HEART FAILURE PREDICTION USING MACHINE LEARNING

Jaya shree[1], Megavarshini[2],Maheshwari[3],Angel Meriba [4]

ABSTRACT- Heart Failure is a major issue in the modern society. It presently ranks among the most dangerous illnesses that can affect people, shortening lifespans. The chance of heart disease can be raised by a number of factors, including smoking, body cholesterol levels, a family history of the disease, obesity, high blood pressure, diabetes, and inactivity. To analysis heart failure dataset here we are using machine learning algorithms. Machine learning is a method to predict the target value using input values. The outcomes of this prediction is based on the Naive Bayes (NB), Logistic Regression (LR) and Support Vector Machine (SVM) are introduced for effective forecasting as these machine learning algorithm achieve the best accuracy.

I INTRODUCTION

The human body's heart is a vital component. Blood is given to each organ in the body by it. If it doesn't operate properly, the brain and several other organs will stop working, and the person will die in a matter of minutes. A variety of heart-related diseases are becoming more common as a result of lifestyle changes, stress, diabetes, and poor eating practises. Our research focuses on predicting the incidence of heart failure prediction using patient datasets and a dataset of patients for whom we need to make such predictions. One application of this machine learning is in prediction. It is a technology that can assist in making a diagnosis of heart illness before a person suffers significant harm. E-commerce is just one of the many uses where machine learning is extensively used.

The primary goal of this discussion is to emphasize the value of using medical data to predict heart disease using machine learning techniques. This investigation is also emphasizing on overcoming the precarious situation, a computerized method was suggested, ensuring that heart consultants would not miss any information due to incorrect reading and comprehension of the data.

The dataset contains 299 patient in medical records. This dataset consists of 13 essential attributes that are important in determining whether or not Heart Disease is present. This project uses a machine learning model and Python, the most widely used computer language. Support Vector Machine (SVM), Naive Bayes (NB), Logistic Regression and are supervised machine learning methods that we use in this situation. As a result, these algorithms are now extremely helpful in correctly forecasting heart failure.

The paper divided into further sections, the next section describes the related works and section III explains the overview of heart failure prediction and followed by IV, V, VI and VII tells about the data analysis, implementation, methodology and result analysis of the given dataset. Finally, section VIII has been concluded.

II RELATED WORKS

1. The author **Prasanta Kumar Sahoo, Pravalika Jeripothula** published in 2020[1]. They used many algorithms, including SVM, Naive Bayes, Logistic Regression, Decision Tree, and others are used to carry out this work. KNN, SVM was determined to have the highest accuracy, reaching 85.2%. This study also uses the model validation technique to create the most appropriate model for the current situation.

2. The author **Fahd Saleh Alotaibi1**,published in 2019[2]. This study uses data from the UCI heart disease dataset to increase the accuracy of HF prediction. For this, a variety of machine learning techniques were utilised to analyse the data and predict the possibility of heart failure in a medical database. The technique can be useful and beneficial for doctors and heart surgeons in

the future for accurately diagnosing a patient's risk of a heart attack, according to the results of the study.

3. The author **Jing Wang,** published in 2021[3]. Using z-score or min-max normalisation approaches and Synthetic Minority Oversampling Technique (SMOTE) for the imbalance class problem that is frequently found in this problem, we compare and contrast 18 common machine learning models for the prediction of heart failure in this research. Our findings show how effective z-score normalisation and SMOTE are at predicting heart failure.

4. The author **K. Karthick ,S. K. Aruna , Ravi Samikannu , Ramya Kuppusamy ,Yuvaraja Teekaraman , and Amruth Ramesh Thelkar , K. Karthick**, published in 2022.This paper helps doctors decide on patients' health with greater accuracy. Those who are diagnosed early can adjust their lifestyles and, if necessary, receive excellent medical therapy. Machine learning (ML) is a feasible solution for minimising and understanding heart disease symptoms. The experiments' results show that the random forest algorithm achieves 88.5% accuracy during validation for 303 data samples with 13 chosen Cleveland HD dataset attributes.

5. The author **Davide Chicco and Giuseppe Jurman,** published in 2020. In this paper they predict the patients' survival and rank the features related to the most significant risk factors, different machine learning classifiers were used. In fact, serum creatinine and ejection fraction may be the key areas of attention for medical professionals trying to determine whether a patient would survive after heart failure.

III PROPOSED ARCHITECTURE

Data gathering is the first step in this project's process. We have gathered the open-source, readily available data collection from kaggle.Data preparation comes after data gathering. The data is cleaned up in this step by getting rid of pointless values. Additionally, it eliminates corrupted, missing, or null values.

After the data has been cleaned, the next step is to divide it into two sets: training data and testing data. Values must be dealt with before we can build the training model. With training data, we create a prediction model.

We choose the SVM algorithm since it is effective and obtained greater accuracy. We must now determine the precision of the model. Predicting the disease is the last step. Numerical

1 = Yes; 0 = No will represent the result.

The proposed architecture is shown in figure1

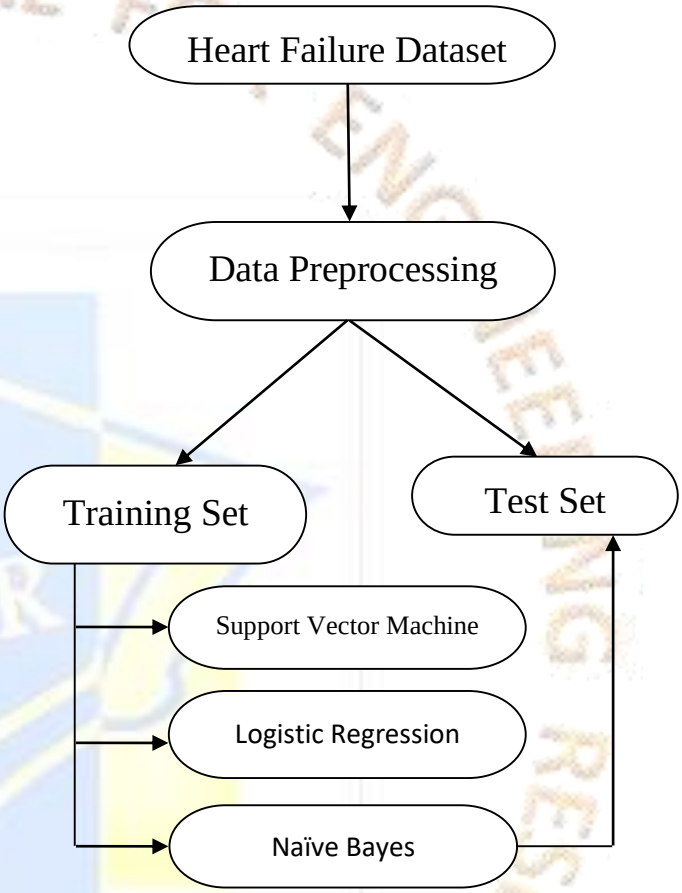

Fig1: Proposed System

1. DATASET

A gathering of data is known as a dataset. A dataset relates to one or more database tables, where each row correlates to a specific record in the corresponding dataset and each column to a specific variable. To forecast accuracy in this case, we are using the Heart Failure Dataset from Kaggle website. The dataset contains 299 patients data with 13 attributes. These 13 clinical attributes have been trained and preprocessed to determine whether or not there is heart failure.This dataset contains 0 and 1 for some attributes to denote 'Yes' or 'No', if Yes the value will be 1 and 0 for No.

- Sex -> Male = 1, Female =0
- Age -> Patient Age
- Diabetes -> 0 = No, 1 = Yes
- Anaemia ->0 = No, 1 = Yes
- High_blood_pressure-> 0 = No, 1 = Yes
- Smoking -> 0 = No, 1 = Yes
- DEATH_EVENT -> 0 = No, 1 = Yes

S.NO	ATTRIBUTES	DESCRIPTION
1	Age	Patient Age(Years)
2	Anaemia	Decrease of red blood cells or hemoglobin (boolean)
3	Creatinine_phosphokinase	Level of the CPK enzyme in the blood (mcg/L)
4	Diabetes	If the patient has diabetes (boolean)
5	Ejection_fraction	Percentage of blood leaving the heart at each contraction (percentage)
6	High_blood_pressure	If the 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	Patient smokes or not
12	Time	Follow-up period
13	DEATH_EVENT	Patient affected by heart failure or not

Table 1: Data Overview

2. DATA PREPROCESSING

Data preprocessing is a process of Preparing the data for use in any experiment involving machine learning or data mining is an essential stage in the process of cleaning the data.it eliminates outliers and extracts features from the data. Moreover, feature noise from the data must be removed, and any missing features must be filled in appropriately.

The dataset used in this paper undergo the following data preprocessing task.

- **Index** of the dataset:

RangeIndex(start=0, stop=299, step=1)

- **Columns** of the dataset:

Index(['anaemia', 'creatinine_phosphokinase','diabetes','ejection_fraction','high_blood_pressure', 'platelets', 'serum_creatinine','serum_sodium','sex','smoking','time','DEATH_EVENT'],dtype='object')

- **Size** of the dataset: 3887
- **Shape** of the dataset: (299, 13)
- **Memory usage** of the dataset:

Index	128
age	2392
Anaemia	2392
creatinine_phosphokinase	2392
diabetes	2392
ejection_fraction	2392
high_blood_pressure	2392
platelets	2392
serum_creatinine	2392
serum_sodium	2392
sex	2392
smoking	2392
time	2392
DEATH_EVENT	2392
dtype	int64

Table 2: Memory Usage

➢ **Info** of the dataset:

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 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: 30.5 KB
```

Fig2: Info

➢ **Describe** of the dataset:

Index	age	anaemia	e phosol	diabetes	tion frac	blood pre	platelets	um creati	um sodi	sex	smoking	time	ATH EVEI
count	299	299	299	299	299	299	299	299	299	299	299	299	299
mean	60.8339	0.431438	581.839	0.41806	38.0836	0.351171	263358	1.39388	136.625	0.648829	0.32107	130.261	0.32107
std	11.8948	0.496107	970.288	0.494067	11.8348	0.478136	97804.2	1.03451	4.41248	0.478136	0.46767	77.6142	0.46767
min	40	0	23	0	14	0	25100	0.5	113	0	0	4	0
25%	51	0	116.5	0	30	0	212500	0.9	134	0	0	73	0
50%	60	0	250	0	38	0	262000	1.1	137	1	0	115	0
75%	70	1	582	1	45	1	303500	1.4	140	1	1	203	1
max	95	1	7861	1	80	1	850000	9.4	148	1	1	285	1

Fig 3: Describe of the Dataset

➢ **Null values** of the dataset:

```
age                         0
anaemia                     0
creatinine_phosphokinase    0
diabetes                    0
ejection_fraction           0
high_blood_pressure         0
platelets                   0
serum_creatinine            0
serum_sodium                0
sex                         0
smoking                     0
time                        0
DEATH_EVENT                 0
dtype: int64
```

Fig4: Null values

An essential phase in the data mining process is data preprocessing. It describes the processes of preparing data for analysis by cleaning, transforming, and integrating it.After these all data preprocessing and data mining process are done, it undergoes training set, validation set, test set to predict the best accuracy.

3. TRAINING SET

The initial set of data needed to train machine learning models is known as training data (or a training dataset).Machine learning algorithms are taught how to generate predictions or complete a specified task using training datasets. The training data will differ according to the supervised or unsupervised learning.

Shape of the training set (209, 13)

	age	anaemia	creatinine_phosphokinase	diabetes	ejection_fraction	high_blood_pressure	platelets	serum_creatinine	serum_sodium	sex	smoking	time	DEATH_EVENT
13	50.0	1	168	0	38	1	276000.0	1.1	137	1	0	11	1
82	50.0	1	249	1	35	1	319000.0	1.0	128	0	0	28	1
63	45.0	1	582	0	35	1	385000.0	1.0	145	1	0	61	1
146	52.0	1	132	0	30	1	218000.0	0.7	136	1	1	112	0
115	58.0	1	400	0	40	1	164000.0	1.0	139	0	0	91	0

Table 3: Trained Dataset

Trained accuracy: 0.8229665071770335

The median for x Train variables is:

Age	60.0
Anaemia	1.0
creatinine_phosphokinase	235.0
diabetes	0.0
ejection_fraction	38.0
high_blood_pressure	1.0
platelets	259000.0
serum_creatinine	1.1
serum_sodium	137.0
sex	1.0
smoking	0.0
time	117.0

dtype: float64

After training the dataset, we get 209 rows and 13 columns in the trained data set and accuracy with 82.3%.It also predict the median of the each attribute in a dataset.

4. TEST SET

A machine learning programme is tested using a test set, which is a secondary (or tertiary) data set used after the programme has been trained on a training set of data**.** If a model that fits the training data set likewise fits the test data set well, there hasn't been much over-fitting. Over-fitting is typically shown by the training data set fitting the model better than the test data set.

Shape of the test set (90,13)

	age	anaemia	creatinine_phosphokinase	diabetes	ejection_fraction	high_blood_pressure	platelets	serum_creatinine	serum_sodium	sex	smoking	time	DEATH_EVENT
73	65.0	1	224	1	30	1	149000.0	1.3	137	1	1	72	0
110	65.0	1	129	0	60	1	306000.0	1.2	132	1	1	90	1
291	60.0	1	320	0	35	1	130000.0	1.4	139	1	0	258	0
168	65.0	1	582	1	40	1	270000.0	1.0	138	0	0	140	0
209	50.0	1	337	0	38	1	390000.0	0.9	144	0	0	256	0

Table 4: Test Dataset

Test accuracy: 0.8111111111111111

The median for x Test variables is:

Age	60.00
Anaemia	1.00
creatinine_phosphokinase	310.00
diabetes	0.00
ejection_fraction	36.50
high_blood_pressure	1.00
platelets	263358.03
serum_creatinine	1.10
serum_sodium	136.50
sex	1.00
smoking	0.00
time	112.50

dtype: float64

After testing the dataset, we get 90 rows and 13 columns in the trained data set and accuracy with 81.1%.It also predict the median of the each attribute in a dataset.

5. VALIDATION SET

A validation set is a collection of data used to train machine learning in order to identify and improve the best model for a certain problem. Development sets or dev sets are another name for validation sets.The number of hidden units in each layer is a hyperparameter for artificial neural networks. It should follow the same probability distribution as the training data set, as well as the testing set.

IV DATA ANALYSIS

Let's examine the age range of those who have the sickness or not is shown in figure4. Hence, target = 1 suggests that the person has heart failure, whereas target = 0 suggests that the person is not affected.

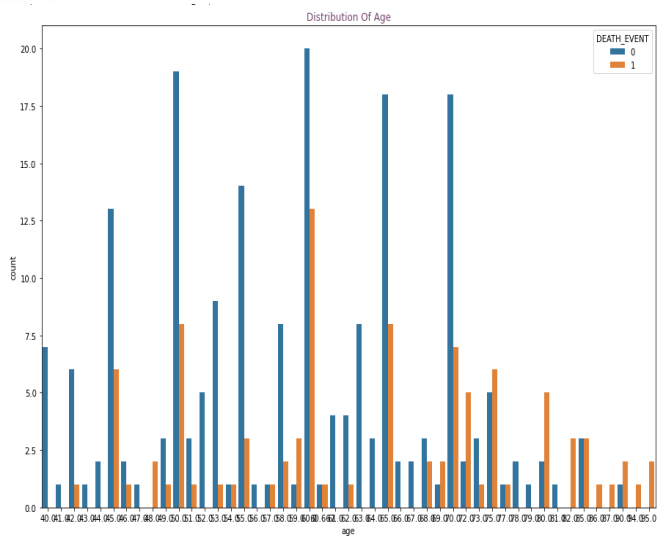

Fig 4: Distribution of Age Graph

.

Here the most people who are affected by heart failure are above the age of 60, followed by 51.Majorly, people belonging to the age group 50+ are suffering from the heart failure.

A statistical measure known as correlation expresses whether closely two variables are related linearly (meaning they change together at a constant rate). It's a typical technique for describing simple relationships without explicitly stating cause and effect.

The correlation visual representation has shown in figure5

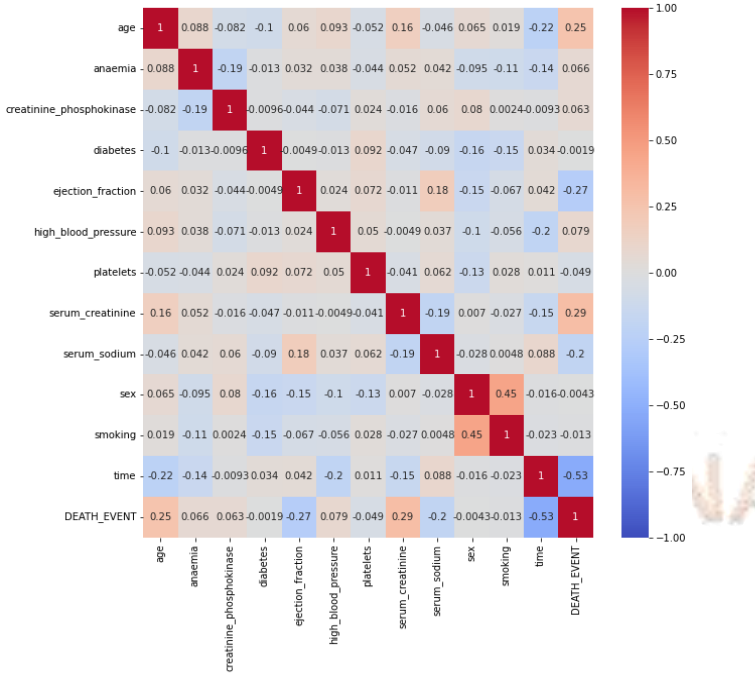

Fig5: Correlation Graph between Columns

V IMPLEMENTATION

Python programming was chosen for this project as it is a high level programming language, has a large library, and automates processes to make them more efficient.

Installing Python is the first step, after which we must import the following libraries:

1. Numpy: Numpy is used to process multi-dimensional arrays, perform element-to-element operations, and utilize a variety of array processing techniques.

2. Pandas: One of the most popular Python libraries, Pandas offers outstanding performance. It manipulates data and makes quick work of data analysis.

3. Sklearn: This is the most helpful library. It has a lot of effective tools, and it is used to create statistical models like clustering, regression, and classification.

4. Matplotlib.pyplot: It is a most popular python library, Matplotlib is a cross-platform for data visualisation and graphical plotting package. The APIs (Application Programming Interfaces) for matplotlib allow programmers to include graphs into GUI applications.

5. Seaborn: Python's Seaborn package allows you to create statistical visuals. It combines closely with Pandas data structures and is built upon Matplotlib. Seaborn helps you explore and understand your data.

VI METHODOLOGY

In this study, machine learning was used to predict heart failure. Machine learning offers high scope for classification and prediction particularly in the medical field. The accuracy of data mining has improved. Here we are using algorithms such as Support Vector Machine, Logistic Regression, Naïve Bayes to achieve the best accuracy.

1. SUPPORT VECTOR MACHINE:

Support Vector Machine is a linear model for classification Problems. SVM can be used to handle both linear and non-linear problems. The Support Vector Machine algorithm generates a line known as a hyper-plane that groups various types of data.

STEPS:

- Import the necessary libraries.
- Import the dataset and segregate X and Y variables for extraction.
- Split the dataset into train and test sets.
- Initializing the SVM classifier model.
- Fitting the SVM classifier model.
- The Final step is making predictions.

$$Accuracy = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FN} + \text{TN} + \text{FP}}$$

SVM Accuracy Score: 0.7666666666666667

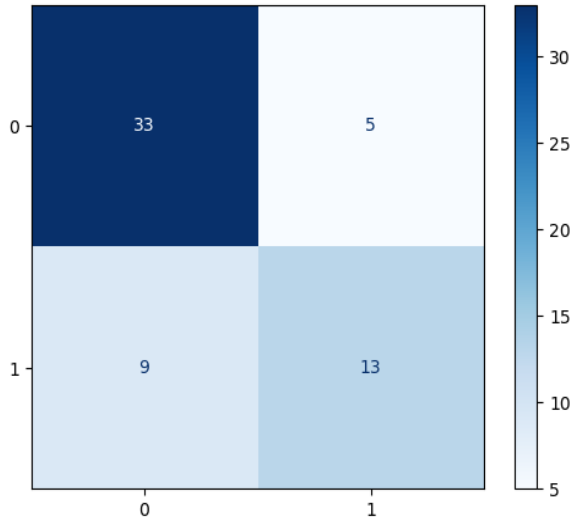

Fig6: Confusion matric for SVM

Using support vector machine we get the accuracy of 76.7% while predicting our heart failure dataset.

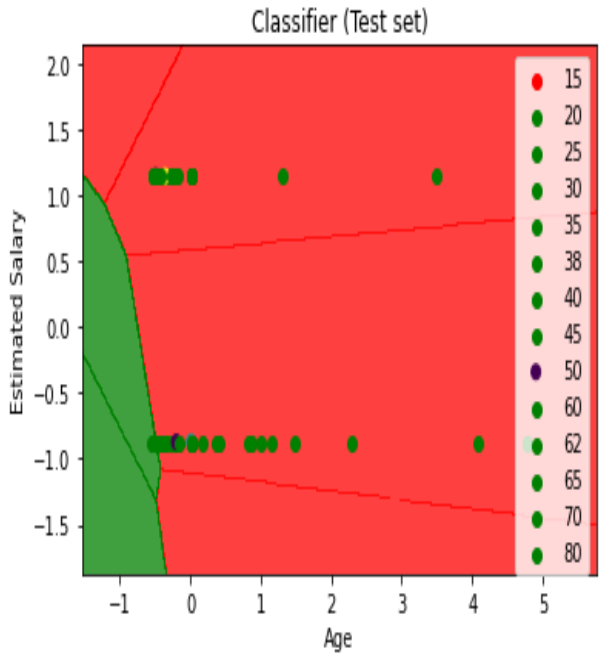

Fig7:Graph for Logistic Regression

2. LOGISTIC REGRESSION

Another classification technique is logistic regression, which uses regression analysis to identify and predict the parameters in the provided dataset. The prediction and learning processes rely on calculating the possibility of a binary classification. The "target" column in this one has two types of binary numbers: "0" for individuals who have no possibility of developing heart failure and "1" for those who have been identified as heart failure patients. The independent variables, on the other hand, can be polynomial, nominal, or binary classified.

STEPS

- ❖ Data preparation step.
- ❖ Logistic regression is applied to the training set.
- ❖ Calculating the test outcome.
- ❖ Test the result's correctness
- ❖ Displaying the results of the test set

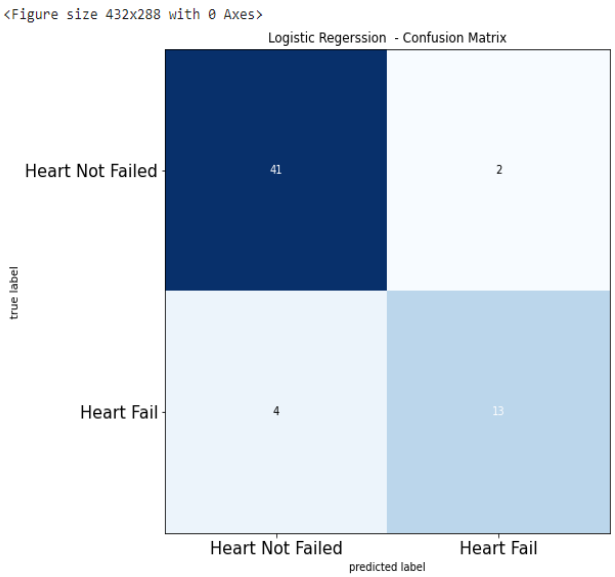

Fig8: Confusion matric for LR

LR Accuracy Score: 90.00%

After using logistics regression we get the accuracy of 90.00% for heart failure prediction.
It has given maximum possible accuracy outcome while prediction.

3. NAIVE BAYES

Naive Bayes classifier uses supervised learning to classify the data by calculating the probability of independent variables. The high probability class is given for the entire transaction once the probabilities for each class have been calculated. For predicting classes for various datasets, such as those used in educational data mining and medical data mining, naive Bayes is a popular approach.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

STEPS

- Import key libraries.
- Importing the dataset.
- Data preprocessing takes place
- Training the model.
- The model is tested and evaluated.
- Visualizing the model.

NB Accuracy Score: 0.7833333333333333

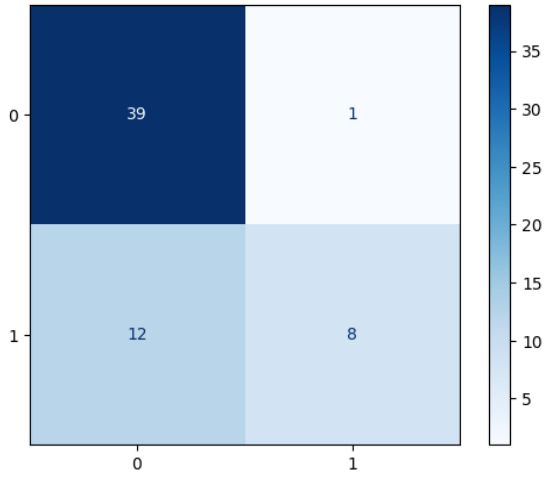

Fig9: Confusion matric for NB

Finally using Naïve Bayes we get the ac curacy of 78.3% for the dataset which we used i n prediction of heart failure.

VII RESULT ANALYSIS

This research uses a data set made up of different attributes like as patient age, anaemia, diabetes, platelets and more. The data set is then split into two sets, with the training set being 82.3% and the testing set being 81.1%.The model is built using a training set, and its accuracy is assessed through testing. The data set is implemented across three different algorithms for this research project, and the results are compared.

ALGORITHMS	ACCURACY
SVM	**76.7%**
NB	**78.3%**
LR	**90.00%**

Table5: Accuracy table for the Dataset

VIII CONCLUSION

More than 550,000 people are impacted by heart failure every year, which is a major issue that negatively affects people's lives. It must forecast this illness and employ various strategies to lessen its effects on the human body. It makes use of information like blood pressure, cholesterol, and chest pain before assisting in the prediction of a patient's future heart attack. As previously mentioned, a family history of heart failure may also contribute to the development of heart disease. So, it may be possible to use this patient knowledge to enhance the model's accuracy.

IX REFERENCES

1. Prasanta Kumar Sahoo, 2 Pravalika Jeripothula., “Heart Failure Prediction Using Machine Learning Techniques”, pp.0000-0002-5164-1010
2. Fahd Saleh Alotaibi1, “Implementation of Machine Learning Model to Predict Heart Failure Disease”, Vol. 10, No. 6, 2019.
3. Jing Wang, “Heart Failure Prediction with Machine Learning: A Comparative Study”,2031 (2021) 012068.
4. K. Karthick ,S. K. Aruna , Ravi Samikannu , Ramya Kuppusamy ,Yuvaraja Teekaraman , and Amruth Ramesh Thelkar , K. Karthick , “Implementation of a Heart Disease Risk Prediction Model Using Machine Learning” , Volume 2022, Article ID 6517716.

5. Davide Chicco and Giuseppe Jurman, ”Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone”, https://doi.org/10.1186/s12911-020-1023-5.
6. Po-Yu Liang, Lee-Jyi Wang, Yang-Sheng Wu, “Prediction of patients with heart failure after myocardial infarction”,DOI:10.1109/BIBM49941.2020.9313253.
7. Xin Sang,Quan Zhu Yao,Ling Ma , Hong Wen Cai, Peng Luo,”. Study on survival prediction of patients with heart failure based on support vector machine algorithm”, DOI 10.1109/ICRIS52159.2020.00160.