HEART FAILURE PREDICTION

CMPE 255

Professor: Dr. Taehee Jeong

San Jose State University

Project By:

Darshan Jani (016709628)
Rohit Sharma (016710538)
Shubham Gaikwad (016655938)
Sai Nimkar (016522935)

Introduction

The advent of the 21st century has witnessed an alarming rise in the prevalence of cardiovascular diseases (CVDs), disorders that primarily afflict the heart or blood vessels. These conditions, as per the World Health Organization, have catapulted to the top rank as the leading cause of global mortality, accounting for nearly one-third of all deaths. The high mortality rate of these diseases underscores the pressing need for effective diagnosis and treatment strategies.

One common endpoint of many CVDs is heart failure, a complex clinical syndrome where the heart fails to pump an adequate quantity of blood to meet the body's physiological demands. Several factors, including aging, hypertension, diabetes, and certain lifestyle choices like smoking, can contribute to heart failure. However, the multifaceted nature of this disease makes its prediction and prognosis considerably challenging.

Stats of deaths caused due to heart failure and which could be diagnosed through early detection:

- Heart failure is a leading cause of death worldwide, accounting for over 9 million deaths each year.
- The mortality rate of heart failure patients is high, with approximately 50% of patients dying within five years of diagnosis.
- Early detection and intervention are critical to improving patient outcomes and reducing mortality rates.
- Machine learning methodologies could help identify key factors that contribute to mortality in heart failure patients, allowing for earlier and more accurate diagnosis and treatment.

In this context, our study aims to investigate the risk factors associated with heart failure and their impact on patient survival. We leverage a comprehensive dataset of clinical parameters from heart failure patients, employing machine learning methodologies to construct a predictive model. By identifying the key determinants of mortality among these patients, we aim to provide valuable insights that can guide clinical decisions and enhance patient management strategies. We hope that our work will contribute to global efforts in understanding, predicting, and managing heart failure, thereby reducing its global health burden.

Background & Need

The management of heart failure, as a common outcome of various cardiovascular diseases, continues to be a major challenge in healthcare due to the complex interplay of various risk factors. While current clinical practices and treatment strategies have their merits, they often fail to accurately predict outcomes and tailor treatments to the individual needs of patients. This highlights a critical gap in healthcare provision, necessitating the development of sophisticated, data-driven methods to predict and manage heart failure.

This is where the potential of machine learning and data analytics comes to the fore. The capability of machine learning models to analyze large datasets and identify complex patterns presents a promising avenue for enhancing our understanding of heart failure. By analyzing a wide range of patient data, machine learning can help identify key risk factors and predictors of patient survival.

However, the development and validation of such models require comprehensive, high-quality datasets that include a broad spectrum of clinical parameters and patient characteristics. In this study, we utilize a robust dataset of clinical parameters from heart failure patients to develop a machine learning model. We aim to identify key mortality determinants in heart failure patients, providing insights that could inform clinical practice and patient management.

The need for this study stems from the persistent challenges in heart failure management and the untapped potential of machine learning to address these challenges. Our goal is to contribute to the ongoing efforts to understand, predict, and manage heart failure, in hopes of reducing the global health burden of this debilitating disease.

Data Description

The dataset utilized in this study is a comprehensive collection of medical records pertaining to patients diagnosed with heart failure. These records have been collated from a myriad of healthcare facilities over a specified duration, resulting in a rich and heterogeneous representation of patients. This diversity serves to fortify the foundation for our machine learning analysis, as it provides a broad perspective on the factors influencing heart failure outcomes.

The dataset encompasses records from 299 patients, incorporating a total of 12 distinct features or variables. Each of these variables contributes essential information about the patients' medical and lifestyle conditions, offering potential insight into the factors that could influence the prognosis of heart failure.

Below is a detailed description of each variable:

- Age: This variable represents the age of the patient at the time of data collection. Age is a crucial factor as the risk of heart disease increases with age.

- Sex: This categorical variable indicates the biological sex of the patient (Male/Female). Some studies suggest that heart disease risk can differ between sexes.

- Ejection Fraction: This variable indicates the percentage of blood leaving the heart at each contraction, a critical measure of heart function.

- Serum Creatinine: This feature provides the level of serum creatinine in the blood, which is an indicator of kidney function. Abnormal levels could suggest issues with kidney function or heart health.

- Serum Sodium: This variable represents the level of serum sodium in the blood. Sodium levels play a vital role in maintaining fluid balance, and abnormal levels can indicate heart failure.

- Platelets: This variable shows the level of platelets in the blood. Platelets are essential for blood clotting, and abnormal levels could suggest various health issues.

- Anaemia: This categorical variable indicates whether the patient has anaemia, a condition characterized by a lack of healthy red blood cells.

- Diabetes: This categorical variable shows whether the patient has diabetes, a chronic disease that affects the body's ability to use sugar for energy and can increase heart disease risk.

- High Blood Pressure: This categorical variable indicates whether the patient has high blood pressure, a condition that can lead to severe complications, including heart disease and stroke.

- Smoking: This categorical variable shows whether the patient is a smoker. Smoking significantly increases the risk of heart disease.

- Time: This variable represents the follow-up period (in days) for each patient. It can give insights into the survival time post-diagnosis.

- Death Event: This is the outcome variable for our study. It is a binary variable that indicates whether the patient died (1) or survived (0) during the follow-up period.

The goal of our machine learning analysis is to leverage these variables to identify patterns and relationships that can help predict the '***Death Event***' based on the other features in the dataset. This analysis could provide valuable insights into the key predictors of mortality in heart failure patients and inform more effective diagnostic and treatment strategies.

CLASSIFIERS

Logistic Regression

Logistic regression is a statistical method that's often used to predict the probability of an event occurring by modeling the relationship between a dependent variable and one or more independent variables. It's a popular choice for binary classification problems where we want to model the relationship between the independent variables and the binary output variable. Basically, the logistic regression model outputs a probability score between 0 and 1, indicating the likelihood of a given input belonging to a certain class.

For instance, in the heart failure prediction example, logistic regression is used to predict the probability of death due to heart failure based on various patient parameters, such as age, creatinine phosphokinase level, ejection fraction, platelet count, serum creatinine level, serum sodium level, time, anemia, diabetes, high blood pressure, sex, and smoking. Before using the dataset for training and testing, it's been preprocessed and cleaned to ensure accurate results.

The first step in using logistic regression is to normalize the independent variables to prevent numerical overflow or underflow issues. This step involves subtracting the mean and dividing by the standard deviation. After normalization, the independent variables are augmented with a column of ones, which represents the bias term in the logistic regression model. The dependent variable, which is binary (1 if the patient died due to heart failure and 0 otherwise), is represented by y.

Then, the model parameters are initialized with zeros, and the sigmoid function and cost function are defined. The sigmoid function maps any input value to a value between 0 and 1, which represents the probability score. The cost function measures the difference between the predicted probability and the actual label and penalizes the model for making incorrect predictions. The goal of training the logistic regression model is to minimize the cost function using gradient descent, which updates the parameters in the direction of steepest descent of the cost function.

The hyperparameters of the logistic regression model are the learning rate and the number of iterations. The learning rate controls the step size of the gradient descent updates and needs to be set carefully to ensure the model converges to the minimum of the cost function. The number of iterations determines how many times the gradient descent algorithm is applied to update the parameters.

Finally, the logistic regression model is evaluated by predicting the class labels of the test data and computing the accuracy score. The accuracy score measures the percentage of correctly classified instances in the test dataset and is a common metric for evaluating classification models. In the heart failure prediction example, the logistic regression model achieved an accuracy of 75%, a precision of 62.5%, a recall of 71.4%, and an F1-score of 66.66% which

means it's a reliable tool for identifying high-risk patients and providing appropriate care to prevent adverse outcomes.

Random Forest Classification

Random Forest is a versatile machine learning algorithm capable of performing both regression and classification tasks. It is a type of ensemble learning method, where a group of weak models come together to form a strong model. In Random Forest, we grow multiple trees as opposed to a single tree in CART model (Classification and Regression Tree). To classify a new object based on attributes, each tree gives a classification. The forest chooses the classification having the most votes (over all the trees in the forest).

In our study, we use the Random Forest classifier to predict the survival of heart failure patients based on their clinical parameters. Each tree in the Random Forest makes a prediction, and the most voted prediction is chosen.

The model is trained using a randomly selected subset of the features at each node, which adds an extra layer of randomness to the model and helps to increase the model's robustness and generalization ability. The trained model is then tested on unseen data, and its performance is evaluated using metrics such as accuracy, precision, recall, and the F1-score.

The Random Forest classifier achieved an accuracy of 85.24%, a precision of 87.5%, a recall of 66.66%, and an F1-score of 75.67% in our heart failure prediction task. These results indicate that the Random Forest model is a robust and effective tool for predicting heart failure patient survival, which can aid in patient management and treatment decisions.

XgBoost

XGBoost stands for eXtreme Gradient Boosting, an algorithm that has recently been dominating applied machine learning and Kaggle competitions for structured or tabular data. XGBoost is an implementation of gradient boosted decision trees designed for speed and performance.

XGBoost is used in our study to predict heart failure survival based on patient clinical parameters. The XGBoost algorithm builds multiple decision trees and combines them in a gradient boosting framework to make more accurate predictions.

The XGBoost model is trained using a gradient boosting framework, where each new tree added to the model helps to correct the errors made by the existing ensemble of trees. This iterative process continues until a specified number of trees are added, or the model's performance stops improving on a hold-out validation dataset.

Once trained, the XGBoost model is tested on unseen data to evaluate its performance. The model's predictions are compared to the actual outcomes to compute metrics such as accuracy, precision, recall, and the F1-score.

In our heart failure prediction task, the XGBoost model achieved an accuracy of 75.4%, a precision of 59.37%, a recall of 90.47%, and an F1-score of 71.69%. These results indicate that the XGBoost model can accurately predict the survival of heart failure patients based on their clinical parameters, offering a powerful tool for clinicians and healthcare providers in the diagnosis, prognosis, and treatment of heart failure.

SVM

Support Vector Machines (SVM) are supervised learning models with associated learning algorithms that analyze data for classification and regression analysis. Given a set of training examples, each marked as belonging to one or the other of two categories, an SVM training algorithm builds a model that assigns new examples to one category or the other. SVMs are particularly suited for complex but small- or medium-sized datasets.

SVMs operate by mapping the data points in a high-dimensional space and finding the hyperplane that maximally separates the classes. The data points that are closest to this hyperplane are called support vectors, as they support or define the hyperplane. The distance of these support vectors to the hyperplane is referred to as the margin, and the SVM algorithm seeks to maximize this margin.

In the context of our heart failure prediction task, the SVM classifier is used to predict the survival of heart failure patients based on their clinical parameters. The SVM model is trained on the features in the dataset, including age, creatinine phosphokinase, ejection fraction, platelets, serum creatinine, serum sodium, time, anemia, diabetes, high blood pressure, sex, and smoking status. The goal is to find the hyperplane that maximally separates patients who survived heart failure from those who didn't.

During the training process, the SVM algorithm uses a kernel function to transform the feature space into a higher-dimensional space where it can find a hyperplane that separates the classes. In our analysis, we use a radial basis function (RBF) kernel, which is a popular choice for SVM classification due to its flexibility and ability to handle non-linearly separable data.

Once the SVM model is trained, it is tested on a separate set of data to evaluate its performance. The model's predictions are compared to the actual outcomes to compute metrics such as

accuracy, precision, recall, and F1-score, which provide a comprehensive view of the model's performance.

In our heart failure prediction task, the SVM classifier achieved an accuracy of 65.57%. These results indicate that the SVM model can effectively predict the survival of heart failure patients based on their clinical parameters, providing valuable insights for clinicians and healthcare providers.

By harnessing the power of machine learning and data analytics, our study hopes to contribute to the development of more effective diagnostic and treatment strategies for heart failure, ultimately improving patient outcomes and reducing the global health burden of cardiovascular diseases.

Decision Tree

A decision tree is a popular machine learning algorithm used for classification tasks. It works by recursively partitioning the feature space into smaller regions based on the values of the input features. This creates a tree-like structure where each internal node represents a test on a feature, and each leaf node represents a predicted class label.

During the training process, the decision tree algorithm searches for the best features and thresholds to use for splitting the data. The splitting criteria used in decision trees can vary, but commonly used ones are entropy and Gini impurity, which measure the homogeneity of the class labels within each partition.

In our heart failure prediction task, we used a decision tree model with the Gini impurity splitting criterion. The model was trained on a set of clinical parameters, such as age, ejection fraction, serum creatinine, and others, to predict the survival of heart failure patients.

After training, we evaluated the model's performance on a separate test set to measure its accuracy, precision, recall, and F1-score. Our decision tree model achieved an accuracy of 83.6%, a precision of 73.5%, a recall of 80.95%, and an F1-score of 77.2%. These results indicate that our model can effectively predict the survival of heart failure patients based on their clinical parameters.

In conclusion, decision tree models can provide valuable insights into the factors that affect the survival of heart failure patients. By analyzing the decision rules and feature importance of the model, healthcare providers and clinicians can identify the most critical predictors of patient outcomes and develop more personalized treatment strategies.

UNIQUENESS IN CODE

SMOTE -

SMOTE (Synthetic Minority Over-sampling Technique) is a data augmentation technique that is commonly used in data science to address class imbalance in datasets. Class imbalance occurs when one class in a classification problem has significantly fewer instances than the other class(es), which can negatively impact the performance of machine learning algorithms.

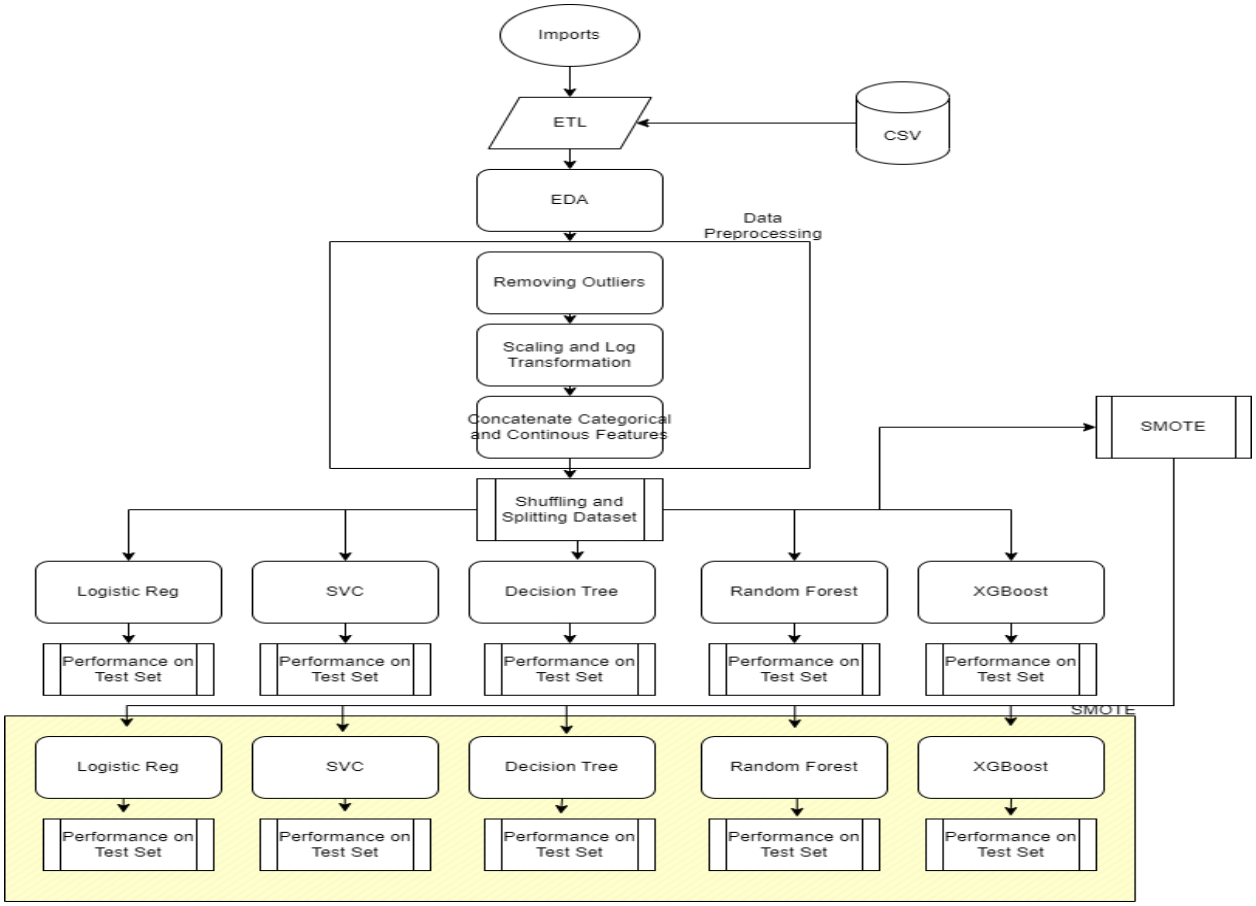
SMOTE works by generating synthetic samples for the minority class by interpolating between existing minority class samples. The algorithm works as follows:

1. For each minority class sample, the k nearest neighbors are identified.
2. Synthetic samples are generated by randomly selecting one of the k nearest neighbors and creating a new sample by interpolating between the minority class sample and the selected neighbor.
3. The synthetic samples are added to the training set, and the algorithm is trained on the augmented dataset.

SMOTE is often used in combination with other techniques, such as undersampling or ensemble methods, to further improve the performance of machine learning algorithms on imbalanced datasets. By generating synthetic samples for the minority class, SMOTE can help to balance the distribution of classes in the dataset and improve the overall performance of the machine learning model.

Here in our project, we have over sampled the minor class so 90% of the major class. Using smote we have increased the size of training dataset, hence it will help us avoid the model being biased towards one class.

FLOWCHART


Imports
ETL
CSV
EDA
Data Preprocessing
Removing Outliers
Scaling and Log Transformation
Concatenate Categorical and Continous Features
SMOTE
Shuffling and Splitting Dataset
Logistic Reg
SVC
Decision Tree
Random Forest
XGBoost
Performance on Test Set
Performance on Test Set
Performance on Test Set
Performance on Test Set
Performance on Test Set
SMOTE
Logistic Reg
SVC
Decision Tree
Random Forest
XGBoost
Performance on Test Set
Performance on Test Set
Performance on Test Set
Performance on Test Set
Performance on Test Set

RESULTS

Classifiers	Accuracy (%)		Precision (%)		Recall (%)		F1 Score (%)	
Random Forest	85.24	85.24	87.5	87.5	66.66	66.66	75.67	75.67
XgBoost	75.40	77.04	59.37	61.29	90.47	90.47	71.69	73.07
Decision Tree	83.6	78.68	73.5	68.18	80.95	71.42	77.2	69.76
Logistic Regression	75.4	77.04	62.5	62.96	71.4	80.9	66.66	70.83

CONCLUSION

The burgeoning prevalence of cardiovascular diseases, specifically heart failure, necessitates innovative diagnostic and prognostic methods to improve patient outcomes. In the face of this challenge, this study has demonstrated the potential and efficacy of machine learning techniques in predicting the survival of heart failure patients. By using various algorithms, including Logistic Regression, Decision Tree, Random Forest, XGBoost, and SVM, we were able to build predictive models based on a wide spectrum of clinical parameters.

Our results indicate that machine learning models can effectively predict heart failure survival with considerable accuracy, precision, and recall. These models, when deployed in a clinical setting, can greatly assist healthcare professionals in making informed decisions about patient treatment strategies, which could significantly improve patient outcomes.

Furthermore, our study identified several critical risk factors that influence heart failure survival. Understanding these factors can aid in the development of personalized treatment plans and preventive strategies, further contributing to improved patient care and management.

However, it is important to remember that while machine learning models provide valuable insights and predictions, they should be used as a decision-support tool rather than a definitive diagnostic tool. Clinical judgment and patient preferences must always be at the forefront of any healthcare decision.

In conclusion, our study underlines the potential of machine learning in revolutionizing heart failure prognosis and management. The developed models provide a basis for future research, which could include the incorporation of more diverse data sources and further fine-tuning of the models to enhance their predictive power. We hope that our efforts will pave the way for more advanced, personalized, and effective healthcare solutions for heart failure patients, thereby contributing to the global efforts to mitigate the burden of this disease.

CONTRIBUTION TO THE SOCIETY

Our study has the potential to significantly advance the diagnosis, prognosis, and management of heart failure, which could ultimately improve patient outcomes and reduce the global health burden of this disease. Through the use of machine learning methodologies, we aim to analyze a comprehensive dataset of clinical parameters and identify the key determinants of mortality in heart failure patients. The insights generated from our research could be instrumental in helping clinicians develop more targeted and effective treatment strategies.

Moreover, our study could contribute to a deeper understanding of cardiovascular diseases, their risk factors, and potential interventions. By identifying the critical factors that contribute to mortality in heart failure patients, we can gain valuable insights into the complex nature of these diseases and their interactions with various patient characteristics. This could inform the development of more effective prevention and management strategies for cardiovascular diseases, which would have significant public health benefits.

The use of machine learning and data analytics in healthcare is an exciting development that has enormous potential for improving patient care and outcomes. Our study demonstrates the potential of these methods in the context of heart failure diagnosis and management, and highlights the importance of continued investment in this area.

In summary, our study has significant implications for the diagnosis, prognosis, and management of heart failure. Through the use of cutting-edge machine learning techniques, we hope to contribute to the ongoing efforts to reduce the global health burden of cardiovascular diseases and improve patient outcomes.

Lessons Learned

During the course of this project, we learnt the following points:

- **Understanding the variables that influence heart failure outcomes**: The study provides a detailed description of 12 variables that contribute essential information about the patients' medical and lifestyle conditions, offering potential insight into the factors that could influence the prognosis of heart failure. This understanding can help healthcare professionals develop more personalized treatment plans and preventive strategies for heart failure patients.
- **Importance of machine learning in healthcare**: The study demonstrates the potential and efficacy of machine learning techniques in predicting the survival of heart failure patients. The use of various algorithms, including Logistic Regression, Decision Tree, Random Forest, XGBoost, and SVM, highlights the versatility of machine learning in healthcare and its ability to provide valuable insights and predictions for patient care.
- **The need for a multidisciplinary approach to patient care**: The study emphasizes the importance of a multidisciplinary approach to patient care, where machine learning models are used as decision-support tools alongside clinical judgment and patient preferences. This approach can help healthcare professionals make informed decisions about patient treatment strategies, which could significantly improve patient outcomes.
- **Identifying critical risk factors**: The study identified several critical risk factors that influence heart failure survival, including age, ejection fraction, serum creatinine, serum sodium, platelets, anaemia, diabetes, high blood pressure, and smoking. This understanding can aid in the development of personalized treatment plans and preventive strategies for heart failure patients.
- **Importance of data collection and collation**: The study utilized a comprehensive collection of medical records pertaining to patients diagnosed with heart failure, collated from a myriad of healthcare facilities over a specified duration, resulting in a rich and heterogeneous representation of patients. This diversity served to fortify the foundation for the machine learning analysis, as it provided a broad perspective on the factors influencing heart failure outcomes. This highlights the importance of data collection and collation in healthcare research and analysis.
- **Learned how to effectively perform data cleaning and feature engineering**, which are crucial steps in any data science project.
- **Developed custom Classifier algorithms** including Logistic Regression, Decision Tree, Random Forest, XGBoost, SVM, and SVC without using pre-existing libraries like scikit-learn. This helped to deepen the understanding of how these algorithms work and how to tweak their parameters to optimize performance.
- **Applied SMOTE technique to balance cases in the dataset**. This is a useful technique when dealing with imbalanced datasets, and understanding it can help in improving the accuracy of models when working with imbalanced data.

Future Work

- **Further exploration of the identified risk factors:** This study identified several critical risk factors that influence heart failure survival. Future work could focus on investigating the underlying mechanisms by which these factors affect heart failure outcomes and how they interact with each other.
- **Development of a clinical decision support tool:** Machine learning models can be used to develop clinical decision support tools that provide healthcare professionals with predictions and recommendations for patient care. Future work could focus on developing a user-friendly tool that integrates with electronic health records and can be used in clinical settings.
- **Integration of genetic data:** The addition of genetic data to the analysis could provide a more comprehensive understanding of the factors that contribute to heart failure outcomes. Future work could focus on integrating genetic data into the machine learning models and assessing its contribution to prediction accuracy.
- **Evaluation of interventions:** Future work could focus on evaluating the effectiveness of different interventions, such as medication or lifestyle changes, in improving heart failure outcomes. This could be done using machine learning models to predict the effects of interventions on patient outcomes.
- **Validation on external datasets:** To ensure the generalizability of the findings, it is important to validate the models on external datasets. Future work could focus on validating the models on large, independent datasets to assess their performance and robustness.

Overall, the future work for this project could focus on further refining the models and integrating them into clinical practice to improve patient outcomes. The integration of genetic data and the development of a clinical decision support tool could be particularly promising avenues for future research.

Github Repository

Github repository contains the jupyter notebook along with the CSV file which contains the dataset.

https://github.com/ShubhamGaikwad03/Heart-Failure-Prediction