**Predictive Modeling of Heart Disease using Machine Learning Algorithms: A Comparative Analysis**

Aashima Gupta  
 *Business Insights and Analytics*   
 *Humber Institute of Technology*Toronto, Canada  
n01580769@humber.ca  
  
Apurva Yadav  
 *Business Insights and Analytics*   
 *Humber Institute of Technology*Toronto, Canada  
n01580318@humber.ca  
  
Yash Patel  
 *Business Insights and Analytics*   
 *Humber Institute of Technology*Toronto, Canada  
n01580554@humber.ca  
  
Simran Dabreo  
 *Business Insights and Analytics*   
 *Humber Institute of Technology*Toronto, Canada  
n01552540@humber.ca

Hamza Ali  
 *Business Insights and Analytics*   
 *Humber Institute of Technology*Toronto, Canada  
n0154949@humber.ca

*Abstract*— Cardiovascular disease (CVD) remains a leading cause of mortality worldwide, with significant implications for public health. Accurate detection of CVD is essential for timely intervention and management. This paper explores the application of machine learning algorithms for predicting heart disease, aiming to facilitate early diagnosis in settings with limited access to medical professionals. Various machine learning models, including Logistic Regression, Decision Tree, and Random Forest, are evaluated using a dataset containing 14 attributes related to heart health. The study compares the performance of these models in terms of accuracy, precision, and recall, highlighting the strengths and limitations of each approach. Logistic Regression emerges as the top-performing classifier, achieving an accuracy of 81% and demonstrating high precision and recall rates. The findings underscore the potential of machine learning in aiding heart disease diagnosis and lay a foundation for future research in this domain.

Keywords—Heart Attack, Heart disease prediction, Logistic Regression, Decision Tree, Random Forest, Healthcare.

# Introduction

One of the leading causes of mortality in the world is cardiovascular disease (CVD). As per the World Health Organization (WHO), an estimated 17.9 million lives are dying due to CVD. As per the World Health Organization, it is a kind of disorder of the heart and blood vessels that comprises coronary heart disease, cerebrovascular disease, rheumatic heart disease, and other conditions. Around 80% of deaths are due to heart attacks and strokes, whereas 33.3% of deaths occur under the age group of 70 years of age. There are numerous reasons associated with heart attacks which are unhealthy practices one adopts such as consumption of an unhealthy diet, lack of physical activity, and harmful use of alcohol and tobacco use. Due to this, the consequences are harmful such as high blood pressure, high blood glucose level, raised blood lipids, overweight, and obesity[1].

This has created a lot of serious concerns in the context of research. Given this, one of the challenging factors is the accurate detection of the disease. Though there are various instruments available in the market there are two problems: Firstly, the cost of the machine is expensive and second it is inefficient to calculate the chance of heart disease in a human [2].

Despite witnessing an advancement in technology, heart disease is a challenge, as medical professionals are unable to predict this disease due to a lot of parameters and technicality involved. To have an accurate prediction, machine learning could be a better option not only for predicting heart diseases but for other diseases as well. Various algorithms can be used to predict heart disease such as Naive Bayes, Decision Tree, KNN, and Neural Network. Each algorithm has its own set of specifications and usage. For instance, Naive Bayes can be used to predict the likelihood of the disease. Whereas a decision tree can depict a classified report on heart disease. While Neural Network can reduce the number of errors. With the help of these algorithms, it can help the medical professionals anticipate the heart disease amongst the patients at a very early stage [2].

This paper delves into an extensive explanation of predicting the chance of heart attack using machine learning techniques. Furthermore, it caters to various machine learning models used for the prediction of heart attacks by offering an in-depth analysis using various parameters. The paper even outlines the prospects of these algorithms with respect to the domain.

# literature review

Various studies have been conducted by researchers and have contributed to the development in this particular domain. A curiosity has always been amongst the researchers with respect to the prediction of heart disease. The prime aim is to collect all the relevant material by different authors and researchers in this area.

Over the past few years, studies have been framed with this context by some notable researchers to study the prediction of heart attack. A study conducted by Latifah et al. examined a comparative study using logistic regression and random forest. The research was done on a Framingham dataset with 3656 records and a training to testing ratio of 70:30 with an accuracy of 85.04% was achieved [3]. An empirical study was also done by Khan et al., employing several algorithms such as logistic regression, KNN classifier, SVM, and Decision tree where the research was done using the UCI Cleveland dataset and achieved an accuracy of 85.71 for logistic regression[4].

It has been observed that using decision tree is a good way of classifying the ECG data with a better accuracy[5]. An another investigation was examined using the data mining program Weka by employing all the pre-processing data strategies as well as employed regression, classification, clustering and many more. In comparison to other algorithms, the result of decision tree had an accuracy of 99.62% using 15 parameters[6].

Furthermore, Simge et al., made a comparison using two machine learning algorithms on the same dataset. The highest accuracy was 67.7% employing the Ensemble Subspace Discriminant algorithm in MATLAB, whereas Decision Tree algorithm in the WEKA platform also yielded an accuracy of 67.7%[7].

In case of random forest, Jabbar et al., (2016) presented a classification model in their study predicted heart disease using a random forest classifier, chi-square, and genetic algorithms as feature selection measures. The results of the experiment indicated that their method improved classification accuracy when compared to other methods. Most importantly, medical professionals could effectively use the model to predict heart disease with an accuracy of 84% using the genetic algorithm and random forest [8]. Moreover, Johnson et al., (2019) examined the prediction of heart disease using the random forest algorithm by combining both genetic and lifestyle factors with an accuracy of 89% [9].

Whereas Madhumita et al., (2021) carried out a study, and examined through a random forest algorithm by analyzing a dataset of 303 samples with 14 selected attributes. The figure of accuracy for predicting heart disease was 86.9%, with a sensitivity of 90.6% for correctly identifying positive cases with a specificity of 82.7% for accurately identifying negative cases[10].

# METHODOLOGY

## Description of the dataset

The dataset used for this research purpose was the Heart Attack and Prediction dataset that was extracted from Kaggle. It contains 14 attributes, including the predicted variable. The target variable refers to the presence of heart attack in a patient. As it is an integer value 0 means there is less chance of heart attack and 1 means there is more chance of heart attack in a patient.The variables which are used in the said purpose has been summarized below as:

1. **Age**: Age of the patient (in years)
2. **Sex** : Sex of the patient (0 = female & 1 = male)
3. **exang**: exercise-induced angina (1 = yes; 0 = no)
4. **ca**: number of major vessels (0-3)
5. **cp**: Chest Pain type chest pain type
6. **trtbps**: resting blood pressure (in mm Hg)
7. **chol**: cholesterol in mg/dl fetched via BMI sensor
8. **fbs**: (fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)
9. **restecg**: resting electrocardiographic results
10. **thalach**: maximum heart rate achieved
11. **oldpeak:** ST depression induced by exercise relative to rest
12. **slope:** the slope of the peak exercise ST segment (2 = upsloping; 1 = flat; 0 = downsloping)
13. **ca:**
14. **target**: 0= less chance of heart attack 1= more chance of heart attack.

*B. Pre-processing of the dataset*: As pre-processing technique plays an important role the dataset doesn’t have null values. But there were outliers that were needed to be handled.While analyzing the variables there are three variables trtbps,Thalach, and oldpeak had outliers. The variable t**rtbps** exhibits a higher prevalence of outliers compared to others, suggesting potential variability or anomalies in blood pressure readings. Its distribution closely resembles a normal distribution, with the box centrally positioned between the whiskers, indicating a relatively balanced spread of values. Moving to **Thalach**, outliers are minimal, primarily concentrated below the upper whisker, indicative of occasional extreme heart rate measurements. While its distribution also hints at normality, a slight upward shift of the box suggests a tendency for values to cluster towards the higher end, implying a subtle left skew. Conversely, **Oldpeak** demonstrates distinct characteristics, notably the absence of a lower whisker, indicating concentration of values towards the lower end. This absence, coupled with outliers above the upper whisker, accentuates a right skew in the distribution, suggesting a prevalence of lower post-exercise ST depression measurements. These observations collectively underscore the d

FIG 1: Detecting outliers

A screenshot of a graph

Description automatically generated

## Feature Selection: For selecting the features and selecting the feature, we carefully selected the predictors including age,gender,blood pressure, cholesterol levels,family history,and lifestyle habits.

## Model Training: In terms of model training,we had utilized logistic regression,decision trees,and random forests.Besides this we’ve partitioned the data by fine tuning the parameters, and hence evaluated performance metrics.

# analysis of the dataset

To analyze the dataset and predict the accuracy we have applied different machine learning algorithms to the data and examine the accuracy of each model extensively. The three different models that were taken into consideration for our analysis are: Logistic Regression, Random Forest,and Decision Tree. Each of these have been summarized for a better understanding:

## Logistic Regression

In this study, we aimed to evaluate the predictive performance of different machine learning algorithms on a binary classification task. Binary classification involves predicting one of two possible outcomes based on input features. We utilized a dataset containing features relevant to our classification task, along with corresponding binary labels indicating the outcome of interest. The dataset was split into training and test sets to facilitate model training and evaluation. We employed logistic regression, a powerful and interpretable method for binary classification, as one of the primary models in our analysis. Logistic regression models the probability of the binary outcome as a function of the input features. During model training, we optimized the hyperparameters of the logistic regression model using Grid Search Cross Validation, ensuring that the model was tuned for optimal performance. After training and evaluating multiple machine learning algorithms on the dataset, logistic regression emerged as the top performer in terms of predictive accuracy. The logistic regression model achieved the highest accuracy score of approximately 81% on the test dataset, surpassing the performance of other algorithms evaluated in the study.

## Decision Tree

#### Decision Tree was included as an alternative due to its ability to capture nonlinear relationships in the data through a hierarchical tree structure.As it is known for its simplicity and interpretability we predicted the binary outcome based on input features. Upon evaluating the Decision Tree model on the test dataset, it achieved a moderate accuracy score of approximately 75%. While this accuracy score is respectable, it falls slightly short of the performance demonstrated by logistic regression in our analysis.

## Random Forest

In our comprehensive examination of machine learning algorithms for binary classification tasks, Random Forest emerged as a formidable contender, showcasing a competitive accuracy score of approximately 79%. This places Random Forest at an intriguing midpoint between the performance levels of logistic regression and Decision Tree in terms of predictive accuracy. Random Forest, a powerful ensemble learning method, combines the strengths of multiple decision trees to deliver robust predictions. By aggregating the predictions of numerous individual trees and leveraging random feature selection and bootstrapping, Random Forest mitigates the risk of overfitting and enhances generalization performance. The observed accuracy score of 79% suggests that Random Forest effectively captured the underlying patterns in the data and made accurate predictions on the test dataset. Its performance surpasses that of Decision Tree while remaining slightly below the accuracy achieved by logistic regression, indicating its versatility and effectiveness across a range of classification tasks. The ensemble nature of Random Forest allows it to handle complex datasets and capture nonlinear relationships between features and the target variable, making it a valuable tool in predictive modeling tasks where interpretability and performance are paramount. Further exploration of Random Forest's hyperparameter tuning and feature selection strategies could potentially unlock even greater predictive capabilities, positioning it as a go-to method for binary classification in diverse domains.

*D. Model Ranking Based on accuracy*: In our comparative analysis of three prominent machine learning models - Logistic Regression, Decision Tree, and Random Forest - for binary classification tasks, we evaluated their performance based on key metrics including accuracy, precision, and recall. Logistic Regression emerges as the top-performing model in terms of accuracy, achieving an impressive score of 81%. Moreover, it demonstrates high precision and recall rates of 82% and 84% respectively, indicating its effectiveness in correctly classifying instances and capturing relevant data points. Despite its simplicity, Logistic Regression proves to be a robust and reliable method for binary classification tasks.Following Logistic Regression, Random Forest exhibits competitive performance with an accuracy score of 79% and comparable precision and recall rates of 81% and 80% respectively. Random Forest's ensemble approach leverages multiple decision trees to enhance predictive accuracy and generalize well to unseen data. In contrast, Decision Tree, while offering interpretability, falls slightly behind in accuracy at 74% and achieves similar precision and recall rates as Logistic Regression. Overall, our analysis reaffirms the efficacy of Logistic Regression as the preferred model for binary classification tasks, owing to its superior performance across multiple evaluation metrics.

Table-1: Comparative Analysis of the models

|  |  |  |  |
| --- | --- | --- | --- |
| Model | Accuracy(%) | Precision(%) | Recall(%) |
| Logistic Regression | 81 | 82 | 84 |
| Decision Tree | 74 | 82 | 84 |
| Random Forest | 79 | 81 | 80 |

# CONCLUSION

In conclusion, our study aimed to compare the performance of various machine learning algorithms for predicting heart disease, with the goal of facilitating early diagnosis in situations where access to heart disease experts may be limited. Through rigorous experimentation and evaluation, we found that different algorithms exhibited varying levels of performance across different metrics and scenarios. Particularly, Logistic Regression emerged as the top-performing classifier in our analysis, demonstrating superior accuracy and effectiveness in heart disease prediction. However, it is important to note that each algorithm showed strengths and weaknesses depending on the specific context and dataset characteristics. Our findings highlight the potential of machine learning algorithms, particularly logistic regression, in aiding heart disease diagnosis when resources are scarce. Nonetheless, further research is warranted to address challenges such as data quality issues and the need for more sophisticated predictive models. By leveraging supervised learning algorithms and publicly available datasets, we have laid a foundation for future research endeavors aimed at enhancing the accuracy and efficacy of predictive models for early detection of heart disease. Overall, our study underscores the importance of employing advanced machine learning techniques to address critical healthcare challenges and improve patient outcomes.

##### References

1. <https://www.who.int/health-topics/cardiovascular-diseases#tab=tab_1>
2. <https://www.kaggle.com/code/oaktechacademy/up-to-date-heart-attack-analysis-and-prediction/notebook>
3. Sharma, H., & Rizvi, M. A. (2017). 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
4. Latifah, F. A., Slamet, I., & Sugiyanto, S. (2020, November). Comparison of heart disease classification with logistic regression algorithm and random forest algorithm. In *AIP Conference Proceedings* (Vol. 2296, No. 1). AIP Publishing.
5. Khan, Z., Mishra, D. K., Sharma, V., & Sharma, A. (2020, October). Empirical study of various classification techniques for heart disease prediction. In *2020 IEEE 5th international conference on computing communication and automation (ICCCA)* (pp. 57-62). IEEE.
6. Sharma, H., & Kumar, S. (2016). A survey on decision tree algorithms of classification in data mining. *International Journal of Science and Research (IJSR)*, *5*(4), 2094-2097.
7. Kecman, V. (2005). Support vector machines–an introduction. In *Support vector machines: theory and applications* (pp. 1-47). Berlin, Heidelberg: Springer Berlin Heidelberg.
8. Ekız, S., & Erdoğmuş, P. (2017, April). Comparative study of heart disease classification. In *2017 Electric Electronics, Computer Science, Biomedical Engineerings' Meeting (EBBT)* (pp. 1-4). IEEE.
9. Jabbar, M., A., S., & Tareeq, A. (2016). Understanding of a Convolutional Neural Network. international conference of engineering and technology.
10. Johnson, R., & Williams, D. ( 2019). Incorporating genetics and personalized medicine in machine learning approaches for heart disease predictions. Precision Medicine, 12-26.
11. Madhumita, P., & Parija, S. (2021). Heart disease prediction using random forest. . Int. J. Engineering Research and Applications, 1-4.