Comparative Study on Heart Disease Prediction Using ML Techniques and Neural Networks

Project submitted to the

SRM University – AP, Andhra Pradesh

for the partial fulfillment of the requirements to award the degree of

Bachelor of Technology

In

Computer Science and Engineering

School of Engineering and Sciences

Submitted by

Omkar Subhash Ghongade

AP20110010090

Under the Guidance of

Dr.Murali Krishna Enduri

SRM University–AP

Neerukonda, Mangalagiri, Guntur

Andhra Pradesh – 522 240

[November, 2022]

Certificate

Date: 15-Dec-22

This is to certify that the work present in this Project entitled "**Comparative Study on Heart Disease Prediction Using ML Techniques and Neural Networks**" has been carried out by **Omkar Subhash Ghongade** under my supervision. The work is genuine, original, and suitable for submission to the SRM University – AP for the award of Bachelor of Technology in **School of Engineering and Sciences**.

Supervisor

Prof. / Dr. Murali Krishna Enduri

Assistant Professor,

Department of CSE,

SRM University – AP,

Andhra Pradesh.

Acknowledgements

We are ever so thankful to all of those who have contributed to the successful completion of this project in any way. It is because of their dedication, hard work, patience and guidance that this has been achieved. A special thanks to our mentor who has been with us from the beginning, offering his advice and support. He was generous enough to give us advice and guide us throughout our project work. We are also immensely grateful to our friends for their support, as well as anyone else who has contributed to this project. Without them, we would not have been able to reach the end goal. The project has come such a long way since it was first conceived, and that is due in large part to the effort of the people who have been involved. We are conscious of the fact that a project of this magnitude could not have been successful without the help of many individuals. We are extremely grateful to all those who have aided us in any way, whether it be directly or indirectly. We thank you all for your support. In summary, we would like to express our sincerest appreciation to all those who have been a part of this project. Without their help, we would not have been able to complete it. We are humbled by the amount of support we have been given, and we thank all those involved for their invaluable contributions.

Table of Contents

Abstract

Heart disease is a major cause of death and disability across the world. Early diagnosis and treatment of heart disease can significantly reduce mortality and morbidity rates. Hence, the development of efficient and accurate methods for early diagnosis of heart disease has become a priority in the medical field. In this study, we did a comparative study of exiting supervised machine learning approaches for predicting heart disease diagnosis and also improved the accuracy of KNN. We used a dataset that consists of a variety of features such as age, gender and other important indicators for heart disease diagnosis. We then explored and evaluated traditional machine learning algorithms such as logistic regression, decision tree, random forest and support vector machine for the predictive analysis. The performance of the models was evaluated using metrics such as accuracy, recall, precision and F1 score. This study provides a proof-of-concept for using machine learning algorithms to predict the diagnosis of heart disease. The results of this study could be used by healthcare providers and medical practitioners for early diagnosis and treatment of heart disease. Future research will focus on exploring and evaluating other machine learning algorithms for improved accuracy and performance.

Index Terms: supervised machine learning; algorithms; K nearest neighbours;

Abbreviations

AS: Accuracy Score

ECG: Electrocardiogram

FS: F1 Score

HDL: High Density Lipoprotein

KNN: K Nearest Neighbours

LDL: Low Density Lipoprotein

PS: Precision Score

RS: Recall Score

SVM: Support Vector Machine

XGBoost: eXtreme Gradient Boosting

List of Tables

List of Figures

List of Equations

1. Introduction

Heart disease is a major public health challenge and a leading cause of death in many countries. It is caused by a variety of factors, including genetics, lifestyle, and environmental influences. Heart disease can affect blood flow to the heart, leading to serious complications such as stroke, heart attack, and heart failure.

Machine learning techniques can be used to analyse large amounts of data quickly and accurately, allowing for more accurate diagnoses and predictions. Machine learning algorithms can be used to identify patterns in data which can help detect and diagnose heart diseases. These algorithms can also be used to predict the future risk of developing heart diseases based on data such as past medical history, lifestyle, and genetics. In addition to being more accurate, machine learning techniques are also more cost-effective and less time consuming than traditional methods. For example, an electrocardiogram (ECG) can be used to measure the electrical activity in the heart, but it is costly and time consuming. On the other hand, machine learning algorithms can be used to analyse ECG data more quickly and accurately. Machine learning algorithms can also be used to identify patterns in other types of data including patient medical history, lifestyle, and genetics. This data can be used to identify individuals who may be at risk of developing heart diseases, allowing doctors to take preventative measures. Machine learning techniques are a more accurate and cost-effective method of diagnosing and predicting heart diseases. This technology can provide doctors with more accurate diagnoses and predictions and can help identify individuals who may be at risk. Machine learning algorithms can also be used to analyse large amounts of data quickly, saving both time and money.

Supervised machine learning is a form of artificial intelligence that uses data to learn how to solve a problem. It is an advanced form of data analysis that can identify patterns in data and to develop models that can accurately predict the outcome of a given situation. Supervised machine learning uses labelled data sets, which means that the data is labelled with the correct answer. This allows the machine learning algorithm to learn from the data and make predictions based on the input data. The algorithm can then be used to make predictions on new data points where the answer is not known. In the context of heart disease, supervised machine learning can be used to develop models that can accurately predict the risk of developing heart diseases. This is done by using data sets that contain information about various risk factors such as age, gender, lifestyle, genetic history and other factors. The data is then used to develop models that can accurately predict the risk of developing heart diseases. Using supervised machine learning, doctors and medical researchers can develop models that will accurately predict the risk of developing heart diseases. This can help to identify those who are at a higher risk of developing heart diseases and help to create appropriate prevention and treatment plans. Supervised machine learning is a powerful tool that can be used to identify patterns in data and to develop models that

can accurately predict the outcome of a given situation. It can be used to develop models that can accurately predict the risk of developing heart diseases, making it a valuable tool in the fight against heart disease.

The use of data and machine learning models to predict the risk of developing heart diseases has become increasingly prevalent in the healthcare industry. These models use a variety of data such as age, lifestyle, risk factors, and medical history to develop a comprehensive risk profile for each patient. This data is then used to train the machine learning models to accurately predict the risk of developing heart diseases. This allows healthcare professionals to identify patients who are at a higher risk of developing heart diseases and to provide them with the necessary preventative measures. The accuracy of these models has been proven in numerous studies and is now being used in clinical settings to identify at-risk patients. By utilizing these models, healthcare professionals can accurately identify those who are at a higher risk of developing heart diseases and provide them with the necessary preventative care. This is important for early detection and treatment of heart diseases, which can significantly improve patient outcomes. In addition to their use in clinical settings, these models are also being used to develop personalized treatment plans for patients. By analysing a patient's risk factors and medical history, the models can determine the best course of treatment for each individual patient. This allows healthcare professionals to provide personalized and effective care to each patient. Ultimately, the use of machine learning models to predict the risk of developing heart diseases has revolutionized the healthcare industry. By utilizing these models, healthcare professionals can identify at-risk patients and provide them with the necessary preventative care. Additionally, these models can be used to develop personalized treatment plans for patients, allowing them to receive the most effective care possible.

Supervised machine learning models have been used to detect early signs of heart diseases. These models analyse data to recognize patterns that can be indicative of the disease, allowing doctors to take timely action and reduce the risk of further complications. The data used by supervised machine learning models can be collected from multiple sources, including patient records, medical scans, and vital signs. By analysing this data, the models are able to detect patterns that may indicate the early stages of heart disease. This includes changes in the heart rate, blood pressure, and other vital signs. The models can also use data from medical scans to detect signs of damage to the heart, such as blocked arteries or signs of inflammation. By combining this data with patient records, the models can accurately detect early signs of heart diseases. The models can also be used to detect conditions that may lead to heart disease, such as high cholesterol or elevated blood sugar levels. By detecting these conditions early, doctors can take the necessary steps to prevent the development of a more serious form of the disease. Overall, supervised machine learning models can be used to accurately detect early signs of heart diseases. By combining data from multiple sources, the models can identify patterns that may indicate the presence of

heart diseases and alert the doctor to take the necessary actions. This can help to reduce the risk of developing a more serious form of the disease.

Supervised machine learning techniques are increasingly being used for diagnosing and predicting heart diseases. This is due to their ability to provide accurate and efficient predictions and diagnoses. These supervised machine learning techniques use data from past cases and patient records to learn patterns and trends related to illnesses. By doing this, they can detect subtle differences in the data that may indicate the presence of a disease. This allows for more precise predictions and diagnoses of heart diseases. Since the data used for predictions and diagnoses is from past cases, supervised machine learning techniques are also able to identify potential risks and warning signs that may be associated with the disease. This allows medical professionals to take precautionary measures to reduce the risk of developing a heart disease. Furthermore, these techniques can also provide insights into the causes and possible treatments of a heart disease. This helps medical professionals to develop personalized treatments for their patients. The use of supervised machine learning techniques for diagnosing and predicting heart diseases is becoming increasingly popular and is the preferred method for many medical professionals. This is because these techniques can provide accurate and efficient predictions and diagnoses. Furthermore, they can also help to identify potential risks and warning signs as well as provide insights into the causes and possible treatments of a heart disease. As a result, supervised machine learning techniques will continue to become increasingly important in helping to tackle the increasing prevalence of heart diseases.

In this paper we will be using various supervised machine learning algorithms to get Accuracy, Precision, Recall and F1 Score, on the basis of which we will be doing comparative study of the various supervised machine learning algorithms used. This comparative study will evaluate the effectiveness of the machine learning algorithms.

1.1 Performance Measures:

Performance measures are essential metrics used to evaluate the effectiveness of a machine learning algorithm. Performance measures are used to compare the performance of different models on a given data set and to determine which model provides the best results. Performance measures can be used to identify areas in which the model can be improved and to determine the overall accuracy of the model.

There are four machine learning classification model performance measures:

1. Accuracy Score
2. Precision Score
3. Recall Score
4. F1-Score

Accuracy Score: Accuracy Score is one of the most commonly used performance measures in machine learning. It is used to measure how accurately a model can predict the expected output. It is calculated by taking the number of correct predictions divided by the total number of predictions. A higher accuracy score indicates that the model is more accurate in predicting the expected output.

$$AS = \frac{TP + TN}{TP + FN + TN + FP}$$

Equation 1

Precision Score: Precision score is a performance measure used in machine learning to evaluate the accuracy of a model's predictions. It is a measure of the ratio of true positives (TP) to the sum of true positives and false positives (FP). Precision score is typically expressed as a percentage, with a higher percentage indicating a better model performance.

$$PS = \frac{TP}{FP + TP}$$

Equation 2

Recall Score: Recall Score is a performance measure used in machine learning that evaluates a model's ability to correctly identify relevant instances from a dataset. It is calculated by dividing the number of relevant instances correctly identified by the total number of relevant instances in the dataset. The higher the recall score, the better the performance of the model.

$$RS = \frac{TP}{FN + TP}$$

Equation 3

F1-Score: F1-score is a performance measure in machine learning that combines precision and recall into a single metric. It is often used to evaluate the performance of a classification model, as it takes both false positives and false negatives into account. The F1-score is the harmonic mean of precision and recall, where the best value is 1.0 and the worst value is 0.0. A model which has a high F1-score is considered to be a better model than one with a low F1-score. The F1-score is often used in conjunction with other performance measures such as accuracy, precision and recall. The F1-score is a good measure of a model's performance when there is an uneven class distribution. This is because it takes both false positives and false negatives into account, and gives more weight to the minority class. The F1-score is also useful when there is a need to weigh precision and recall equally.

$$F1 = \frac{2 * PS * RS}{PS + RS}$$

Equation 4

2. Methodology

In this paper, the authors used some models like Logistic Regression, Decision Tree, XGBoost, KNeighbors Classifier, Multinomial Naive Bayes, Bernoulli Naive Bayes, Random Forest Classifier & Neural Networks.

Logistic Regression model: Logistic Regression is a supervised learning methodology which is used to predict the output of a categorical dependent variable. It is used to solve classification problems and generates values that range from 0 to 1. It can be imported through the SKlearn library. In the real world, it can be used in areas such as movie reviews. For example, Logistic Regression can be used to predict if a movie is likely to be watched or not based on its review rating. It can also be applied to predict the likelihood of a certain audience being active for a given issue in the marketing world. Logistic Regression is a powerful tool for decision-making and can be used in a variety of applications.

$$P = \frac{e^{a+bx}}{1 + e^{a+bx}}$$

Equation 5

KNeighbors Classifier: K-Nearest Neighbour (KNN) is a supervised learning algorithm for both classification and regression tasks. It is considered to be a non-parametric algorithm, which means it does not assume any underlying distributions of the data. KNN works by storing the training dataset and applying the classification or regression on it during the time of prediction. In other words, the algorithm is "lazy", meaning it does not learn from the given training set. KNN is most effective when dealing with data sets which contain a large number of reviews, as it is robust to noisy training data. KNN is also useful when there is a need to predict values in regions where no data is available, as it is based on the closest data points. KNN can be used for both regression and classification tasks, and is considered to be one of the simplest yet most powerful machine learning algorithms.

$$d(x, y) = \sqrt{\Sigma_{i=1}^{n}(x_i - y_i)^2}$$

Equation 6

Naive Bayes Model: Naive Bayes is a supervised machine learning algorithm based on applying Bayes' theorem with strong (naive) independence assumptions between the features. It is a probabilistic classification algorithm used in various applications such as text classification, spam filtering, sentiment analysis and medical diagnosis. Naive Bayes model works on the basis of Bayes' theorem, which states that the probability of an event occurring is equal to the probability of the event given the prior knowledge, multiplied by the probability of prior knowledge. It is a simple yet powerful algorithm that can be used for various classification tasks.

$$P(c|x) = \frac{P(x|c)P(c)}{P(x)}$$

Equation 7

Support Vector Machine: Support Vector Machines (SVMs) are a type of supervised machine learning algorithm used for classification and regression tasks. SVMs are based on the concept of finding a hyperplane that best divides a dataset into two classes. The optimal hyperplane is determined by maximizing the distance between the two classes, known as the margin. SVMs are powerful algorithms that can be used for both linear and non-linear classification. SVMs are used for many different tasks. One of the most common applications is to classify data points into two classes. This can be done by finding the optimal hyperplane that best separates the two classes. Another common application is regression, where the goal is to predict a continuous value. This is done by training an SVM to find the function that best fits the data.

$$f(x) = w^T x + b$$

Equation 8

Decision Tree: Decision trees are a type of supervised machine learning algorithms that are used to classify and categorise data. They are based on a set of rules that are used to determine the outcome of a given input. A decision tree is a visual representation of the conditions and possible outcomes of a decision-making process. It consists of nodes and branches, where each node represents a certain decision or outcome, and each branch represents a certain condition that must be true in order for the decision to be taken. The process of building a decision tree involves setting up a set of rules that will guide the decision-making process. These rules are based on data that has been collected, such as historical data, or data from surveys or experiments. The rules are then used to predict the outcome of a given input.

$$Gini = 1 - \Sigma_{i=1}^{c}(p_i)^2$$

Equation 9

XGBoost: Xgboost stands for extreme gradient boosting, and it is an advanced implementation of the gradient boosting algorithm. It is a powerful machine learning algorithm used for predictive modeling and classification problems. Xgboost has become popular due to its superior performance in terms of accuracy and speed compared to other machine learning algorithms. Xgboost works by building an ensemble of weak learners, each of which is a decision tree. The decision trees are trained using gradient descent to minimize the error of the model. The model uses a regularization technique known as regularization, which helps to prevent overfitting. The model also uses a tree-based learning algorithm called boosting, which helps to improve the accuracy of the model by combining the prediction of multiple weak learners.

$$f(x) = f(a) + f'^{(a)}(x-a) + 0.5(f``(a)(x-a)^2)$$

Equation 10

Random Forest: Random Forest is an ensemble machine learning algorithm that utilizes multiple decision trees to make predictions. It is an effective method for both classification and regression tasks. The algorithm works by randomly selecting a subset of features from the training dataset and then building individual decision trees using these features. The individual trees are then combined to create a "forest" of trees that are then combined to make a prediction. The predictions are made by averaging the predictions of each individual tree. The main benefit of random forest is that it is not prone to the overfitting problem found in other algorithms. This is because of its ensemble nature, in which it combines the results of multiple models to make the final prediction. Additionally, it is able to handle large datasets with high dimensionality. Random Forest is also a fast algorithm, requiring only a few lines of code to implement. This makes it highly suitable for large-scale tasks as it can be trained quickly and efficiently.

$$RFfi_i = \frac{\sum_j normfi_{ij}}{\sum_{j \in all\ features, k \in all\ trees} normfi_{jk}}$$

Equation 11

Neural Networks: Neural networks are a form of artificial intelligence that simulates the way the human brain works. They are composed of neurons, or nodes, that are connected together in layers and process information through a series of mathematical equations. Each neuron takes in input from other neurons in the network and processes it according to a set of weights and biases. The output of each neuron is then sent to the next layer of neurons. Based on the input, each neuron's output is updated,

creating a chain of interconnected “neurons” that work together to process information. By “training” the network with large amounts of data, the neurons can learn to recognize patterns and make predictions. Neural networks are used in a variety of applications, from image recognition and natural language processing to autonomous driving and medical diagnosis. They are useful for complex tasks that require the processing of large amounts of data and are often more accurate than traditional algorithms.

3. Datasets

The given data is in text format in csv files. While extraction, a comma-separated values (CSV) file saved in a tabular format of the data in a plain text. Every line in the given file is a data information. For this Project, we trained on datasets containing 303 people's information about their heart disease issues. Each Data Set contain 14 attributes. They are

1.Age: The duration of something from its beginning up to the present. This attribute contains age of each person which ranges from 0 to 100. It plays a major role in heart disease prediction because most of the people who die due to heart disease are 60 and older.

2.Sex: Sex refers to the social and cultural distinctions between people who are male or female. This attribute contains sex of each person where 1 indicates male and 0 indicates female. It attributes plays a role in 2 different ways.

- Diabetes: Females are more likely to have heart problems than males.
- No Diabetes: Males are more likely to have heart problems than females.

3.Chest Pain: Chest pain is discomfort or pain felt in the chest area. It can be a symptom of various medical conditions, such as heart attack, angina, or pneumonia. The pain may be sharp or dull, and may be felt in the center of the chest or in a specific area.

This attribute contains the type of chest pain experienced by every person. There are 4 types of chest pains.

1. Typical Angina - indicates 0
2. Atypical Angina - indicates 1
3. Non-Anginal pain - indicates 2
4. Asymptotic - indicates 3

So, these attribute ranges from 0 to 3.

4.Resting Blood Pressure: Having high blood pressure can be detrimental, as it can damage the arteries that supply blood to the heart. If high blood pressure is accompanied by other health conditions such as obesity, high cholesterol or diabetes, the risk is even higher.

This attribute contains the resting blood pressure value of every person in mmHg(unit).

5.Serum Cholesterol: Having a high amount of LDL cholesterol can lead to clogged arteries, whereas having a high amount of triglycerides, which are associated with diet, increases the risk of a heart attack. On the other hand, having a high level of HDL cholesterol, which is beneficial, reduces the risk of a heart attack.

This attribute contains the serum Cholesterol of every person in mg/dl(unit).

6.Fasting Blood Sugar: Having insufficient insulin secretion or an inability to effectively utilize insulin from your pancreas can cause your blood sugar levels to rise, putting you at greater risk of having a heart attack.

This attribute compares Fasting Blood Sugar value of every person with 120mg/dl. This attribute contains 0's and 1's, where 1 indicates fasting blood sugar > 120mg/dl and 0 indicates fasting blood sugar <= 120mg/dl.

7.Resting ECG: The U.S. Preventive Services Task Force determined that there is moderate certainty that the potential risks of screening with resting or exercise ECG outweigh the potential benefits for people at low risk of cardiovascular disease. For people at intermediate to high risk, the evidence is not sufficient to assess the balance of benefits and harms of screening.

This attribute contains the results of Resting ECG of every person which range from 0 to 2 where

1. Normal - indicates 0
2. having ST-T wave abnormality - indicates 1
3. left ventricular hypertrophy - indicates 2

8.Max heart rate achieved: An increase in heart rate by 10 beats per minute was associated with a 20% rise in the risk of cardiac death, which is similar to the increase in risk associated with a 10 mmHg rise in systolic blood pressure. This attribute contains the maximum heart rate achieved by every person.

9.Exercise induced angina: Angina is a type of chest pain that is usually experienced as a tight, gripping or squeezing sensation and can range from mild to severe. It typically occurs in the center of the chest, but may also spread to the shoulders, back, neck, jaw or arms and even the hands. There are four main types of anginas: stable angina/angina pectoris, unstable angina, variant (Prinz metal) angina, and microvascular angina. This attribute says whether person has Exercise induced angina or not. It contains 0's and 1's were

1. Having Exercise induced angina - indicates 1
2. No Exercise induced angina - indicates 0

10.ST depression induced by exercise relative to rest: ST depression induced by exercise refers to a decrease in the ST segment of the ECG (electrocardiogram) compared to the ST segment of the ECG at rest. This decrease is typically seen during exercise stress tests and is a sign of myocardial ischemia, which is a decrease in blood flow to the heart caused by an occlusion in the coronary arteries.

11.Peak exercise ST segment: Peak exercise ST segment is a measurement of electrical activity in the heart during exercise. It is measured by an electrocardiogram and is used to detect signs of ischemia, or reduced blood flow to the heart muscle. During exercise, the ST segment may become elevated, flattened, or depressed, which can indicate a decrease in blood flow to the heart.

1. Upsloping – indicates 0
2. flat – indicates 1
3. downsloping – indicates 2

12.Number of major vessels (0–3) colored by fluoroscopy: Fluoroscopy is a type of medical imaging that uses X-rays to obtain real-time moving images of the internal structures of a patient. It is used to diagnose and treat diseases and medical conditions. It is also used to guide and monitor the progress of medical procedures such as catheter placement and biopsies.

1. Normal – indicates 0
2. Mild Abnormality – indicates 1
3. Moderate Abnormality – indicates 2
4. Severe Abnormality – indicates 3

13.Thal: Thalassemia is an inherited blood disorder. It affects the body's ability to produce hemoglobin, a protein in red blood cells that carries oxygen to other parts of the body. People with thalassemia make either no hemoglobin or too little hemoglobin, which can lead to anemia (low red blood cell count) and other serious health problems.

1. No Thalassemia – indicates 0
2. Mild Thalassemia – indicates 1
3. Moderate Thalassemia – indicates 2
4. Severe Thalassemia – indicates 3

14.Target: This attribute contains 0's and 1's where 1 indicates a person is suffering with heart disease and 0 indicates absence of heart disease.

As we are applying supervised Machine Learning models, here my Target attribute acts as a dependent variable and rest of the attributes are independent variables. Here, we are making Comparative study on each performance measure i.e.,

Accuracy Score, Precision Score, Recall Score and F1-Score by taking different Train-Test Split Ratios on Data Sets. With this, we can apply all the proposed seven machine learning algorithms and Neural Network on the given data set to get the accuracy of each and every model.

Model	Performance Measure	Train and Test Split		
		80 : 20	**67 : 33**	**50 : 50**
Logistic Regression	Accuracy Score	85.25	83.00	81.58
	Precision Score	88.24	88.46	91.57
	Recall Score	85.71	80.70	78.35
	F1-Score	86.96	84.40	84.44
Naïve Bayes	Accuracy Score	85.25	81.00	80.26
	Precision Score	91.18	86.54	86.75
	Recall Score	83.78	78.95	79.12
	F1-Score	87.32	82.57	82.76
SVM	Accuracy Score	81.97	81.00	78.95
	Precision Score	88.24	86.54	86.75
	Recall Score	81.08	78.95	77.42
	F1-Score	84.51	82.57	81.82
KNeighbors Classifier	Accuracy Score	67.21	68.00	67.11
	Precision Score	67.65	73.08	73.49
	Recall Score	71.88	67.86	68.54
	F1-Score	69.70	70.37	70.93
Decision Tree	Accuracy Score	81.97	80.00	75.66
	Precision Score	82.35	78.85	77.11
	Recall Score	84.85	82.00	78.05
	F1-Score	83.58	80.39	77.58
	Accuracy Score	90.16	87.00	83.55

Random Forest Classifier	Precision Score	94.12	90.38	91.57
	Recall Score	88.89	85.45	80.85
	F1-Score	91.43	87.85	85.88
XGBoost	Accuracy Score	85.25	81.00	75.66
	Precision Score	88.24	86.54	85.54
	Recall Score	85.71	78.95	73.96
	F1-Score	86.96	82.57	79.33
Neural Network	Accuracy Score	80.33	81.00	76.32
	Precision Score	82.35	90.38	84.34
	Recall Score	82.35	77.05	75.27
	F1-Score	82.35	83.19	79.55

Table 1

4. Conclusion:

In this Project, after training the seven machine learning algorithms and Neural Network, we observed that Random Forest Classifier is giving the high values for all the performance measures, i.e., Accuracy Score, Precision Score, Recall Score and F1-Score.

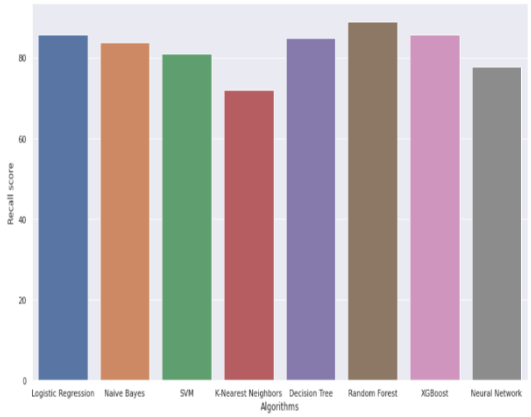

Figure 1

We also observed that, when train ratio decreases or test ratio increases, then the scores of all performance measures decreases. In Random Forest, when constructing trees, a randomly chosen subset of features is used instead of searching for the most important feature when splitting a node. This adds a level of randomness to the model, which reduces the overfitting problem of decision trees and decreases the variance, thereby boosting the accuracy.

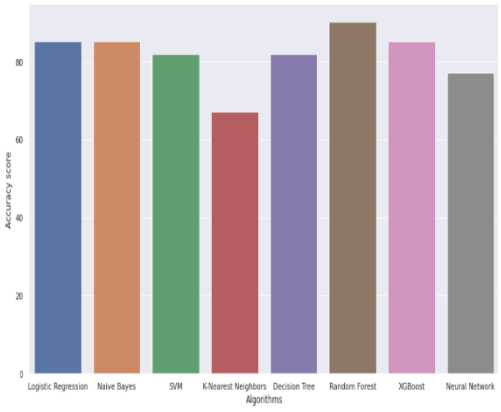

Figure 2

When the training data decreases, the model will have less information to learn from, resulting in a lower accuracy on unseen data (the test set). On the other hand, when the test set increases, the model will need to be more precise in its predictions, resulting in a lower accuracy score. This is because the model will now have more data to make predictions on, meaning that it needs to be more precise in its predictions in order to correctly classify the new data.

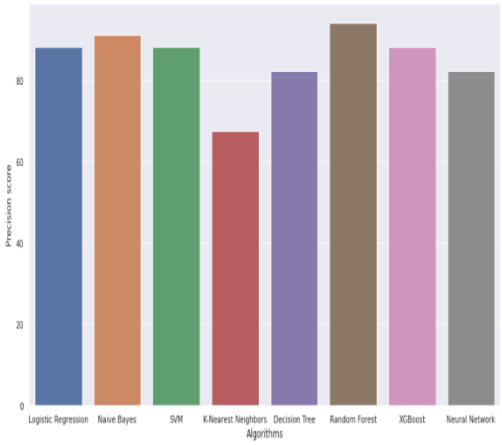

Figure 3

After Rainforest, XGBoost and Naïve Bayes are giving better accuracy compared to Neural Networks. For the predicting heart disease, KNN is giving the poor accuracy compared to all other Machine Learning Models and Neural Networks. Finally, we can conclude that Random Forest Classifier is giving the best Scores(Accuracy, Precision, Recall, F1) for the Heart Disease Prediction compared with other Machine Learning Models and Neural Network.

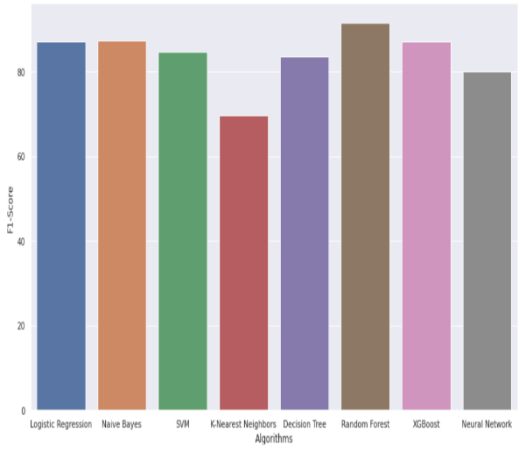

Figure 4

5. Future Work:

As a next step in this field of research, some promising directions could be explored. Investigating the impact of incorporating different types of data sources on the accuracy of heart disease prediction. For instance, incorporating environmental and lifestyle factors, such as air pollution, physical activity, and diet, into the predictive models could lead to more accurate predictions. Developing more sophisticated machine learning techniques, such as ensemble methods, to improve the accuracy of heart disease predictions. Using transfer learning techniques to leverage existing models and datasets to build more accurate and efficient predictive models. Developing deep learning architectures for heart disease prediction. Enhancing the interpretability of predictive models by using techniques, such as feature importance and sensitivity analysis. Investigating the potential of using natural language processing techniques to extract and analyse medical records for heart disease prediction. Exploring the use of reinforcement learning techniques for heart disease prediction. Investigating the impact of utilizing large-scale medical records and electronic health records on the accuracy of heart disease predictions. It would also be beneficial to consider the use of unsupervised learning techniques to identify hidden patterns in the data, such as clustering or anomaly detection. Another possibility would be to compare the results of different neural network architectures, such as convolutional neural networks, recurrent neural networks, and generative adversarial networks, in order to determine which is the most effective for heart disease prediction. Additionally, it would be interesting to explore the use of transfer learning, which could enable the model to leverage knowledge from existing models in the domain to improve its prediction accuracy. Finally, it would be beneficial to investigate the impact of incorporating additional sources of data, such as medical images and patient records, into the model to further improve its accuracy.

References :

[1] M. Akhil Jabbar, Priti Chandrab and B.L Deekshatuluc, "Heart Disease Prediction System using Associative Classification and Genetic Algorithm", *ICECIT*, 2012.

[2] Syed Umar Amin, Kavita Agarwal and Rizwan Beg, "Genetic Neural Network Based Data Mining in Prediction of Heart Disease Using Risk Factors", *ICT*, 2013.

[3] Amma N.G. Bhuvaneswari, Cardiovascular Disease Prediction System using Genetic Algorithm and Neural Network, IEEE, 2012.

[4] Aditi Gavhane, Gouthami Kokkula, Isha Pandya and Kailas Devadkar, "Prediction of Heart Disease using Machine Learning", *ICECA*, 2019.

[5] Nilakshi P. Waghulde and Nilima P. Patil, *Genetic Neural Approach for Heart Disease Prediction*, vol. 4, no. 3, sep 2014.

[6] William G. Bat, "Use of an Artificial Neural Network for Data Analysis in Clinical Decision-Making: The Diagnosis of Acute Coronary Occlusio", *MITP*, vol. 2, no. 4, 1990.

[7] V. Krishnaiah, G. Narsimha and N. Subhash Chandra, "Heart Disease Prediction System Using Data Mining Technique by Fuzzy K-NN Approach", *Springer International Publishing*, vol. 1, 2015.

[8] Yanwei Xing, Jie Wang, Zjijong and Yonghong Gao, Combination Data Mining Methods with New Medical Data to Predicting Outcome of Coronary Heart Disease, IEEE, 2007.

[9] Krishnan. J Santhana and S. Geetha, Prediction of Heart Disease Using Machine Learning Algorithms, IEEE, 2019.

[10] C. Kalaiselvi, Diagnosing of heart diseases using average k-nearest neighbor algorithm of data mining, IEEE, 2016.

[11] Jagdeep Singh, Amit Kamra and Harbhahg Singh, Prediction of heart diseases using associative classification, IEEE, 2016.