HEART FAILURE PREDICTION

Soham Govardhan Navale
PRN- 23070149021

INTRODUCTION

With an estimated 17.9 million fatalities annually, or 31% of all deaths worldwide, cardiovascular diseases (CVDs) are the leading cause of death worldwide. Heart failure is a prevalent CVD-related occurrence, and this dataset includes 12 variables that are useful in predicting heart failure-related mortality. By utilizing population-wide methods to address behavioral risk factors like tobacco use, unhealthy food and obesity, physical inactivity, and harmful alcohol consumption, most cardiovascular diseases can be averted. Early detection and treatment are essential for those who have cardiovascular disease or are at high risk for developing it because they have one or more risk factors, such as diabetes, hypertension, hyperlipidemia, or pre-existing conditions. With assistance of a machine learning model which can accurately predict heart failure using the significant parameters would help the patients effectively.

PROBLEM STATEMENT

Heart failure is a dangerous cardiovascular disease that affects people all over the world and is a major public health concern. It is essential to promptly identify those who are at risk of heart failure in order to put preventive measures and therapies into place. Machine learning has demonstrated potential in forecasting an individual's risk of heart failure due to its capacity to evaluate intricate data patterns. In order to create an accurate predictive model for early heart failure risk identification, this study will make use of a variety of machine learning methods.

OBJECTIVES

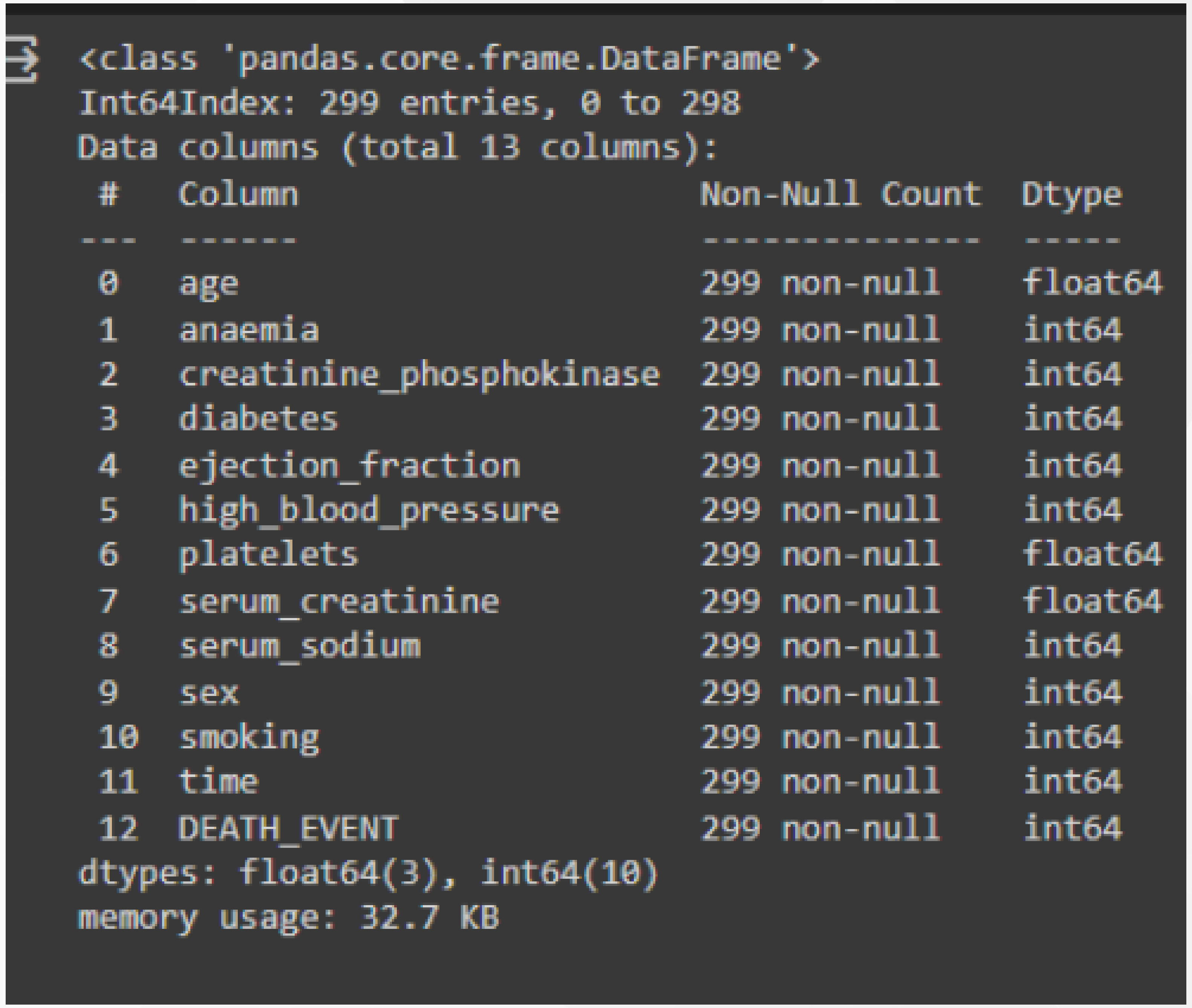
- To analyse various important parameters that lead to heart failure.
- To train various machine learning models/algorithms to predict heart failure.
- To present a best training model for predicting heart failure.

LITERATURE REVIEW

Sr.No	Title	Publication	Year	Accuracy	Models	Dataset
1	Heart Disease Prediction using Machine Learning	Journal of Engineering Sciences Vol 14 Issue 04,2023	2023	81%	extreme gradient boosting classifier	The information gathered for this project is UCI Heart Disease. The dataset has 76 properties, of which the system uses 14 for its operation.
2	Machine Learning-Based Automated Diagnostic Systems Developed for Heart Failure Prediction Using Different Types of Data Modalities: A Systematic Review and Future Directions. Computational and Mathematical Methods in Medicine	Alquran, M., Ghafoor, K. Z., & Jabbar, W. A. W. A. (2022). 2022.	2022	75.1-90%	Various machine learning algorithms	Various datasets (a few hundred to tens of thousands of patients)
3	Heart disease prediction using machine learning algorithms	IOP Conference Series: Materials Science and Engineering, Volume 1022	2021	88.52%	K nearest neighbors (KNN)	UCI repository
4	Machine learning techniques for early heart failure prediction	Malaysian Journal of Computing (MJoC)	2021	Average Performance Score under Random forest = 0.88	Random Forest, Support Vector Machine, Naive Bayes, Logistic Regression	The dataset is obtained from the Kaggle heart disease dataset consisting of 303 patients.
5	Improving risk prediction in heart failure using machine learning	Adler, A. J., & Fonarow, G. C. (2020). European Journal of Heart Failure, 22(11), 1782-1787.	2020	The machine learning model achieved an area under the curve (AUC) of 0.88	Support vector machine	ADHERE registry, which is a large-scale observational study of patients with heart failure.

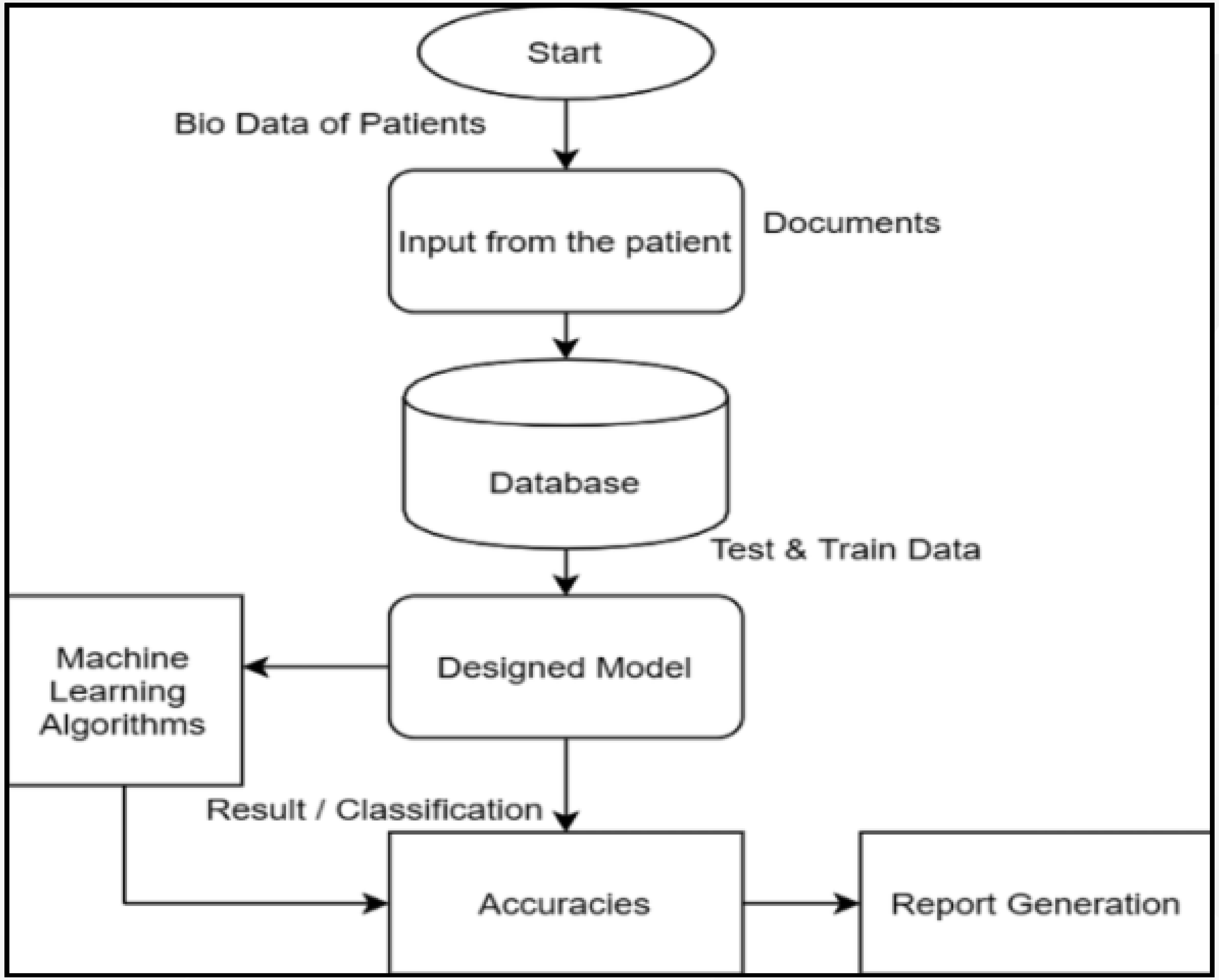
DATASET

This dataset provides a person's medical information. There are 13 columns and 300 rows (patients/persons). There are some factors that affect the death event of the person. This dataset contains parameters like age, anaemia ,sex , blood pressure, smoke, diabetes, ejection fraction, creatinine phosphokinase, high blood pressure, serum creatinine, serum sodium, time and their death event. Based on the above parameters the predictions are made in case of heart failure.



```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 299 entries, 0 to 298
Data columns (total 13 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   age                       299 non-null    float64
 1   anaemia                   299 non-null    int64
 2   creatinine_phosphokinase  299 non-null    int64
 3   diabetes                  299 non-null    int64
 4   ejection_fraction         299 non-null    int64
 5   high_blood_pressure       299 non-null    int64
 6   platelets                 299 non-null    float64
 7   serum_creatinine          299 non-null    float64
 8   serum_sodium              299 non-null    int64
 9   sex                       299 non-null    int64
 10  smoking                   299 non-null    int64
 11  time                      299 non-null    int64
 12  DEATH_EVENT               299 non-null    int64
dtypes: float64(3), int64(10)
memory usage: 32.7 KB
```

PROPOSED ARCHITECTURE WITH METHODOLOGY OF IMPLEMENTATION

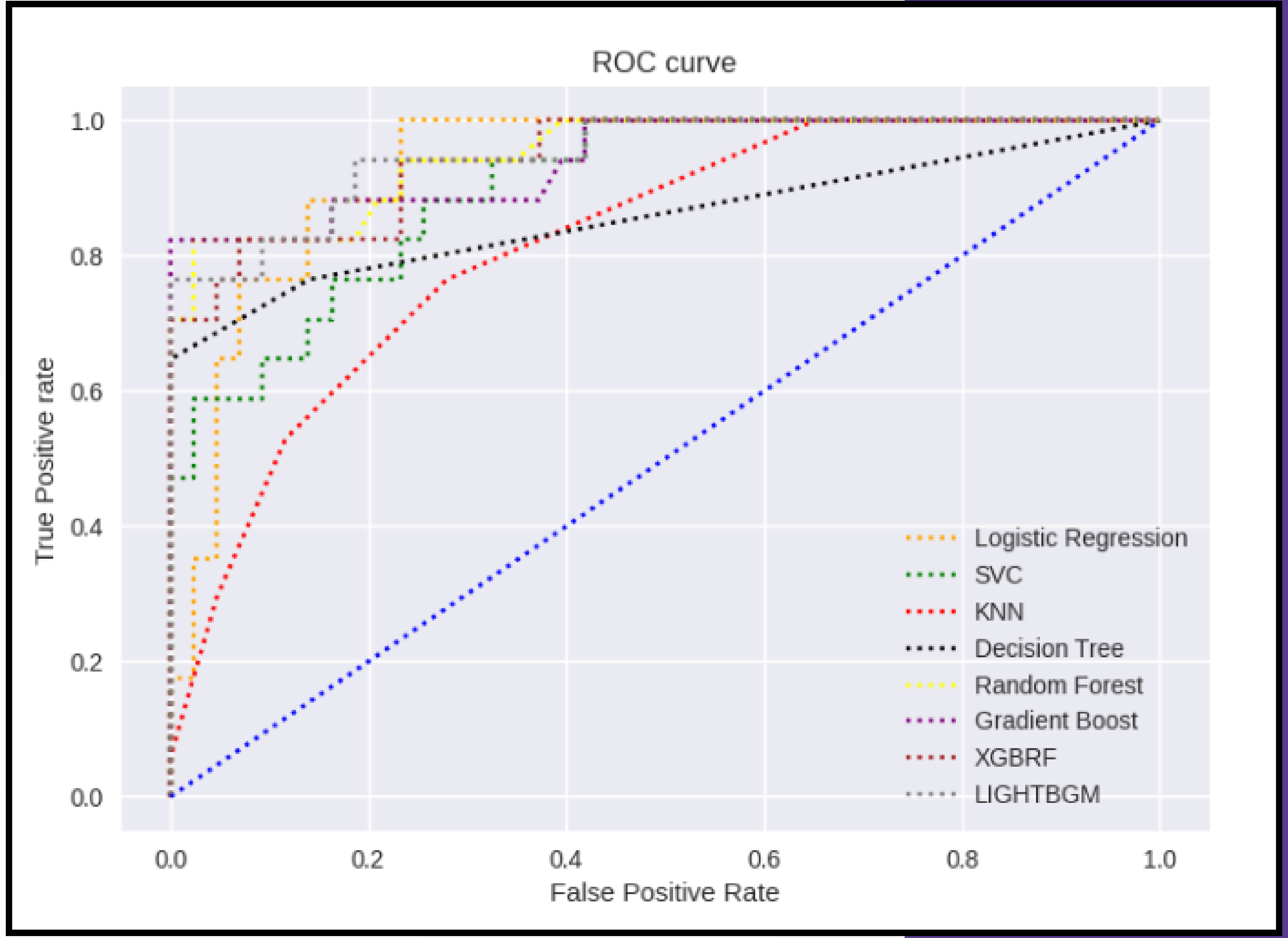


EXPLANATION OF FLOWCHART

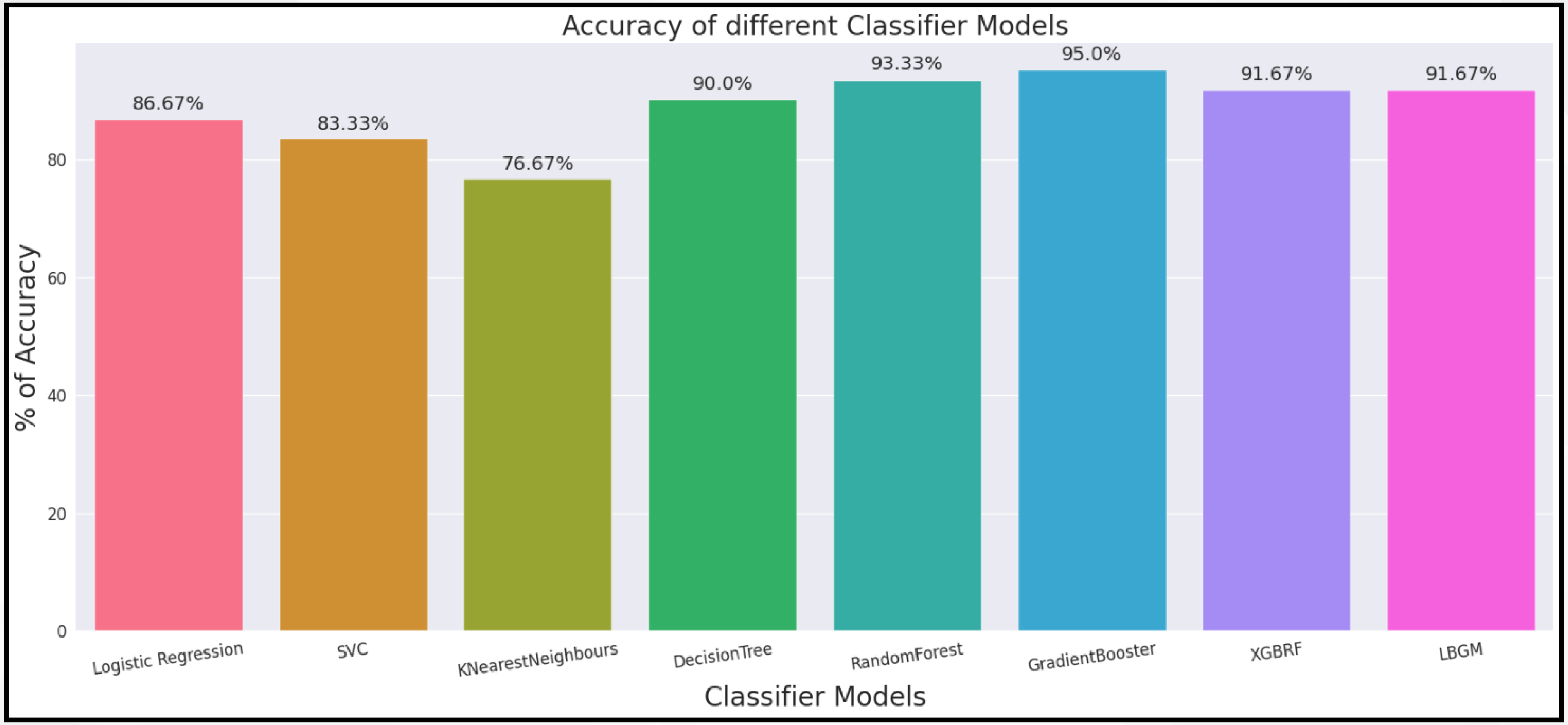
- First the data is collected, then it is pre-processed handling null values, imbalancement of dataset etc.
- Then, the process data is used for Exploratory Data Analysis where significant features are to be extracted using certain visualizations.
- Based on the features and requirements appropriate models are to be selected and trained.
- Based on the performance of the model fine tuning is to be done so that it can enhance the performance resulting in good/better performance.
- Based on the results, the best model is to be selected and conclusions are to be made.

RESULT ANALYSIS

The ROC curve is produced by calculating and plotting the true positive rate against the false positive rate for a single classifier at a variety of thresholds. Here various classifiers are used as shown below which represents their respective performances. The models used here are Logistic Regression, SVC, KNN, Decision Tree, Random Forest, Gradient Boost, XGBRF, LIGHTGBM.

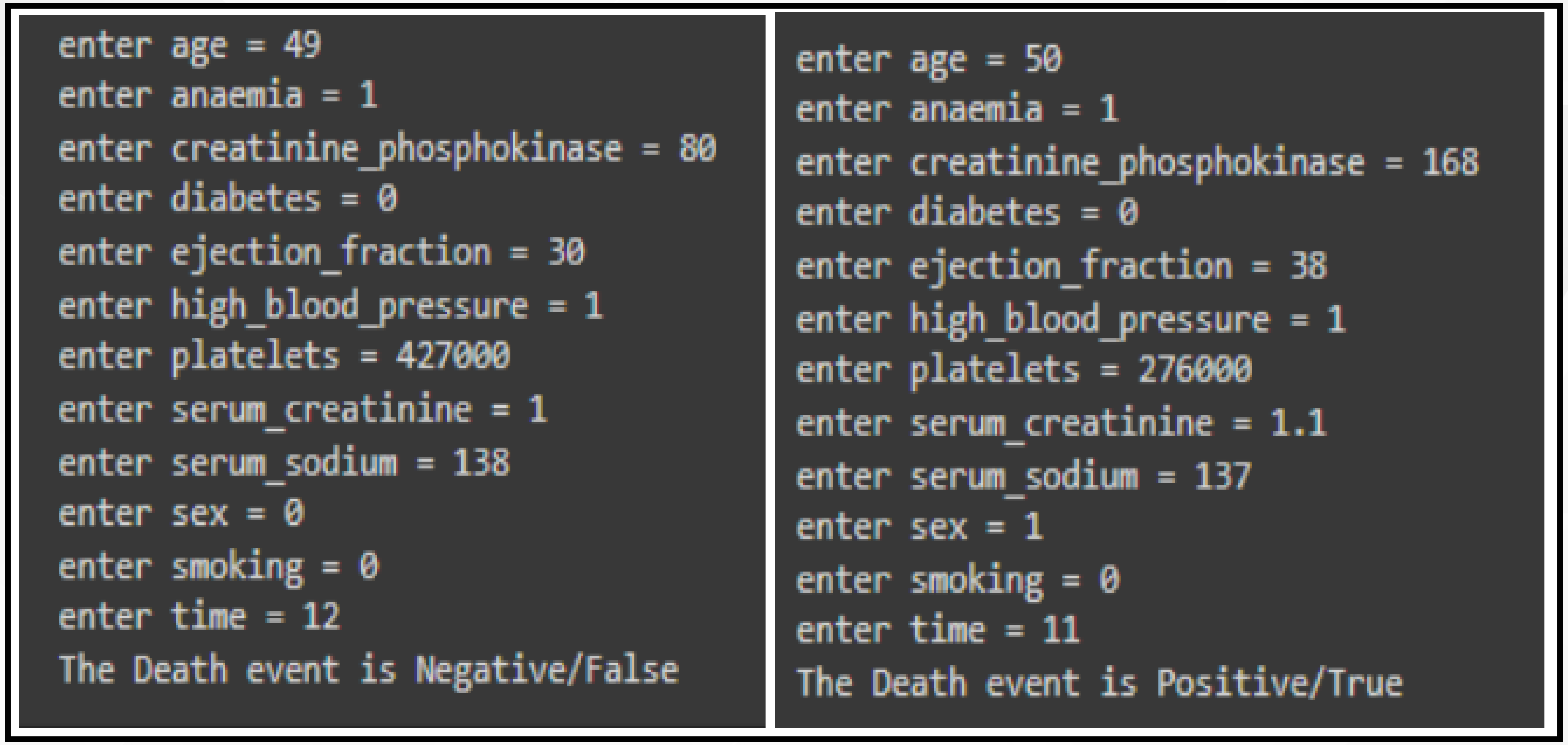


RESULT ANALYSIS



The above chart shows the accuracy results for the models used and clearly the Gradient boost classifier has the highest accuracy hence that model is to be selected for prediction purposes.

PREDICTION SAMPLES



```
enter age = 49
enter anaemia = 1
enter creatinine_phosphokinase = 80
enter diabetes = 0
enter ejection_fraction = 30
enter high_blood_pressure = 1
enter platelets = 427000
enter serum_creatinine = 1
enter serum_sodium = 138
enter sex = 0
enter smoking = 0
enter time = 12
The Death event is Negative/False
```

```
enter age = 50
enter anaemia = 1
enter creatinine_phosphokinase = 168
enter diabetes = 0
enter ejection_fraction = 38
enter high_blood_pressure = 1
enter platelets = 276000
enter serum_creatinine = 1.1
enter serum_sodium = 137
enter sex = 1
enter smoking = 0
enter time = 11
The Death event is Positive/True
```

Above figures represent the different outcomes for different test samples used predicting Negative/False and Positive/True respectively.

CONCLUSION

Medical officers can make early predictions for the purpose of managing healthcare more quickly and with less effort when they use machine learning techniques. A machine learning approach that can help forecast heart failure precisely and effectively as the number of deaths caused by heart failure rises. This study demonstrates how the healthcare management system may be enhanced by using machine learning approaches to predict cardiac failure early on. When compared to other methods in this experiment, the Gradient Boost Classifier appears to obtain the highest performance score. It may result in a viable disease management plan that slows the illness's course. In order to improve accuracy, a hybrid of optimisation methods and machine learning approaches with additional data will be investigated in future study.

THANKYOU