Heart Diseases Prediction

FORHAD ALI EMON

20-44178-2 @student.aiub.edu

*dept of Computer Science & Engineering*

*American International University of Bangladesh*

Dhaka, Bangladesh

***Abstract*— Predicting heart disease is one of the most challenging challenges in the medical industry today. About one person dies from heart disease every minute in the contemporary era. Processing a vast amount of data in the healthcare industry requires data science. Since predicting cardiac illness is a complicated undertaking, it is necessary to automate the process in order to reduce risks and warn the patient well in advance. The heart disease dataset from the UCI machine learning repository is used in this study. The suggested study uses a variety of data mining approaches, including Logistic Regression, Decision Tree, Support Vector Machine (SVM), and Naive Bayes algorithm, to forecast the likelihood of heart disease and categorize patient risk level. In order to compare the effectiveness of various machine learning algorithms, this article will do so. In comparison to other ML algorithms used, the trial results show that logistic regression has the greatest accuracy of 65.5%.**

1. INTRODUCTION

The human body's most important organ is the heart. Human Life is entirely dependent on the heart's effective operation. A healthy life starts with a healthy heart. But in today's environment, heart disease has emerged as a major factor in both male and female fatalities. The Corona Virus during Covid-19 causes several physical problems in people. Heart failure results from inflammation of the heart muscle caused by the corona virus. Despite respiratory symptoms, experimental research indicates that 1 in 5 patients had cardiac damage from the Corona virus. The most prevalent kind of heart disease is coronary heart disease. One in four fatalities, or around 630,000, are caused by heart disease. The patient's medical and family histories, a physical examination, and test findings are often the foundation of the diagnostic process for heart failure [1]. Due to several risk factors, including diabetes, high blood pressure, high cholesterol, an irregular pulse rate, and many other conditions, it can be challenging to diagnose heart disease [2]. As the most prevalent disease, heart disease requires urgently precise and early identification in order to save the lives of many individuals. There are various scanners available to find heart illness, but many lives can be saved by forecasting a heart problem before it manifests itself. By employing a tool that enables the admin to visually assess the patient's data, we are giving auxiliary information to the admin. [4] In order to save people from developing cardiac issues, it is crucial to spot and examine the occurrence of arrhythmia as soon as possible. In many instances, the existence of a stroke or heart failure may be caused by the small levels of cardiac rhythm. The healthcare sector has a lot of promise for machine learning to help health systems assess and diagnose diseases utilizing data through data mining. Habits , and high cholesterol; we improve this second group. Data mining's subsection of machine learning, which deals with huge, well formatted datasets, does so effectively. Machine learning may be used in the medical industry to diagnose, detect, and forecast a variety of illnesses. This paper's major objective is to give clinicians a tool for early heart disease detection. As a result, patients will receive appropriate care and serious repercussions will be avoided. The importance of machine learning (ML) in identifying hidden discrete patterns and analyzing the provided data is crucial. Following data analysis, machine learning approaches aid in the early detection and prediction of cardiac disease. In order to predict cardiac disease at an early stage, this research analyzes the performance of many ML approaches, including Naive Bayes, Decision Trees, Logistic Regression, and SVM.

1. CLASSIFICATIONS

# A. Logistic Regression:

Based on a given dataset of independent variables, logistic regression calculates the likelihood that an event will occur, such as voting or not voting. Given that the result is a probability, the dependent variable's range is 0 to 1. In logistic regression, a logit transformation is applied on the odds—that is, the probability of success divided by the probability of failure. This is also commonly known as the log odds, or the natural logarithm of odds, and this logistic function is represented by the following formulas:

Logit(pi) = 1/(1+ exp(-pi))

ln(pi/(1-pi)) = Beta\_0 + Beta\_1\*X\_1 + … + B\_k\*K\_k

# B. Decision Tree:

Decision Tree is a Supervised learning technique that can be used for both classification and Regression problems, but mostly it is preferred for solving Classification problems. It is a tree-structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome. A decision tree algorithm always tries to maximize the value of information gain, and a node/attribute having the highest information gain is split first. It can be calculated using the below formula:

Information Gain= Entropy(S)- [(Weighted Avg) \*Entropy

(each feature)

Entropy(s)= -P(yes)log2 P(yes)- P(no) log2 P(no)

# C. Support Vector Machine(SVM):

One of the most well-liked supervised learning algorithms, Support Vector Machine, or SVM, is used to solve Classification and Regression issues. However, it is largely employed in Machine Learning Classification issues. Classification and regression problems are resolved using Support Vector Machine, or SVM, one of the most used supervised learning techniques. It is mostly used, nevertheless, in Machine Learning Classification problems.

# D. Naïve Bayes Algorithm:

The Naive Bayes classification algorithm is a probabilistic classifier. It is based on probability models that incorporate strong independence assumptions. The independence assumptions often do not have an impact on reality. Therefore, they are considered as naive. It is mainly used in text classification that includes a high-dimensional training dataset. The following equation is,

P(A|B) = {P(B|A) \* P(A)}/P(B) (1)

Where,

P(A|B) is Posterior probability: Probability of hypothesis A on the observed event B.

P(B|A) is Likelihood probability: Probability of the evidence given that the probability of a hypothesis is true.

III. RELATED WORKS

Using the UCI Machine Learning dataset, much research has been done to predict cardiac disease. Varied data mining approaches have been used to achieve various accuracy levels, which are detailed below. Avinash Gelande and colleagues investigate several ML algorithms that can be used to categorize cardiac disease. An investigation was conducted to examine the accuracy of the classification algorithms Decision Tree, KNN, and K-Means[1]. The study found that Decision Trees had the best accuracy, and it was concluded that by combining various methodologies and fine-tuning its parameters, it might be made more effective.

A system that combined the MapReduce algorithm with data mining techniques has been suggested by T. Nagamani et al. For the 45 instances in the testing set, the accuracy obtained according to this article was higher than the accuracy obtained using a traditional fuzzy artificial neural network. Here, the usage of dynamic schema and linear scaling increased the algorithm's accuracy.

A system that employs NB (Naive Bayesian) approaches for dataset categorization and the AES (Advanced Encryption Standard) algorithm for safe data transport was proposed by Anjan Nikhil Repaka, et al., in.

After reading the mentioned studies, the fundamental concept behind the suggested system was to build a heart disease prediction system based on the inputs. In order to determine the best classification algorithm that may be used to predict heart disease, we compared the classification algorithms Logistic Regression, Decision Tree, Support Vector Machine (SVM), and Naive Bayes based on their Accuracy, Precision, Recall, and f-measure scores.

IV. METHODOLOGY

We examined four machine learning methods that are extremely accurate to this specific forecasting model: Logistic Regression, Decision Tree, Support Vector Machine (SVM), and Naive Bayes classification algorithm. The suggested method is structured as follows: data collection comes first, considerable value extraction comes next, and data exploration comes third. Data preparation addresses missing values, data cleansing, and standardization, depending on the techniques utilized. After the pre-processed data has been preprocessed, the classifier utilized in the suggested models is employed to identify the pre-processed data. Later, we tested the suggested model and assessed its performance and accuracy using a range of performance indicators. To test the model, we used 50% of the whole dataset.

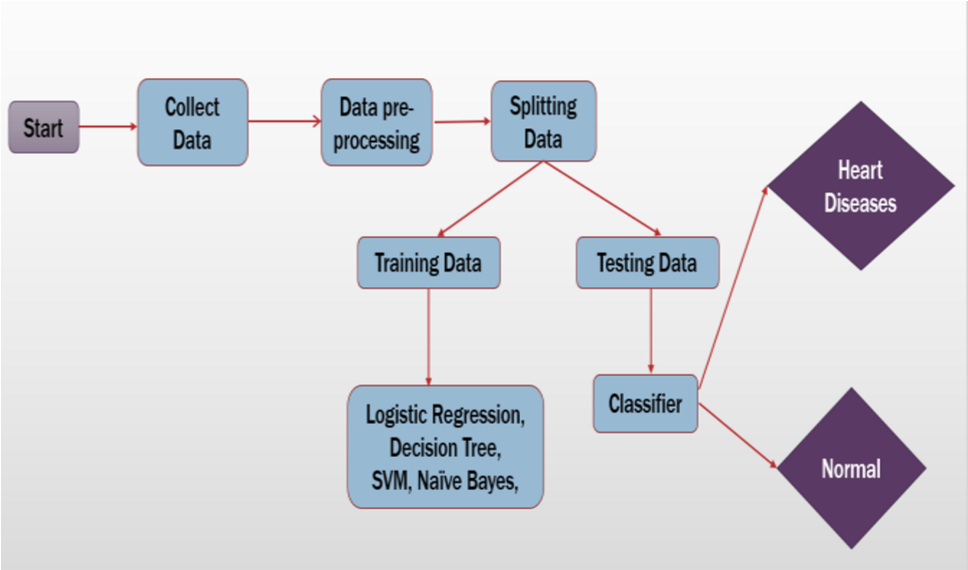


Fig 1: Model of Methods

# A. Data Source

To train the model, we imported the UCI dataset [4]. The information was taken from many sources. The UCI repository's database is used [5]. A subset of 14 attributes were utilized instead of the original 76 attributes. The collection contains information on a wide range of people, including their medical histories and histories of heart disease. The dataset comprises of the medical histories of 303 distinct people with a variety of features. The patient's medical features, including age, the type of chest discomfort experienced, blood pressure, sugar levels, angina, and other factors, are well-detailed in this dataset, allowing us to determine whether or not the patient has been given a heart disease diagnosis. The following characteristics are listed:

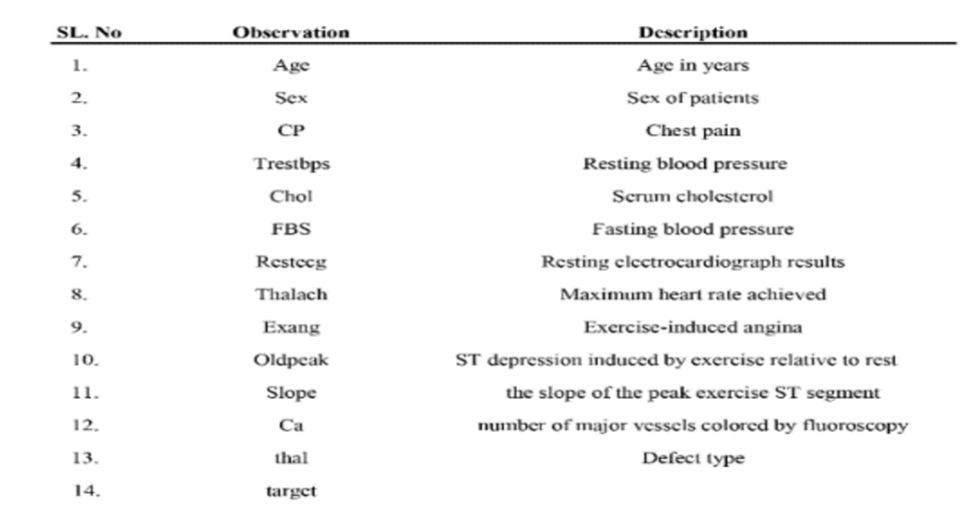


Fig : Feature in dataset

V. RESULT AND ANALYSIS

This section displays the outcomes of the applications of Logistic Regression, Decision Tree, Support Vector Machine (SVM), and Naive Bayes. The metrics Accuracy score, Precision (P), Recall (R), and F-measure are used to analyze the algorithm's performance. The accurate measure of positive analysis is provided by the precision metric (stated in equation (2)). Recall [stated in equation (3)] specifies the quantity of genuine correct positives. Equation (4)'s F-measure is an accuracy test.

Precision = (TP) / (TP +FP) (2)

Recall = (TP) / (TP+FN) (3)

F- Measure = (2 \* Precision \* Recall) / (Precision +Recall) (4)

TP True positive: the patient has the disease and the test is positive.

FP False positive: the patient does not have the disease but the test is positive.

TN True negative: the patient does not have the disease and the test is negative.

FN False negative: the patient has the disease but the test is negative.

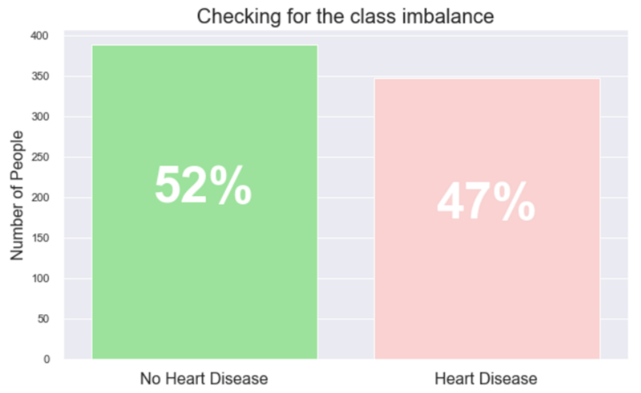
In our result we obtained,

Actual 0 = True negative

Actual 1= True positive

Prediction 0 = False negative

Prediction 1 = False positive



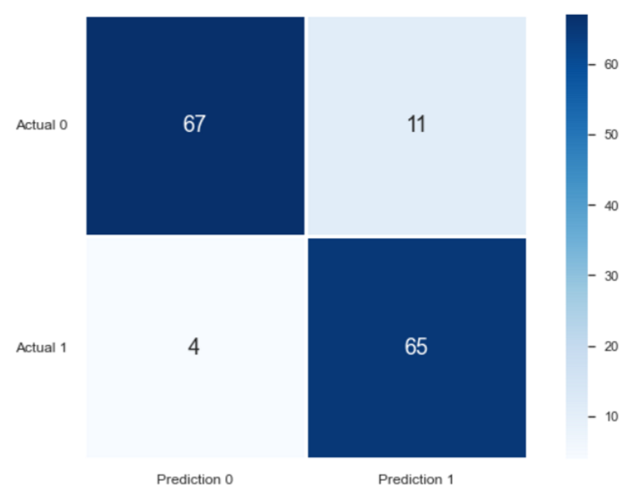


Fig 2: Confusion Matix Fig 3: Class imbalance

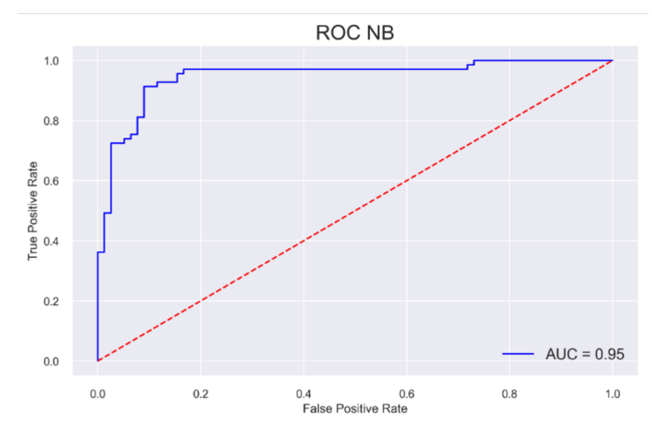
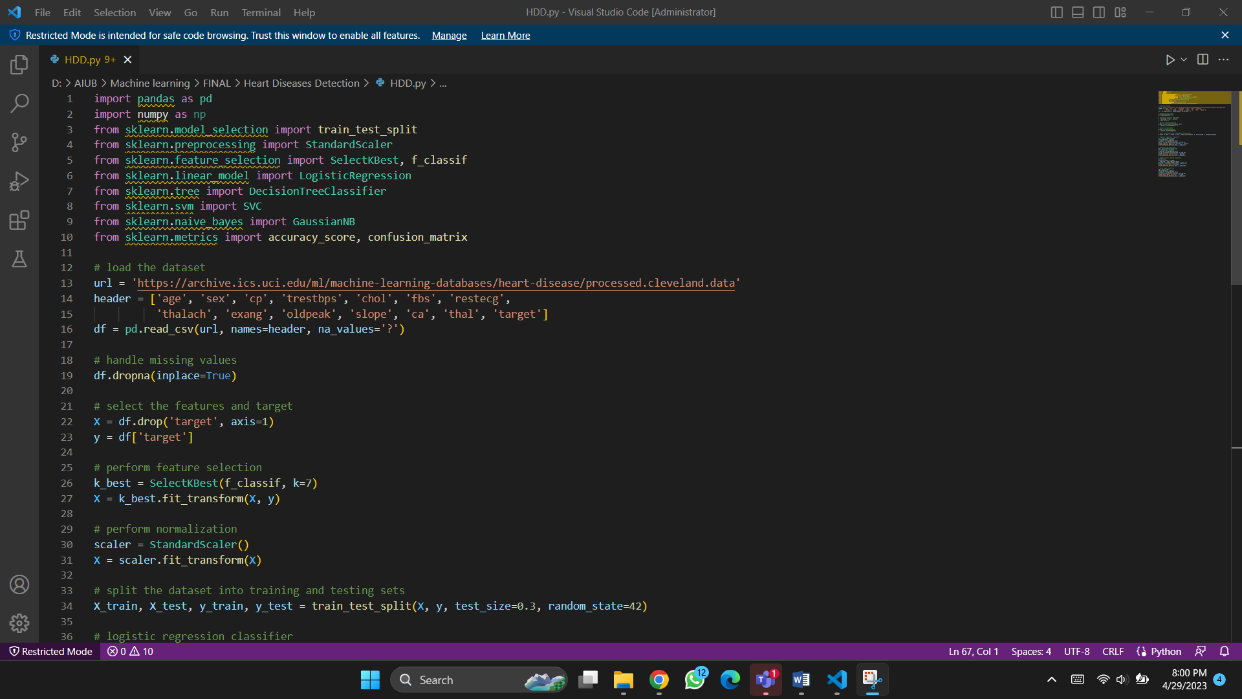


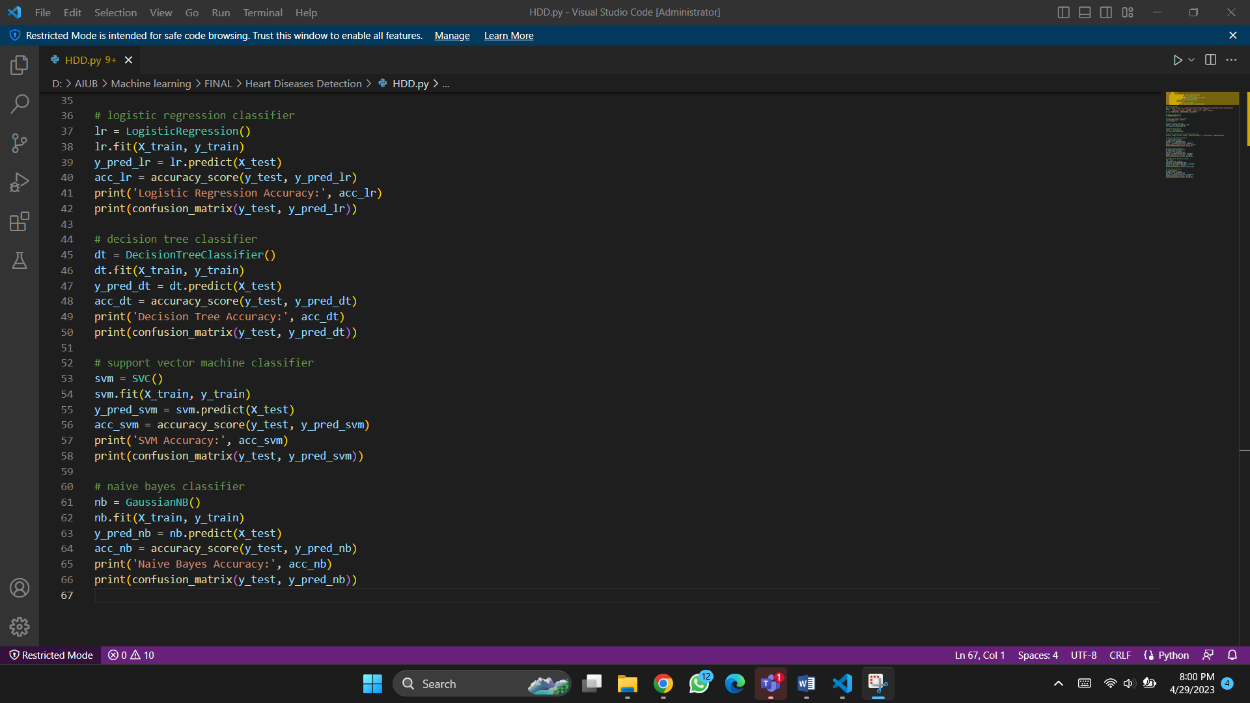
Fig 4: ROC NB

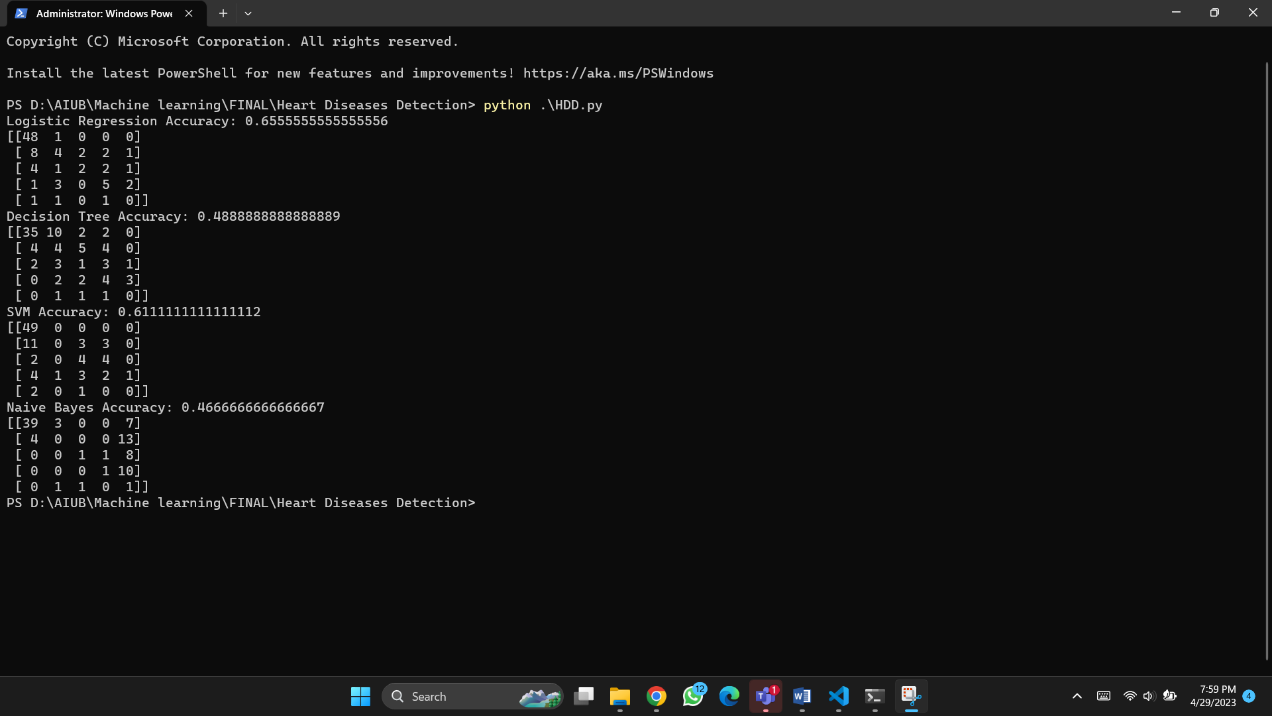
The above-mentioned performance metrics are obtained using the confusion matrix. Confusion Matrix describes the performance of the model. The confusion matrix obtained by the proposed model for different algorithms is shown above.

The accuracy score obtained for Logistic Regression, Decision Tree, Support Vector Machine (SVM) and Naive Baves classification techniques is shown above in Fig 2.

1. CODE IMPLEMENTED







1. CONCLUSION

It is essential to create a system that can forecast heart illnesses precisely and effectively given the rise in fatalities caused by heart diseases. Finding the best effective ML algorithm for heart disease diagnosis was the study's driving force. The UCI machine learning repository dataset is used in this work to examine the accuracy scores of the Logistic Regression, Decision Tree, Support Vector Machine (SVM), and Naive Bayes algorithms for predicting heart disease. The outcome of this study shows that the Support Vector Machine (SVM) algorithm is the most effective one for predicting heart disease, with an accuracy score of 65.5%.

1. FUTURE WORK

The study may be improved in the future by creating a web application based on the Support Vector Machine (SVM) and utilizing a larger dataset than the one used in this research. This would assist to deliver better findings and aid medical professionals in accurately and efficiently predicting cardiac disease.

1. ACKNOWLEDGMENT

First of all, we want to express our gratitude to our esteemed course instructor, PROF. DR. MD. ASRAF ALI of American International University Bangladesh (AIUB), for his ongoing direction and assistance with this project work.

1. REFERENCES
2. Avinash Golande, Pavan Kumar T, ”Heart Disease Prediction Using Effective Machine Learning Techniques”, International Journal of Recent Technology and Engineering, Vol 8, pp.944-950,2019.

1. T.Nagamani, S.Logeswari, B.Gomathy,” Heart Disease Prediction using Data Mining with Mapreduce Algorithm”, International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-8 Issue-3, January 2019.
2. World Health Organization. (n.d.). Cardiovascular diseases. World

Health Organization. https://www.who.int/healthtopics/cardiovasculardiseases/.

1. https://papers.ssrn.com/sol3/papers.cfm?abstract\_id=3759562
2. Ronit, R. (2018, June 25). Heart Disease UCI. Kaggle. https://www.kaggle.com/ronitf/heart-disease-uci.
3. C. S. Dangare and S. S. Apte, “Improved study of heart disease predictionsystem using data mining classification techniques,”

InternationalJournal of Computer Applications, vol. 47, no. 10, pp. 44–

48, 2012

1. Sayali Ambekar, Rashmi Phalnikar,“Disease Risk Prediction by Using Convolutional Neural Network”,2018 Fourth International Conference on Computing Communication Control and Automation.
2. C. B. Rjeily, G. Badr, E. Hassani, A. H., and E. Andres, ―Medical
3. Data Mining for Heart Diseases and the Future of Sequential Mining Medical Field,‖ in Machine Learning Paradigms, 2019, pp. 71–99.
4. Jafar Alzubi, Anand Nayyar, Akshi Kumar. "Machine Learning from Theory to Algorithms: An Overview", Journal of Physics: Conference Series, 2018
5. Palaniappan, S., & Awang, R. (2008). Intelligent heart disease prediction system using data mining techniques. 2008 IEEE/ACS International Conference on Computer Systems and Applications.

https://doi.org/10.1109/aiccsa.2008.4493524