



Heart Failure prediction using machine learning

Submitted By
Farjana Akter Papri
Student ID: 171-35-216

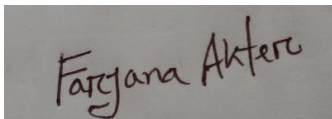
Submission Date: Date/Month/Year

Department Of Software Engineering
Daffodil International University

Declaration

I declare that thesis paper title "is a unique proposition done by me under the supervisor of Bikash Kumar Paul Lecturer of Mawlana bhashani Science and Technology University. towards the fractional satisfaction of the prerequisite for the honor of the level of Bachelor of Science in Software Engineer during the time of 2017-2021. I additionally express that this has not been submitted in some other spot.

Submitted by:



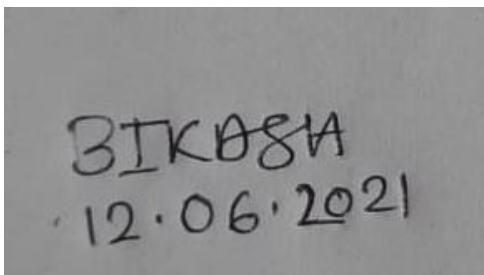
Farjana Akter Papri

ID: 171-35-216

Department of Software Engineer

Daffodil International University

Certified by:



Bikash Kumar Paul

Lecturer

Mawlana Bhashani Science and Technology

Acknowledgement

To turn into a successful person, we must be persevering and have an enthusiastic person at work. That is the reason I generally attempt to my best level. For finishing my graduation, I accept the proposal is the best answer for increment my ability. Since the thesis is an extension of theoretical and practical work.

First of all, I would like to thank my ALLAH who always guide me for work on the right path of the life. I must thankful to my parents and my family for giving me the opportunity and always be to myself.

I'm uncommonly committed to the deferential instructor and uniquely to my supervisor "Bikash Kumar Paul" for giving me essential data for finishing my thesis.

I'm appreciative to my specialization staff part, Lab experts, and non-showing staff part for their limited help all through my undertaking.

And finally I would like to express my love to my batch mate for their co-operation, consolation and motivation which help me to finish my thesis and project.

Table of content

THESIS DECLEARATION:	ii
ACKNOWLEDGEMENT:	iii
TABLE OF CONTENT:	iv
LIST OF TABLE:	v
LIST OF FIGURE:	v
LIST OF ABBREVIATION:	v
ABSTRACT:	vi
CHAPTER 01: INTRODUCTION:	1
1.1 Background:	1
1.2 Motivation of the Research:	2
1.3 Problem Statement:	2
1.4 Research Question:	3
1.5 Research Objectives:	3
1.6 Research Scope:	3
1.7 Thesis Organization:	3
CHAPTER 02: DataSET.....	4
2.1 Introduction of dataset:	4
CHAPTER03: METHODOLOGY.....	5
3.1 process of Description:	5
3.2Random Forest Algorithm :	7
3.3Decision Tree Algorithm:	7
3.4 Logistic Regression Algorithm:	8
3.5 GaussianNB :.....	9
3.6 Model Performance	9
3.7 Evaluation of classifier algorithm	9
CHAPTER 04: RESULT AND DISCUSSION.....	10
4.1 Summary:	10
4.2 Confusion Matrix:	11
4.3 Classification report:	11
4.4 Correlation:	12
CHAPTER 05: CONCLUSION.....	13
5.1 Findings and contribution:	13
5.2 Recommendation for Future works:	13
REFERENCES:	14

List of Table

Table 01: dataset description

Table 02: confusion matrix

Table 03: classification report

List of Figure

Figure 01: Work flow of heart failure prediction

Figure 02: Decision tree

Figure 4.3: correlation

List of ABBREVIATION

HF = Heart Failure

SNPs, = single-nucleotide polymorphisms

TP = True Positive

TN = True Negative

FP = False Positive

FN = False Negative

RF = Random Forest

DT = Decision Tree

LR = Logistic Regression

GNB = Gaussian Naive Bayes

CR = Classification Report

CM = Confusion Matrix

Abstract

Introduction: Heart failure kill around 14-15 million individuals universally consistently, and they for the most part show as myocardial areas of localized necrosis and heart failure. HF happens when a heart can't siphon sufficient the blood to address the issues in body. Nowadays HF patients are so common and have lots of reason which causes HF. For research here used clinical information of HF patients and tried to get the best results. Every patient has all medical records in this dataset. this is the process that declares prediction of heart failure.

Methodology: This research study focuses on that, we take apart a dataset of 299 patients with HF accumulated in 2015. Here utilize a few algorithms as well as machine learning to forecast the patient's perseverance and classify the attributes identifying with the primary a potential risk. After running the model, we get accuracy and then using accuracy we get confusion matrix result and the final step is the classification report.

Results: The current work predicts experiencing a pace of a patient HF using Decision Tree, Random Forest, Logistic Regression, and GaussianNB algorithm calculation. Confusion matrix representing TP, FP, TN, FN value. From the confusion matrix, we get more developed metrics that can assist us in making a classification decision. On the other side classification reports show the accuracy of the model with precision, recall, and F1 result.

Conclusion: This disclosure can possibly affect clinical work, turning into another supporting apparatus for doctors while predicting if heart failure patients will endure or not. We can additionally grow this examination by consolidating other Machine Learning. Thinking about the limitation of this study, there is a need to execute more perplexing and mix of models to get higher exactness for early prediction of heart failure.

Chapter 01: Introduction

1.1 Background:

The primary part of heart failure is a blockage in arteries. It has many different names, for example, cardiovascular illness and blood vessel hypertension. Around, there are just about 25-26 million individuals all throughout the planet influencing by heart illness. The stressing point is, this proportion is (Fahd Saleh Alotaibi1, 2019)required to increment quickly in the coming years if safeguards are not taken proficiently. Aside from making way of life sound and diet control, the opportune time diagnosing and exhaustive analysis are other fundamental components, which can eventually save the lives. Thusly, this paper has made a little stride towards saving the existences of HF patients and depicts an approach to improve the exhibition of diagnosing the patients based on their clinical history. More often than not patients go for a few tests, which can overburden them with extra proactive tasks, time, and for sure extra monetary charges. According to previous research, the most common causes of heart disease are poor diet, tobacco use, excessive sugar, being overweight, or being obese. Fat versus muscle despite the fact that the regular manifestations can be painful, in the upper arms and chest. These reasons are distinct from one another; an appropriate evaluation of this type of reasoning is required. The dataset can help to improve the patient care. Heart specialists are also available. In this paper Machine learning applied to clinical records, specifically, can be a successful device both to anticipate the endurance of every tolerant having Heart Failure side effects, also, to recognize the main clinical highlights (Kazem Rahimi, Derrick Bennett, Nathalie Conrad, Timothy M. Williams, Joyee Basu, Jeremy Dwight, Mark Woodward, Anushka Patel, John McMurray, and Stephen MacMahon, 2014) (risk factors) that may prompt HF. Researchers can exploit machine learning not just for clinical expectation, yet additionally for highlight positioning. Given the significance of an essential organ like the heart, anticipating HF has become a need for clinical specialist doctors. However, to date deciding heart frustration major information in clinical care have attempted to show a high result. Demonstrating endurance for HF is as yet difficult these days. Both as far as accomplishing high forecast exactness and recognizing the driving variables. The majority among the modeling techniques created for this reason arrive at just unobtrusive exactness, with restricted interpretability from the anticipating factors. Later models show enhancements, particularly if the endurance result is combined with extra focuses (for instance, hospitalization). In this specific situation, electronic wellbeing data sets (EHRs, moreover known as clinical datasets) can still be assessed a helpful asset of data to divulge covered up and non-self-evident connections and connections between patients' information, not just for research yet in addition for clinical practice and for exposing customary fantasies on risk factors.

To this point, a few screening considers have been led somewhat recently, covering various circumstances and socioeconomics as well as various information origin to extend the information on the risk ratio. It is important to mention one of them. referencing the PLIC study], (Davide Chicco¹ and Giuseppe Jurman², 2020) where EHRs, blood test, single-nucleotide polymorphisms (SNPs), carotid ultrasound imaging, and metagenomics information have been gathered in a four-visit longitudinal screening all through 15 years a long time in Milan (Italy, EU) to help a superior appraisal of Heart failure disease . This paper discusses, analyze a group of clinical information from patients who have suffered a heart failure. Here have 0-299 patient medical records. Using dataset main target is a prediction of heart failure. recognizing patients who are in danger and do not react to at present suggested treatments for HF may prompt customized medication focused on directed medicines for patients. At long last, Improved forecast danger level could aid in the organization of fundamentals by selecting credits with greater event rates. Several studies on predictive markers for patients with HF (Wouter Ouwerkerk 1, Adriaan A Voors 2, Aeilko H Zwinderman 3, 2014), as well as a few overviews on forecast models, have been done. Most of these models focused on the assumption for HF hospitalization and were found to perform insufficiently or only reasonably on a specific patient people surveyed 4 prognostic models expecting mortality. They assumed that the figure models used were adequate in isolating patients, notwithstanding, they may barely care about the incomparable threat of mortality in old patients. All of these reviews focused on unmistakable assessments to explain the farsighted power of the models.

1.2 Motivation of the Research:

heart failure is a blockage in arteries. It has many different names, for example, cardiovascular illness and blood vessel hypertension. Day by day it will increase and it is danger for human.in this research using different algorithm find out heart failure prediction. Hope that this research will help to detect the risk factors of Heart Failure. The main motivation of the research is to improve the survival rate and predict the death rate as using information.

1.3 Problem Statement:

1.4 Research Question:

1. **Question:** How does a dataset work on disease identification?
2. **Question:** How can this work be effective in this field?
3. **Question:** How to add references?
4. **Question:** How to analyses accuracy

1.5 Research Objectives:

By utilized the patients' medical record get Prediction of heart failure. give different accuracy using different algorithm. Attributes such as age, cpk, blood-pressure, smoking etc are fed into the algorithms which is used to predict risk of heart failure in a person.

1.6 Research Scope:

The analysis of Heart Failure is hard to make in light of the fact that there is no particular test or recognizing lab finding that obviously separates the problem from comparative issues. Along these lines, it is critical to find out how to discover in the beginning phases with the goal that the patients can get by through taking an appropriate treatment ideal.

1.7 Thesis Organization:

The paper hasbeen furnished with five chapters which is described below:

Chapter 1: In this chapter, research background, motivation, problem statement, objectives andresearch scope are given.

Chapter 2: This chapter includes discussion of the dataset.

Chapter 3: This chapter contains the research methodology and approaches as it follows for the Research.

Chapter 4: This chapter is about discussion and result.

Chapter 5: In this chapter declare about contribution and the direction of the future work of the research has also been guided.

CHAPTER 02: Dataset

Introduction to dataset:

Heart Failure clinical records data Set involves the clinical data of 299 patients who suffered from Heart Failure. (Davide Chicco¹ and Giuseppe Jurman², 2020) This data set have 299 rows and 13 columns. The dataset contains 11 clinical highlights (some of them are parallel, others are mathematical), the subsequent period, and the name DEATH_EVENT that shows whether the patient has kicked the bucket. We can discover a few highlights stringently identified with clinical viewpoints like levels of enzymes, sodium, creatinine, and platelets in the blood and others that are more normal like age, sex, or smoking.

The dataset is gathered in 2015 at the Allied Hospital in Faisalabad (Punjab, Pakistan). This dataset collects from Kaggle.

Name	Type	Description
Age	Years	Age of the patient
Anemia	Boolean	Red blood cell or hemoglobin levels drop
creatinine_phosphokinase	mcq/L	CPK catalyst concentration in the blood
Diabetics	Boolean	In the event that the patient is diabetic,
Ejection_fraction	Percentage	The quantity of blood away from the heart is measured.
High_blood_pressure	Boolean	If a patient has high blood pressure,
Platelets	Kiloplatelets/ml	Platelets in the blood
Serum_creatinine	Mg/dL	Creatinine level in the blood
Serum_sodium	mEq/L	Sodium level in the blood
SEx	Binary	Man or women
Smoke	Boolean	If the patient smoker
Time	Days	Follow-up time
DEATH_EVENT	Boolean	During the follow-up period, if the patient died,

CHAPTER 03: METHODOLOGY

3.1 process of Description:

This paper shows the examination of different machine learning algorithms, the algorithm that is utilized in this paper are GaussianNB, Decision Tree, Logistic Regression, and Random Forest Classifiers which can be useful for forecast or clinical experts for accurately analyze Heart Failure. (Harshit Jindal¹, Sarthak Agrawal¹, Rishabh Khera¹, Rachna Jain² and Preeti Nagrath², 2021)The methodology gives a structure for the proposed model. The strategy is a cycle that includes steps that change given data into perceived data patterns for the knowledge of the users. The proposed procedure (Figure 3.1: workflow of heart failure prediction) includes steps, where the initial step is referred to as the collection of the data than in the second stage it separates significant values than the third is the preprocessing stage where we explore the data. After the pre-preparing of data, the classifier is utilized to classify the pre-processed data. The classifier utilized in the proposed model is GaussianNB, Decision Tree, Logistic Regression, and Random Forest Classifier. At last, the proposed model is attempted, where we evaluated our model based on accuracy and performance utilizing different confusion metrics and give classification reports. As a result, an appropriate Heart Disease Prediction is provided in the model.

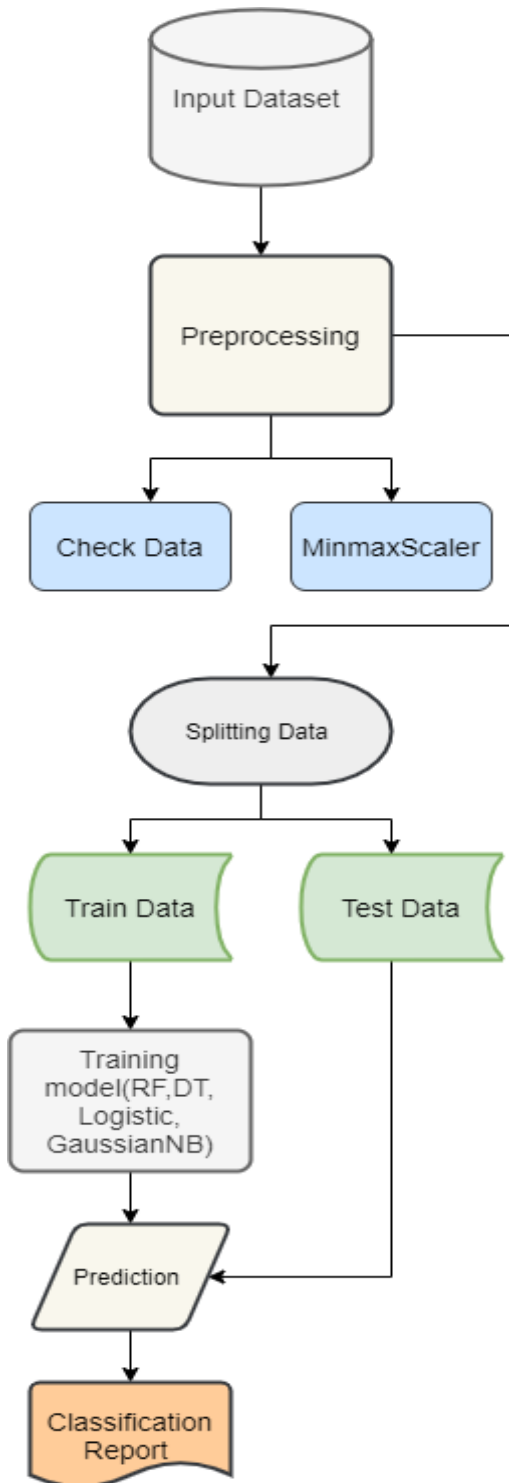


Figure 3.1: Work flow of heart failure prediction

3.2 Random Forest Algorithm:

RF is a machine learning technique that make a few choice trees. An official conclusion is settled on dependent on most of the choice trees. Choice trees experience the ill effects of low predisposition and high fluctuation. The random forest changes from high fluctuation over to low difference. (Jehad Ali¹, Rehanullah Khan², Nasir Ahmad³, Imran Maqsood⁴, 2012)

RF created by Leo Breiman is a collection of standard of health and safety characterization or relapse trees produced using the arbitrary choice of tests of the preparation information. Arbitrary highlights are chosen in the acceptance interaction. The prediction is made by amassing (dominant part approve of relapse arrangement or averaging) the organization's forecasts. Every tree is developed as portrayed in.:

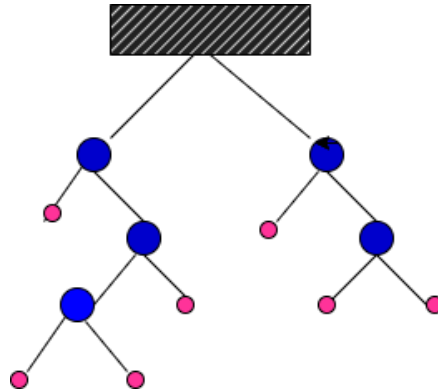
- Using Analysis N haphazardly, on the off chance that the amount of cases in the planning set is N yet with replacement, from the main data. This model will be used as the planning set for fostering the tree.
- The variable m is chosen for M information factors. To such an extent that $m \ll M$ is indicated at every hub, m factors are chosen indiscriminately out of the M and the best part on this m is utilized for parting the hub.

During the woodland development, the worth of m is held consistent. Each tree is developed to the biggest conceivable degree. No pruning is utilized. Random Forest by and large shows a huge exhibition improvement when contrasted with a single tree classifier.

3.3 Decision Tree Algorithm:

DT is typifying an administered characterization approach. The thought came from the common tree structure which is comprised of a root and hubs (the positions where spots branch separate), branches, and leaves. Likewise, a Decision Tree is built from hubs that address circles, and the branches are addressed by the fragments that interface the hubs (Jehad Ali¹, Rehanullah Khan², Nasir Ahmad³, Imran Maqsood⁴, 2012). A Decision Tree begins from the root, moves descending, and for the most part, is attracted from left to right. The hub from where the tree begins is known as a root hub. The hub where the chain closes is known as the "leaf" hub. At least two branches can be reached out from each inward hub for example a hub that isn't a leaf hub. A hub addresses

a specific trademark while the branches address a scope of qualities. These scopes of qualities go about as a parcel focuses on the arrangement of upsides of the given trademark. in Figure portrays the construction of a tree.



3.4 Logistic Regression Algorithm:

Another type of classification model is LR, (Fahd Saleh Alotaibi1, 2019) which uses relapse investigation to learn and predict the boundaries in a given dataset. Estimating the probability of paired attempt is required for the learning and expectation measures. The logistic regression model necessitates the use of class factors that should be double characterized. Similarly, in this dataset, the objective section contains two types of double numbers: 0 for patients who have no opportunities for HF and 1 for patients who have been predicted to be heart failure patients. The independent variables, on the other hand, can be double characterized, ostensible, or polynomial in nature.

The condition of calculated relapse is as per the following:

$$\text{Logit}(p) = \ln\left(\frac{p}{1-p}\right) = \frac{\text{prob.of presence of characteristics}}{\text{prob.of absence of characteristics}} \dots\dots\dots (1)$$

3.5 GaussianNB :

The GNB calculation is the supervised learning method. (Manjula C. Belavagi* and Balachandra Muniyal, 2016) The probabilities of each quality that has a place with each class is considered for a forecast. This algorithm accepts that the likelihood of each property having a place with a given class value doesn't rely upon any remaining ascribes. On the off chance that the worth of the character is known the likelihood of class esteem is known as the contingent probabilities. Information examples provability can be discovered by increasing all credits contingent probabilities together. Prediction can be made by computing each class occurrence probabilities and by choosing the most elevated likelihood class value.

$$P(M|N) = P(N|M) * P(M)/P(N)..... (2)$$

3.6 Model Performance:

CR have been used to calculate precision, recall, and F1 to evaluate pattern oppression. (SuveenAngraalMDab*Bobak J.MortazaviPhDc*AakritiGuptaMDdRohanKheraMDeTariqAhmadMD, MPHfNihar R.DesaiMD MPHafDaniel L.JacobyMDfFrederick A.MasoudiMD, MSPHgJohn A.SpertusMD, MPHhHarlan M.KrumholzMD, SMafi, 2020)The best performance as determined by the greatest Accuracy was selected and further analyzed. The score was used to assess the accuracy of the best-performing model's possibility. Which is defined as the mean squared difference between the observed and predicted results. The accuracy score selection from 0 to 1.00 where 1.00 predicts the best performance.

3.7 Evaluation of classifier algorithm:

The implementation of the Classification calculation is largely broken down by evaluating the precision, recall, F1, and precision of the arrangement. The Recall is the extent of positive occurrences that are effectively named positive (for example the extent of patients known to have the infection, who test positive for it). The precision is the extent of significant occasions in the outcomes returned. The accuracy is the extent of examples that are effectively arranged. To

measure the reliability of the implementation of the proposed model, the data is detached into planning and testing data. (Patel, Jaymin & Tejalupadhyay, Samir & Patel, Samir., 2016)

$$\text{Recall} = \text{TP} / (\text{TP} + \text{FN}) \dots\dots\dots (3)$$

$$\text{Precision} = \text{TP} / (\text{TP} + \text{FP}) \dots\dots\dots (4)$$

$$\text{Accuracy} = (\text{TP} + \text{TN}) / (\text{TP} + \text{TN} + \text{FN} + \text{FP}) \dots\dots\dots (5)$$

Everything measures can be determined centered around 4 characteristics explicitly TP, FP, FN, TN where,

- TP is different adequately grouped that cases positive.
- FP is various mistakenly ordered that an example is positive.
- FN is various erroneously arranged that an occasion is negative.
- TN is various accurately ordered that an occasion is negative.
- . • Recall is the proportion of applicable cases found in the query item to the completion of every single pertinent occasion.
- Precision is the extent of pertinent occasions in the outcomes returned.

CHAPTER 04: RESULT AND DISCUSSION

4.1 Summary:

The current work predicts experiencing a pace of a patient HF using Random Forest, Decision Tree, Logistic Regression, and GNB algorithm calculation. Complete 299 information Sample with 13 clinical highlights feature have been used to predict heart disease Eighty percent of the dataset was (Madhumita Pal1 and Smita Parija2, 2021) used for preparation, and twenty percent was used for testing. After training the model to predict the test dataset result and this result show in the CM. (figure4.1) and the CR (figure 4.2). in confusion matrix presenting TP, TN, FP, FN value. From the confusion matrix, we get more sophisticated metrics that can assist us in making

a classification decision. On the other side the classification report show accuracy of the model with precision, recall, and F1 result.

4.2 Confusion Matrix:

Random Forest	Decision Tree	Logistic Regression	GaussianNB
0 1 0 45 1 1 4 10	0 1 0 41 5 1 4 10	0 1 0 45 1 1 5 9	0 1 0 45 1 1 6 8

Table 02: confusion matrix

4.3 Classification report:

Name	Accuracy	precision	Recall	F1
Random Forest	0.92	0 0.92 1 0.91	0 0.98 1 0.71	0 0.95 1 0.80
Decision Tree	0.85	0 0.91 1 0.67	0 0.89 1 0.71	0 0.90 1 0.69
Logistic Regression	0.9	0 0.90 1 0.90	0 0.98 1 0.64	0 0.94 1 0.75
GaussianNB	0.88	0 0.88 1 0.89	0 0.98 1 0.57	0 0.93 1 0.70

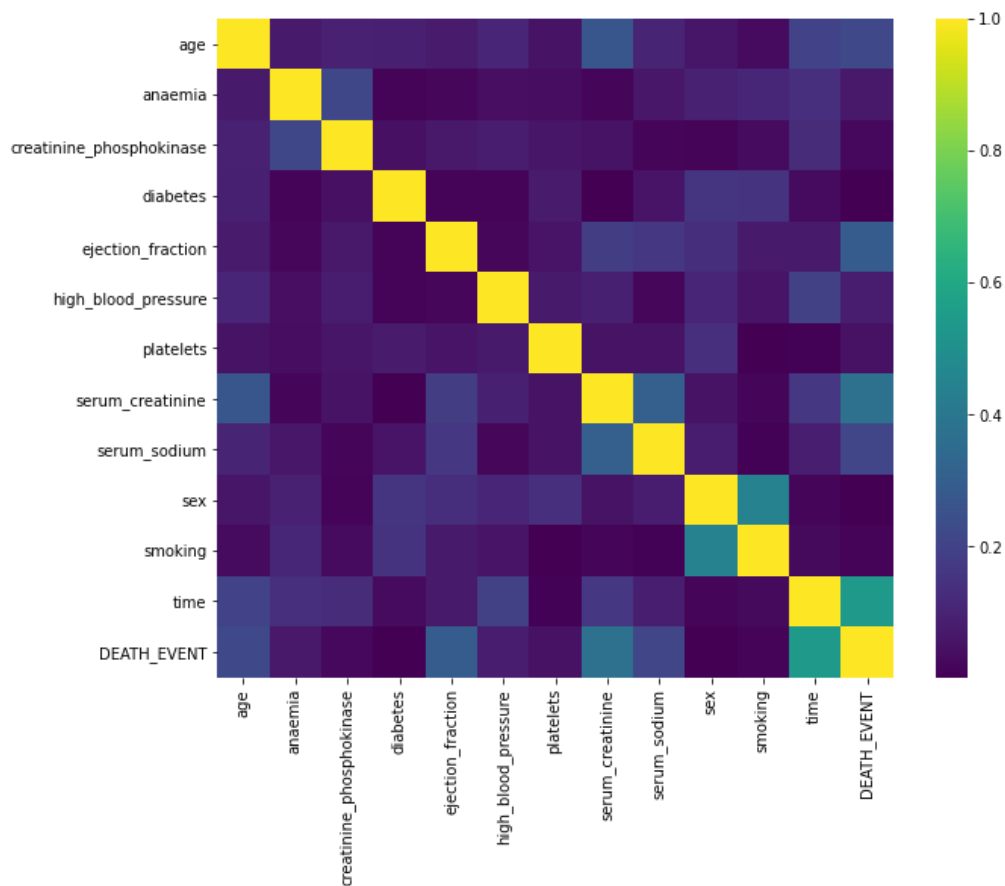
Table 03: classification report

According to the table (table: 03) of accuracy we can see here Random forest result is better than others. On the others side Decision Tree is the lowest result. Here accuracy of random forest is 0.92%,

Logistic Regression is 0.90%, GaussianNB is 0.88% and Decision Tree is 0.85. According to the accuracy confusion matrix give his result and classification report also.

4.4 Correlation:

Correlation is the matrix who is showing the relations between feature.in this figure (figure 4.3: correlation) every row feature has relation with column feature where maintain a correlation value and deciding which feature correlation are highest.



.Figure 4.3: correlation

CHAPTER 05: CONCLUSION

5.1 Findings and contribution:

Four machine learning classification modeling techniques were used to create a heart failure prediction model. (Harshit Jindal¹, Sarthak Agrawal¹, Rishabh Khera¹, Rachna Jain² and Preeti Nagrath², 2021) This project predicts people who will develop heart disease by extracting the patient medical history that leads to a resulting in death heart disease from a dataset that includes patients' medical history such as CPK, diabetes, ejection_fraction ,blood pressure, anemia and so on. Gaussian, Decision Tree, Logistic regression, Random Forest Classifier are the algorithms used to build the given model. Using more training data increases the model's chances of accurately predicting whether a given person has heart disease or not. There have been a number of medical databases on which we can work because machine learning techniques are better and can predict better than humans, which benefits both patients and doctors. In conclusion, this project guides us in predicting patients with heart diseases by cleaning the dataset and applying Gaussian, Decision Tree, logistic regression and Random Forest to achieve accuracy of different percent on our model. Furthermore, it is reached the conclusion that the accuracy of Random forest classification is the highest among the three algorithms that we have used.

5.2 Recommendation for Future works:

It is theoretically shown that Heart Failure happens when a heart can't siphon sufficient the blood to address the issues in body. Nowadays HF patients are so common and have lots of reason which causes HF. For research here used clinical information of HF patients and tried to get the best results. Every patient has all medical records in this dataset .this is the process that declares prediction of heart failure.

REFERENCES

- Davide Chicco¹ and Giuseppe Jurman². (2020). *Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone*. Chicco and Jurman BMC Medical Informatics and Decision Making .
- Devansh Shah, Samir Patel & Santosh Kumar Bharti . (2020). *Heart Disease Prediction using Machine Learning Techniques*. SN Computer Science.
- Fahd Saleh Alotaibi¹. (2019). *Implementation of Machine Learning Model to Predict Heart Failure Disease*. Jeddah, Saudi Arabia: (IJACSA) International Journal of Advanced Computer Science and Applications.
- Harshit Jindal¹, Sarthak Agrawal¹, Rishabh Khera¹, Rachna Jain² and Preeti Nagrath². (2021). *Heart disease prediction using machine learning algorithms*. india: IOP Conference Series: Materials Science and Engineering.
- Jehad Ali¹, Rehanullah Khan², Nasir Ahmad³, Imran Maqsood⁴. (2012). *Random Forests and Decision Trees*. pakistan: JCSI International Journal of Computer Science Issues.
- Kazem Rahimi, Derrick Bennett, Nathalie Conrad, Timothy M. Williams, Joyee Basu, Jeremy Dwight, Mark Woodward, Anushka Patel, John McMurray, and Stephen MacMahon. (2014). *Risk Prediction in Patients With Heart Failure*. J Am Coll Cardiol HF.
- Madhumita Pal¹ and Smita Parija². (2021). *Prediction of Heart Diseases using Random Fores*. licence by IOP Publishing Ltd.
- Manjula C. Belavagi* and Balachandra Muniyal. (2016). *Performance Evaluation of Supervised Machine Learning Algorithms for Intrusion Detection*. india: ELSEVIER.
- Patel, Jaymin & Tejalupadhyay, Samir & Patel, Samir. (2016). *Heart Disease prediction using Machine learning and Data Mining Technique*. Journal - IJCSC.
- SuveenAngraalMDab*Bobak
 J.MortazaviPhDc*AakritiGuptaMDdRohanKheraMDeTariqAhmadMD, MPHfNihar
 R.DesaiMD MPHafDaniel L.JacobyMDfFrederick A.MasoudiMD, MSPHgJohn
 A.SpertusMD, MPHhHarlan M.KrumholzMD, SMafi. (2020). *Machine Learning Prediction of Mortality and Hospitalization in Heart Failure With Preserved Ejection Fraction*. EL SEVIER.
- Wouter Ouwerkerk 1, Adriaan A Voors 2, Aeilko H Zwinderman 3. (2014). *Factors influencing the predictive power of models for predicting mortality and/or heart failure hospitalization in patients with heart failure*. Elsevier Inc.