VISVESVARAYA TECHNOLOGICAL UNIVERSITY

JNANA SANGAMA, BELAGAVI – 590 018

An Internship Project Report on

“HEART FAILURE ANALYSIS”

Submitted in partial fulfilment of the requirements as a part of the

AI/ML INTERNSHIP

(NASTECH)

For the award of degree of

Bachelor of Engineering
in
Information Science and Engineering

Submitted by

Karthik Raj R — **1RN19IS068**

M Balakrishna Kamath — **1RN19IS074**

Internship Project Coordinators

Dr. R Rajkumar
Associate Professor
Dept. of ISE, RNSIT

Mr. T Bhagavath Singh
Assistant Professor
Dept . of ISE, RNSIT

Department of Information Science and Engineering

RNS Institute of Technology

Channasandra, Dr. Vishnuvardhan Road, RR Nagar Post,
Bengaluru – 560 098

2021 - 2022

RNS Institute of Technology

Channasandra, Dr. Vishnuvardhan Road, RR Nagar Post,

Bengaluru – 560 098

DEPARTMENT OF INFORMATION SCIENCE & ENGINEERING

CERTIFICATE

This is to certify that the project report entitled **HEART FAILURE ANALYSIS** has been successfully completed by **M BALAKRISHNA KAMATH** bearing USN **1RN19IS074** and **KARTHIK RAJ R** bearing USN **1RN19IS068** , presently VI semester students of **RNS Institute of Technology** in partial fulfilment of the requirements as a part of the ***AI/ML Internship (NASTECH)*** for the award of the degree of ***Bachelor of Engineering in Information*** Science and Engineering under Visvesvaraya Technological University, Belagavi during academic year **2021 – 2022** . It is certified that all corrections/suggestions indicated for Internal **Assessment have** been incorporated in the report and deposited in the departmental library. The project report has been approved as it satisfies the academic requirements as a part of Internship.

Dr. R Rajkumar
Coordinator
Associate Professor

Mr. T Bhagavath Singh
Guide
Assistant Professor

Dr. Suresh L
Professor and HoD

External Viva

Name of the Examiners **Signature with date**

1.

2.

ABSTRACT

Heart failure is one of the major cause of mortality in the world today. Heart failure is a condition in which the heart can't pump enough blood to meet the body's needs. Heart failure does not mean that your heart has stopped or is about to stop working. It means that your heart is not able to pump blood the way it should. It can affect one or both sides of the heart.

Prediction of cardiovascular disease is a critical challenge in the field of clinical data analysis. With the advanced development in machine learning (ML), artificial intelligence (AI) and data science, it is proved to be effective in assisting in decision making and predictions from the large quantity of data produced by the healthcare industry. ML approaches has brought lot of improvements and broadens the study in medical field which recognizes patterns in the human body by using various algorithms and correlation techniques.

There are many features/factors that lead to heart failure like age, blood pressure, sodium creatinine, ejection fraction etc. In this paper we propose a method to finding important features by applying machine learning techniques. The work is to design and develop prediction of heart disease by feature ranking machine learning. Hence ML has huge impact in saving lives and helping the doctors, widening the scope of research in actionable insights, drive complex decisions and to create innovative products for businesses to achieve key goals.

ACKNOWLEDGMENT

The fulfilment and rapture that go with the fruitful finishing of any assignment would be inadequate without the specifying the people who made it conceivable, whose steady direction and support delegated the endeavours with success.

We would like to profoundly thank **Management** of **RNS Institute of Technology** for providing such a healthy environment to carry out this AI/ML Internship Project.

We would like to express our thanks to our Principal **Dr. M K Venkatesha** for his support and inspired us towards the attainment of knowledge.

We wish to place on record our words of gratitude to **Dr. Suresh L,** Professor and Head of the Department, Information Science and Engineering, for being the enzyme and master mind behind our Internship Project Work.

We like to express our profound and cordial gratitude to my Internship Project Coordinators**, Dr. R Rajkumar,** Associate Professor, Department of Information Science and Engineering for their valuable guidance, constructive comments, continuous encouragement throughout the Project Work and guidance in preparing report.

We would like to thank all other teaching and non-teaching staff of Information Science & Engineering who have directly or indirectly helped us to carry out the Project Work.

Also, we would like to acknowledge and thank our parents who are source of inspiration and instrumental in carrying out this Project Work.

KARTHIK RAJ R	**M BALKRISHNA KAMATH**
USN:1RN19IS068	**USN:1RN19IS074**

TABLE OF CONTENTS

LIST OF FIGURES

Chapter 1

INTRODUCTION

1.1 ORGANIZATION/INDUSTRY

1.1.1 COMPANY PROFILE

NASTECH is formed with the purpose of bridging the gap between Academia and Industry Nastech is one of the leading Global Certification and Training service providers for technical and management programs for educational institutions. We collaborate with educational institutes to understand their requirements and form a strategy in consultation with all stakeholders to fulfil those by skilling, reskilling and upskilling the students and faculties on new age skills and technologies.

1.1.2 DOMAIN/TECHNOLOGY

The domain chosen for our project is AI/ML. Machine learning, the fundamental driver of AI, is possible through algorithms that can learn themselves from data and identify patterns to make predictions and achieve your predefined goals, rather than blindly following detailed programmed instructions, like in traditional computer programming. This technology allows the machine to perceive, learn, reason and communicate through observation of data, like a child that grows up and acquires knowledge from examples. Machines also have the advantage of not being limited by our inherent biological limitations. With machine learning, manufacturing companies have increased production capacity up to 20%, while lowering material consumption rates by 4%.

Nowadays, the revolutionary AI technology evolved from rule-based expert systems to machine learning and more advanced subcomponents such as deep learning (learning representations instead of tasks), artificial neural networks (inspired by animal brains) and reinforcement learning (virtual agents rewarded if they made good decisions).

The AI can master the complexity of the intertwining industrial processes to enhance the whole flow of production instead of isolated processes. This enormous cognitive capacity gives the AI the ability to consider the spatial organization of plants and the timing constraints of live production. Another key advantage is the capability of AI algorithms to think probabilistically, with all the subtlety this allows in edge cases, instead of traditional rule based methods that require rigid theories and a full comprehension of problems.

1.1.3 Department

R.N.Shetty Institute of Technology (RNSIT) established in the year 2001, is the brain-child of the Group Chairman, Dr. R. N. Shetty. The Murudeshwar Group of Companies headed by Sri. R. N. Shetty is a leading player in many industries viz construction, manufacturing, hotel, automobile, power & IT services and education. The group has contributed significantly to the field of education. A number of educational institutions are run by the

R. N. Shetty Trust, RNSIT being one amongst them. With a continuous desire to provide quality education to the society, the group has established RNSIT, an institution to nourish and produce the best of engineering talents in the country. RNSIT is one of the best and top accredited engineering colleges in Bengaluru.

1.2 PROBLEM STATEMENT

1.2.1 Existing System and their Limitations

A manual method is currently used in the market to analyse the cardiovascular condition of a patient. These methods include blood test, ECG and many more. These methods are both time consuming and costly.

1.2.2 Proposed Solution

In order to analyse the heart condition of the patient we take various parameters into consideration and prediction the patient’s heart condition based on ML models. Hence it is cost effective and almost perfectly accurate.

1.2.3 Program formulation

The proposed project will help in studying and analyzing various models to predict if a person is prone to heart failure or not based on his or her medical report.
The analysis has been done with the help of multiple ML algorithms namely Logistic Regression, K-nearest neighbors, Random forest Regression, Decision tree, SVC to compare and choose the optimum result.

Chapter 2

REQUIREMENT ANALYSIS, TOOLS &TECHNOLOGIES

2.1 Hardware and Software Requirements

2.1.1 Hardware Requirements:

- Processor: i5/ryzen 5 or above
- RAM: 512Mb or more
- Hard Disk: 4GB or more

2.1.2 Software Requirements:

- Operating System: Windows 7 or above
- IDE: Google Colab or Jupyter Notebook

2.2 Tools/Languages/Platforms

- Python
- Pandas
- Sklearn
- Streamlit
- Matplotlib
- Seaborn
- Numpy

CHAPTER 3

3.1 Architecture of classification models

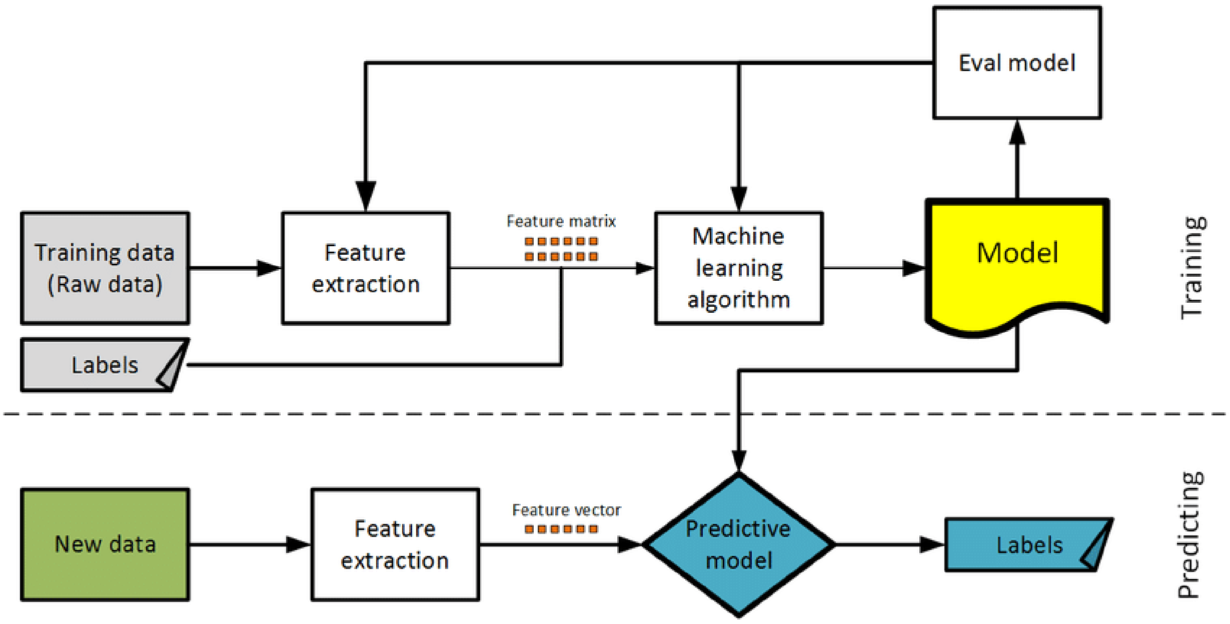


Figure 3.1 Architecture of classification model

The Classification algorithm is a Supervised Learning technique that is used to identify the category of new observations on the basis of training data. In Classification, a program learns from the given dataset or observations and then classifies new observation into a number of classes or groups. Such as, **Yes or No, 0 or 1, Spam or Not Spam, cat or dog,** etc. Classes can be called as targets/labels or categories.

Unlike regression, the output variable of Classification is a category, not a value, such as "Green or Blue", "fruit or animal", etc. Since the Classification algorithm is a Supervised learning technique, hence it takes labelled input data, which means it contains input with the corresponding output.

The main goal of the Classification algorithm is to identify the category of a given dataset, and these algorithms are mainly used to predict the output for the categorical data.

3.2 Problem Statement

- A manual method is currently used in the market to analyse the cardiovascular condition of a patient. The main objective is to study and analyze various models to predict if a person is prone to heart failure or not based on his or her medical report.

- A better prediction for this disease is one of the key approaches of decreasing its impact. Both linear and machine learning models are used to predict heart failure based on various data as inputs, e.g., clinical features.

- The following features have been used for analysis:

1. Age : age of the patient [years]
2. Anaemia : Decrease of red blood cells or hemoglobin (boolean)
3. creatine_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_creatine : 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)

3.3 Algorithm

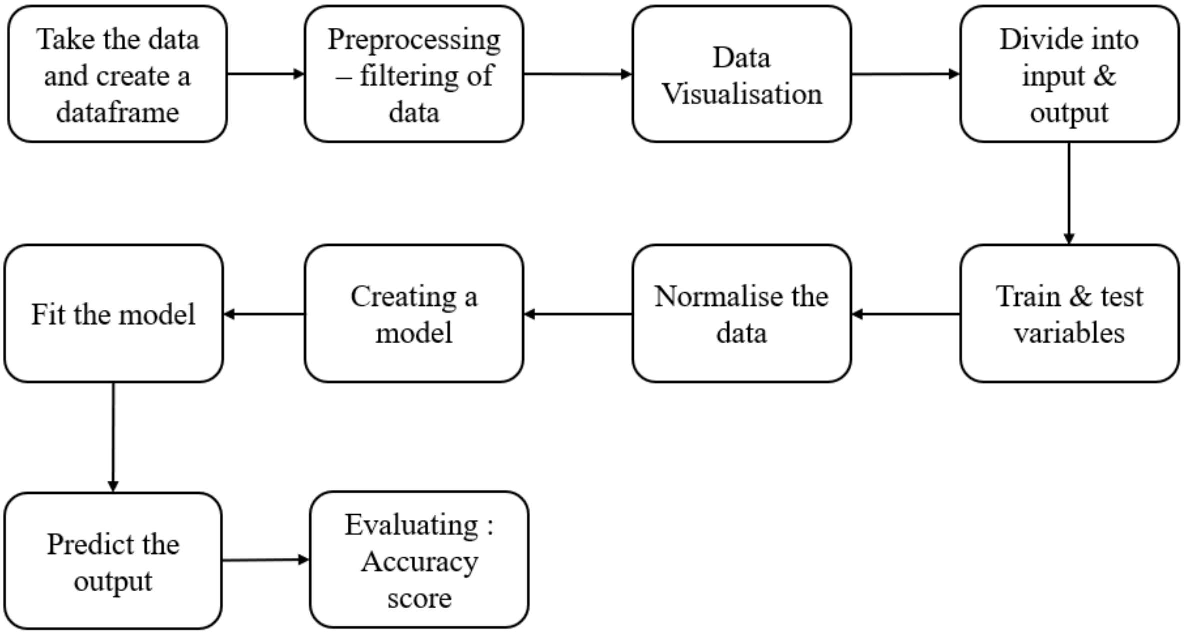

Take the data and create a dataframe
Preprocessing – filtering of data
Data Visualisation
Divide into input & output
Train & test variables
Normalise the data
Creating a model
Fit the model
Predict the output
Evaluating : Accuracy score

Figure 3.2 Steps to build ML models

1. Retrieve the data and create a data frame.
2. Filtering the data – Data cleaning, encoding, dropping values, missing values.
3. Data Visualisation – Plotting different graphs to compare the different fields in the dataset.
4. Divide the dataset into inputs and outputs by selecting the required columns.
5. Splitting the dataset into train and test set.
6. Normalise (Scaling) the input data (if required).
7. Run the classifier/regressor.
8. Fitting the model (map inputs with output).
9. Predict the output.
10. Evaluation: Accuracy score (shows how accurate the result of the model is).

3.3.1 Classification models

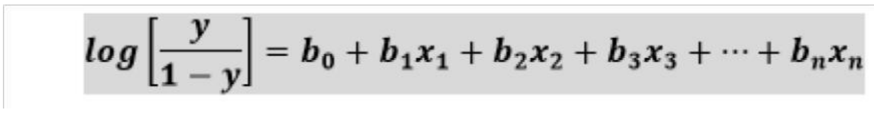
1. Logistic Regression

Logistic regression is one of the most popular Machine Learning algorithms, which comes under the Supervised Learning technique. It is used for predicting the categorical dependent variable using a given set of independent variables.

Logistic regression predicts the output of a categorical dependent variable. Therefore the outcome must be a categorical or discrete value. It can be either Yes or No, 0 or 1, true or False.

The sigmoid function is a mathematical function used to map the predicted values to probabilities.

It maps any real value into another value within a range of
0 and 1 The formula for logistic regression is as follows



$$log\left[\frac{y}{1-y}\right] = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + \cdots + b_nx_n$$

Figure 3.3 Logistic regression formula

2. K-Nearest Neighbours

K-NN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories. K-NN algorithm stores all the available data and classifies a new data point based on the similarity.

Suppose there are two categories, i.e., Category A and Category B, and we have a new data point x1, so this data point will lie in which of these categories.

To solve this type of problem, we need a K-NN algorithm. With the help of K-NN, we can easily identify the category or class of a particular dataset.

Consider the below diagram:

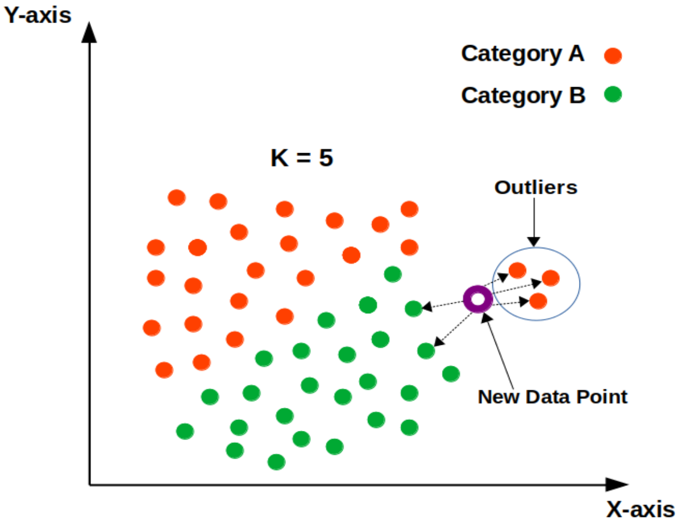


Figure 3.4 KNN

3. Support Vector Machine

The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane.

SVM chooses the extreme points/vectors that help in creating the hyperplane. These extreme cases are called as support vectors, and hence algorithm is termed as Support Vector Machine. Consider the below diagram in which there are two different categories that are classified using a decision boundary or hyperplane.

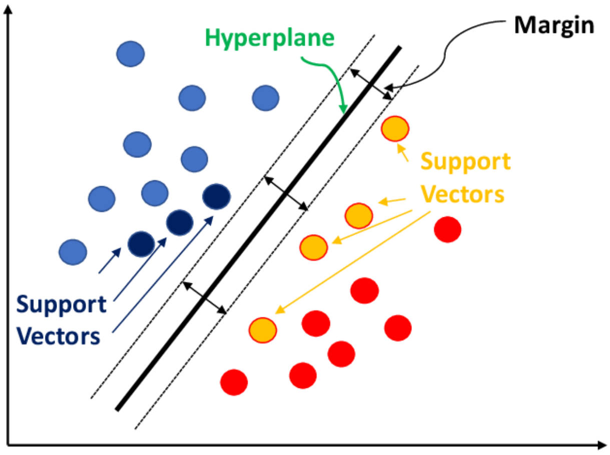


Figure 3.5 SVM

4. Decision Tree Classifier

It is a tree-structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome.

In a Decision tree, there are two nodes, which are the Decision Node and Leaf Node. Decision nodes are used to make any decision and have multiple branches, whereas Leaf nodes are the output of those decisions and do not contain any further branches.

The decisions or the test are performed on the basis of features of the given dataset. It is a graphical representation for getting all the possible solutions to a problem/decision based on given conditions.

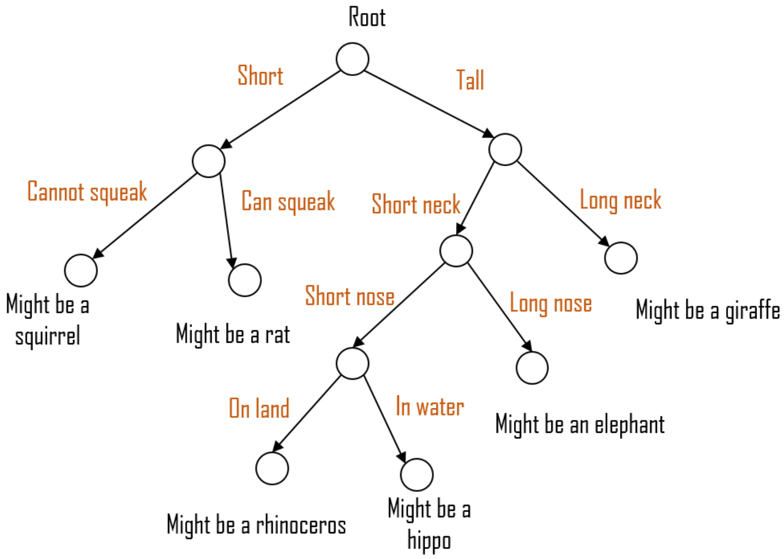


Figure 3.6 Decision Tree

5. Random Forest

It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model.

As the name suggests, "Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset." Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output.

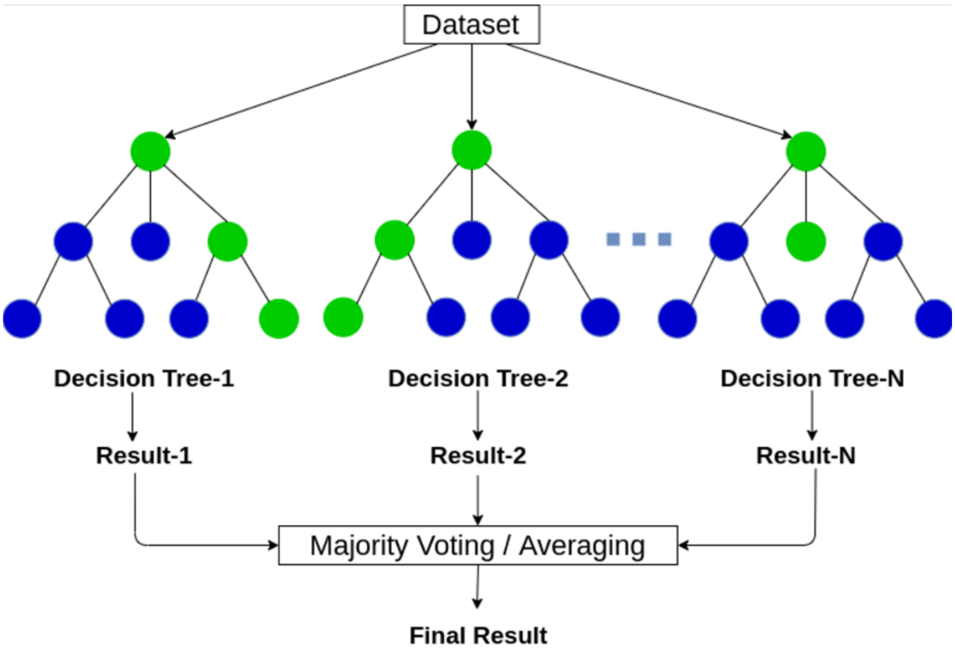


Figure 3.7 Random Forest

Libraries

- Python
- Pandas
- Sklearn
- Streamlit
- Matplotlib
- Seaborn
- Numpy

3.4 IMPLEMENTATION

- **Creation of data frame**

```
dataset = pd.read_csv('/content/heart_failure_clinical_records_dataset.csv’)
```

- **Pre-processing – filtering of data**

```
dataset = dataset[dataset['ejection_fraction']<70]
```

- **Data visualization**

```
import plotly.express as px
fig = px.histogram(dataset, x="ejection_fraction", color="DEATH_EVENT",
marginal="rug", hover_data=dataset.columns,
title ="Distribution of EJECTION FRACTION vs DEATH_EVENT",
labels={"ejection_fraction": "EJECTION FRACTION"},
template="plotly_dark",
color_discrete_map={"0": "RebeccaPurple", "1": "MediumPurple"})
fig.show()
```

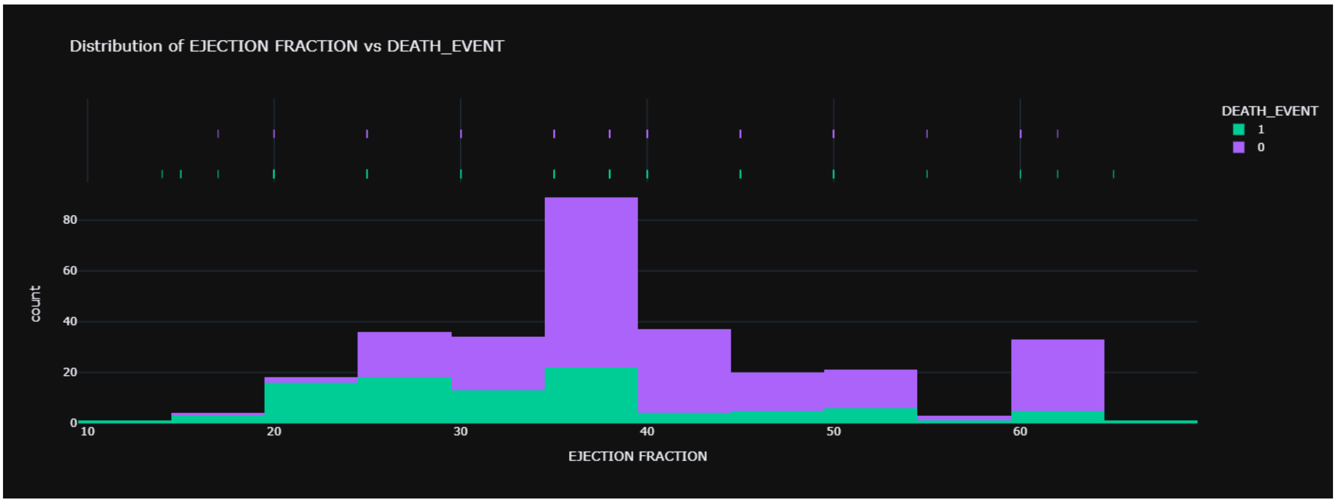


Figure 3.8 Ejection Fraction vs Death_Event

PIE CHART

```
import plotly.graph_objects as go
from plotly.subplots import make_subplots
d1 = dataset[(dataset["DEATH_EVENT"]==0) &
(dataset["high_blood_pressure"]==0)]
d2 = dataset[(dataset["DEATH_EVENT"]==1) &
(dataset["high_blood_pressure"]==0)]
```

```
d3 = dataset[(dataset["DEATH_EVENT"]==0) &
(dataset["high_blood_pressure"]==1)]
d4 = dataset[(dataset["DEATH_EVENT"]==1) &
(dataset["high_blood_pressure"]==1)]
label1 = ["No High BP","High BP"]
label2 = ['No High BP - Survived','No High BP - Died',
"High BP –  Survived", "High BP  - Died"]
values1 = [(len(d1)+len(d2)), (len(d3)+len(d4))]
values2 = [len(d1),len(d2),len(d3),len(d4)]
```

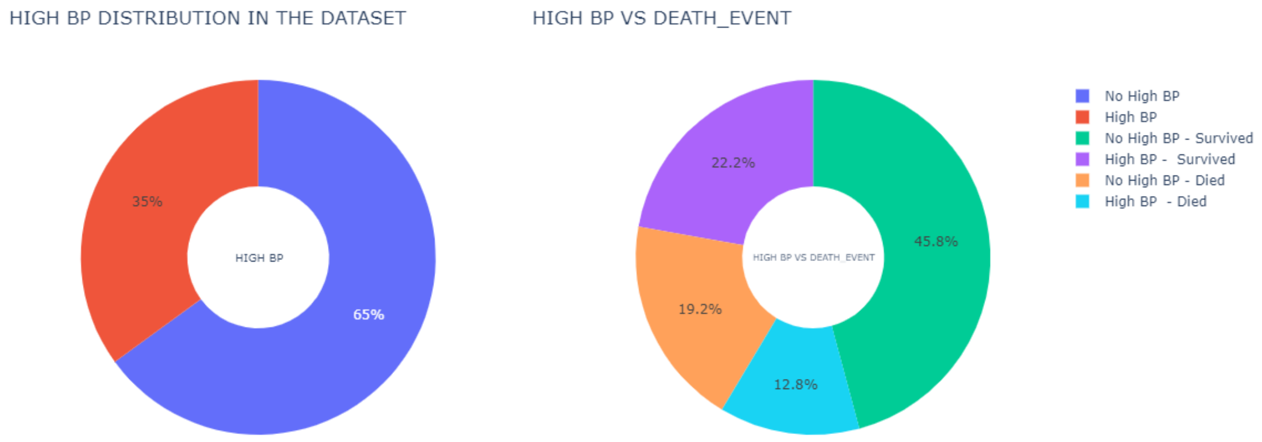


Figure 3.9 High BP vs Death_Event

❖ Divide into input and output

```
x = dataset.iloc[:, [4,7,11]].values
y = dataset.iloc[:,-1].values
```

❖ Splitting the dataset into train and test sets

```
from sklearn.model_selection import train_test_split
x_train, x_test, y_train, y_test = train_test_split(x, y, test_size = 0.2,
random_state =0)
```

❖ Normalising the inputs (Scaling for KNN & SVC)

```
from sklearn.preprocessing import StandardScaler
sc = StandardScaler()
X_train1 = sc.fit_transform(x_train)
X_test1 = sc.transform(x_test)
```

❖ Creating the models

```
clf = LogisticRegression()
clf = KNeighborsClassifier(n_neighbors=params['K'], metric='minkowski')
clf = SVC(C=params['C'],random_state=0, kernel = 'rbf')
clf = DecisionTreeClassifier(max_leaf_nodes = params['M'],
      random_state=0, criterion='entropy')
clf = RandomForestClassifier(n_estimators = params['E'],
      criterion='entropy',random_state=0)
```

❖ Fit the model

```
clf.fit(X_train, y_train)    //Logistic, Decision Tree, Random Forest
clf.fit(X_train1, y_train)   //KNN & SVC
```

❖ Predict the output

```
y_pred=clf.predict(x_test)   //Logistic, Decision Tree, Random Forest
y_pred=clf.predict(x_test1)  //KNN & SVC
print(y_pred)
```

❖ Evaluation : Accuracy Score of all models

```
res=[]
for all models:
        acc = accuracy_score(y_test, y_pred)
        res.append(acc)
```

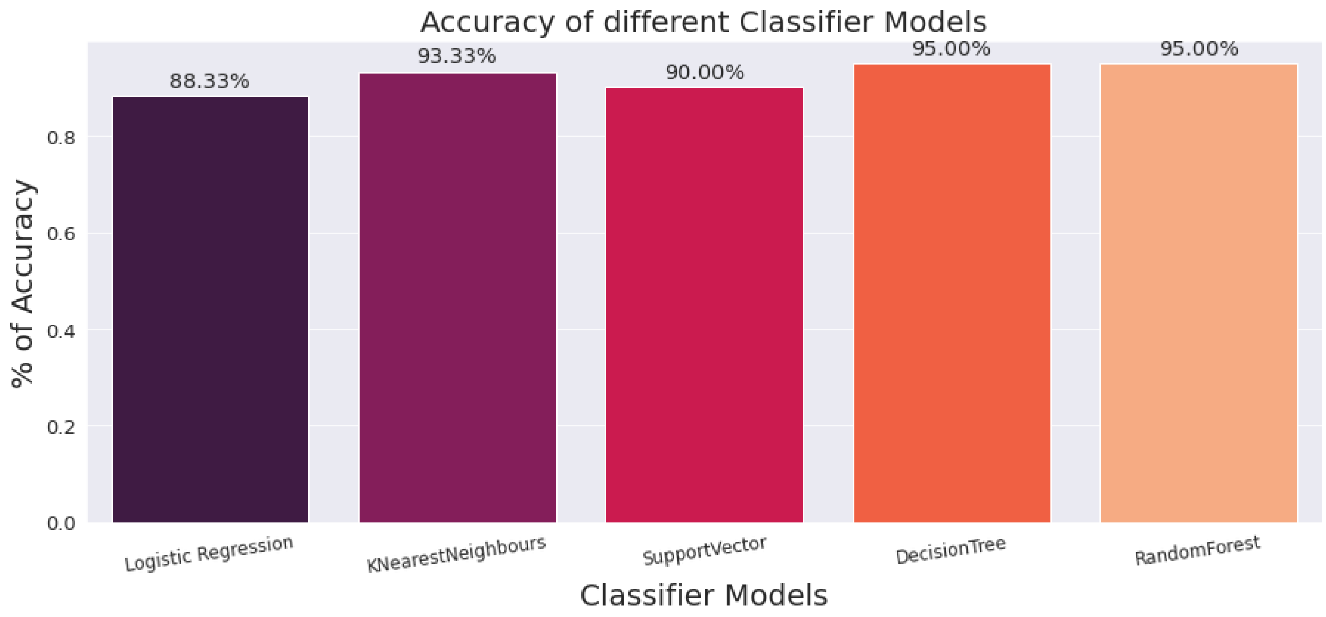


Figure 3.10 Accuracy Graph

Chapter 4

RESULTS & SNAPSHOTS

- The below conclusions can be made from the accuracy score graph:

  The accuracy of Logistic Regression is 88.33%.

  The accuracy of K-Nearest Neighbours is 93.33%.

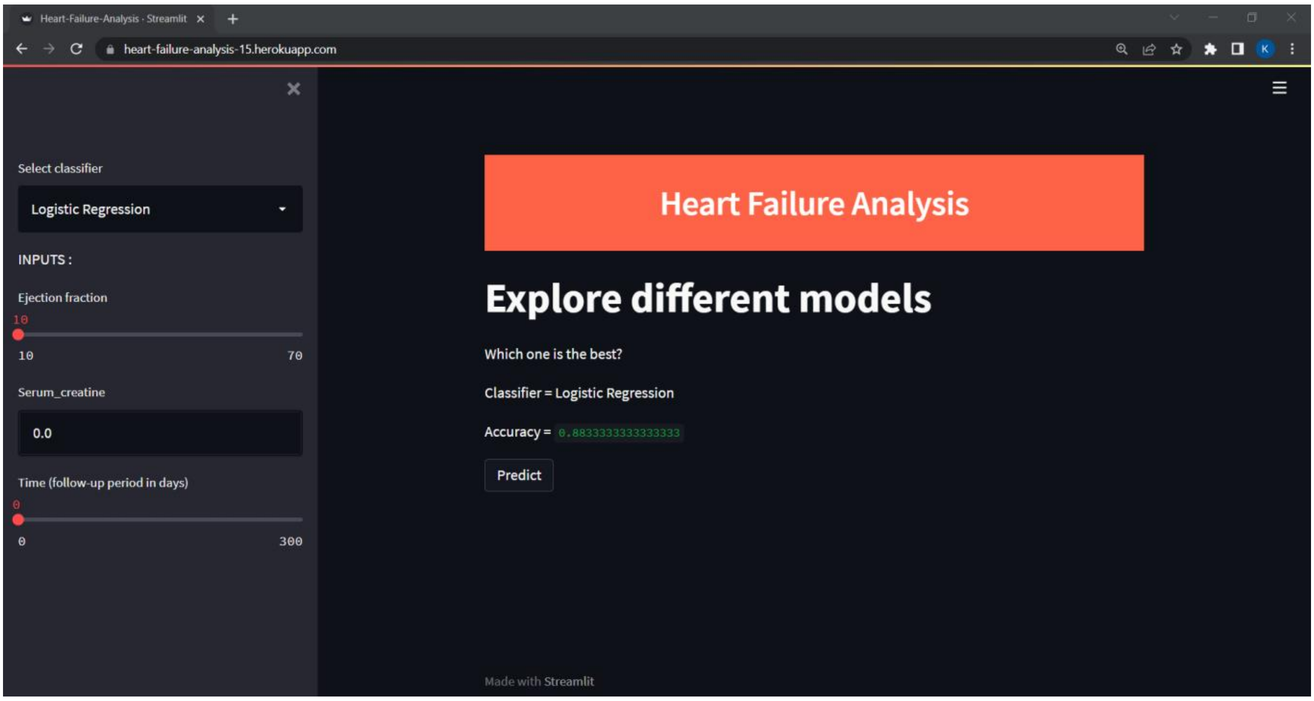
  The accuracy of Support Vector Classifier is 90%.

  The accuracy of Decision Tree is 95%.

  The accuracy of Random Forest Regression is 95%.

- It can be observed that Random forest and Decision tree gives the best accuracy score amongst all the tested models.

- The UI implemented using Streamlit is shown in the diagrams below.



Heart-Failure-Analysis · Streamlit

heart-failure-analysis-15.herokuapp.com

Select classifier

Logistic Regression

INPUTS :

Ejection fraction

10

10 70

Serum_creatine

0.0

Time (follow-up period in days)

0

0 300

Heart Failure Analysis

Explore different models

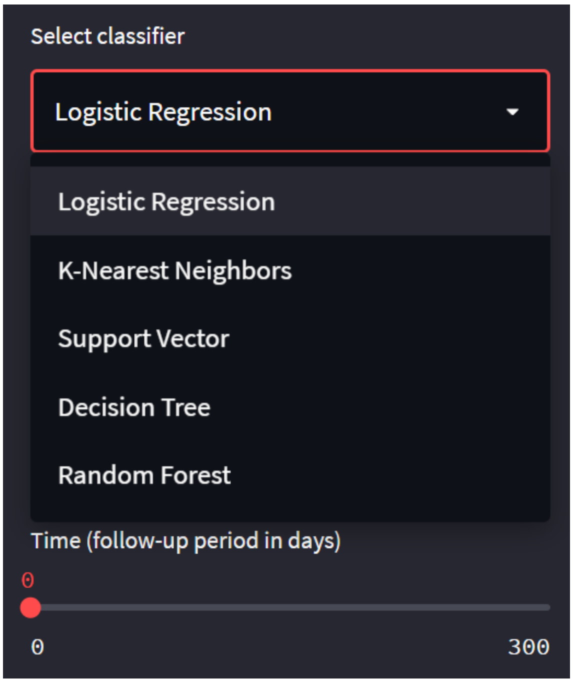
Which one is the best?

Classifier = Logistic Regression

Accuracy = 0.8833333333333333

Figure 4.1 Home Page

This is the landing page consisting of drop down list to select the models and parameters of it.



Select classifier

Logistic Regression

Time (follow-up period in days)

0

Figure 4.2 Drop-list

- Select classifier provides a drop down list from which we can choose a desired model to predict the likeliness of a heart failure occurring during the follow-up period.
- Based the selected model we get the appropriate input parameters to provide for the prediction.
- We get the accuracy score of the selected model and its prediction (deceased or not deceased).

ACCURACY GRAPH

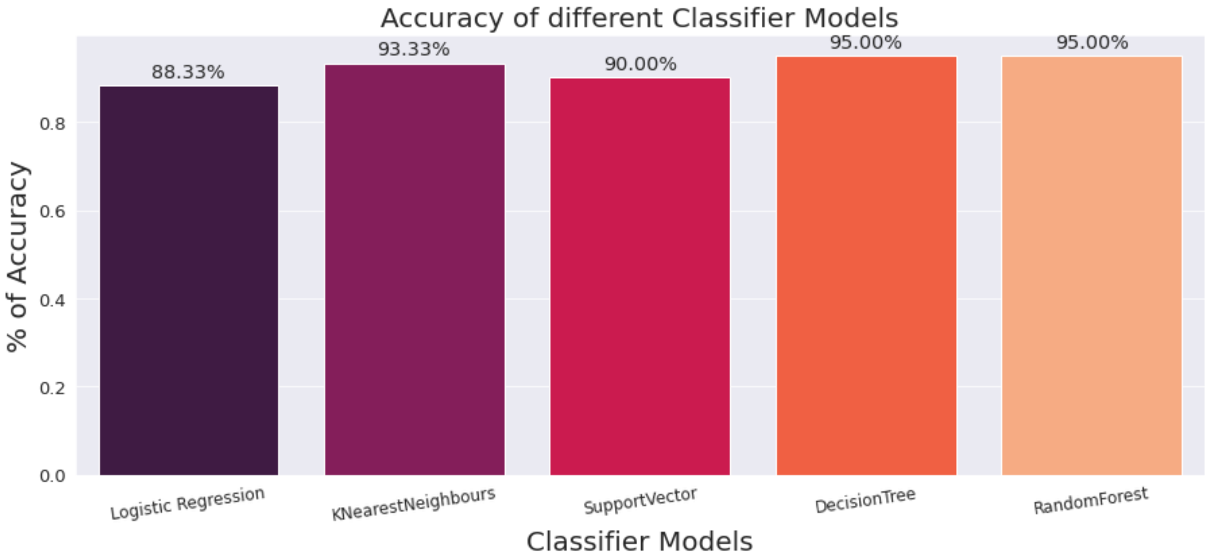
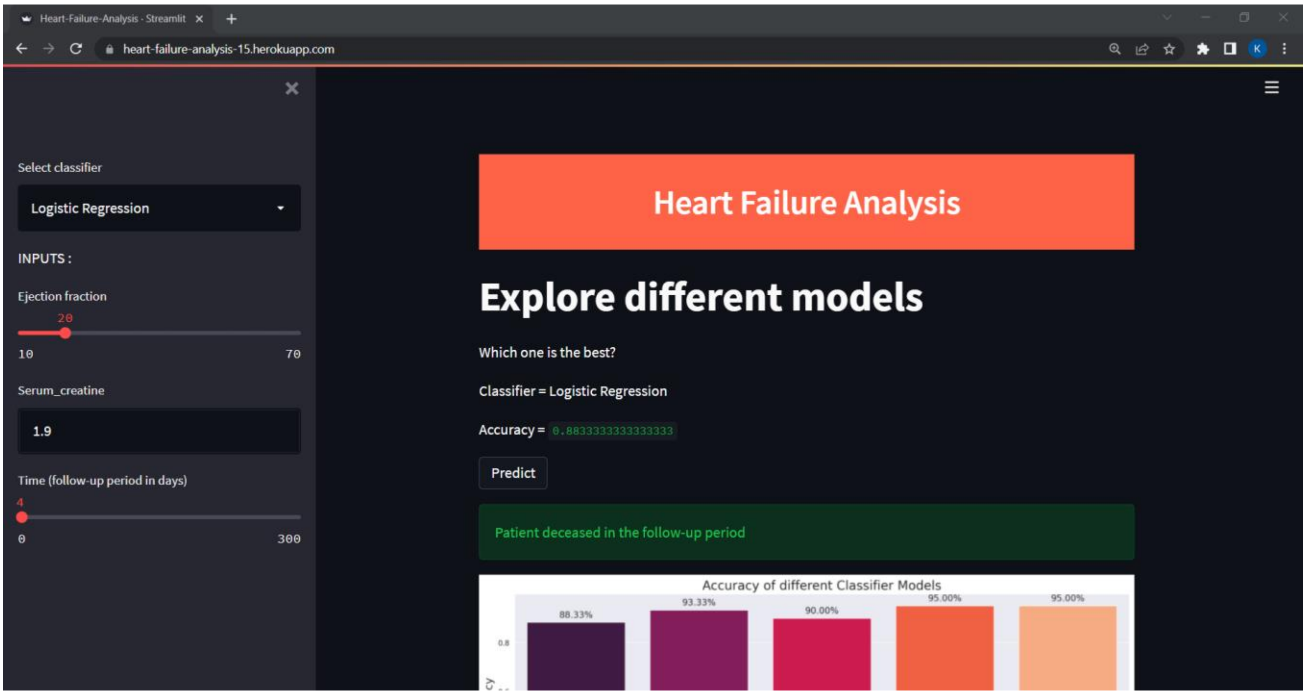


Figure 4.3 Accuracy Graph

1. LOGISTIC REGRESSION


Heart-Failure-Analysis · Streamlit
heart-failure-analysis-15.herokuapp.com
Predict
Accuracy of different Classifier Models
88.33%
93.33%
90.00%
95.00%
95.00%
0.8

Select classifier

Logistic Regression

INPUTS :

Ejection fraction: 20 (10–70)

Serum_creatine: 1.9

Time (follow-up period in days): 4 (0–300)

Heart Failure Analysis

Explore different models

Which one is the best?

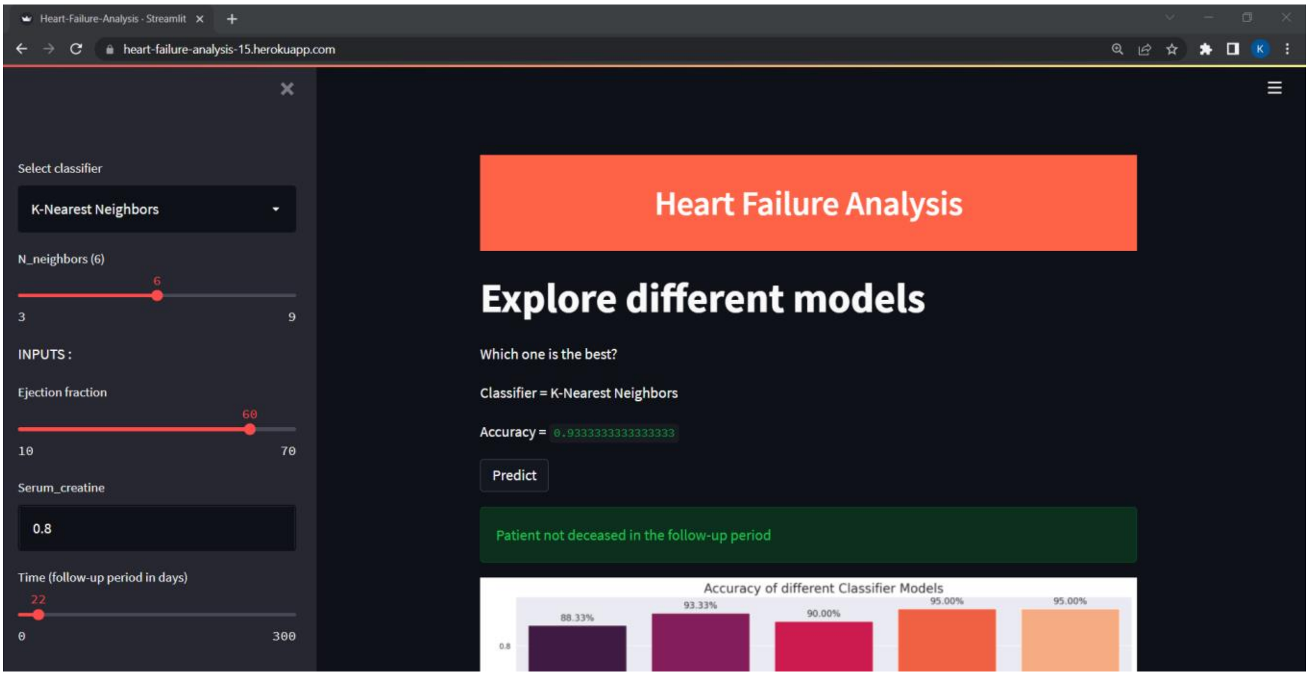
Classifier = Logistic Regression

Accuracy = `0.8833333333333333`

Patient deceased in the follow-up period

Figure 4.4 Logistic Regression UI

2. K-NEAREST NEIGHBOURS


Heart-Failure-Analysis · Streamlit
heart-failure-analysis-15.herokuapp.com
Predict
Accuracy of different Classifier Models
88.33%
93.33%
90.00%
95.00%
95.00%
0.8

Select classifier

K-Nearest Neighbors

N_neighbors (6): 6 (3–9)

INPUTS :

Ejection fraction: 60 (10–70)

Serum_creatine: 0.8

Time (follow-up period in days): 22 (0–300)

Heart Failure Analysis

Explore different models

Which one is the best?

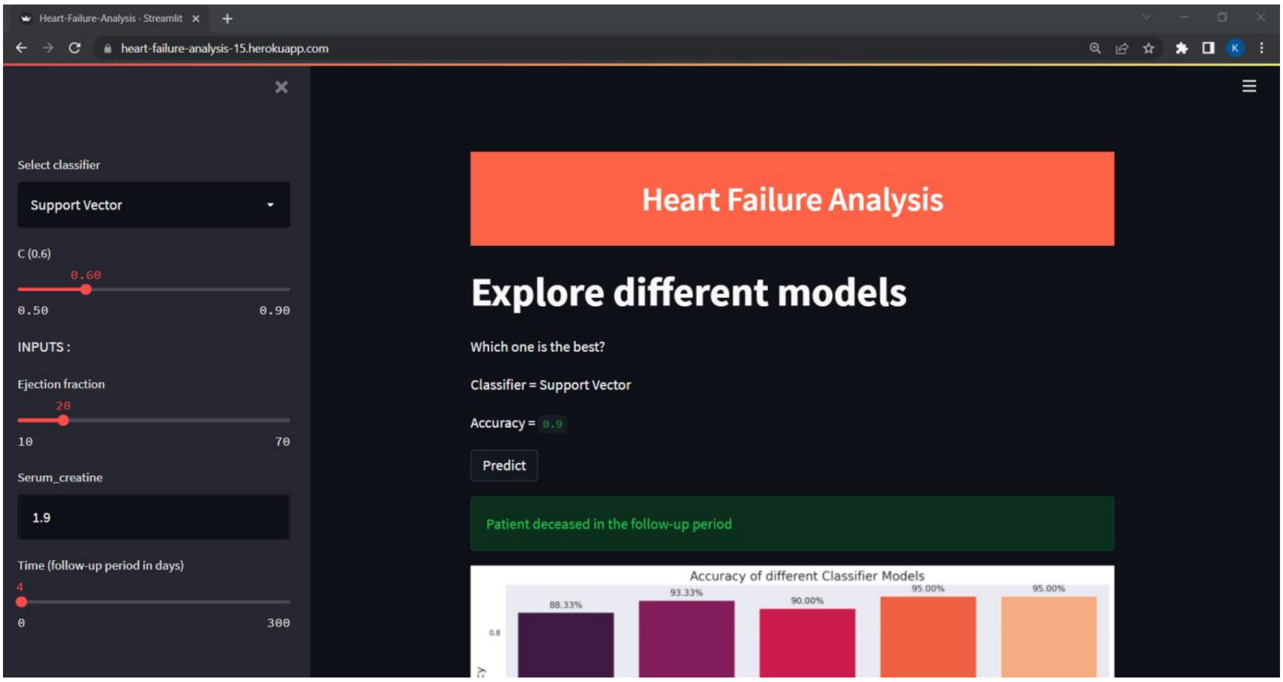
Classifier = K-Nearest Neighbors

Accuracy = `0.9333333333333333`

Patient not deceased in the follow-up period

Figure 4.5 KNN UI

3. SUPPORT VECTOR


Heart-Failure-Analysis · Streamlit
heart-failure-analysis-15.herokuapp.com
Accuracy of different Classifier Models
88.33%
93.33%
90.00%
95.00%
95.00%
0.8

Select classifier

Support Vector

C (0.6)

0.60

0.50 0.90

INPUTS :

Ejection fraction

20

10 70

Serum_creatine

1.9

Time (follow-up period in days)

4

0 300

Heart Failure Analysis

Explore different models

Which one is the best?

Classifier = Support Vector

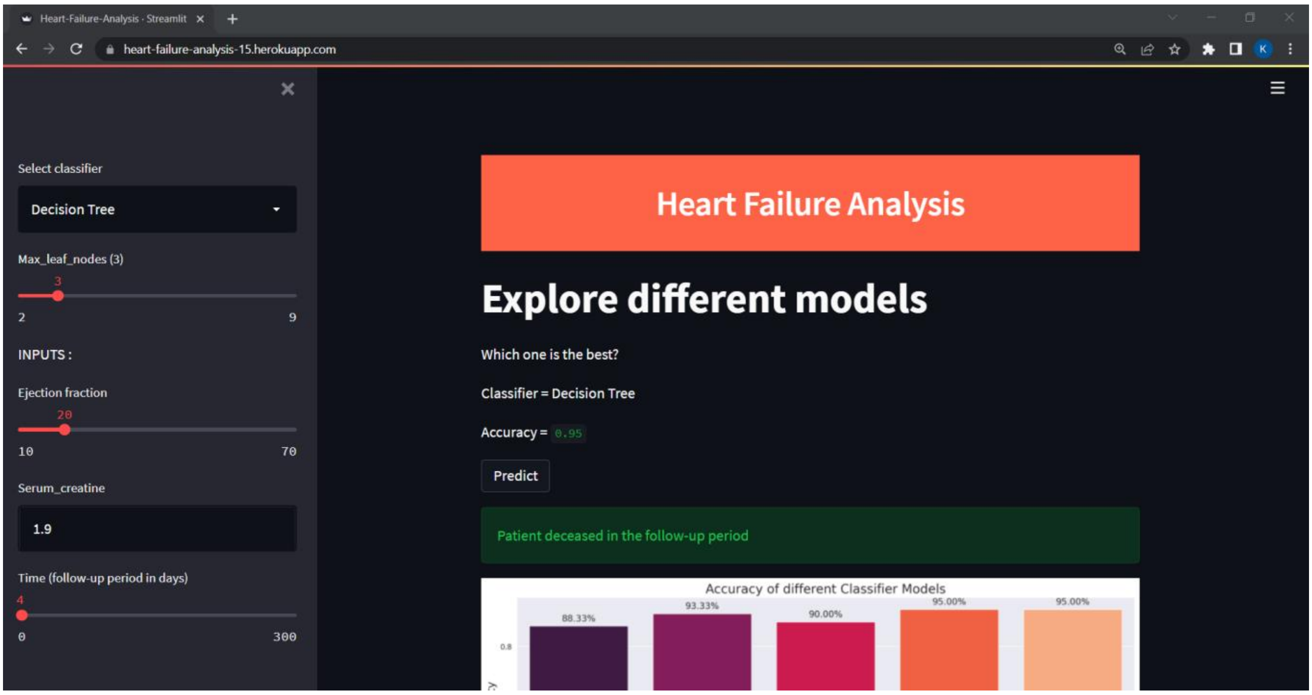
Accuracy = 0.9

Predict

Patient deceased in the follow-up period

Figure 4.6 SVM UI

4. DECISION TREE


Heart-Failure-Analysis · Streamlit
heart-failure-analysis-15.herokuapp.com
Accuracy of different Classifier Models
88.33%
93.33%
90.00%
95.00%
95.00%
0.8

Select classifier

Decision Tree

Max_leaf_nodes (3)

3

2 9

INPUTS :

Ejection fraction

20

10 70

Serum_creatine

1.9

Time (follow-up period in days)

4

0 300

Heart Failure Analysis

Explore different models

Which one is the best?

Classifier = Decision Tree

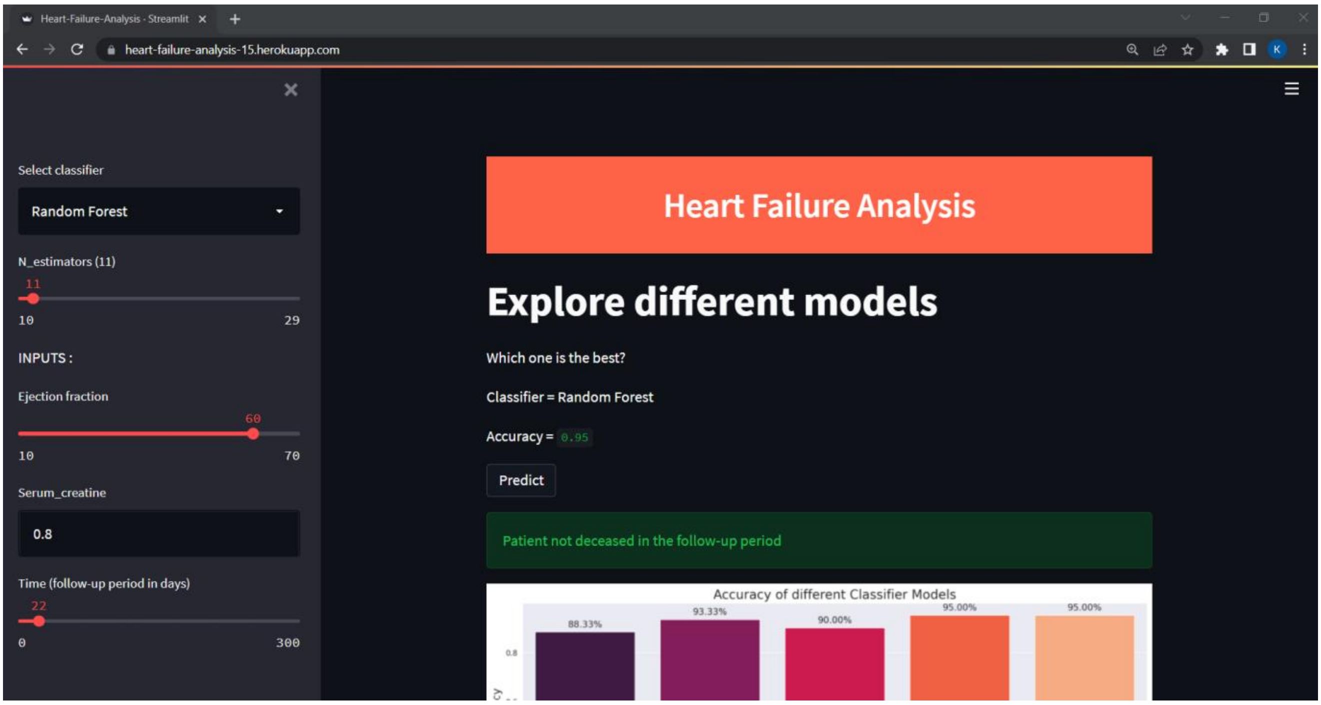
Accuracy = 0.95

Predict

Patient deceased in the follow-up period

Figure 4.7 Decision Tree UI

5. RANDOM FOREST


Heart-Failure-Analysis · Streamlit
heart-failure-analysis-15.herokuapp.com
Accuracy of different Classifier Models
88.33%
93.33%
90.00%
95.00%
95.00%
0.8

Select classifier

Random Forest

N_estimators (11)

11

10 29

INPUTS :

Ejection fraction

60

10 70

Serum_creatine

0.8

Time (follow-up period in days)

22

0 300

Heart Failure Analysis

Explore different models

Which one is the best?

Classifier = Random Forest

Accuracy = 0.95

Predict

Patient not deceased in the follow-up period

Figure 4.8 Random Forest UI

Chapter 5

CONCLUSION AND FUTURE ENHANCEMENT

5.1 Conclusion

- Given the present scenario where heart failures are becoming common, the project was an attempt to create a prediction system for the same.
- To increase the exactness of the result, different models have been used for thorough analysis and prediction.
- The system has 2 models that give an accuracy score of 95% but the other models give a considerable score as well.
- Medical science will be helped with similar analysis or prediction systems that can help save lives.

5.2 Future Enhancement

- An improved analysis system can be made after a thorough research in medical science wherein the most important factors can be considered to predict the chances of a heart failure occuring.
- An improved system can be made by applying same analysis to other cardiac diseases datasets and the performance of these classifiers to classify and predict these diseases.
- An future work, considering the more attributes as input data can expand this study. Furthermore, this could be done on early detection of heart failure by processing family’s historical data.
- There are many possible improvements that could be explored to improve the scalability and accuracy of this prediction system.

Chapter 6

REFERENCE

[1] Chicco, Davide, and Giuseppe Jurman. "Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone." BMC medical informatics and decision making ..

[2] https://www.kaggle.com/datasets

[3] https://docs.streamlit.io/

[4] https://scikit-learn.org/stable/user_guide.html

[5] https://matplotlib.org/stable/users/getting_started/