HEART FAILURE PREDICTION

TEAM MEMBERS

Name: Akarsh S Nair

Roll number: AM.EN.U4AIE21008

Name: Nayan MK

Roll Number : AM.EN.U4AIE21048

Name: Sabharinath B

RollNumber: AM.EN.U4AIE21080

Name : Alfy Alex

Roll number: AM.EN.U4AIE21011

Name: Shyamdev Krishnan J

Roll Number: AM.EN.U4AIE21060

Abstract-

Heart failure is a serious medical condition that affects millions of people worldwide and leads to high mortality rates. Early prediction of heart failure can greatly improve patient outcomes and reduce healthcare costs. Machine learning techniques have shown promising results in predicting heart failure. This paper presents a review of various machine learning algorithms that have been used to predict heart failure, including decision trees, random forests, k-nearest neighbors, artificial neural networks, and support vector machines. The performance of each algorithm is compared and evaluated using various metrics such as accuracy, sensitivity, and specificity. The results show that machine learning algorithms can effectively predict heart failure and can be used as a valuable tool for healthcare professionals in the early diagnosis and treatment of this condition. The paper concludes with a discussion of the limitations of the current studies and the future directions for the use of machine learning in heart failure prediction. The authors emphasize the need for further research to validate the results and to improve the performance of machine learning algorithms in predicting heart failure. It is also important to consider the ethical and privacy implications of using machine learning algorithms in clinical practice.

Index Terms : Heart failure prediction, Machine learning, dataset,SVM,Random Forest,KNN

I. INTRODUCTION

Heart failure is a growing public health concern, affecting millions of people worldwide. Early detection and diagnosis of heart failure can greatly improve patient outcomes and reduce healthcare costs. The traditional methods of detecting heart failure rely on physician interpretation of clinical symptoms and signs, which can be subjective and prone to error. To address these limitations, machine learning methods have been applied in the field of heart failure prediction. These methods involve collecting data on medical history, risk factors, and biomarkers and using machine learning algorithms to analyze this data and make predictions about a patient's risk for heart failure. The goal of heart failure prediction is to identify individuals at risk early, so that interventions can be implemented to prevent or slow the progression of the disease. This paper provides an overview of the use of machine learning algorithms in heart failure prediction.

In the past, several researchers have employed various techniques to gather and evaluate data with the goal of forecasting heart failure. Electronic health record

(EHR) information about patients with The Cleveland Heart Disease dataset, UCI biomedical science datasets, and heart failure in various hospitals from various nations are just a few examples. On the basis of these data, a variety of techniques are being used, such as using classifiers for machine learning to predict patient survival, supervised deep learning and machine learning algorithms, training boosted decision tree algorithms, using machine intelligence-based statistical models, the random under-sampling method, and deep neural network models, as well as the bioinformatic explainable deep neural network (BioExpDNN), etc.

In this paper, we will explore the various machine learning algorithms that have been used for heart failure prediction, as well as their strengths, weaknesses, and performance metrics. We will also discuss the importance of feature selection and the role of clinical and demographic variables in improving the accuracy of heart failure prediction. Ultimately, this paper aims to provide a comprehensive overview of the use of machine learning algorithms in heart failure prediction, highlighting the potential of these methods for improving patient outcomes and reducing the burden of heart disease and heart failure on public health.

II. OBJECTIVE

Machine learning has become an important tool in the prediction and diagnosis of heart disease and heart failure. Different machine learning models such as decision trees, random forests, and neural networks have been applied in this field with varying degrees of success. The advantage of these models lies in their ability to analyze vast amounts of data and identify patterns that may be missed by human observation. However, limitations such as the potential for biased results, limited interpretability, and the need for large amounts of high-quality data must be considered. In recent years, there have been significant developments in the use of machine learning for heart disease and heart failure prediction, including the use of electronic health records and imaging data. The use of machine learning also raises ethical and privacy concerns, particularly in the handling of sensitive medical information. To further improve patient outcomes and reduce healthcare costs, future research directions should focus on improving the accuracy and reliability of these models while ensuring that privacy and ethical considerations are fully addressed.

III. PROBLEM DEFINITION

Given clinical parameters about a patient, can we predict whether or not they have heart disease?
Heart failure prediction using machine learning (ML) is a task of developing a model that can accurately predict whether an individual is at risk of developing heart failure in the future, based on demographic and clinical data. We use various prediction models to determine that.

IV. DATASETS AND RELATED WORK

We have used 5 dataset in our project.

Dataset 1

The datasets that we are using for the project has been taken from kaggle ("Heart.csv")

Link:https://www.kaggle.com/fedesoriano/heart-failure-prediction.

The dataset was created by combining different datasets already available independently but not combined before. In this dataset, 5 heart datasets are combined over 11 common features which makes it the largest heart disease dataset available so far for research purposes.

Attribute Information:

1. Age: age of the patient [years]
2. Sex: sex of the patient [M: Male, F: Female]
3. ChestPainType: chest pain type [TA: Typical Angina, ATA: Atypical Angina, NAP: Non-Anginal Pain, ASY: Asymptomatic]
4. RestingBP: resting blood pressure [mm Hg]
5. Cholesterol: serum cholesterol [mm/dl]
6. FastingBS: fasting blood sugar [1: if FastingBS > 120 mg/dl, 0: otherwise]
7. RestingECG: resting electrocardiogram results
8. MaxHR: maximum heart rate achieved [Numeric value between 60 and 202]
9. ExerciseAngina: exercise-induced angina [Y: Yes, N: No]
10. Oldpeak: oldpeak = ST [Numeric value measured in depression]
11. ST_Slope: the slope of the peak exercise ST segment [Up: upsloping, Flat: flat, Down: downsloping]
12. HeartDisease: output class [1: heart disease, 0: Normal]

Dataset 2

The dataset that we are using for the project has been taken from Kaggle ("heart_failure_clinical_records_dataset.csv")

Link: [Heart Failure Prediction | Kaggle]

The dataset was created by taking different attributes altogether , It consist of a total of 13 attributes all in correlation with factors leading to heart disease

Attribute Information:

1. Age : age of the patient [years]
2. Anemia:decrease of red blood cells or hemoglobin
3. Creatinine_phosphokinase : Level of CPK enzyme in the blood(mcg/L)
4. Diabetes : Level of insulin (boolean)
5. Ejection_fraction : percentage of blood leaving the heart at each contraction (percentage)
6. High_blood_pressure : If patient has hypertension (boolean)
7. Platelets : Platelets in the blood (kiloplatelets/mL)
8. Serum_creatinine : Level of serum creatinine in the blood (mg/dL)
9. Serum_sodium : Level of serum sodium in the blood (mEq/L)
10. Sex : Woman or man (binary)
11. Smoking : If the patient smokes or not (boolean)
12. Time : Follow-up period (days)
13. DEATH_EVENT : If the patient deceased during the follow-up period (boolean)

Dataset 3

The dataset has been taken from Kaggle **(heart_cleveland_upload.csv)**
Link:[https://www.kaggle.com/datasets/cherngs/heart-disease-cleveland-uci]. There are 14 attributes involving different features to assess the possibility of Heart disease.

Attribute Information:

1. age: age in years
2. sex (1 = male; 0 = female)
3. cp: chest pain type
   -- Value 0: typical angina
   -- Value 1: atypical angina
   -- Value 2: non-anginal pain
   -- Value 3: asymptomatic
4. trestbps: resting blood pressure (in mm Hg on admission to the hospital)
5. chol: serum cholesterol in mg/dl
6. fbs: (fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)
7. restecg: resting electrocardiographic results
   -- Value 0: normal
   -- Value 1: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV)
   -- Value 2: showing probable or definite left ventricular hypertrophy by Estes' criteria
8. thalach: maximum heart rate achieved
9. exang: exercise induced angina (1 = yes; 0 = no)
10. oldpeak = ST depression induced by exercise relative to rest
11. slope: the slope of the peak exercise ST segment
    -- Value 0: upsloping
    -- Value 1: flat
    -- Value 2: downsloping
12. ca: number of major vessels (0-3) colored by fluoroscopy
13. thal: 0 = normal; 1 = fixed defect; 2 = reversable defect and the label
14. condition: 0 = no disease, 1 = disease

Dataset 4

The dataset has been taken from Kaggle **(HEART.CSV)**

Link:[https://www.kaggle.com/datasets/rashikrahmanpritom/heart-attack-analysis-prediction-dataset]

There are 11 different attributes involving different features to find the possibility of Heart disease.

Attribute Information:

1· Age : Age of the patient
2· Sex : Sex of the patient
3· exang: exercise induced angina (1 = yes; 0 = no)
4· ca: number of major vessels (0-3)
5· cp : Chest Pain type chest pain type
   o Value 1: typical angina
   o Value 2: atypical angina
   o Value 3: non-anginal pain
   o Value 4: asymptomatic
6· trtbps : resting blood pressure (in mm Hg)
7· chol :cholestoral in mg/dl fetched via BMI sensor
8· fbs : (fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)
9· rest_ecg : resting electrocardiographic results
   o Value 0: normal
   o Value 1: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV)
   o Value 2: showing probable or definite left ventricular hypertrophy by Estes' criteria
10· thalach : maximum heart rate achieved
11· target : 0= less chance of heart attack 1= more chance of heart attack

Dataset 5

The dataset has been taken from Kaggle **(Heart_Disease_Prediction.csv)**
Link:[https://www.kaggle.com/datasets/rishidamarla/heart-disease-prediction]

There are 12 attributes involving different features to assess the possibility of Heart disease.

1. Age: age of the patient [years]

2. Sex: sex of the patient [M: Male, F: Female]
3. Chest pain type: chest pain type [TA: Typical Angina,ATA: Atypical Angina, NAP: Non-Anginal Pain, ASY: Asymptomatic]

4. BP: blood pressure [mm Hg]
5. Cholesterol: serum cholesterol [mm/dl]
6. FBS over 120 : fasting blood sugar [1: if FastingBS > 120 mg/dl, 0: otherwise]

7. EKG results
8. Max HR: maximum heart rate achieved [Numeric value between 60 and 202]

9. Exercise angina: exercise-induced angina [Y: Yes, N: No]

10. ST depression
11. Heart Disease: output class [1: heart disease, 0: Normal]

V. METHODS

The working of the system starts with the collection of data and selecting the important attributes. Then the required data is preprocessed into the required format. The data is then divided into two parts: training and testing data. The algorithms are applied and the model is trained using the training data. The accuracy of the system is obtained by testing the system using the testing data. This system is implemented using the following modules.

1.) Collection of Dataset
2.) Selection of attributes
3.) Data Pre-Processing
4.) Balancing of Data
5.) Disease Prediction

In this study, we explore, evaluate and analyze dataset characteristics using different machine learning algorithms like KNN, Random Forest, Decision Tree, Logistic Regression , and SVM.

A . K-Nearest Neighbor

KNN is a non-parametric method used for classification and regression, where an object is classified by a majority vote of its k nearest neighbors. It is a lazy learning method, where the computation is delayed until classification.It can be used to predict heart failure by analyzing patient data such as age, blood pressure, cholesterol levels, and medical history.

In KNN, the algorithm classifies a new patient based on the nearest neighbors in the training dataset. The nearest neighbors are determined based on a distance metric such as Euclidean distance. The majority class among the k nearest neighbors is then assigned to the new patient. To apply KNN for heart failure prediction, a training dataset with labeled patient data must be prepared. The algorithm is then trained on this dataset to determine the nearest neighbors for new patients. The prediction performance can be evaluated by comparing the predicted outcomes with the actual outcomes of the test dataset. It is important to note that KNN is just one of the many algorithms that can be used for heart failure prediction. The performance of KNN will depend on the quality of the training dataset and the selection of the optimal k value.

B . Random Forest

Random Forest is an ensemble learning method for classification and regression that constructs a multitude of decision trees during training, and outputs the class or mean prediction of the individual trees.

The Random Forest algorithm works by creating a large number of decision trees and combining the predictions from each tree to form a final prediction. In a decision tree, each internal node represents a feature and each leaf node represents a prediction. Each tree in the forest is trained on a different subset of the training data, and the prediction is made by taking the majority vote of the trees.

To apply Random Forest for heart failure prediction, a training dataset with labeled patient data must be prepared. The algorithm is then trained on this dataset to generate a forest of decision trees. The prediction performance can be evaluated by comparing the predicted outcomes with the actual outcomes of the test dataset. Random Forest has several advantages, such as the ability to handle non-linear relationships, the ability to handle missing values, and improved prediction performance compared to a single decision tree. It is important to note that the performance of Random Forest will depend on the quality of the training dataset and the number of trees in the forest.

C . Decision Tree

Decision Tree is a tree-based algorithm that uses a flowchart-like tree structure to make a prediction by recursively partitioning the data into subsets and selecting a feature that provides the most information gain.

The Decision Tree algorithm works by recursively splitting the data into smaller and smaller subsets based on the feature that provides the largest information gain. Each split results in a new internal node in the tree, and the leaves of the tree represent the final predictions. The prediction for a new patient is made by following the path from the root of the tree to a leaf node. To apply a Decision Tree for heart failure prediction, a training dataset with labeled patient data must be prepared. The algorithm is then trained on this dataset to generate a decision tree. The prediction performance can be evaluated by comparing the predicted outcomes with the actual outcomes of the test dataset. Decision Tree has several advantages, such as the ability to handle both categorical and numerical features, the ability to handle missing values, and the interpretability of the results. However, it is important to note that Decision Trees can easily overfit the data and lead to poor generalization performance, especially when the tree is deep and has many splits. Pruning techniques or using an ensemble of decision trees can address this issue.

D . Logistic Regression

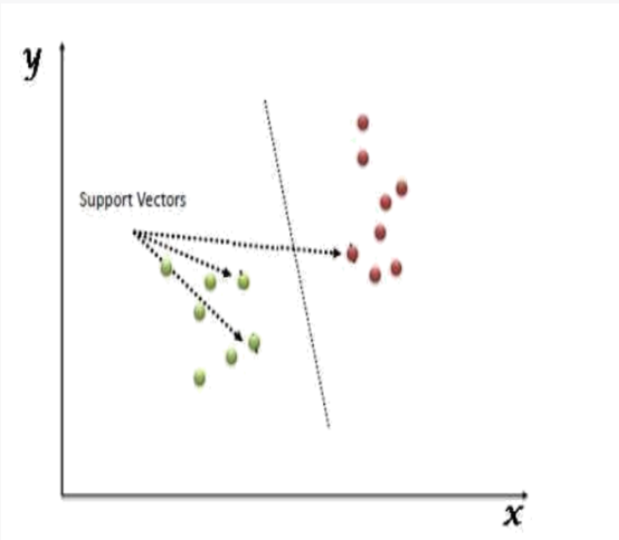
Logistic Regression is a statistical method that uses a logistic function to model a binary dependent variable and make predictions based on input features. It can analyze patient data such as age, blood pressure, cholesterol levels, and medical history to predict the probability of a patient developing heart failure.

In Logistic Regression, a linear model is fit to the input features to predict a probability value between 0 and 1. This probability value is then thresholded to produce a binary prediction (e.g. heart failure vs. no heart failure). The coefficients of the linear model are estimated using a maximum likelihood estimate. To apply Logistic Regression for heart failure prediction, a training dataset with labeled patient data must be prepared. The algorithm is then trained on this dataset to estimate the coefficients of the linear model. The prediction performance can be evaluated by comparing the predicted probabilities with the actual outcomes of the test dataset. Logistic Regression has several advantages, such as the ability to handle both categorical and numerical features, the ability to handle missing values, and the interpretability of the results. It is important to note that Logistic Regression assumes a linear relationship between the input features and the target variable, which may not always be the case in practice. Additionally, Logistic Regression may not perform well when there is a complex non-linear relationship between the input features and the target variable.

E . Support Vector Machines

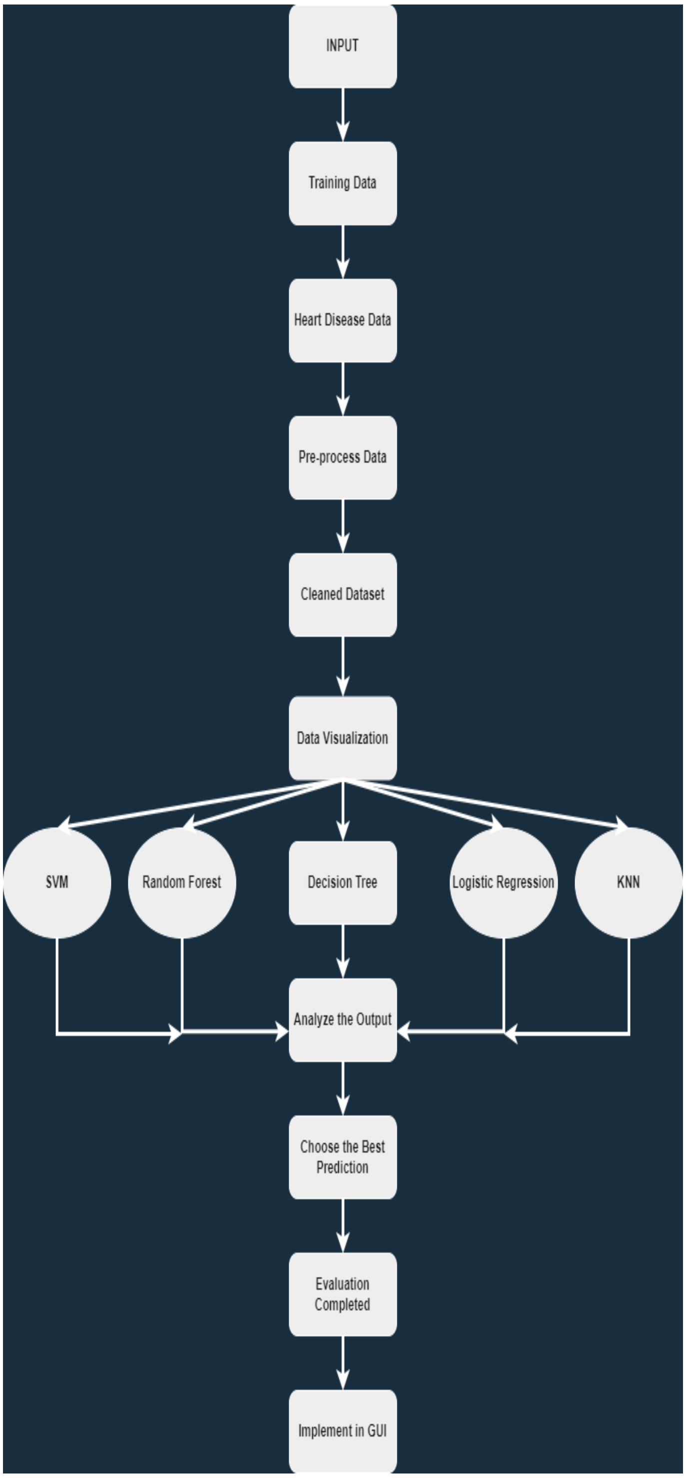
Support Vector Machines (SVM) is a machine learning algorithm that can be used for both classification and regression problems. It can also be used to predict heart failure by analyzing patient data such as age, blood pressure, cholesterol levels, and medical history.

In SVM, the algorithm constructs a boundary (called a hyperplane) that separates the data into different classes. The boundary is chosen in such a way that it maximizes the margin between the closest data points of the two classes, which are called support vectors. SVM can handle non-linearly separable data by transforming the input features into a higher-dimensional space where a linear boundary can be found. To apply SVM for heart failure prediction, a training dataset with labeled patient data must be prepared. The algorithm is then trained on this dataset to find the optimal hyperplane that separates the two classes. The prediction performance can be evaluated by comparing the predicted outcomes with the actual outcomes of the test dataset.



SVM has several advantages, such as the ability to handle both linear and non-linear boundaries, the ability to handle high-dimensional data, and the ability to handle imbalanced datasets. However, it is important to note that SVM can be computationally expensive and may not scale well to large datasets. Additionally, SVM may not perform well when the data is noisy or has overlapping classes.

PROGRAM FLOW

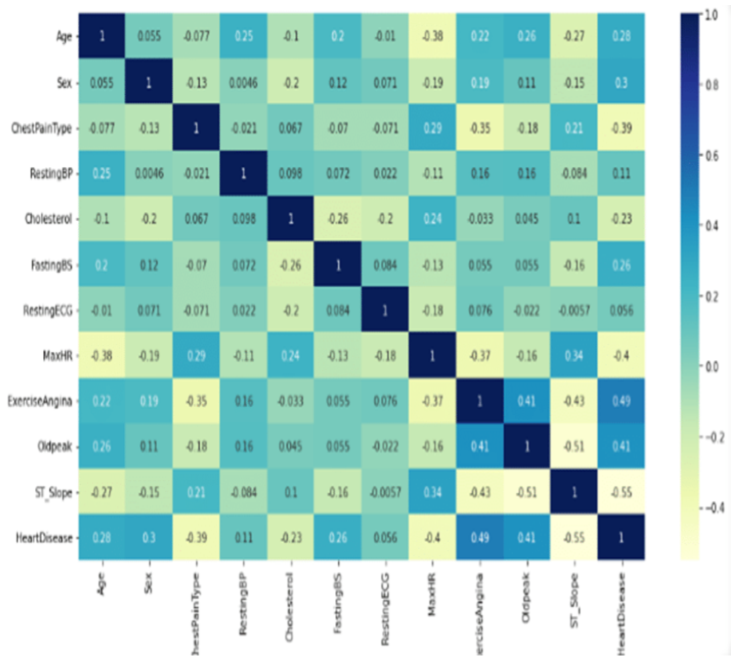


VI. EVALUATIONS AND DISCUSSION

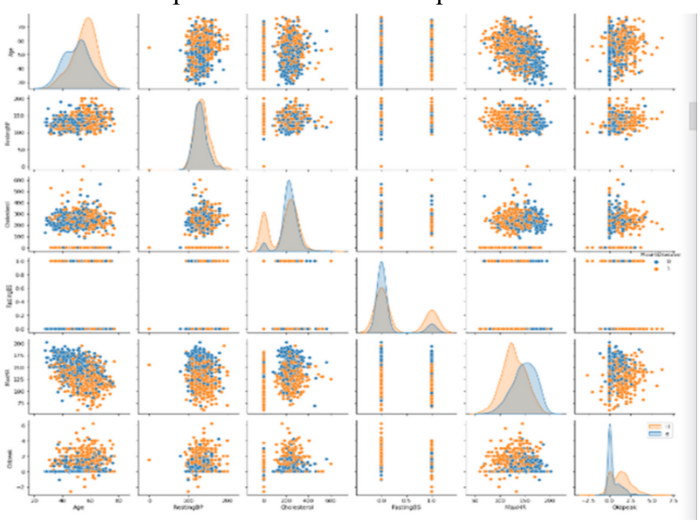
The attributes for each of the dataset are different and there were null values in the dataset which had to be removed, so we have to remove those null values for our better prediction.

Selection of attributes

Attribute or Feature selection includes the selection of appropriate attributes for the prediction system. This is used to increase the efficiency of the system. Various attributes of the patient like gender, chest pain type, fasting blood pressure, serum cholesterol, exang, etc are selected for the prediction. The Correlation matrix is used for attribute selection for this model

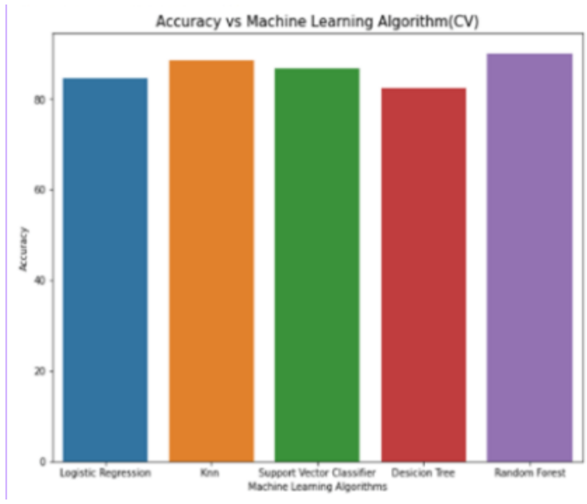


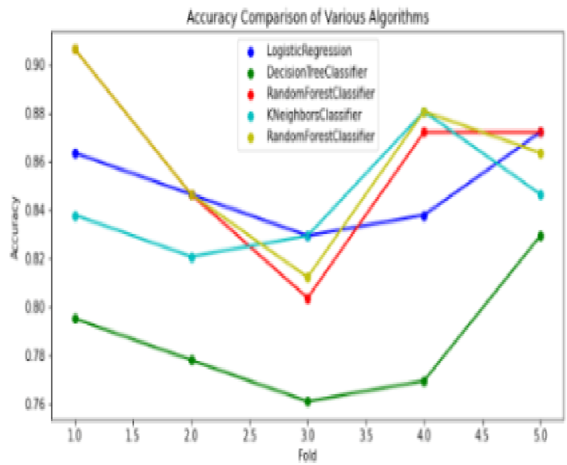
By looking at the pair plot we could also find the relationship between each of the parameters.



Comparison of the performance of different ML models, including supervised learning models, unsupervised learning models, and deep learning models, in predicting heart failure.Examination of the ethical and privacy concerns related to using machine learning for heart disease and heart failure prediction and diagnosis, and how these can be addressed.Analysis of the current state of machine learning in heart disease and heart failure prediction and diagnosis, including the challenges and opportunities in this field.Exploration of the potential for machine learning to improve patient outcomes and reduce healthcare costs, including the use of personalized treatment plans and early diagnosis.

For dataset 1:





The accuracy for logistic regression is 84.2

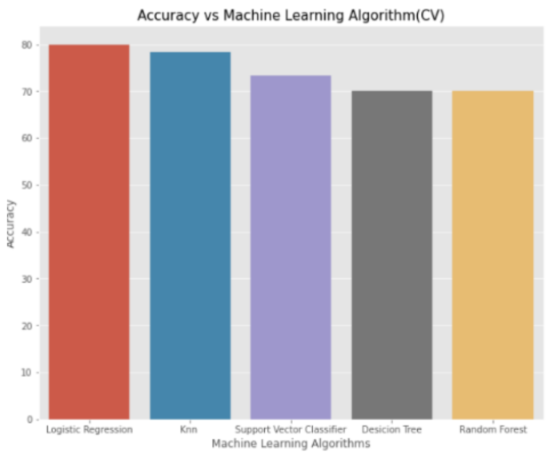
The accuracy for knn is 88.96

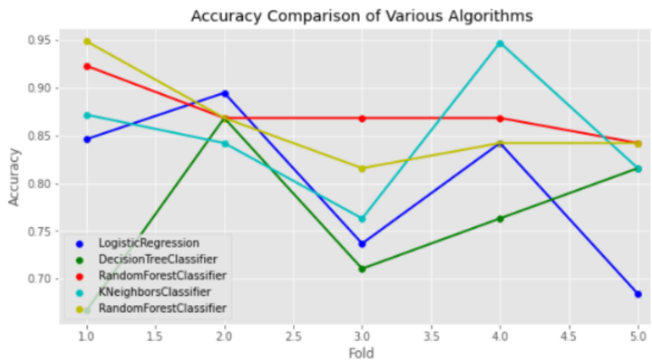
The accuracy for svc is 85.89

The accuracy for Decision tree is 82.63

The accuracy for Random forest is 94.97

For dataset 2:





The accuracy for logistic regression is 80

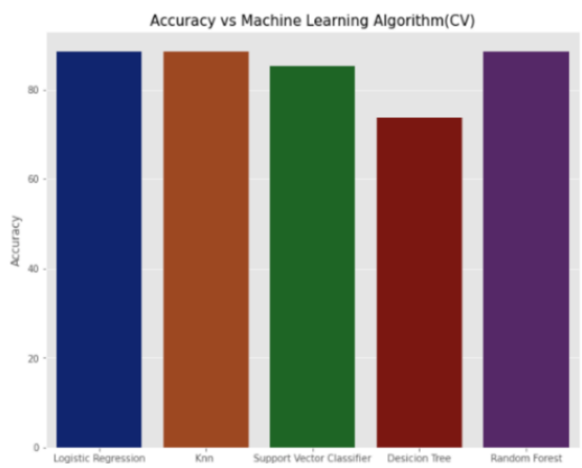
The accuracy for knn is 78.21

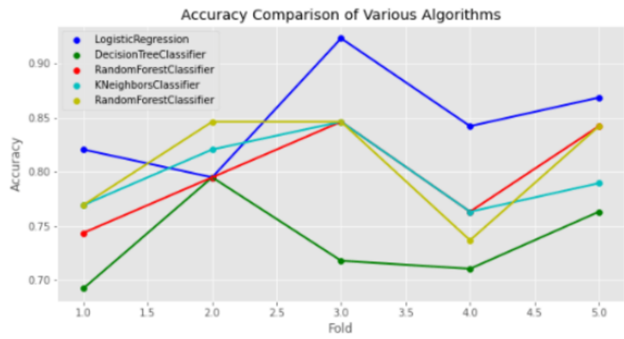
The accuracy for svc is 73.4

The accuracy for Decision tree is 70

The accuracy for Random forest is 82

For dataset 3:





The accuracy for logistic regression is 92.42

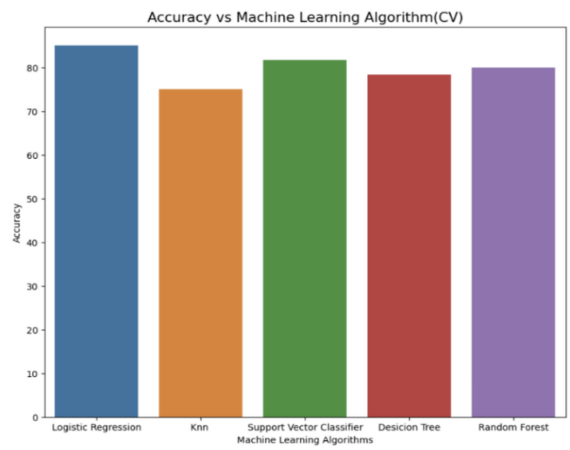
The accuracy for knn is 92.39

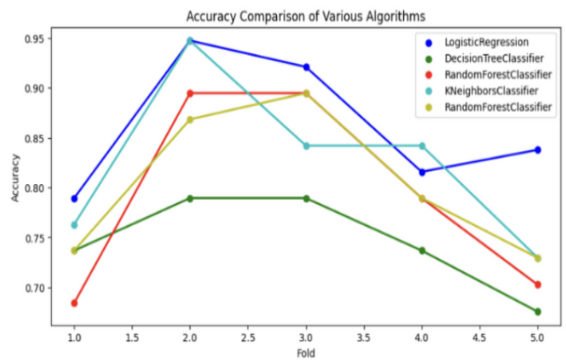
The accuracy for svc is 87.5

The accuracy for Decision tree is 74.2

The accuracy for Random forest is 93.4

For dataset 4:





The accuracy for logistic regression is 93.72

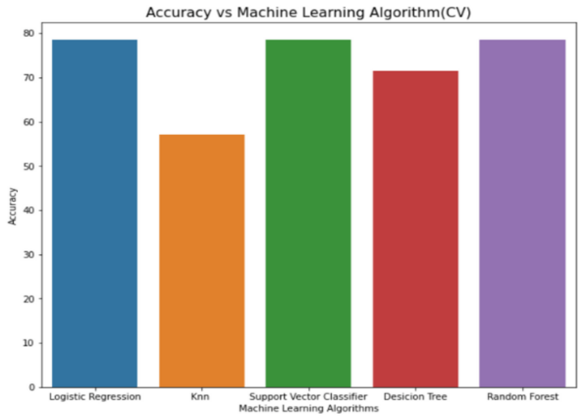
The accuracy for knn is 74

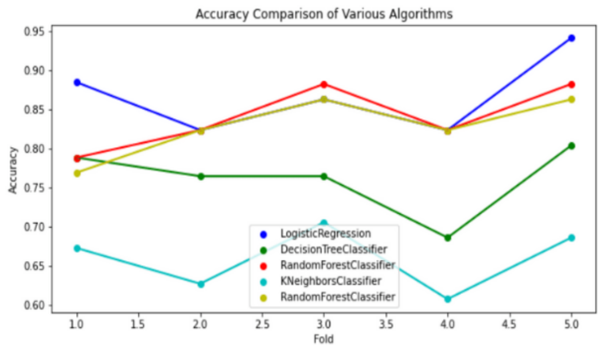
The accuracy for svc is 82

The accuracy for Decision tree is 78

The accuracy for Random forest is 81

For dataset 5:





The accuracy for logistic regression is 78.2

The accuracy for knn is 56.6

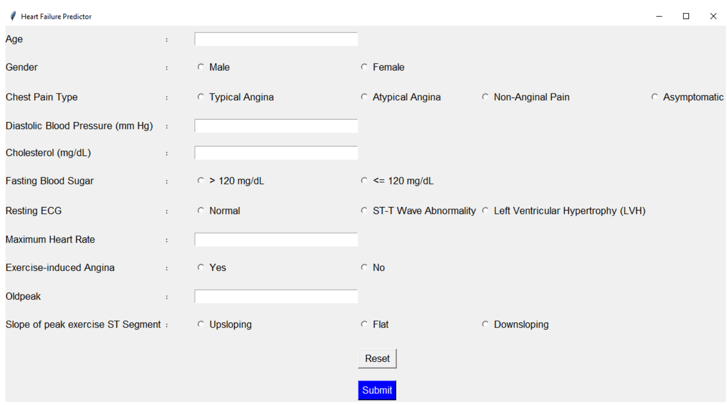
The accuracy for svc is 79.2

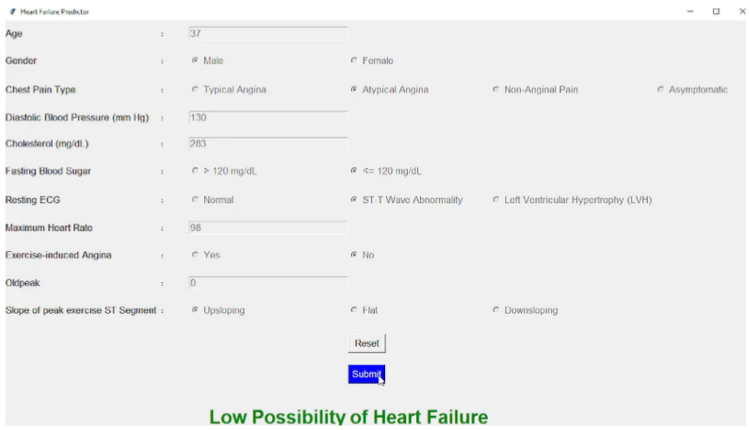
The accuracy for Decision tree is 71

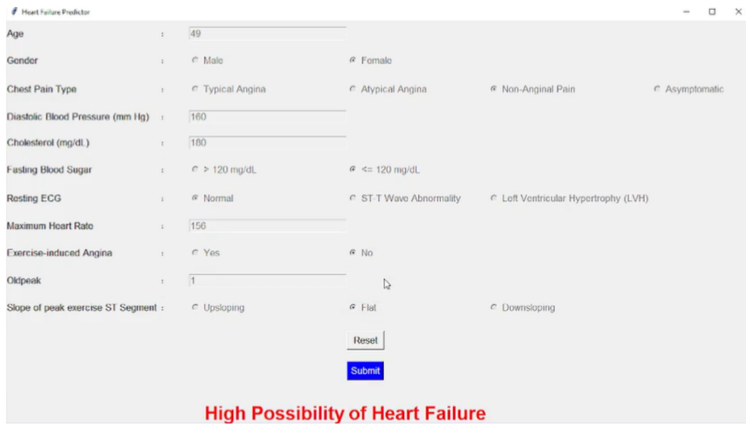
The accuracy for Random forest is 79.23

VII. DEMO

Here we are asking the user to input the various parameters and will predict the output, ie whether he/she has heart disease or not based on the corresponding input taken.







VIII. FINDINGS AND FUTURE DIRECTIONS

Comparison of the performance of different ML models, including supervised learning models, unsupervised learning models, and deep learning models, in predicting heart failure.From all the ML model used , it seems that the random forest performs well in many of the dataset which gives us an accuracy of more than 85% for almost all of the data

Examination of the ethical and privacy concerns related to using machine learning for heart disease and heart failure prediction and diagnosis, and how these can be addressed.

Analysis of the current state of machine learning in heart disease and heart failure prediction and diagnosis, including the challenges and opportunities in this field.

Exploration of the potential for machine learning to improve patient outcomes and reduce healthcare costs, including the use of personalized treatment plans and early diagnosis.

IX. CONCLUSION

Initially, we compare 5 different datasets based on Heart Disease with various attributes. We came to know that Dataset 1 (with 11 attributes) is the best dataset out of the 5 datasets, with it's precise attributes and consistent data. Next, we train the model using various Machine Learning algorithms and find the most accurate one. We'll then be moving forward with that particular algorithm. Here we're comparing the following algorithms: Logistic Regression, KNN, Support Vector Classifier, Decision Tree, Random Forest. Random Forest is found to be the most accurate algorithm after comparison. We'll now be using this algorithm for prediction**.** Finally, we compare the algorithms for all 5 datasets using visualization graphs like bar graphs and scatter plots. With the help of K-Fold Cross Validation, we split the data into different groups which makes the comparison of algorithms easier. The procedure has a single parameter called k that refers to the number of groups that a given data sample is to be split into. After comparing all 5 dataset's accuracy, Dataset 1 has the best overall accuracy among them. Therefore, Dataset 1 is used in the GUI.

X. APPENDIX

NumPy: NumPy is a Python library for working with large, multi-dimensional arrays and matrices of numerical data, as well as a large collection of high-level mathematical functions to operate on these arrays.

Pandas:Pandas is a library for data manipulation and analysis in Python, providing efficient data structures and powerful manipulation capabilities.

Matplotlib:Matplotlib is a plotting library for creating static, animated, and interactive visualizations in Python.

Seaborn:Seaborn is a data visualization library for creating attractive and informative statistical graphics in Python, built on top of Matplotlib.

Scikit-learn: Scikit-learn is a machine learning library for Python, providing simple and efficient tools for data mining and data analysis.

X. IMPLEMENTATION

- https://colab.research.google.com/drive/1w3l0bEX2kuPuvuBvY3aZGpqkT5_B5e91
- https://colab.research.google.com/drive/1i_TMGf6v1aFxumJLif9VkHpcJTe7V-OB#scrollTo=8c4qNdqDCFMk
- https://colab.research.google.com/drive/1CczO9imEesNsgI4DY_5iwtyJTLdFXp1C#scrollTo=36zUvvCibKL0
- https://colab.research.google.com/drive/1qMStZ4hz3kX6GljvHvzaY8BICja8gcI_?usp=sharing

ACKNOWLEDGEMENT

We wish to show our appreciation to Dr .Manjusha Nair M. and Dr Athira who guided us throughout this project. We would like to thank our friends and family for providing us with deep insights into this study. We would also like to extend our special thanks to Amrita university and Chancellor Amma for providing us with her grace and presence in guiding us through this study.

REFERENCES

[1] f1000research. Background referred from:

https://f1000research.com/articles/11-1126

[2] Kaggle. Dataset referred from: https://www.kaggle.com/fedesoriano/heart-failure-prediction.

[3]`tkinter` — Python interface

https://docs.python.org/3/library/tkinter.html

[4] ANITS.EDU - Methodology referred from:

https://cse.anits.edu.in/projects/projects2021C3.pdf

[5] https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction

[6] https://scholar.google.co.in/scholar?q=heart+failure+prediction+using+machine&hl=en&as_sdt=0&as_vis=1&oi=scholart