Predicting Heart Disease Using Machine Learning

Adebukola Catherine Aladeyelu
Computer Science and Quantitative Methods
Austin Peay State University
Clarksville, Tennessee, United States
aaladeyelu @my.apsu.edu

Gabriel Temidayo Adekunle
Computer Science and Quantitative Methods
Austin Peay State University
Clarksville, Tennessee, United States
gadekunle@my.apsu.edu

***Abstract*—Heart disease, also referred to as cardiac arrest, is a serious health condition that has emerged as a major concern for human beings.**

The five algorithms that were compared in this project include Logistic Regression, Decision Tree, Naive Bayes, K-Nearest Neighbors, and Support Vector Machine. These algorithms were evaluated based on their accuracy scores, as well as other metrics such as sensitivity, specificity, and F1 score. After analyzing the results, the Support Vector Machine had the highest accuracy score and outperformed all other models in all evaluation metrics.

It can be deduced that the Support Vector Machine algorithm is the best model for forecasting the existence of heart disease.

Keywords—Logistic Regression, Decision Tree, Naive Bayes, K-Nearest Neighbors, and Support Vector Machine

1. INTRODUCTION

Heart disease, which is also known as a cardiac arrest has become alarming among human being. According to statistics by WHO (World Health Organization), it shows that 17.9 million people die from heart disease which 31% of all global death (WHO, 2017). Most people that are diagnosed with this illness are not aware because of irregular checkup practice. This illness becomes life-threatening when not monitored and not diagnosed on time by the patient.

More so, the evolvement in technology has brought diverse opportunities in the different spheres where the knowledge of computer science and technology can be utilized to solve problems. It is very difficult to

diagnosis a patient with a high risk of heart disease because of the various factors contributing to the causes of it. This is where machine learning is useful to the medical field and its relevance is proven (Shah, Patel, & Bharti, 2020).

In this project, we will be comparing various machine learning algorithms such as Logistic Regression, Support Vector Machine, Decision Tree, KNN, Neural Network, XG Boost algorithms to predict the heart disease rate. The Python programming language will be used, which has become very popular. Programming language and is being used by different industries like Technology companies, Finance, Government, Schools, and many others. And we will conduct Exploratory Data Analysis (EDA). The following are the main objectives of the project:

- To determine the factor that causes the chances of being at more risk of heart disease.
- To classify those with less heart disease and those with more risk of heart disease.
- Comparing different machine learning algorithms to predict the heart attack rate.

1.1 MOTIVATION

Machine Learning is an advanced tool which is commonly used in various fields because of its importance, and it solves a lot of problems. The existence of machine learning innovation, creativity, and strength gives room for solving problems in the medical field. Also, advancement in medical data has been unique, but the information provided is not better compared to the old paper charts they replaced. The medical records can be proficient by using a Machine learning algorithm to perform the analysis. Making use of this machine learning technology, better information can be provided to doctors at the point of patient care. Where the patient can easily go for a checkup where their blood pressure and other important signs are being checked and hopefully get the information accurately. The main issue faced by the medical field is heart disease because a lot of parameters and technicalities are involved for accurate prediction of the disease (Sharma & Rizvi, 2017). Machine learning has various algorithms which help in detecting heart disease. Machine learning algorithms make use of old patient records to predict a new patient with any occurrence of heart disease. If machine learning algorithms can make an accurate prediction, then lives can be saved with appropriate medication. Advancement in technology and application of machine learning to the medical field and effective research conducted has helped whole a lot. This has influenced the medical field accept the use of machine learning models for heart disease detection.

2. LITERATURE REVIEW

Machine learning makes use of artificial intelligence that helps systems to learn on their own and excel from experience without being explicitly programmed (Kalali et al, 2019). Artificial intelligence (AI) helps to produce smart machines that can solve difficult problems and interact with humans. It uses algorithms to enable computers to acquire and show intelligent behavior. It can emulate human abilities and perform intelligent behaviors by learning from data patterns. Since data changes dynamically and AI can start autonomously, it can collect data efficiently and accurately without human intervention (Arogundade, 2023). It has become one of the useful fields in our modern society and has caused a great advancement in technology. Machine learning algorithms can be applied in various spheres of life. There are three different machine learning techniques (VanderPlas, 2016):

- supervised learning methods use training data to build a model, which is subsequently applied to additional data.
- unsupervised methods seek relationships among data points that can be leveraged to construct a model that is subsequently applied to the data of interest.
- While reinforcement learning lies in between supervised and unsupervised learning.

This project aims to compare different machine learning algorithms on a classification dataset and choose the best to predict the existence of heart disease. The new development in the field of medical science with the help of various machine learning models, algorithms, and discoveries has been accomplished in these recent years releasing the important papers. A paper by (Sultana, Haider, & Uddin, 2016) proposed that data mining techniques such as WEKA play a great role in heart disease prediction.

Various strategies have their benefits and negative marks in work done by (Jabbar, Deekshatulu, & Chandra, 2013), and optimization of features has been done to achieve higher classification efficiency in the Decision Tree. It is a method for the early diagnosis of heart disease by using various highlights.

3. DATA

The data set was obtained from the Kaggle website, the aim is to compare and predict whether the patient has heart disease. This dataset contains 14 features, 303 records. The "target" feature refers to the existence of heart disease or none in the patient. It is represented by 1 and 0 respectively.

Features:

1) Age

2) Sex

3) Chest pain type

4) Resting blood pressure

5) Serum cholesterol

6) Fasting blood sugar

7) Resting electrocardiographic results

8) Maximum heart rate achieved

9) Exercise-induced angina

10) Old peak = ST depression induced by exercise relative to rest [6]

11) The slope of the peak exercise ST segment [6]

12) Number of major vessels (0-3) colored by fluoroscopy

13) Thalassemia

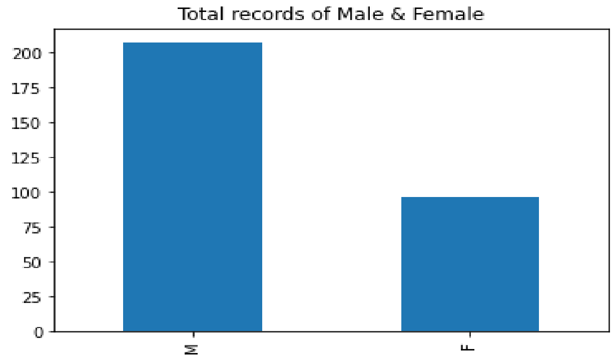
14) target: 0= less chance of heart attack 1= more chance of heart attack

4. METHODS

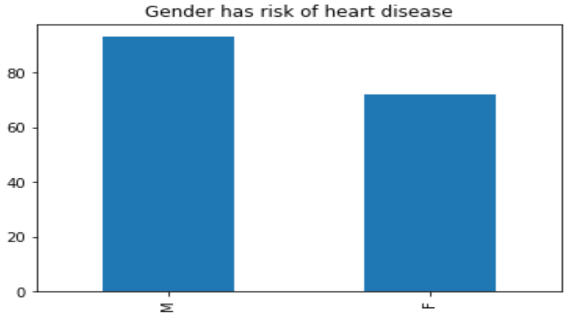
Dealing with a medical dataset needs to be categorized by age to give us a better insight and understanding. I will be categorizing the age based on seniority, middle age, adult, young adult, child, a teenager so we can understand the affected age groups.

- 'S' - Senior Citizen (Age 60 and above)
- 'MA' - Middle Aged (45 - 60)
- 'A' - Adult (30 - 45)
- 'YA' - Young Adult (20-30)
- 'T' - Teenager (12-19)
- 'C' - Child (1-12)

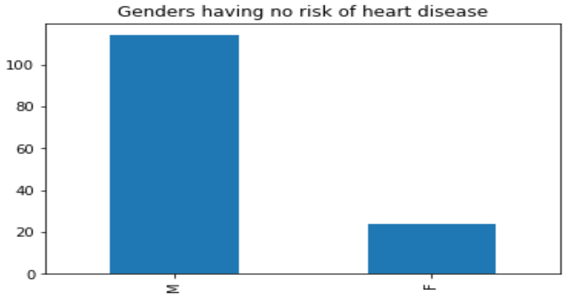
The dataset acquired from Kaggle needs to be cleaned so that the data set can be used for building the models. I perform data visualization and check the data description and statistics but there was no missing value or null. So, I went ahead to carry out my analysis and model training. Exploratory of data analysis are as follows:



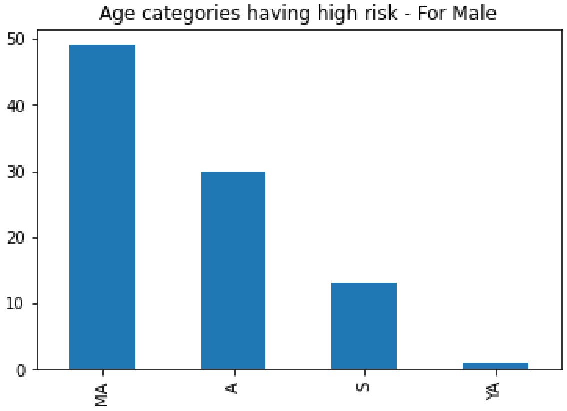
Conclusion: From the data visualization above, it shows the male gender has more records than the female.



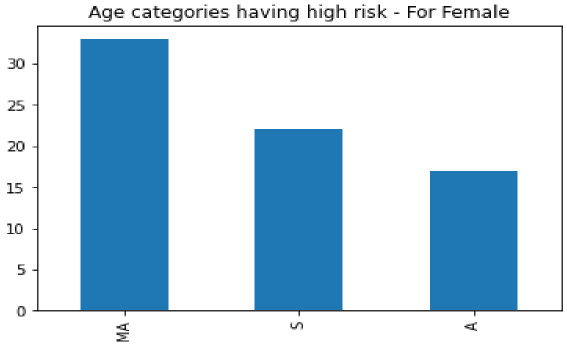
Conclusion: From the diagram above it deduce that the female has a greater chance of heart attack when compared to the total numbers of people.



Conclusion: From the diagram above it shows the male has less chance of heart disease.

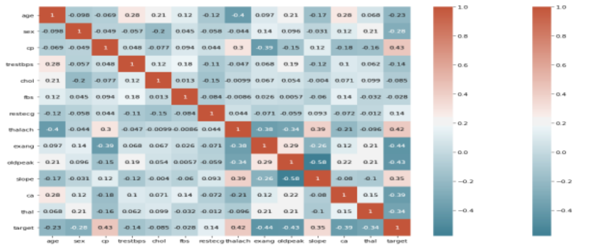


Conclusion: It can be deduced that the middle-aged male has the greater chance of having heart disease.

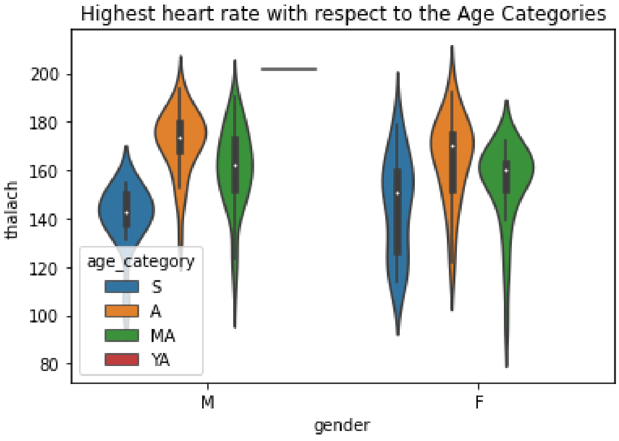


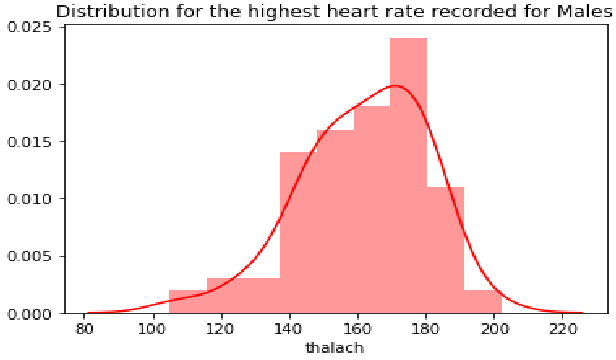
Conclusion: It can be deduced that the middle-aged female has the upper chance of having heart disease.

CORRELATION MATRIX

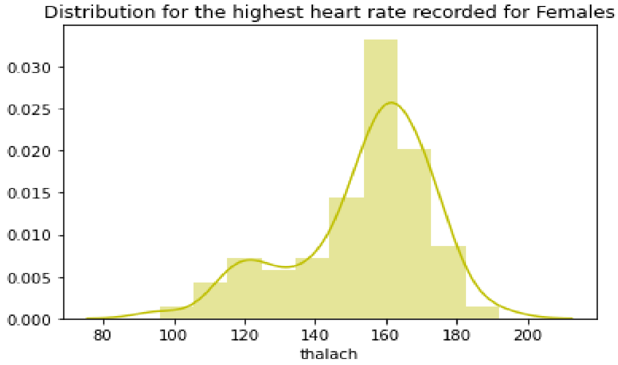


We can deduce that between chest pain (cp) & target (our predictor) there exists a positive correlation. The increase in chest pain led to an increase in the chance of having heart disease. Also, we can deduce that exercise-induced angina (exang) & our predictor are negatively correlated.





The diagram above shows the distribution is somehow skewed to the left.

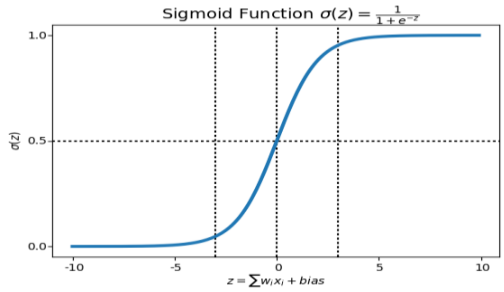


The diagram above shows the distribution is more to the left skewed.

5. IMPLEMENTATIONS AND EXPERIMENTS

Five models were utilized which are Logistic Regression, K-Nearest Neighbors, Decision Trees Support Vector Machine, and Naive Bayes. These models are explained as follows:

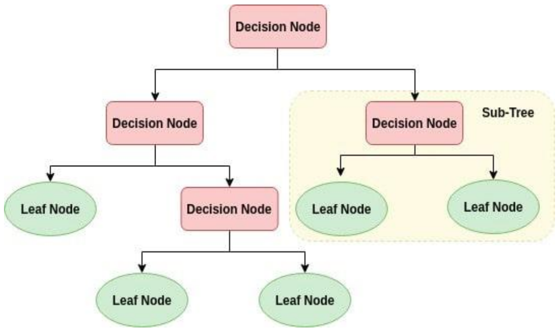
(1) Logistic Regression: It uses assumptions based on Bayes' Theorem. It is a classification machine learning algorithm used to allocate perceptions to a discrete dataset (Manandhar, Srestha, & Tandukar, 2020). Classification problems are Email spam or not spam, Fraud or not Fraud, Tumor, etc. It has a cost function that can be termed as the '**Sigmoid function**' (Manandhar, Srestha, & Tandukar, 2020).



(2) K-Nearest Neighbors: is a non-parametric classification method that classifies a data point based on its nearest neighbor and is biased by the value of k (Guo, Wang, Bell, Bi, & Greer, 2004).

(3) Decision Trees: A decision tree is a tree structure that looks like a flowchart, with each node of the tree representing a test on an attribute, where each of the tree branches representing a test result, and the leaves representing the class distributions (Poojari, 2019). The aim is to produce a model that predicts the value of a target variable by learning easy decision rules deduced from the data features. The tree is produced by repeated dividing a dataset into a new grouping based on a statistical measure of the data along each different dimension. The terminal nodes in the tree are known as leaf nodes and provide the final predictions. In the simplest form, the leaf node simply provides the final answer, however, the values in the leaf node can also be combined to form a probabilistic classification or regression estimate.

Below is the diagram of the Decision Tree (Poojari, 2019).



(4) Support Vector Machine: It is a supervised machine learning technique that is based on the concept of statistical learning developed by Vapnik et al (Zhang, 2012). This statistically learning method aims at finding a hyperplane that separates data points into two dimensions (Zhang, 2012).

(5) Naive Bayes: It is a classification method found on Bayes' Theorem with an assumption of independence among predictors. It shows a Naive Bayes classifier await that the presence of a specific characteristic in a class is not identified with the presence of some other component. It is easy to make and especially useful for extremely huge informational indexes, alongside efficiently, Naive Bayes is known to outperform exceptionally complex classification techniques (Zang, 2004).

Bayes theorems contribute to the method of calculating posterior probability P(c|x) from P(c), P(x), and P(x|c). The equation is as shown below (Rajesh, 2018):

$$p(c\backslash x) = \frac{p(x\backslash c)P(c)}{p(x)}$$

Where Posterior Probability = $(c\backslash x) = P(x_1\backslash c) * P(x_2\backslash c) * P(x_3\backslash c) * \ldots P(x_n\backslash c) * P(c)$ Predictor Prior Probability = P(x)

Class Prior Probability = P(c)

Likelihood = P (x/c)

6. RESULTS AND DISCUSSIONS

The dataset is divided into a training set and testing set, making use of 75% and 25% respectively. The data set is scaled to span a defined range, such as [0,1] using a normalizing scale. The models were evaluated using the following:

Accuracy: this measures the ratio of correctly predicted observations to the total observations.

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

Recall: this measures the ratio of correctly predicted observation to the sum positive observation.

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

Precision: this measures the ratio of correctly predicted positive observation to the sum predicted positive observation.

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

F1 Score: This is a weighted mean of precision and recall. It ranges from 0 to 1. A model with an F1 score close to 1 indicates a good model while 0 shows the model is not sufficient.

MODELLING

We are going to be comparing five models on the trained data set as earlier described in the previous chapter and deduce the one with the highest accuracy. The following are the result:

Model	Accuracy Score	Precision	Recall	F1-Score
Logistic Regression	0.75	0.75	0.80	0.78
KNN	0.75	0.76	0.78	0.77
SVM	0.80	0.78	0.88	0.83
Naïve Bayes	0.78	0.79	0.83	0.81
Decision Tree	0.72	0.79	0.66	0.72

7. CONCLUSION

In conclusion, the Support Vector Machine performs better than other methods compared in the prediction of heart disease. The finding aligns with the research of Yamala Sandhya in his paper (Sandhya, 2020) that SVM produces the most accurate and reliable results in comparison to other machine learning algorithms.

REFERENCE

(1) G. Guo, H. Wang, D. Bell, Y. Bi, K.R. Greer, "KNN Model-Based Approach in Classification", *OTM*, 2003.

[2] M. Jabbar, B. Deekshatulu, P. Chandra, "Classification of Heart Disease Using K- Nearest Neighbor and Genetic Algorithm", *ArXiv, abs/1508.02061*, 2015.

[3] A. Kalali, S. Richerson, E. Ouzunova, R. Westphal, B. Miller, "Chapter 16 – Digital

Biomarkers in Clinical Drug Development. In George G. Nomikos, Douglas E. Feltner (Eds.), Handbook of Behavioral Neuroscience", (pp. 229-238). Elsevier. https://doi.org/10.1016/B978-012-803161-2.00016-3,(https://www.sciencedirect.com/science/article/pii/B9780128031612000163), 2019.

[4] N. Manandhar, S. Srestha, R. Tandukar, "A Mini Project on Heart Disease Prediction", 2020.

[5] D. Poojari, "Machine learning basics: decision tree from scratch. Theoretical framework". https://towardsdatascience.com/machine-learning-basics-descision-tree-from-scratchpart-i-4251bfa1b45c, (August 2019).

[7] Y. Sandhya, "Prediction of Heart Diseases using Support Vector Machine", International Journal for Research in Applied Science and Engineering Technology. 8. 126-135.10.22214/ijraset, 2020-2021.

[8] D. Shah, S.B. Patel, S. Bharti, "Heart Disease Prediction using Machine Learning Techniques", *SN Computer. Sci., 1*, 345, 2020. [9] H. Sharma, M. Rizvi, "Prediction of Heart Disease using Machine Learning Algorithms", A Survey. *International Journal on Recent and Innovation Trends in Computing and Communication*, *5*(8), 99 - 104. https://doi.org/10.17762/ijritcc.v5i8.1175, 2017.

[10] M. Sultana, A. Haider, M.S. Uddin, "Analysis of data mining techniques for heart disease prediction", *3rd International Conference on Electrical Engineering and Information Communication Technology (ICEEICT)*, 1-5, 2016.

[11] J. VanderPlas, "Chapter 5- Machine Learning in Schanafelt D. (Ed), Python for Data Science", (pp 332). O'Reilly Media Inc, 2016.

[12] WHO, World Health Organization, Cardiovascular diseases (cvds) fact sheet

https://www.who.int/en/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds), 2017.

[13] H. Zhang, "The Optimality of Naive Bayes", Proceedings of the Seventeenth International

Florida Artificial Intelligence Research Society Conference, FLAIRS 2004. 2, 2004.

[14] Y. Zhang, "Support Vector Machine Classification Algorithm and Its Application", 2012.

[15] C. Liu, L. Wang, A. Yang (eds), "Information Computing and Applications", (ICICA) Communications in Computer and Information Science, vol 308. Springer, Berlin, Heidelberg, https://doi.org/10.1007/978-3-642-34041-3_27, 2012.