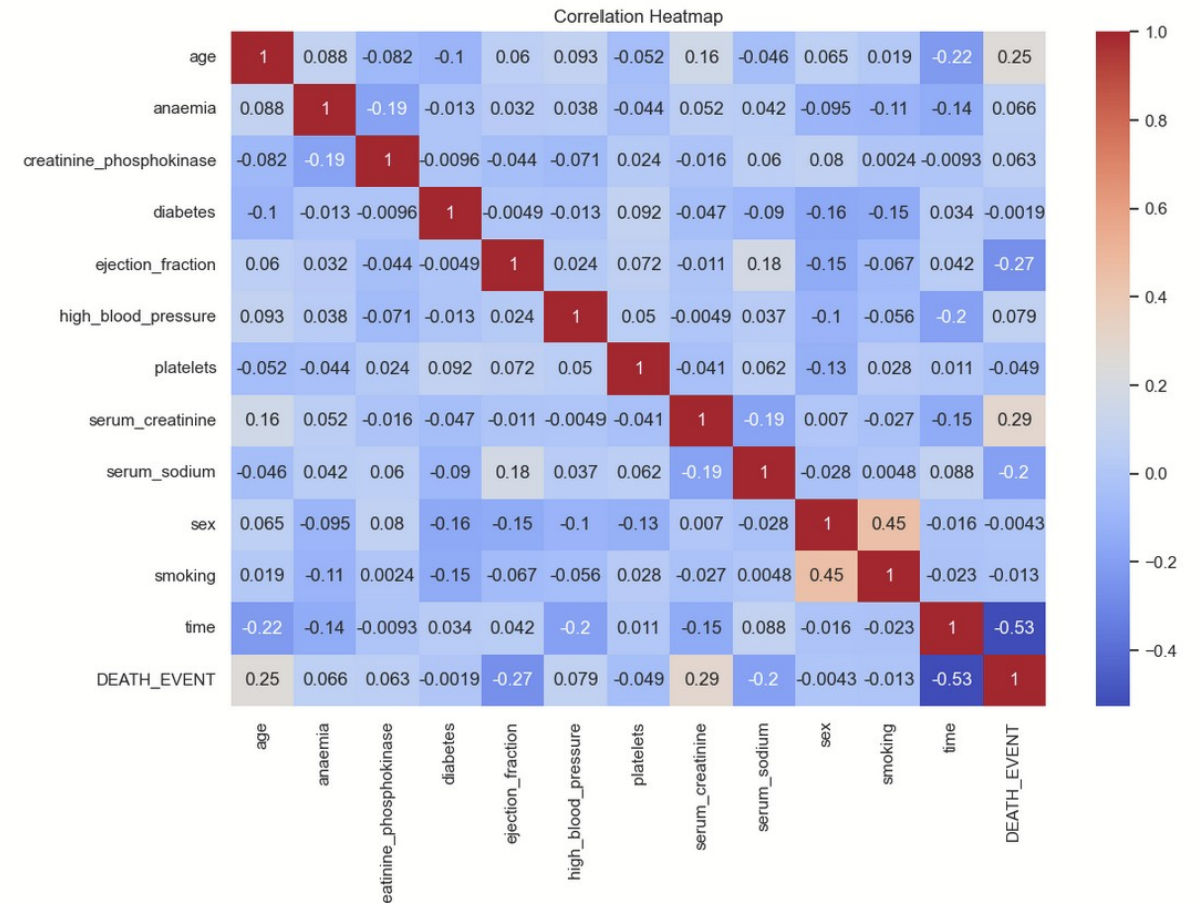


*Fig 3: Correlation Matrix*

The distribution of gender in relation to death events is seen in this countplot. It displays the number of instances for each gender where death events occurred or did not occur. To distinguish between the two outcomes, the colour parameter is set to "DEATH_EVENT" and the "sex" column is used as the x-axis.

We knew that men experienced death experiences at a rate of two times that of women. Even the death rate is twice as high for men as it is for women.

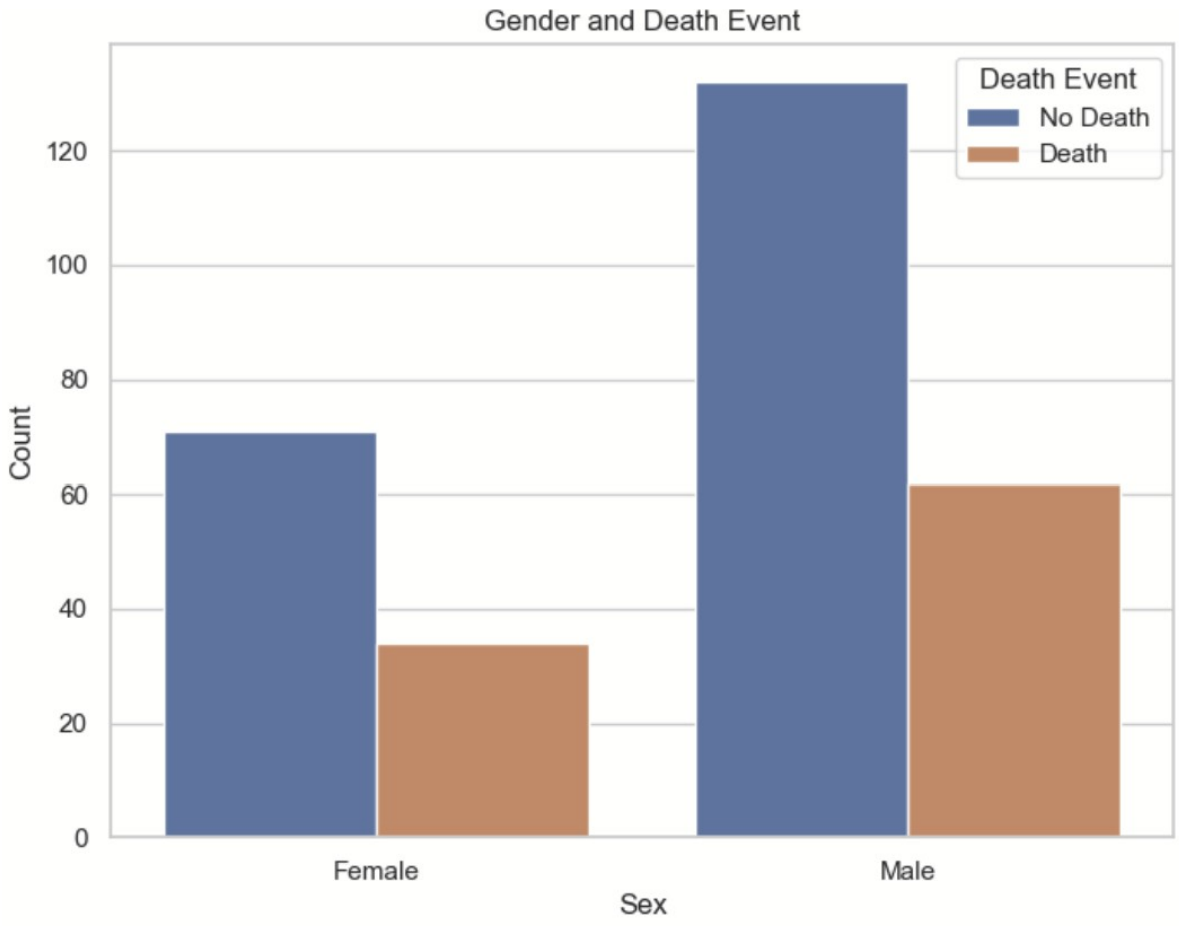


*Fig 4: Gender & Death Event*

The distribution of diabetes and high blood pressure in relation to death events is shown by this countplot. It displays counts of instances where diabetes or high blood pressure are present or missing, categorised by the occurrence of fatal events. To use a particular colour palette, the palette parameter is set to 'Set2'. We learned from this visualisation to our

surprise, that persons without high blood pressure had experienced more fatal incidents than those who have. But if we compare the actual deaths, they almost seem similar.

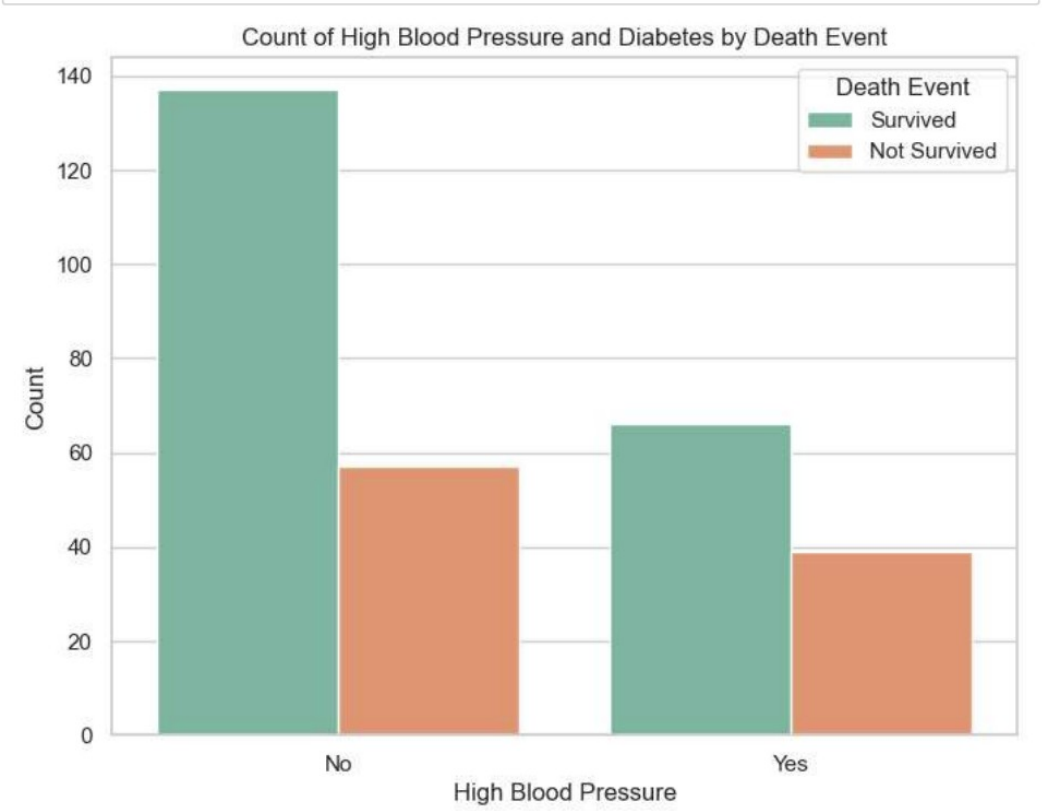


*Fig 5: Count Plot of High Blood Pressure & Diabetes by Death Events*

Following a thorough analysis of all the visualisation, we began using prediction models to expand upon our problem definition. Therefore, we preprocessed our data by removing the target variable (y) and the characteristics (X) from the dataset. The data is divided into training and validation sets using the train_test_split function from the sklearn library, with test_size set to 0.2 (20% of the data is reserved for validation) and random_state set to 42 for reproducibility. This division is necessary for developing and testing machine learning models.

**RANDOM FOREST**

With a predetermined random seed (random_state=1), a Random Forest classifier (rf) is first created. This guarantees the training procedure is repeatable. Then, the training features (train_X) and target labels (train_y) are used to train the classifier. The classifier is used to generate predictions on both the training data and the validation data after it has been trained. The variables train_pred (for training data) and valid_pred (for validation data) contain these predictions. Accuracy, precision, recall, and F1-score are among the calculated metrics for the validation data. These measures offer perceptions on how well the classifier works with unobserved data. Based on our random forest prediction model, the accuracy for training and validation was 1 and 0.7, respectively.

```
Random Forest Metrics:
Training Accuracy: 1.0
Training Precision: 1.0
Training Recall: 1.0
Training F1 Score: 1.0

Validation Accuracy: 0.7
Validation Precision: 0.7058823529411765
Validation Recall: 0.48
Validation F1 Score: 0.5714285714285713
```

*Fig 6: Evaluation Metrics of Random Forest*

**NEURAL NETWORK**

The training features (train_X) and target labels (train_y) for the initialised MLP classifier are used. For training, the.fit() function is employed. Following training, predictions on both the training data and the validation data are made using the trained MLP classifier. Both y_pred_train and y_pred_valid include the predictions for the training data and validation data, respectively. The derived classification metrics are then displayed in a comprehensible manner. The measurements are printed out separately for the training data and the validation data. This enables you to evaluate the MLP classifier's performance on the validation set, which it has never seen before, as well as the training set, which it has already seen during training. Based on our neural network prediction model, the accuracy for training and validation was 0.7 and 0.58, respectively.

```
Training Performance:
Accuracy: 0.702928870292887
Precision: 0.0
Recall: 0.0
F1-score: 0.0
Validation Performance:
Accuracy: 0.5833333333333334
Precision: 0.0
Recall: 0.0
F1-score: 0.0
```

*Fig 7: Evaluation Metrics of Neural Network*

**LOGISTICS REGRESSION**

The training data (train_X and train_y) are used to design a logistic regression model with a maximum number of iterations (max_iter) and then to train the model. Both the training data (train_X) and the validation data (valid_X) class labels are predicted using the trained Logistic Regression model. For the forecasts based on the practise data, evaluation metrics are calculated. Accuracy, precision, recall, and F1-score are among the measured metrics. The average parameter is set to 'binary' to compute binary-class metrics, and zero_division is set to 1 to prevent division by zero. Based on our regression model prediction model, the accuracy for training and validation was 0.84 and 0.80, respectively.

```
Training Metrics:
Training Accuracy: 0.84
Training Precision: 0.77
Training Recall: 0.66
Training F1 Score: 0.71

Validation Metrics:
Validation Accuracy: 0.80
Validation Precision: 0.88
Validation Recall: 0.60
Validation F1 Score: 0.71
```

*Fig 7: Evaluation Metrics of Logistic Regression*

To forecast the result for a fresh patient's data, we employed a trained Logistic Regression model. Using the Pandas library, a new DataFrame with the name new_patient_data is produced. The

columns have names that specifically correspond to the features that were utilised to train the logistic regression model. Within the data parameter, the information for the new patient is provided as a hierarchical list.The new_patient_data DataFrame contains the data for the new patient, and the trained Logistic Regression model (logreg) is used to forecast the outcome. The prediction variable contains the anticipated result.The anticipated value serves as the class label, with 0 often denoting a bad outcome (for instance, no death event) and 1 typically denoting a good outcome (for instance, a death event).For clinical judgement calls or additional research, this forecast may offer insightful information.

# BENEFITS

The following advantages will result from the predictive modelling we utilized to enhance heart failure patient diagnosis, care, and outcome prediction: -

**1. Early detection and intervention:** The predictive models created as part of this project can assist in identifying patients who are more likely to suffer negative outcomes, such as heart failure-related death. Early detection enables medical practitioners to act quickly, administer suitable care, and maybe avert harmful effects.

**2. Resource Optimization:** Healthcare institutions can more effectively deploy resources by correctly forecasting patient outcomes. This entails maximising staffing levels, the use of medical technology, and the accessibility of hospital beds to boost overall operational effectiveness.

**3. Making Informed Clinical judgements:** Physicians can use the information offered by the predictive models to make knowledgeable clinical judgements. This can make it easier to select the best interventions, drugs, and therapies for each patient.

**4. Reduced Mortality:** By identifying patients who are more likely to experience fatal events, the medical staff can concentrate on putting preventative measures into place to lower mortality rates. This may result in a considerable decline in the number of heart failure-related fatalities.

**5. Savings:** The research may result in cost savings for both patients and healthcare organizations by offering targeted interventions to high-risk individuals and avoiding wasteful treatments for low-risk patients.

**6. Better Patient Outcomes:** In the end, this project's greatest advantage is its ability to raise patient outcomes. Healthcare professionals can improve patient quality of life, save lives, and promote the general wellbeing of people with heart failure by utilizing data and analytics.

## CHALLENGES

Here are some challenges that could be encountered during the implementation of the proposed Business Intelligence (BI) solution for improving the diagnosis and treatment of heart failure:

- **Data Privacy and Security**: Healthcare data is sensitive and subject to strict privacy regulations. Proper data anonymization and encryption must be implemented to protect patient confidentiality.

- **Clinical Integration:** Implementing predictive models into the clinical workflow requires coordination between data scientists, healthcare professionals, and IT departments. Ensuring that the model's predictions align with clinical decision-making and that the model can be easily integrated into existing systems is essential.

- **Feature Selection and Relevance:** Selecting the right features (variables) for the predictive models is important. Some features might not have a significant impact on predicting outcomes, while others could be highly relevant. We had to correctly identify and include the most relevant features, which would help us reach a reasonable conclusion and provide valuable insights.

- **Imbalanced Data:** If the dataset contains an imbalanced distribution of outcomes (e.g., more non-death events than death events), the predictive model may be biased

towards the majority class. Techniques such as resampling, oversampling, or using different evaluation metrics are needed to address this issue.

Addressing these challenges requires collaboration between data scientists, healthcare professionals, legal experts, and IT specialists. Additionally, a well-defined plan for data collection, model development, validation, and integration is necessary to ensure the success of the BI solution in improving the diagnosis and treatment of heart failure.

# CONCLUSION

In view of the increased occurrence of heart failure patients, there is a compelling need for a proactive approach to forecasting heart failure survival. This proactive posture is crucial to limiting the rising risk of heart failure-related mortality. Addressing this crucial circumstance, we sought to create effective prediction models that may assist in anticipating heart failure outcomes. Our objective was to construct models that may help medical practitioners identify people at risk and design appropriate therapies to boost survival rates.

In our research, we evaluated a variety of three (3) unique machine learning algorithms. By systematically putting our models through the cross-validation method, we tried to completely test their prediction capabilities. This detailed review allowed us to conduct informed comparisons among the models, simplifying the selection of the most effective forecasting tool.

Upon rigorous investigation, one specific machine learning model came out as the most adept at addressing our purpose. When contrasted with the array of other machine learning methods, the Logistic Regression model demonstrated higher performance. It earned the maximum validation accuracy of 80%, paired with remarkable precision, recall, and F1 score metrics. The Logistic Regression model not only outperformed other algorithms but also displayed a considerable upgrade over earlier study outputs. This surge in accuracy emphasizes the promise of our method in expanding the field's knowledge and forecasting powers.

In summary, our suggested technique has potential for forecasting the risk of heart failure occurrences. By integrating machine learning, we hope to offer medical practitioners a vital tool for assessing patient risk and creating personalized treatment strategies. Although issues related to dataset size were solved to a certain extent, future research might benefit from bigger and more diversified datasets, perhaps leading to even more accurate prediction models.

This solution can be used across a variety of industries to forecast sales and predict the price of items, especially houses, and estimate their value based on historical data. By doing so, it helps solve a crucial finance problem that would potentially impact many people and would help them make better and informed decisions on health management and when suggesting therapies and diagnosis to the new patients.

# RECOMMENDATION

- **Feature Engineering:** Experiment with feature engineering strategies to boost the predictive potential of the models. Consider developing new features, scaling current ones, or changing specific features to capture more relevant information.

- **Model Tuning:** Fine-tune hyperparameters of the chosen Logistic Regression model to improve its performance. Adjust regularization strength, solver type, and other parameters to attain better results.

# APPENDIX

**TEAM MEMBERS:**

PROJECT MANAGER - **AMARACHI EZEJI**

DATA ANALYST - **AYESHA KARADIA**

DATA SCIENTIST - **APARAJITA ROY**

HEALTHCARE SPECIALIST - **IVIN ALEXANDER**

IT SPECIALIST - **GBENGA ADEBOWALE**

LEGAL AND COMPLIANCE EXPERT - **NAVDEEP CHAUHAN**

**Work Breakdown Structure (WBS):**

**Project Initiation**

Define project scope and objectives.

Identify project stakeholders and their roles.

Assign team members and responsibilities.

Create project schedule and timeline.

**Data Collection and Preprocessing**

Collect heart failure patient dataset from Kaggle.

Clean and preprocess the dataset (handling missing values, data type conversion)

Scale data if necessary

**Exploratory Data Analysis (EDA)**

Visualize distribution of patient attributes

Analyze correlation between attributes.

Identify patterns and trends in the data.

**Feature Engineering**

Select relevant features for prediction models.

Encode categorical variables (e.g., gender, smoking) using one-hot encoding.

Split data into training and validation sets

**Model Development**

Random Forest Model:

Train Random Forest classifier

Evaluate model accuracy, precision, recall, and F1-score.

Neural Network Model:

Design and train MLP classifier

Evaluate model performance metrics.

Logistic Regression Model:

Train Logistic Regression model

Evaluate model accuracy, precision, recall, and F1-score.

**Model Comparison and Selection**

Compare performance of different models

Select the most suitable model for heart failure prediction.

**Integration and Clinical Testing**

Integrate selected model into clinical workflow.

Test model predictions on real patient data

Collaborate with healthcare specialists for feedback and validation.

**Data Privacy and Ethics**

Ensure data privacy and security compliance.

Implement data anonymization and encryption techniques.

Obtain patient consent for using their data in predictive modeling.

**Change Management and Training**

Provide training to healthcare professionals on using predictive model.

Facilitate cultural shift towards incorporating model predictions in clinical decisions.

**Legal and Regulatory Compliance**

Ensure compliance with healthcare regulations (HIPAA, GDPR, etc.)

Address legal and ethical considerations related to model predictions.

**Model Validation and Monitoring**

Continuously validate and monitor model performance

Update the model as needed based on changing patient data.

**Communication and Reporting**

Prepare regular progress reports for stakeholders.

Communicate model insights and recommendations to medical team.

Provide patient communication about the benefits and limitations of predictive modeling.

**Project Conclusion and Recommendations**

Summarize project outcomes and achievements.

Provide recommendations for further improvements and future work.

# REFERENCES

*Heart failure prediction*. (2020, June 20). Retrieved from www.kaggle.com: https://www.kaggle.com/datasets/andrewmvd/heart-failure-clinical-data

World Health Organization: WHO. (2019). Cardiovascular diseases. www.who.int. https://www.who.int/health-topics/cardiovascular-diseases#tab=tab_1

M. Mamun, A. Farjana, M. A. Mamun, M. S. Ahammed and M. M. Rahman, "Heart failure survival prediction using machine learning algorithm: am I safe from heart failure?," 2022 IEEE World AI IoT Congress (AIIoT), Seattle, WA, USA, 2022, pp. 194-200, doi: 10.1109/AIIoT54504.2022.9817303.

Harshit Jindal et al 2021 IOP Conf. Ser.: Mater. Sci. Eng. 1022 012072