Heart Failure Prediction Using Machine Learning

Md. Asad Chowdhury Dipu [1a], Kazi Mostaq Hridoy [1b], Adri Saha [1c]
Dr. Md. Golam Rabiul [1d]
[1]*Department of Computer Science and Engineering*
East West University
A/2, Jahrul Islam City,
Aftabnagar, Dhaka-1212
{([a] 2019-1-60-093,[b] 2019-1-60-098,[c] 2019-1-60-024)@std.ewubd.edu,
[d] golam.rabiul@ewubd.edu}

***Abstract*—The heart is an important organ in all surviving organisms. It circulates blood throughout our bodies, supplying oxygen and nutrients to our body cells and eliminating waste. Heart disease indicates to a group of illnesses that impact the cardiovascular system. Heart disease prediction is a critical issue in the field of clinical data analysis. Initial techniques of predicting heart conditions helped in doing important decisions regarding changes that have happened in high-risk patients, resulting in primary prevention. To address the issue, a prediction system for disease awareness is required. Ml algorithms are necessary to make appropriate decisions in the prediction of heart diseases in the healthcare organization because there is a lot of medical information. We will use that data to predict the heart failure. Machine learning (ML) is a form of Artificial Intelligence (AI) that offers renowned help in predicting any type of event using real situations as training data. We will preprocess the data and calculate the accuracy of machine learning algorithms for predicting heart disease in this research. These algorithms are random forest, decision tree, knn, naive baye, support vector machine, logistic regression, gradient boosting. Google Colab Notebook is an excellent tool for learning Python programming since it includes a variety of libraries, header files, and computronium that make the process more precise and efficient.**

***Keywords*— Machine Learning, Artificial Intelligence, Prediction, Accuracy, Algorithms, Training Data, Testing Data**

1. INTRODUCTION

The heart is a vital and significant organ in the human body. As the heart is such an important part of the body, it deserves special attention. Heart failure is a condition when the heart has become unable to supply enough blood to reach the body requirements. This is caused primarily by a lack of blood supply to the heart or a weak heart surface. Heart or cardiovascular disease doesn't mean the heart has stopped beating or seems to stop. It is considered a serious heart condition. Heart disease is difficult to detect due to high blood pressure, cholesterol, diabetes, an irregular pulse, and for many other risks. Since most diseases have been connected to the heart, it is essential to estimate cardiovascular or heart disease, which demands a comparison analysis. Most patients die today even though their diseases are identified at an advanced stage due to instrument lack of accuracy, so more efficient disease prediction algorithms are needed.
Nowadays, health costs are rising every day because of changing lifestyles and hereditary factors. And so over time, it generates a large amount of data. Consequently, the data generated by the health or survey is thrown away. However, with the invention of data analytics, this is no longer the case. The data is being used by hospitals and non-governmental organizations to generate useful information. Cardiovascular disease is the deadliest enemy of the modern world because it affects people in this manner that they cannot be cured as painlessly as possible. As a result, the most difficult task in medicine is diagnosis and treatment patients correctly at the right time. The hospital's bad reputation arises from misunderstandings and incorrect diagnosis. Everyone's blood pressure,

pulse rate, and cholesterol levels are different. However, according to medical evidence, normal blood pressure, pulse rate and cholesterol levels all are 120/80. According to WHO reports, over 12 million people die each year from cardiovascular disease.

We plotted some pair-plots, histograms, box and whisker plots, bar-charts, heatmaps to show differences of dataset variables.

Machine Learning is a training and testing-based evaluation technology that is extremely effective. Artificial Intelligence includes machine learning as a subset (AI), a broad learning field in which computers simulate human abilities.

In this project, we used biological parameters as testing data, such as age, sex, cholesterol, blood pressure, and thalach, and compare accuracy of regression algorithms such as linear regression, polynomial regression, Ridge regression and used classification algorithms such as logistic regression, random forest, SVM, Gradient boosting, Naive Bayes, KNN.

We also discovered that decision trees can be used to accurately predict events related to heart disease. A classification strategy is a decision tree. It uses a tree structure to represent each node's feature and the value of a branch's feature. The leaves are used to understand the class at the same time. The decision tree's root is a single top, also the branches are the possible outcomes. [1] Many aspects of our daily lives involve decision trees. Furthermore, it is very simple to explain something using it.

We have calculated the accuracy of 10 different machine learning approaches in our paper and use the accuracy results to find out which one is the best.

2. RELATED WORK

The heart is just a little organ that circulates blood throughout your body. It is your cardiovascular system's main organ. There is continually a demand for cardio assessment, whether for identification, prediction, or heart-disease treatment. This research benefited from artificial intelligence, machine learning, and data mining, among other topics. Because the structure of the heart is complicated, it must be treated with care, or the patient will die.

Data mining has been used by some of the researchers to predict heart disease. Among the most popular data mining techniques is the decision tree. According patel er al. [2], The goal of this study is to analyze alternative Decision Tree classification algorithms to improve performance in the detection of heart problems using WEKA. Using existing datasets of heart disease patients from the UCI repository's Cleveland database, the usefulness of decision tree algorithms is examined and validated. This collection has 303 entries and 76 features. The goal of this study is to employ data mining techniques to identify hidden patterns related to heart problems and predict the existence of heart illness in patients, ranging from no presence to reminders. They compare the accuracy of several machine learning and data mining techniques. Kaur et al. [2] has been working on this and has defined how the fascinating pattern and knowledge are produced from the vast dataset. They evaluate the results of various machine learning and data mining algorithms to see which one is the best, and they find the result is in favor of SVM

In paper [6], It simply refers to the progress of Support Vector Machines (SVM), which are used to examine and, as a result, isolate learning from large data sets.

Krishnan et al. [4] built a multi-layer perceptron prototype for predicting cardiac illnesses in humans and the algorithm's efficiency utilizing CAD technology. If more people use the prediction method to predict their conditions, disease awareness will rise, and the death rate of heart patients would decrease.

In this paper [5] Parthiban thinks, although most analysts use various classification techniques in the conclusion of cardiac diseases, such as SVM, KNN, Neural system, and two-fold discretization with Gain Ratio Decision Tree, it is assumed that using Nave Bayes and Decision tree with data pick up counts gives better results in the finding of cardiovascular events and better exactness when compared to other classification models.

Machine learning algorithms are used to predict a lot of diseases, and so many researchers are working on this. In paper [7] Arora, Research was carried on heart disease diagnosis using logistic regression, diabetes prediction

using support vector machines, and breast cancer prediction using Adaboot classifier, with the results showing that logistic regression has an accuracy of 87.1 percent, support vector machines have an accuracy of 85.71 percent, and Adaboot classifier has an accuracy of 98.57 percent, which is nice for prediction.

3. DATASETS AND DESCRIPTION

Data source

The term "heart disease" refers to a variety of conditions that harm the human heart. Heart disease, as well recognized as cardiovascular disease, is among the deadliest in nature. In this study. To collect data, we were selected 10 hospitals but responded only 2 hospitals. Based on these records were obtained from Heart foundation, Labaid cardiac hospital ‘s collection of databases which is called UCI machine Learning Repository. The disease-related patterns are separated by datasets. We provided 14 attributes in our dataset heart testing, even though we obtained 920 records and 76 medical attributes. The following table (Table 1) lists the 14 attributes that the system is based on.

Analysis of Data

To begin, we attempted to process the dataset. Our data set had 303 rows and 14 columns. The dataset contained strings in one column (target attribute), float numbers in one column (old peak), and integers in the remaining columns. Since the dataset contains categorical values and redundant data, this phase's first tasks were data pre-processing that is encoding the categorical data and data normalization to remove redundant data. Then we used PCA to reduce the dimension of our dataset to make it easy in finding classifiers. This is how we preprocessed the dataset. After that, we split the dataset into train and test data to find regression and classifications. Finally, we ran the dataset through classifiers and compared the results.

Table 1: Attributes that has been used to predict heart diseases

Attributes	Description	Value	Statistical description	Type
1. Age	Person’s completed age	Age: 29 to 77	Mean: 54.37 Median: 55	Numeric
2. Sex	Gender of the individual.	1 for male. 0 for female	Mean: 0.683 Median: 1	Nominal
3. CP	Types of chest pain	1: typical type1 angina 2: typical type2 angina 3: non-angina pain 4: Asymptomatic	Mean: 0.97 Median: 1	Numeric
4. Trestbps()	Resting blood pressure of a person	Blood pressure: 94-200 mm Hg at the time of admission to the hospital.	Mean: 131.623 Median: 130	Numeric
5. Chol	Cholesterol in serum	126-564 mg/dl	Mean: 246.26 Median: 240	Numeric

6. Fbs	Fasting sugar levels on blood	1: Fasting blood sugar levels >120 md/dl; 0: Fasting blood sugar levels <120 mg/dl	Mean: 0.15 Median: 0	Nominal
7. Restecg	The following three values represent the electrographic resting results: 0 represents the normal state, Value 1 represents abnormalities in the ST-T wave (which include T-wave inversions, depression, or elevation of ST of > 0.05 mV), and 2 represents any probability or certainty of LV hypertrophy according to Estes' criteria.	0: normal. 1: having st-t wave abnormality. 2: Showing probable or definite left ventricular hypertrophy	Mean: 0.53 Median: 1	Nominal
8. Thalach	achieved maximum heart rate	71-202	Mean: 149.646 Median: 153	Numeric
9. Exang	Activity of angina by exercise	1: yes; 0: No	Mean: 0.326 Median: 0	Nominal
10. Oldpeak	In comparison to the resting state, exercise-induced ST depression	0-6.2	Mean: 1.039 Median: 0.800	Numeric
11. Slope	Refers slope of the peak exercise ST segment	0: unsloping. 1: flat. 2: down sloping	Mean: 1.399 Median: 1	Nominal
12. Ca	Fluoroscopy-colored major vessels; categorical: five levels	0-4	Mean: 0.73 Median: 1	Numeric

13. thal	State of heart which is defined by three different numerical values. Normal defects are numbered 1, fixed defects are numbered 2, and reversible defects are numbered 3.	1 for normal. 2 for fixed defect Value 3 for reversible defect Categorical: 3 levels	Mean: 2.31 Median: 1	Nominal
14. Target	Defines whether patient has heart disease or not	Value 0: No, Value 1: Yes	Mean: 0 Median: 1	Nominal

The table has been showed to describe the attributes. Like, blood pressure, pulse rate and cholesterol rate are all different for everyone [2]. According to medical evidence, normal blood pressure, pulse rate and cholesterol are all 120/80.

Operating Environment

We calculated the accuracy of machine learning algorithms for predicting heart disease, including decision tree, logistic regression, support vector machine (SVM), gradient boosting, Naive Bayes, and k-nearest neighbor, using the UCI repository dataset for training and testing (KNN). We used Google Colab Notebook to implement Python programming, which is considered the best tool as it includes a variety of libraries and header files which makes the work more precise and accurate.

Process Model

4. PROPOSED SYSTEM

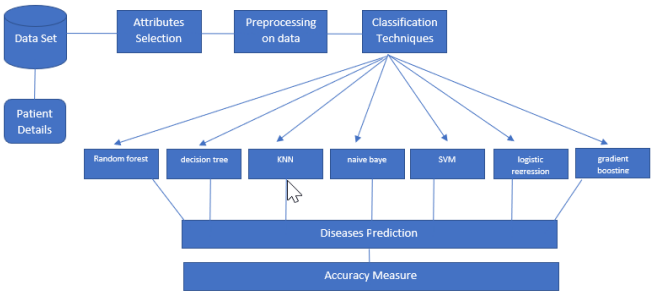

Figure 1: Architecture of Prediction System

5. VARIOUS MACHINE LEARNING ALGORITHMS

To cluster datasets, Decision Tree (DT) features variables and criteria are used. The classifiers are then used to estimate every clustered dataset's performance. Based on their low rate of error, the best and most consistent models are identified from the above results. On this data set, the classifier's performance is evaluated for error optimization. Generally, three types of machine learning algorithms we see, these are supervised algorithm, Unsupervised algorithm, and Reinforced algorithm.

Unsupervised learning algorithm: Machine learning algorithms are used in unsupervised learning to evaluate and cluster unlabeled sets of data. These algorithms uncover hidden patterns or data groupings without the need for human intervention. For example: Clustering, Associative, hidden Markov model. [9]

Supervised learning algorithm: supervised learning algorithms attempt to model dependency relationships between both the target fulfilled and the input attributes to predict the output values for new data using relationships learned from previous data sets. [10] For example: Regression, decision tree, random forest, classification.

Reinforcement learning algorithm: Reinforcement learning is a type of machine learning in which desirable behaviors are rewarded while undesirable behaviors are punished. A reinforcement learning agent, in general,

can comprehend and interpret its environment, act, and learn through trial and error. [11]

In our project, above these learnings we used supervised machine learning algorithm to predict the heart diseases.

Regression model:

Classifier models:

A. Logistic regression: The Supervised Learning technique is used in the Logistic Regression algorithm. [12] It's used to figure out or predict the probability of a binary (yes/no) event happening. [13] It is a method for forecasting a categorical dependent variable from a set of independent variables. For example, to determine whether a person is prone to infection with COVID-19.

The logistic function is written as,

$$f(x) = L/\ [1 + e]\ \hat{}(-k(x - xo))$$

F(x) = function's output.

L = the curve's maximum value.

K = steepness of the curve or rate of logistic growth

Xo = sigmoid midpoint's x value.

X= real number

A. Decision tree: Decision Trees are indeed a type of supervised machine learning and graphical representation of data in which the data is constantly split according to a parameter.

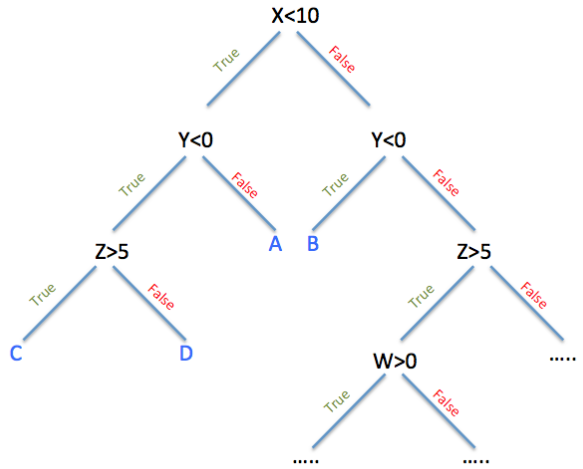

Figure 2: Decision tree

B. Naive Bayes Algorithm

A naive Bayes classifier is an algorithm that classifies objects using Bayes' theorem. Naive Bayes classifiers assume powerful, or naive, independence among data point attributes. Text analysis, spam filters, and medical diagnosis all use Naive Bayes classifiers. Because they are simple to implement, these classifiers are widely used in machine learning. [14]

It identifies the patients' specialties in relation to the disease. It displays the probability of each input attribute being in the predictable state, as well as the probability of events occurring. P (A|B) denotes the probability of event A, P(B) denotes the probability of event B, and P(B|A) denotes the probability of event B if event A occurs.

$$P(A|B) = ((P(B|A) * P(B)))/(P(A))$$

C. Random Forest Algorithms

Random forest (RF) is a regression and classification supervised machine learning algorithm. To create decision trees from various samples, it uses the simple majority for classification and the average for regression. [15]

It corrects their overfitting training set. It also avoids outliers and missing values by following the steps of data pre-processing and data analysis. It's a machine learning algorithm for creating a dynamic model by combining weak models. The random forest tree displays the multiple decision trees associated with the system (figure 3).

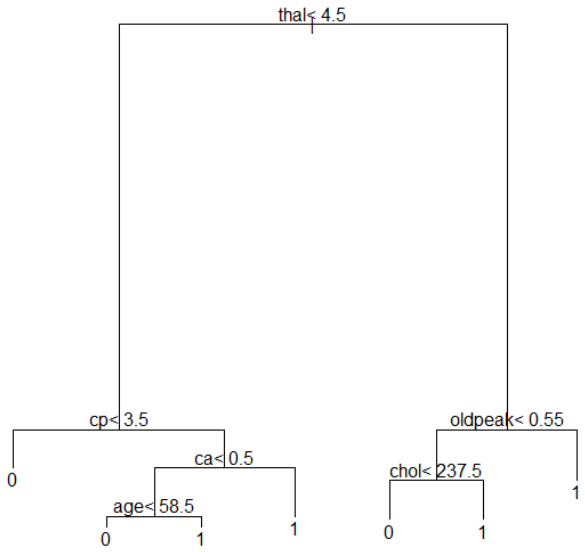

Figure 3: Random Forest tree

D. Support Vector Machine

Support vector machine (SVM) is a regression and classification advanced algorithms for data analysis. After data has been analyzed, Svm is a supervised method that divides it into two groups. An SVM creates a map of such sorted data with the widest possible margins. [16]

E. Gradient boosting

Gradient boosting classifiers are a collection of machine learning algorithms that combine multiple weak learning models into a single strong predictive model. Gradient boosting applies decision trees. [17]

In Figure 4, the boosted tree is used to show the variable importance of heart failure prediction.

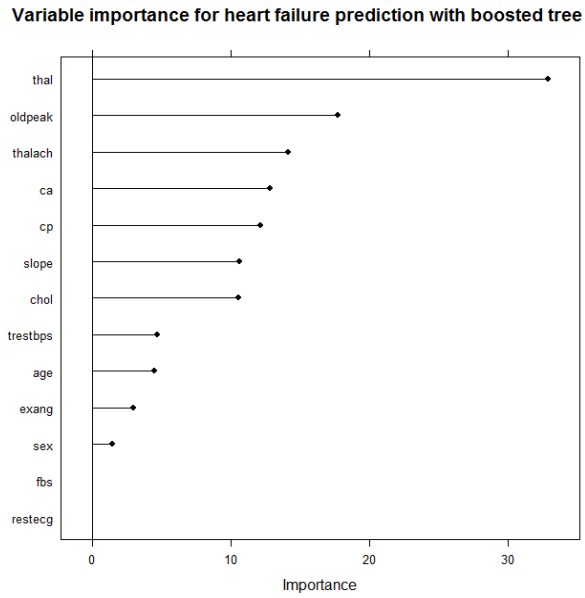

Figure 4: Showing Importance of variables

F. KNN (K-Nearest Neighbor)

The KNN algorithm is a method for solving regression and classification problems using supervised machine learning. It's easy to set up and use, but it has the disadvantage of becoming noticeably slower as the amount of data used grows. [18]

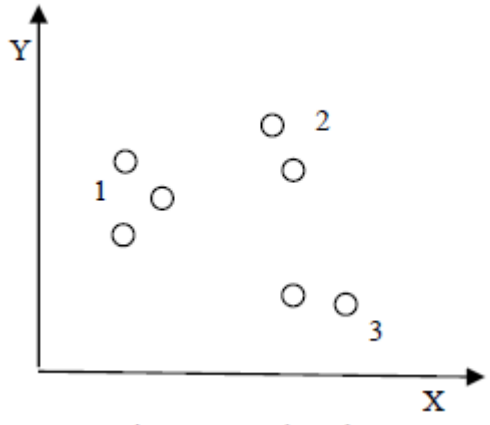

Figure 5: KNN where k=3

In the example above, k=3 indicates that there are three different types of data. Each cluster is represented by coordinates (Xi,Yi), where Xi

represents the x-axis, Y represents the y-axis, and i=1,2,3,....n.

Regression model:

A. Linear regression: The supervised Machine Learning model Linear Regression finds the best fit linear line between the independent and dependent variables. It determines the dependent and independent variables' linear relationship. [19]

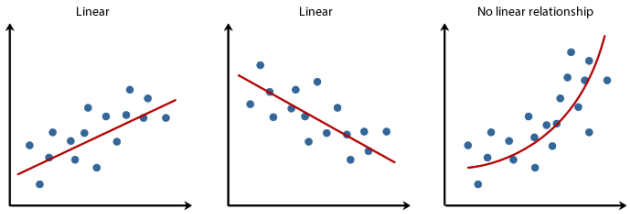

Figure 6: Types of linear regression

B. Ridge regression:

Ridge regression is a model tuning technique that can be used to analyze data with multicollinearity. L2 regularization is achieved using this method. When there is a problem with multicollinearity, least-squares are unbiased, and variances are large, resulting in predicted values that are far from reality. [20]

C. Polynomial regression:

The Cost Function evaluates the performance of a Machine Learning model for a given set of data. The difference between anticipated and expected values is calculated using the Cost Function, which is a single real number. [21]

6. DATA PRE-PROCESSING

Data preprocessing is basically the process of converting raw data into a comprehensible format. Data in the real world is frequently partial, inconsistent, redundant, and loud. Data preprocessing entails a number of stages that aid in the conversion of raw data into a processed and usable state.

Prepossessing is an important part of the data analysis and researching as it is the key to have the most correct accuracy. Therefore, implementing the machine learning algorithms and output the best results, we had pre-processed the dataset. We had a dataset 'heart.csv' and we performed several data pre-processing steps.

The followings depict the numerous phases involved in data preprocessing-

1) Filling Null Values:

'isnull ()' function was used for checking the missing values of columns and 'print(df.isnull().sum())' this code provided a list where number of missing values have in columns could be seen. If any column contained missing values, then they were replaced by 0 with using 'fillna ()' function.

```
df.isnull().sum()

age         0
sex         0
cp          0
trestbps    0
chol        0
fbs         0
restecg     0
thalach     0
exang       0
oldpeak     0
slope       0
ca          0
thal        0
target      0
dtype: int64

[ ] df.isnull().sum().sum()

0
```

Figure 7: Null values

2) Encoding the categorical data:

In the dataset, the target column had categorical data type. Because most machine learning models only take numerical variables, categorical variables must be preprocessed. These category variables were transformed to numbers so that the model could comprehend and extract useful information.

```
[ ] df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 303 entries, 0 to 302
Data columns (total 14 columns):
 #   Column    Non-Null Count  Dtype  
---  ------    --------------  -----  
 0   age       303 non-null    int64  
 1   sex       303 non-null    int64  
 2   cp        303 non-null    int64  
 3   trestbps  303 non-null    int64  
 4   chol      303 non-null    int64  
 5   fbs       303 non-null    int64  
 6   restecg   303 non-null    int64  
 7   thalach   303 non-null    int64  
 8   exang     303 non-null    int64  
 9   oldpeak   303 non-null    float64
 10  slope     303 non-null    int64  
 11  ca        303 non-null    int64  
 12  thal      303 non-null    int64  
 13  target    303 non-null    object 
dtypes: float64(1), int64(12), object(1)
memory usage: 33.3+ KB
```

Figure 8: Dataset info

There are several approaches to take. 1) label-encoding; 2) one-shot encoding are examples of categorical variables. Label-encoding was chosen to perform encoding operation on the categorical data.

Label Encoding is a well-known encoding technique for dealing with categorical information. Based on alphabetical sorting, each label is provided a unique integer in this approach.

To perform label encoding we first had to import the module named "preprocessing" from sklearn. Then the following was executed-

```
from sklearn import preprocessing
```

We will use label-encoding to encode

```
[ ] label_encoder = preprocessing.LabelEncoder()
    df['target']=label_encoder.fit_transform(df['target'])
```

```
[ ] df
```

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
0	63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
1	37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
2	41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
3	56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
4	57	0	0	120	354	0	1	163	1	0.6	2	0	2	1

Figure 9: Data preposing

3) Data Standardization:

When comparing measurements with various units, standardizing the characteristics around the center and 0 with a standard deviation of 1 is significant. Variables recorded at various scales do not contribute evenly to the analysis and may result in the formation of a bias.

The characteristics will be rescaled as a consequence of standardization such that they have the attributes of a standard normal distribution with $\mu = 0$ and $\sigma = 1$, where μ is the mean (average) and σ is the standard deviation from the mean.

The sci-kit-learn StandardScaler () function eliminates the mean and adjusts the data to unit variance. Sci-kit learn was used to import the StandardScalar algorithm and applied it to the dataset.

	age	trestbps	chol	thalach	oldpeak
0	0.952197	0.763956	-0.256334	0.015443	1.087338
1	-1.915313	-0.092738	0.072199	1.633471	2.122573
2	-1.474158	-0.092738	-0.816773	0.977514	0.310912
3	0.180175	-0.663867	-0.198357	1.239897	-0.206705
4	0.290464	-0.663867	2.082050	0.583939	-0.379244
...	...	...	...	...	...

Figure 10: Dataset after Standardization

4) **Dropping unnecessary columns:** In our dataset, there were some unnecessary columns which were not having any impact on the result. So those columns were dropped.

```
dropped= df.drop (columns = ['trestbps','chol','fbs','restecg','thalach'],axis=1)
dropped
```

	age	sex	cp	exang	oldpeak	slope	ca	thal	target
0	63	1	3	0	2.3	0	0	1	1
1	37	1	2	0	3.5	0	0	2	1
2	41	0	1	0	1.4	2	0	2	1
3	56	1	1	0	0.8	2	0	2	1
4	57	0	0	1	0.6	2	0	2	1

Figure 11: Dropped unnecessary columns

7. DATA VISUALIZATION

The graphical display of information and data is known as data visualization. Data visualization tools, which include visual components like as charts, graphs, and maps, make it easy to view and comprehend trends, outliers, and patterns in data. This representation is a mapping between the underlying data (often numerical) and visual components (for example, lines or points in a chart). The mapping defines how these components' characteristics vary based on the data.

In the dataset, there were 303 rows × 14 columns. Several types of plotting like histplot, distplot, boxplot and heatmap were used for visualization purpose.

Distplot: Seaborn module was used to draw the displots. The total distribution of continuous data variables is shown by the Seaborn Distplot. The Seaborn module, in conjunction with the Matplotlib module, is used to represent the distplot with many modifications. The data is represented by a histogram and a line in the Distplot.

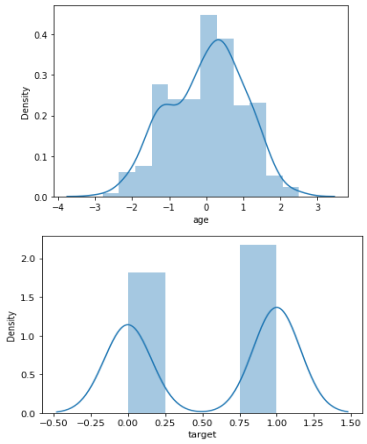

Figure 12: Distplot of two columns- 'age' and 'target'

Histplot: pandas. hist () function was used to plot the histograms. A histogram is a graph that divides a set of data points into user-specified ranges. In pandas, the hist () method allows to plot distinct histograms for various categories of data. The column name can be defined for which separate groups should be created by using the 'by' argument. This will generate distinct histograms for each group.

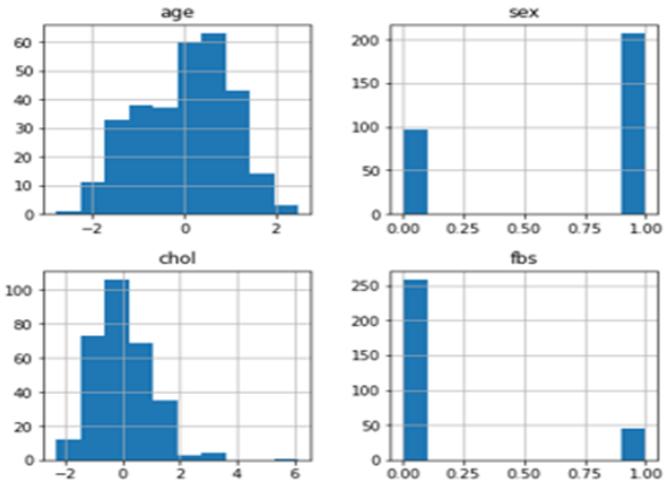

Figure 14: Histogram of 'age', 'sex','chol' and 'fbs' columns

Boxplot: plot () was used to plot the boxplots. A box and whisker plot, often known as a box plot, shows a five-

number summary of a collection of data. The minimum, first quartile, median, third quartile, and maximum are the five-number summary.

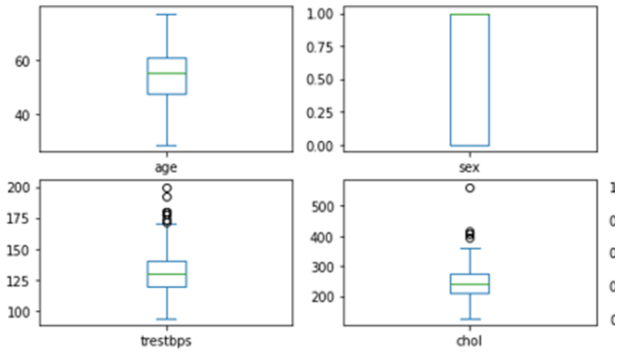

Figure 15: Boxplot of age, sex, trestbps and chol columns

Heatmap: seaborn. heatmap () was used to plot the heatmap. A correlation heatmap is a graphical depiction of a correlation matrix that depicts the relationship between several variables. Correlation can have any value between -1 and 1.

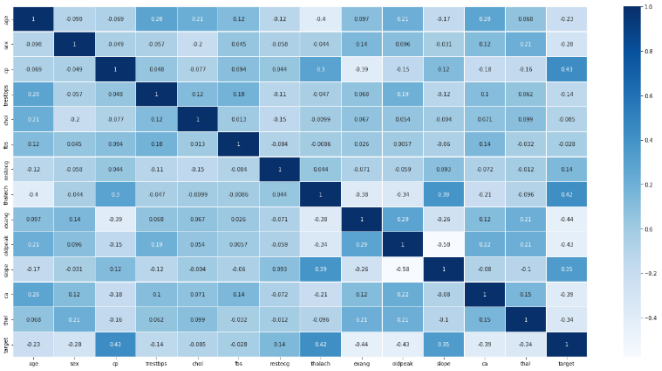

Figure 16: Correlation heatmap of the features.

8. RESULTS AND DISCUSSION

The outputs and accuracies generated are examined and the results are displayed in this part. Both classification and regression algorithm were implemented. Classification algorithms include- random forest, decision tree, k-nearest-neighbor, Naïve Bayes, logistic Regression, Support Vector Machine and Gradient Boosting. And for regression algorithms- linear regression, ridge/L1 regression and polynomial regression were chosen.

For classification algorithms evaluation metrics are- accuracy score, precision score, recall score, f1 score, auc score etc.

Accuracy Score- It is the count of correct projections Total number of projections.

Table 2: Accuracy of the models

Classification Models	Accuracy Score
Random Forest	0.84
Decision Tree	0.75
Logistic Regression	0.82
Support Vector Machine	0.81
K-Nearest-Neighbors	0.80
Naïve Bayes	0.82
Gradient Boosting	0.81

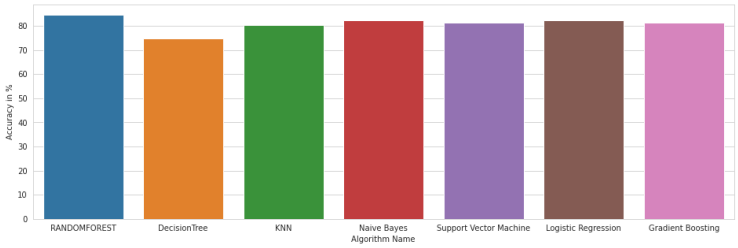

Figure 17: Graphical Representation of the used models Accuracy Scores.

R2-Score- It is a handy 0–100% scale that assesses the strength of the association between the model and the dependent variable.

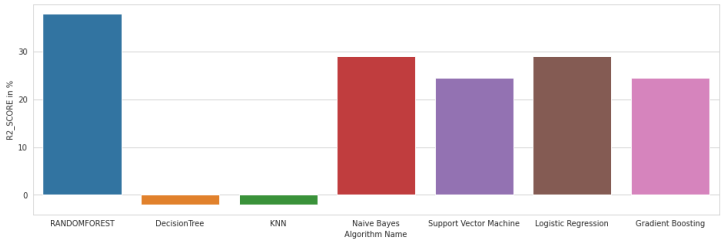

Figure 18: Graphical Representation of the used models R2-Score

Precision Score- It is defined as the number of true positives divided by the total number of positive predictions.

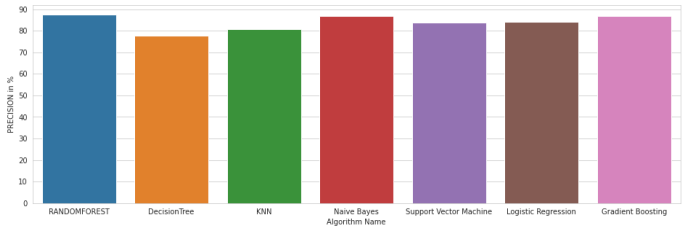

Figure 19: Graphical Representation of the used models Precision Score

Recall Score- It shows how many accurate hits were discovered.

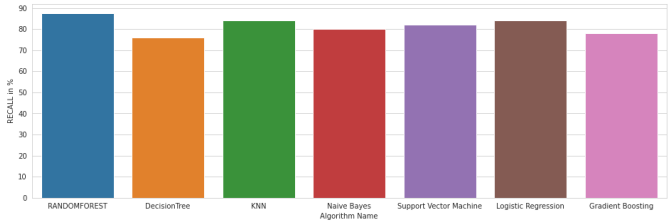

Figure 20: Graphical Representation of the used models Recall-Score

F1-Score- The F1 score is the harmonic mean of precision and recall and is a more accurate metric than accuracy.

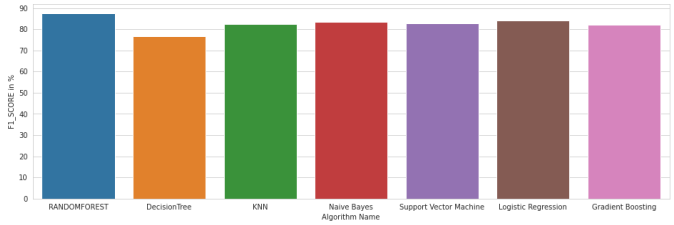

Figure 21: Graphical Representation of the used models F1-Score

It could be seen that random forest classifier is showing the best result. It is showing a good result because of feature randomness. The number/proportion of features may be varied and utilized to discover the optimal split for each node while creating a Random Forest model.

For regression algorithms the evaluation metrics are- mean absolute error, mean squared error, root mean squared error and r2-score. The evaluation metrics of regression models are as follows-

Table 3: Error score

Model	R2_score	mean absolute error	mean squared error	root mean squared error
Linear Regression		0.3145	0.1561	0.561
Polynomial Regression		1.4602	9.2355	1.2084
Ridge Regression		0.3145	0.1561	0.561

9. LIMITATONS

One of the significant shortcomings of these research is that the primary emphasis has been on the use of classification algorithms for heart disease prediction, rather than investigating various data cleaning and pruning strategies that prepare and make a dataset acceptable for mining. A properly cleaned and trimmed dataset outperforms an unclean dataset with missing values in terms of accuracy. The use of appropriate data cleaning procedures in conjunction with appropriate classification algorithms will result in the creation of prediction systems with increased accuracy.

10. CONCLUSION

With the rising number of fatalities from heart disease, it has become necessary to design a system that can forecast heart disease effectively and reliably. The study's goal was to identify the best efficient ML algorithm for detecting cardiac problems. Using the UCI machine learning repository dataset, this study analyzes the accuracy score of the Decision Tree, Logistic Regression, Random Forest, and Naive Bayes algorithms for predicting heart disease. According to the findings of this study, the Random Forest algorithm is the most

effective algorithm for predicting heart disease, with an accuracy score of 84 percent. In the future, the study may be improved by creating a web application based on the Random Forest method and employing a larger dataset than the one used in this analysis, which would assist to deliver better findings and aid health professionals in successfully and efficiently forecasting cardiac disease.

REFERENCES

[1] "What is a Decision Tree Diagram", Lucidchart, 2022. [Online]. Available: https://www.lucidchart.com/pages/decision-tree. [Accessed: 19- May- 2022]

[2]. S. Patel, J. Patel and S. Tejalupadhyay, "Heart Disease prediction using Machine learning and Data Mining Technique", *Journal - IJCSC*, vol. 7, 2022. Available: 10.090592/IJCSC.2016.018 [Accessed 15 May 2022].

[3]. A. Kaur and J. Arora, "HEART DISEASE PREDICTION USING DATA MINING TECHNIQUES: A SURVEY", *International Journal of Advanced Research in Computer Science*, vol. 9, no. 2, pp. 569-572, 2022. Available: IJARCS 2015-2019 [Accessed 15 May 2022].

[4]. S. Krishnan J and K. J, "Prediction of Heart Disease Using Machine Learning Algorithms", *2019 1st International Conference on Innovations in Information and Communication Technology (ICIICT)*, 2022. Available: 10.1109/ICIICT1.2019.8741465 [Accessed 15 May 2022].

[5]. S.K. Srivasta G. Parthiban, "Applying Machine learning methods in Diagnosing Heart disease for Diabetic Patients" International Journal of Applied Information Systems (IJAIS) – ISSN: 2249-0868 Foundation of Computer Science FCS, New York, USA Volume 3– No.7, August 2012.

[6]. Geert Meyfroidt, Fabian Guiza, Jan Ramon, Maurice Brynooghe" Machine learning echniques to examine large patient databases"-Best practice &ReasearchClinicalAnaesthesiology, ElsevierVolume 23 (1) – Mar 1, 2009.

[7]. Shriya Arora and Pahulpreet Singh Kohli, "Application of Machine Learning in Diseases prediction", 4th International Conference on Computing Communication And Automation (ICCCA), 2018

[8] B. D. C. N. Prasad, P. E. S. N. Krishna Prasad, and Y. Sagar, "A Comparative Study of Machine Learning Algorithms as Expert Systems in Medical Diagnosis (Asthma)," Advances in Computer Science and Information Technology Communications in Computer and Information Science, pp. 570–576, 2011.

[9] Education, "What is Unsupervised Learning?", Ibm.com, 2022. [Online]. Available: https://www.ibm.com/cloud/learn/unsupervised-learning. [Accessed: 20- May- 2022]

[10] "Types of Machine Learning Algorithms You Should Know", Medium, 2022. [Online]. Available: https://towardsdatascience.com/types-of-machine-learning-algorithms-you-should-know-953a08248861#:~:text=Supervised%20learning%20algorithms%20try%20to,from%20the%20previous%20data%20sets. [Accessed: 20- May- 2022]

[11] "What is Reinforcement Learning? A Comprehensive Overview", SearchEnterpriseAI,2022. [Online]. Available: https://www.techtarget.com/searchenterpriseai/definition/reinforcement-learning. [Accessed: 20- May- 2022]

[12] "Logistic Regression in Machine Learning - Javatpoint", www.javatpoint.com, 2022. [Online]. Available: https://www.javatpoint.com/logistic-regression-in-machine-learning. [Accessed: 20- May- 2022]

[13] "Logistic Regression for Machine Learning | Capital One", Capital One, 2022. [Online]. Available: https://www.capitalone.com/tech/machine-learning/what-is-logistic-regression/. [Accessed: 20- May- 2022]

[14] "What is Naive Bayes? - Definition from Techopedia", Techopedia.com, 2022. [Online]. Available: https://www.techopedia.com/definition/32335/naive-bayes#:~:text=A%20naive%20Bayes%20classifier%20is,text%20analysis%20and%20medical%20diagnosis. [Accessed: 20- May- 2022]

[15] "Random Forest | Introduction to Random Forest Algorithm", Analytics Vidhya, 2022. [Online]. Available: https://www.analyticsvidhya.com/blog/2021/06/understanding-random-forest/. [Accessed: 20- May- 2022]

[16] "What is a Support Vector Machine (SVM)? - Definition from Techopedia", Techopedia.com, 2022. [Online]. Available: https://www.techopedia.com/definition/30364/support-vector-machine-svm#:~:text=A%20support%20vector%20machine%20(SVM)%20is%20machine%20learning%20algorithm%20that,as%20far%20apart%20as%20possible. [Accessed: 20- May- 2022]

[17] "Gradient Boosting Classifiers in Python with Scikit-Learn", Stack Abuse, 2022. [Online]. Available: https://stackabuse.com/gradient-boosting-classifiers-in-python-with-scikit-learn/. [Accessed: 20- May- 2022]

[18] Medium. 2022. Machine Learning Basics with the K-Nearest Neighbors Algorithm. [online] Available at: <https://towardsdatascience.com/machine-learning-basics-with-the-k-nearest-neighbors-algorithm-6a6e71d01761#:~:text=Summary-,The%20k%2Dnearest%20neighbors%20(KNN)%20algorithm%20is%20a%20simple,that%20data%20in%20use%20grows.> [Accessed 20 May 2022].

[19] "Linear Regression | Introduction to Linear Regression for Data Science", Analytics Vidhya, 2022. [Online]. Available: https://www.analyticsvidhya.com/blog/2021/05/all-you-need-to-know-about-your-first-machine-learning-model-linear-regression/. [Accessed: 21- May- 2022].

[20] "What is Ridge Regression?", 2022. [Online]. Available: https://www.mygreatlearning.com/blog/what-is-ridge-regression/#:~:text=Ridge%20regression%20is%20a%20model,away%20from%20the%20actual%20values. [Accessed: 21- May- 2022]

[21] "Understanding Polynomial Regression Model - Analytics Vidhya", Analytics Vidhya, 2022. [Online]. Available: https://www.analyticsvidhya.com/blog/2021/10/understanding-polynomial-regression-model/. [Accessed: 21- May- 2022]