ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Vol 10, Issue 9, September 2023

Automated Diagnosis and Prediction of Cardiovascular Diseases: An Essential Approach for Efficient Management

[1] Biyyapu Sri Vardhan Reddy, [2] Dagumati Harshavardhan, [3] Birudavolu Vishnudheeraj Reddy, [4] Biyyapu Manvitha Reddy

[1] School of Computer Science and Engineering, Vellore Institute of Technology, Chennai, India
[2] School of Computing Science and Engineering, Vellore Institute of Technology, Vellore, India
[3] Department of Computer Science and Engineering Amrita School of Computing Amrita Vishwa Vidyapeetham, Coimbatore, Vellore, India
[4] Department of Computer Science and Engineering, Saveetha Engineering College, Chennai, India
Corresponding Author Email: [1] biyyapu.srivardhan2019@vitstudent.ac.in, [2] dagumati.harsha2021@vitstudent.ac.in, [3] cb.en.u4cse20013@cb.students.amrita.edu, [4] Manvitha1919@gmail.com

Abstract*— In recent decades, heart diseases have become the leading cause of mortality worldwide. These diseases encompass a range of cardiovascular conditions, including blood vessel diseases, heart rhythm problems, and congenital heart defects. Early diagnosis plays a crucial role in effective management and improved patient outcomes, highlighting the need for precise and reliable methods of early detection through automation currently, physicians rely on clinical tests and their knowledge of patients' symptoms to diagnose cardiovascular disorders. However, patients with heart disease require early diagnosis, effective treatment, and ongoing monitoring. Data mining techniques have been utilized in the past to identify and predict cardiac diseases to address these needs. However, previous research has primarily focused on identifying major contributing factors for heart disease prediction, with less attention given to assessing the strength of these factors. As the incidence of heart diseases continues to rise, it becomes increasingly important to predict and detect these conditions at an early stage. However, accurate diagnosis presents challenges that call for precise and efficient procedures. Automation provides a potential solution by leveraging advanced technology and data mining techniques to analyze large datasets and identify patterns. Automating the diagnostic process can improve the speed and accuracy of diagnosis, leading to better patient outcomes and more effective management of cardiovascular diseases. This study focuses on estimating the risk of heart disease in patients based on various medical characteristics. A heart disease prediction system was developed using the patients' medical information. Machine learning algorithms, such as logistic regression, were employed to predict and categorize patients with heart disease. The suggested model demonstrated sufficient strength in predicting the presence of heart disease, showing higher accuracy compared to other classifiers like naive Bayes. The accurate identification of heart disease adds significant value to medical care, reducing costs and improving patient outcomes. The heart disease prediction system used in this study improves medical care by accurately identifying individuals at risk of heart disease. Automation and data mining techniques enhance the precision and efficiency of diagnosis, leading to better management of cardiovascular diseases.

Index Terms*— Heart diseases, Naive Bayes, Logistic regression, Data analysis.

I. INTRODUCTION

Cardiovascular disease has been a significant cause of mortality worldwide for the past decade. According to the World Health Organization (WHO), more than 17.9 million people die each year from cardiovascular disease, with 80% of these deaths attributed to cardiovascular disease and stroke. Machine learning plays a pivotal role in predicting heart disease using artificial intelligence. It utilizes past experiences to determine whether a patient is likely to have a specific type of disease. Our model specifically employs supervised learning techniques to predict the early stages of heart disease.

According to statistics from the World Health Organization, heart disease causes approximately 12 million deaths worldwide every year, making it a significant global health concern. The prediction of cardiovascular disease is an important area within data analysis, considering the increasing burden of heart disease worldwide. Extensive research has been conducted to identify influential factors and accurately predict the overall risk of heart disease. Heart disease is often referred to as a "silent killer" due to its ability to cause fatalities without obvious symptoms. Early diagnosis plays a vital role in making informed decisions regarding lifestyle modifications for high-risk individuals and reducing the potential complications associated with heart disease.

Machine learning techniques have proven to be effective in analyzing the vast amount of healthcare data generated and assisting in decision-making and predictions. This project aims to predict future occurrences of heart disease by analyzing patient data and employing machine learning algorithms to classify individuals as having heart disease or not. By leveraging these machine learning techniques, we can

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Vol 10, Issue 9, September 2023

tap into the potential of data-driven predictions.

Heart disease can manifest in various forms, but there are common core risk factors that significantly influence an individual's likelihood of developing heart disease. By collecting data from diverse sources, organizing it into relevant categories, and analyzing it, we can harness the power of machine learning to accurately predict heart disease presents a substantial global health challenge, with millions of deaths occurring each year. The application of machine learning algorithms and data analysis techniques offers a promising approach to predict and tackle heart disease. By utilizing patient data and identifying core risk factors, we can develop effective prediction models that aid in early detection and prevention strategies for heart disease.

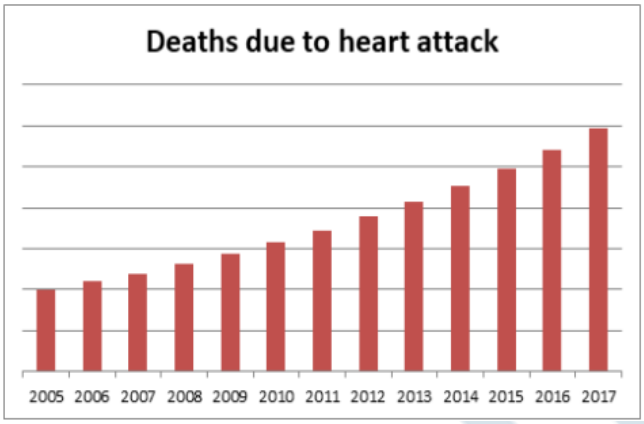


Fig.1. Deaths due to heart attack

Heart diseases are a growing concern globally, with an alarming rise in cases. This emphasizes the importance of early detection and prediction to effectively manage these conditions. In India, the situation is particularly worrisome, with heart disease being the leading cause of death. The Global Burden report highlights those 1.7 million deaths occur annually in the country due to heart disease. Shockingly, the mortality rate related to heart disease has seen a significant 53% increase since 2005. These statistics clearly demonstrate the escalating nature of heart disease cases, with an annual growth rate of 9.5%.

To tackle the challenges posed by heart disease, machine learning, a branch of artificial intelligence, has emerged as a valuable tool. Machine learning systems are designed to learn from experience and make predictions based on that knowledge. By training machine learning algorithms on extensive databases of past cases, predictive models can be created. These models can then utilize new data inputs to assess the likelihood of heart disease in individuals. By leveraging machine learning algorithms, researchers and healthcare professionals can develop predictive models that analyze various risk factors and patterns associated with heart disease. These models enable early detection of heart disease and provide valuable insights for timely intervention and treatment.

The integration of machine learning in heart disease prediction represents a significant advancement in healthcare. Machine learning algorithms can analyze vast amounts of data, including patient information, medical records, and lifestyle factors, to uncover hidden patterns and correlations. This approach enhances the accuracy of heart disease prediction and allows for the development of personalized treatment plans tailored to individual patients. The increasing prevalence of heart diseases calls for proactive measures and early detection. Machine learning offers a powerful tool for predicting and managing heart disease by leveraging past experiences and generating predictive models. By utilizing machine learning algorithms, healthcare professionals can make informed decisions that lead to improved patient outcomes and a better understanding of heart disease patterns and risks.

Diagnosing heart disease accurately and effectively is a complex task that requires precision and efficiency. Our model focuses on predicting the likelihood of a patient having heart disease based on various medical conditions. We have developed a cardiovascular prediction system that utilizes different machine learning algorithms, such as logistic regression, to differentiate between patients with and without heart disease. Logistic regression is a widely used algorithm in medical prediction tasks, particularly for binary classification problems. By training our model on a dataset containing relevant medical conditions and their corresponding outcomes, we can create a predictive model capable of assessing the probability of heart disease in an individual.

Our cardiovascular prediction system takes into account a range of medical conditions and factors that are known to be associated with heart disease. These may include demographic information, medical history, lifestyle choices, family history, and diagnostic test results. By analyzing these factors, our model can identify patterns and relationships that contribute to the prediction of heart disease. By utilizing machine learning algorithms, our cardiovascular prediction system automates the analysis of large amounts of patient data, enabling more efficient and accurate predictions. The algorithms can learn from past examples and continuously improve their predictive capabilities as more data becomes available.

The primary objective of our cardiovascular prediction system is to provide healthcare professionals with a valuable tool for early detection and intervention in heart disease cases. By accurately identifying patients at a higher risk of heart disease, medical practitioners can prioritize preventive measures, implement appropriate treatment strategies, and ultimately improve patient outcomes. Our cardiovascular prediction system employs various machine learning algorithms, including logistic regression, to predict and differentiate patients with heart disease. By considering a wide range of medical conditions and factors, the system aims to accurately identify individuals at risk, enabling

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Vol 10, Issue 9, September 2023

healthcare professionals to intervene at an early stage and provide timely and targeted care.

II. MOTIVATION FOR THE WORK

The objective of this study is to develop a heart disease prediction model that accurately forecasts the occurrence of heart disease. Additionally, the research aims to determine the most effective classification algorithm for assessing an individual's risk of developing heart disease. Three commonly used classification algorithms, namely Naïve Bayes, Decision Tree, and Random Forest, are compared in this study. Given the critical nature of heart disease prediction and the need for high accuracy, the algorithms are evaluated using various assessment methodologies at different levels. The findings from this research will have practical implications for medical professionals and researchers, enabling the development of an improved method for anticipating cardiac illness, enhancing diagnostic accuracy, and improving patient care.

A. Problem Statement

A dataset consisting of information from 303 individuals is utilized to tackle the task of predicting the presence of heart disease. The primary objective is to determine whether an individual is likely to suffer from cardiovascular ailments based on the available data. Effective management of cardiovascular diseases can be achieved through lifestyle modifications, regular medication, practicing yoga, and in some cases, surgical interventions. By accurately predicting the presence of heart disease, the implementation of preventive measures can reduce the need for costly surgical treatments and associated expenses. The global burden of cardiovascular diseases is on the rise, with approximately 17.9 million lives lost annually, as reported by the World Health Organization (WHO). Recording and monitoring patients' health data have become crucial in combating this trend. Existing diagnostic procedures, such as ECG, blood pressure, cholesterol, and blood sugar tests, can be time-consuming and resource-intensive.

While healthcare systems possess vast amounts of data, extracting meaningful insights and uncovering hidden patterns from this information remains a challenge. Machine learning, driven by the increasing availability of data, offers a promising solution to handle and extract valuable knowledge from large datasets that would be impractical or impossible for manual analysis.In this study, the focus is on utilizing a subset of relevant attributes to accurately predict the risk of heart disease. The goal is to achieve efficient and timely assessments by leveraging machine learning algorithms. By shifting from intuitive decision-making to evidence-based predictions, biases, omissions, and excessive costs in healthcare can be mitigated, leading to improved patient care.Through the application of various machine learning algorithms, the objective is to provide accurate predictions regarding the presence of heart disease using a limited number of tests and attributes. This approach aims to enhance diagnostic efficiency, facilitate proactive interventions, and optimize patient care in the context of cardiovascular health.

III. LITERATURE SURVEY

Nayab Akhtar's et al. has proposed a compartive study on machine learning algorithms and data mining techniques were employed to predict heart disease, with the naive Bayes algorithm achieving an impressive accuracy rate of 88%. These findings have significant implications for automated diagnosis in healthcare, enabling more accurate predictions and interventions for individuals at risk of heart disease. By leveraging the power of machine learning, healthcare professionals can improve patient outcomes and optimize healthcare delivery in the context of heart disease.[1]

Armin Yazdani conducted a study proposing an algorithm to address the existing gap in the literature regarding heart disease prediction. The algorithm focuses on measuring the strength of significant features contributing to the prediction of heart disease using Weighted Associative Rule Mining. Through experimentation on the UCI open dataset, the study confirms the effectiveness of the algorithm, achieving a high confidence score of 98% in predicting heart disease. This research by Armin Yazdani provides valuable insights into the computation of strength scores for significant predictors in heart disease prediction, contributing to the field of heart disease research and offering potential advancements in the accurate identification and management of this critical health condition.[2]

Harshit Jindal's research paper focuses on predicting the likelihood of individuals having heart disease based on various medical attributes. A heart disease prediction system was developed using the patients' medical history. Machine learning algorithms, including logistic regression and K-nearest neighbors (KNN), were utilized to classify and predict the presence of heart disease. The proposed model demonstrated satisfactory strength and accuracy, outperforming other classifiers like naive Bayes. This model significantly reduces the burden by accurately identifying the probability of heart disease, thus enhancing medical care and reducing costs. The insights gained from this project offer valuable knowledge for predicting and identifying patients with heart disease.[3]

Chintan M. Bhatt conducted research to develop a model aimed at accurately predicting cardiovascular diseases and reducing associated fatalities. The study proposes a k-modes clustering method with Huang starting to enhance classification accuracy. Various models, including random forest (RF), decision tree classifier (DT), multilayer perceptron (MP), and XGBoost (XGB), were employed. GridSearchCV was used to optimize the models by tuning their parameters. The proposed model was applied to a real-world dataset of 70,000 instances from Kaggle. Accuracy rates were evaluated, with the decision tree

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Vol 10, Issue 9, September 2023

achieving 86.37% (with cross-validation) and 86.53% (without cross-validation), XGBoost achieving 86.87% (with cross-validation) and 87.02% (without cross-validation), random forest achieving 87.05% (with cross-validation) and 86.92% (without cross-validation), and multilayer perceptron achieving the highest accuracy of 87.28% (with cross-validation) and 86.94% (without cross-validation). The models also demonstrated strong performance with AUC values, with the decision tree, XGBoost, random forest, and multilayer perceptron achieving AUC values of 0.94, 0.95, 0.95, and 0.95, respectively. The study concludes that the multilayer perceptron model with cross-validation outperforms other algorithms in terms of accuracy, achieving the highest accuracy rate of 87.28%.[4]

R. Indrakumari conducted a research study that emphasizes the importance of Exploratory Data Analysis (EDA) in various aspects of data analysis. The study specifically focuses on predicting risk factors for heart disease using the K-means algorithm. In healthcare, analytics plays a vital role in improving patient care by facilitating preventive measures and responding to emergencies. The research paper highlights the significance of EDA in identifying errors, selecting relevant data, and understanding the correlations among explanatory variables. By employing data analytics and visualization tools, the study utilizes a publicly available dataset to analyze attributes such as age, chest pain type, blood pressure, blood glucose level, ECG readings, and heart rate. The paper discusses preprocessing methods, classifier performance, and evaluation metrics as part of the analysis process. This research contributes to the understanding and prediction of heart disease risk factors through data analysis techniques.[5]

Rohit Bharti's research paper explores the application of different machine learning algorithms and deep learning techniques in analyzing the UCI Machine Learning Heart Disease dataset. The objective is to compare the results and analysis obtained from these methods. The dataset contains 14 key attributes, and the study focuses on achieving promising outcomes by addressing irrelevant features using the Isolation Forest algorithm and implementing data normalization techniques. Additionally, the paper discusses the potential integration of this study with multimedia technologies, particularly mobile devices. By leveraging deep learning approaches, an impressive accuracy rate of 94.2% is achieved. The research findings demonstrate the effectiveness of machine learning and deep learning algorithms for analyzing heart disease datasets and highlight the possibilities of utilizing multimedia technology for improved analysis and predictions.[6]

Xin Qian conducted a research study focusing on the prediction of cardiovascular disease (CVD) in a population in Xinjiang. The study involved collecting data from a cohort in two stages: a baseline survey conducted from 2010 to 2012 and a second-phase baseline survey conducted from September to December 2016. The follow-up period lasted until December 2017 and August 2021. A total of 12,692 participants, including Uyghur and Kazak individuals, were included in the study. The research utilized various techniques, such as Lasso regression, logistic regression, random forest, and feature importance analysis, to identify predictive factors and establish subsets of variables for the CVD prediction model. Over a follow-up period of 4.94 years, 1,176 participants were diagnosed with CVD, resulting in a cumulative incidence of 9.27%. The prediction model based on L1 regularized logistic regression demonstrated the best performance, with important predictors including age, systolic blood pressure, lipoprotein levels, triglyceride blood glucose index, body mass index, and body adiposity index. These findings contribute valuable insights into predicting CVD in the rural population of Xinjiang.[7]

Pooja Anbuselvan's research explores the effectiveness of different supervised learning models in predicting outcomes. In this study, Logistic Regression, Naïve Bayes, Support Vector Machine, K-Nearest Neighbors, Decision Tree, Random Forest, and the ensemble technique of XGBoost are analyzed and compared. The goal is to determine the most accurate algorithm among these models. The findings reveal that Random Forest demonstrates the highest accuracy, with a score of 86.89%, surpassing the performance of other algorithms. This research provides valuable insights into the comparative performance of various algorithms, emphasizing the superiority of Random Forest in predicting outcomes.[8]

IV. SYSTEM ARCHITECTURE

The Dataset collection involves the gathering of patient data from various sources, such as hospitals, clinics, or research studies. This data includes information about the patients, such as their age, gender, blood pressure, cholesterol levels, family history, and other pertinent medical details. During the attribute selection process, the most relevant attributes are chosen to aid in the prediction of heart disease. This step is crucial in identifying the features that will contribute the most to accurate predictions. Collaboration between domain experts and data scientists is typically necessary to determine the most informative attributes while disregarding irrelevant or redundant ones.

Once the attributes are selected, the collected data undergoes a thorough cleaning process to eliminate inconsistencies, errors, and missing values. This data cleaning stage ensures that the dataset is reliable and accurate for analysis. Techniques such as imputation for missing values, outlier handling, and resolving inconsistencies are employed to enhance the quality of the data.

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Vol 10, Issue 9, September 2023

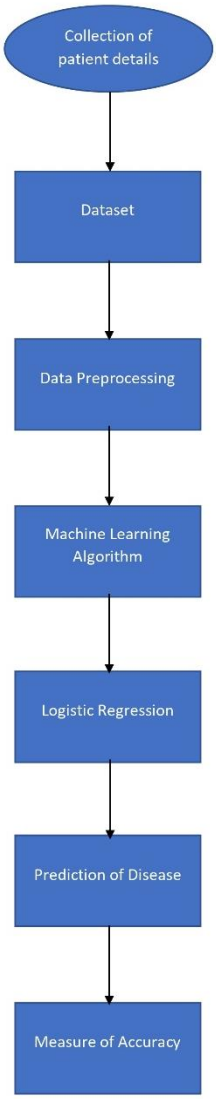


Fig.2. Flow chart

The transformed dataset is then prepared for analysis by applying appropriate techniques such as normalization, scaling, or encoding categorical variables.Various classification techniques, including logistic regression, decision trees, random forests, and support vector machines, can be employed on the preprocessed data. Each algorithm learns patterns from the data and constructs a model for predicting the presence or absence of heart disease. The accuracy of these models is evaluated using performance metrics like accuracy, precision, recall, and F1 score.

By comparing the accuracy of different classifiers, data scientists can gauge their performance. This comparison aids in selecting the most effective classifier for predicting heart disease with the given dataset. Consequently, this information facilitates informed decisions and future predictions.The process of dataset collection, attribute selection, data cleaning, and transformation is crucial for accurate heart disease prediction. Employing various classification techniques and evaluating their accuracy allows data scientists to identify the most effective model for predicting heart disease based on the available data.

V. METHODOLOGY

A. Description of the Dataset

The dataset utilized for this research is known as the Public Health Dataset, which dates back to 1988. It encompasses four distinct databases, namely Cleveland, Hungary, Switzerland, and Long Beach V. This dataset consists of a total of 76 attributes, including the predicted attribute. However, in published experiments, researchers typically focus on a subset of 14 attributes that are deemed most relevant for their analysis.

The primary focus attribute in the dataset is referred to as the "target" field, indicating the presence or absence of heart disease in a patient. It is represented by integer values, where 0 signifies the absence of the disease, and 1 represents the presence of the disease.

```
# getting information about the data
data.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    int64
dtypes: float64(1), int64(13)
memory usage: 33.3 KB
```

Fig.3. Dataset Info

Table 1 provides a glimpse of the dataset, displaying the first four rows along with all the dataset features without undergoing any preprocessing. To gain a comprehensive understanding of the dataset, let's delve into the significance of the attributes used in this research:

Age: This attribute denotes the age of the patient. It serves as a fundamental demographic feature as age plays a crucial role in assessing the risk of heart disease. Advancing age is often associated with an increased likelihood of developing heart-related conditions.

Sex: The sex attribute indicates the gender of the patient. It is typically represented by the values 0 for female and 1 for male. Gender is an essential factor in understanding the prevalence and impact of heart disease, as it can influence various physiological and lifestyle-related aspects.

CP (Chest Pain Type): CP describes the type of chest pain experienced by the patient. It is categorized into four distinct levels, providing insights into the nature and severity of the

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Vol 10, Issue 9, September 2023

pain. Different types of chest pain can be indicative of specific cardiac conditions and assist in diagnosing the underlying cause.

Trestbps: Trestbps represents the resting blood pressure of the patient, measured in millimeters of mercury (mm Hg). Blood pressure is a critical physiological measurement used to assess the risk of cardiovascular diseases. Elevated blood pressure can be an indication of hypertension and other heart-related issues.

Chol: This attribute represents the serum cholesterol level in milligrams per deciliter (mg/dL). Cholesterol plays a significant role in the development of heart disease, with high levels being associated with an increased risk. Monitoring cholesterol levels is crucial for identifying individuals who may be susceptible to heart-related conditions.

FBS (Fasting Blood Sugar): FBS indicates the fasting blood sugar level of the patient. A value greater than 120 mg/dL is generally considered high, suggesting elevated blood sugar levels. High blood sugar can be indicative of conditions such as diabetes, which in turn increases the risk of heart disease.

Restecg: Restecg describes the resting electrocardiographic results of the patient. It is classified into three categories, providing information on the electrical activity of the heart during rest. Abnormal resting ECG patterns can be indicative of underlying cardiac issues and aid in diagnosing specific conditions.

Thalach: Thalach represents the maximum heart rate achieved during exercise. Maximum heart rate is often used as an indicator of cardiovascular fitness and can provide insights into the overall health of the patient's heart. Abnormalities in the maximum heart rate response may be associated with underlying heart conditions.

Exang (Exercise Induced Angina): Exang indicates whether the patient experiences angina (chest pain) during exercise. Angina during physical exertion is often a symptom of underlying coronary artery disease. Assessing the presence or absence of exercise-induced angina helps in evaluating the patient's cardiac health.

Oldpeak: Oldpeak represents the ST depression induced by exercise relative to rest. ST depression is an electrocardiographic finding that can be indicative of reduced blood flow to the heart during exercise, suggesting the presence of coronary artery disease or other cardiac abnormalities.

Slope: Slope describes the slope of the peak exercise ST segment. It provides additional information about the ST segment, which aids in assessing cardiac function and potential abnormalities. Different slope patterns can indicate various cardiac conditions and assist in diagnosis.

CA (Number of Major Vessels): CA indicates the number of major vessels colored by fluoroscopy. It provides insights into the presence and severity of coronary artery disease, as the number of affected vessels can indicate the extent of arterial blockages. Evaluating the number of major vessels helps in determining the severity of the disease.

Thal: Thal represents the results of thallium stress tests and is classified into three categories. Thallium stress testing is often performed to assess blood flow to the heart and identify areas of reduced perfusion. The Thal attribute provides information on the results of this important diagnostic test.

Target: The target attribute is the predicted attribute indicating the presence (1) or absence (0) of heart disease. It serves as the main focus of analysis and prediction in this research, as the objective is to accurately predict the presence or absence of heart disease based on the available dataset and the chosen attributes.

These attributes have been selected based on their relevance to predicting heart disease and have been widely used in related research studies. By analyzing the relationships between these attributes and their impact on the target variable, researchers can gain valuable insights into the factors associated with the presence or absence of heart disease in patients.

Table.1. Data set

Attribute	Description	Range
Age	Age of person in years	29-79
Sex	Gender of person (1-M 0-F)	0,1
Cp	Chest pain type	1,2,3,4
Trestbps	Resting blood pressure in mm Hg	94-200
Chol	Serum cholesterol in mg/dl	126-564
Fbs	Fasting blood sugar in mg/dl	0,1
Restecg	Resting Electrocardiographic results	0,1,2
Thalach	Maximum heart rate achieved	71-202
Exang	Exercise Induced Angina	0,1
OldPeak	ST depression induced by exercise relative to rest	1-3
Slope	Slope of the Peak Exercise ST segment	1,2,3
Ca	Number of major vessels colored by fluoroscopy	0-3
Thal	3 – Normal, 6 – Fixed Defect, 7 – Reversible Defect	3,6,7
Result	Class Attribute	0,1

B. Preprocessing of the Dataset

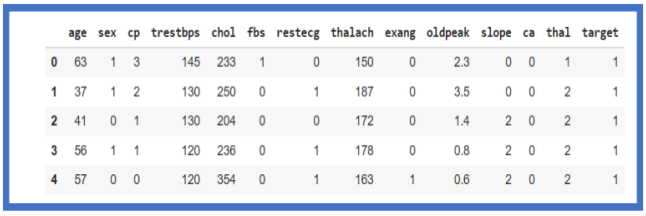
The dataset used for this research is devoid of any missing values. However, it does exhibit the presence of outliers that required proper handling. Additionally, the dataset does not possess a proper distribution, which can adversely affect the performance of machine learning algorithms. Two distinct approaches were employed to address these challenges and improve the outcomes.In the first approach, the dataset was

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Vol 10, Issue 9, September 2023

directly utilized for machine learning algorithms without addressing the outliers or performing feature selection. Unfortunately, this initial approach did not yield promising results. It underscores the significance of dealing with outliers and achieving a proper distribution in the dataset, as these factors can have a detrimental impact on the accuracy and effectiveness of the models.



	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

Fig.4. first 5 rows of the dataset

```
# print last 5 rows of the dataset
data.tail()
```

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
298	57	0	0	140	241	0	1	123	1	0.2	1	0	3	0
299	45	1	3	110	264	0	1	132	0	1.2	1	0	3	0
300	68	1	0	144	193	1	1	141	0	3.4	1	2	3	0
301	57	1	0	130	131	0	1	115	1	1.2	1	1	3	0
302	57	0	1	130	236	0	0	174	0	0.0	1	1	2	0

Fig.5. Last 5 rows of the dataset

To overcome the issue of overfitting caused by the dataset's non-normal distribution, the second approach involved transforming the dataset to adhere to a normal distribution. This step was crucial in enhancing the performance of the models. Additionally, the Isolation Forest algorithm was implemented to detect and appropriately handle the outliers present in the dataset. Isolation Forest, a well-known technique for outlier detection, effectively identifies and addresses outliers within the data.

Various plotting techniques were employed to assess the skewness of the data, detect outliers, and visualize the distribution of the dataset. These plotting techniques played a significant role in gaining insights into the data's characteristics and guiding informed decisions during the preprocessing stage. Preprocessing techniques such as outlier handling, transforming the data distribution to a normal form, and utilizing visualization plots play a pivotal role when utilizing the data for classification or prediction purposes. By adequately preprocessing the data, researchers can improve the quality of the dataset, mitigate the impact of outliers, and ensure that the data is suitable for accurate analysis and reliable predictions.

C. Checking the Distribution of the Data

The distribution of data plays a crucial role in the prediction or classification of a problem. In this case, it is observed that heart disease occurs approxi-mately 54.46% of the time in the dataset, while the remaining 45.54% repre-sents instances without heart disease. This imbalance in the dataset can lead to overfitting issues if not properly addressed.

To overcome this challenge, it is necessary to balance the dataset by equaliz-ing the number of instances for each class. This balance enables the model to find meaningful patterns and contribute to more accurate predictions for heart disease. By addressing the class imbalance, the model can avoid being biased towards the majority class and ensure that both classes are adequately repre-sented during the training process.

When a dataset is imbalanced, the model tends to prioritize the majority class, making it difficult to identify patterns and characteristics associated with the minority class. By balancing the dataset, the model is exposed to a fair repre-sentation of both classes, enabling it to capture important features and make accurate predictions for heart disease, regardless of the class imbalance. By addressing the data distribution and achieving a balanced dataset, the model can effectively learn from both classes, reducing the risk of biased predic-tions, and improving its overall performance in accurately identifying the presence or absence of heart disease.

D. Feature Selection

In the feature selection process, the Logistic Regression algorithm is employed as an embedded technique to select and rank important features. Compared to filter methods, Logistic Regression is known for its superior predictive accuracy and efficiency in feature selection.Each feature is assigned a weight by the Logistic Regression algorithm, indicating its importance. The goal of the feature selection process is to choose a subset of features that are most relevant to the selected algorithm using Logistic Regression. This selected subset significantly enhances the model's ability to predict outcomes.

To further refine the feature subset, the "select from the model" strategy is utilized. This approach, available in the scikit-learn library, assists in selecting features based on their impact on the model's predictions. By considering the model's predictions, this strategy narrows down the selected features to those that have the greatest influence on the model's performance.By combining the Logistic Regression algorithm with the "select from the model" approach, the feature selection process identifies a subset of highly informative features that greatly improve the performance of the selected algorithm. This strategy aids in reducing the dimensionality of the dataset, improving computational efficiency, and enhancing the overall predictive accuracy of the model.

E. Checking Duplicate Values in the Data

To ensure the model's generalization is not affected, it is crucial to handle duplicate data properly. Mishandling duplicates can lead to issues when the same data appears in both the training and test datasets.Addressing duplicates should be done during the data preparation phase to avoid bias and maintain the model's predictive performance. Duplicates in the training dataset can result in overfitting,

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Vol 10, Issue 9, September 2023

where the model becomes overly sensitive to repeated patterns and struggles to generalize to new, unseen data.

Moreover, if duplicates are present in both the training and test datasets, it can lead to an inflated assessment of the model's performance. The model may already have learned from these duplicated instances during training, causing it to perform well on the test data that it is already familiar with. To handle duplicates effectively, several methods can be employed. This includes identifying identical rows or comparing feature values to detect and remove duplicates. Advanced techniques such as hashing or clustering algorithms can also be utilized for duplicate detection and elimination.

By properly addressing duplicates, the model can be trained and evaluated on a clean and representative dataset. This ensures accurate performance evaluation, unbiased predictions, and improved generalization capabilities of the model.

```
data.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
```

Fig.6. checking Duplicate Values

F. Machine Learning Classifiers Proposed

Supervised learning is a fundamental concept in machine learning where machines are trained using labeled data. Labeled data consists of input samples paired with their corresponding correct output values. The main objective of supervised learning is to find a mapping function that can accurately predict the output variable based on the given input variables.

Throughout the supervised learning process, the labeled training data serves as a guide and instructor for the machine learning model. It provides the necessary information and guidance for the model to learn and make accurate predictions. By analyzing the patterns and relationships in the labeled data, the model generalizes its understanding to make predictions on new, unseen data.The ultimate goal of supervised learning algorithms is to discover an effective mapping function that can correctly map the input variables to the output variable. This is achieved by minimizing the discrepancy between the predicted output and the true output in the training data. Once the model is trained, it can be used to make predictions on new input data where the output is unknown.

Various supervised learning algorithms are employed, including neural networks, support vector machines, decision trees, and linear regression. Each algorithm utilizes its own approach to determine the optimal mapping function based on the characteristics of the data and the specific task at hand.Supervised learning finds applications in diverse industries such as image classification, natural language processing, sentiment analysis, and predictive modeling. It is commonly utilized in tasks that require accurate predictions or classifications based on previously labeled data.

Logistic regression is a widely-used supervised learning algorithm that is specifically designed for predicting categorical dependent variables using a set of independent variables.Unlike linear regression, which is used for predicting continuous numerical values, logistic regression focuses on forecasting discrete outcomes. The dependent variable in logistic regression should have categorical values, such as Yes/No, 0/1, True/False, and so on. Rather than providing exact 0 or 1 values, logistic regression generates probabilistic results that fall between 0 and 1.

Although the underlying concept of logistic regression is similar to linear regression, their applications differ. Linear regression is used to address regression problems where the objective is to predict a continuous numerical value. On the other hand, logistic regression is employed in classification problems where the goal is to categorize data into distinct groups.In logistic regression, instead of fitting a straight regression line, an "S"-shaped logistic function is utilized. Predictions from this function are limited to two maximum values: 0 and 1. The curve of the logistic function represents the likelihood of a specific outcome, such as determining whether cells are cancerous or non-cancerous based on certain characteristics or classifying a mouse as obese or not based on its weight.

Logistic regression is a powerful machine learning algorithm because it can provide probabilities and accurately categorize new data, regardless of whether the datasets are continuous or discrete. This makes it highly useful in various domains, including medical diagnosis, customer churn prediction, fraud detection, and sentiment analysis, as it enables accurate prediction of outcomes with a certain level of confidence.

VI. PROPOSED WORK

Logistic Regression is a straightforward and efficient machine learning algorithm that offers several advantages. Its simplicity and ease of implementation make it accessible, and it can achieve good training efficiency without requiring high computational power. One of the key benefits of Logistic Regression is that it provides insights into the importance and direction of association for each feature. By analyzing the learned parameters or weights, we can understand how each

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Vol 10, Issue 9, September 2023

feature contributes to the prediction. This makes Logistic Regression a useful tool for uncovering relationships between features and gaining insights into their impact on the outcome.

Unlike some other algorithms such as Decision Trees or Support Vector Machines, Logistic Regression allows for easy model updates when new data becomes available. It can be updated using techniques like stochastic gradient descent, enabling the model to adapt and incorporate new information effectively. This flexibility is particularly valuable in dynamic environments where the data distribution may change over time. In addition to its interpretability and updateability, Logistic Regression is capable of providing well-calibrated probabilities alongside classification results. This means that instead of solely providing the final predicted class, it offers a probability estimate for each class. This is advantageous as it allows for a more nuanced understanding of the certainty associated with each prediction. By comparing the probabilities, we can gauge the relative accuracy and confidence of different predictions.

The ability to obtain interpretable insights, easily update the model, and generate calibrated probabilities makes Logistic Regression a popular choice in various domains. It finds applications in areas such as healthcare, finance, marketing, and social sciences, where understanding feature relationships and obtaining reliable probability estimates are critical for decision-making and analysis.

Logistic Regression is a statistical modeling approach that aims to predict precise probabilities of outcomes based on independent features. However, when dealing with high-dimensional datasets, there is a risk of overfitting the model to the training set. Overfitting occurs when the model performs well on the training data but fails to generalize accurately to unseen test data. This is particularly likely when the training data is limited and there are numerous features present. To address this, regularization techniques can be applied to prevent overfitting, although it can increase the complexity of the model. It is essential to strike a balance with the regularization factor, as excessively high values can result in underfitting the training data.Logistic regression is limited in its ability to handle non-linear problems since it assumes a linear decision boundary. In practice, it is rare to encounter linearly separable data in real-world scenarios. Therefore, when confronted with non-linear relationships between features and outcomes, it becomes necessary to transform the data to achieve linear separability in higher dimensions. This can be achieved through techniques such as feature engineering or applying non-linear transformations.

In cases where the data exhibits complex relationships and non-linear separability, logistic regression may not be the most suitable algorithm. More sophisticated and powerful algorithms, such as Neural Networks, are capable of capturing and modeling intricate relationships in the data. Neural Networks excel in scenarios where the relationships between features and outcomes are highly non-linear and require more flexibility in the decision boundaries. It is important to consider the complexity of the problem and the nature of the data when selecting an appropriate algorithm. While logistic regression is a valuable tool in many applications, it is important to understand its limitations and explore alternative algorithms when dealing with non-linear relationships and complex datasets.

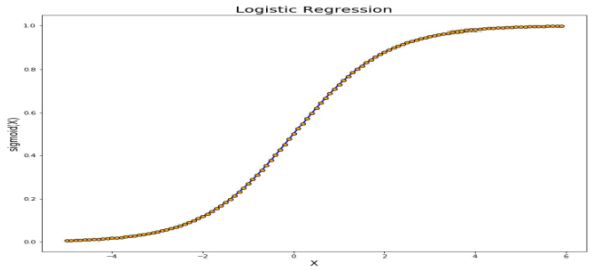


Fig.8. Logistic Regression

The code begins by importing the necessary libraries for the implementation. The numpy library is imported for array manipulation, pandas is imported for reading CSV files, train_test_split from sklearn.model_selection is imported for splitting the dataset, confusion_matrix from sklearn.metrics is imported for visualizing classification results, and seaborn is imported for enhancing the visualization of the confusion matrix.

The main body of the code defines a class called Logistic Regression and initializes its constructor with two parameters: the learning rate and the number of iterations. The constructor assigns the provided values to the instance of the class.

The code includes a method named fit within the LogitRegression class. This method is used to train the logistic regression model. It takes X_train and Y_train as input parameters, representing the training features and labels, respectively. Within the fit method, the shape of the training data is determined, and the values are assigned to m (number of rows) and n (number of columns). Two variables, W and b, are created to represent the weights and bias of the model, respectively. Initially, W is set to a numpy array of zeros with size n, and b is set to 0.

```
# separating X and Y attributes
X = data.drop(columns='target', axis=1)
Y = data['target']
print(X)
```

```
     age  sex  cp  trestbps  chol  ...  exang  oldpeak  slope  ca  thal
0     63    1   3       145   233  ...      0      2.3      0   0     1
1     37    1   2       130   250  ...      0      3.5      0   0     2
2     41    0   1       130   204  ...      0      1.4      2   0     2
3     56    1   1       120   236  ...      0      0.8      2   0     2
4     57    0   0       120   354  ...      1      0.6      2   0     2
..   ...  ...  ..       ...   ...  ...    ...      ...    ...  ..   ...
298   57    0   0       140   241  ...      1      0.2      1   0     3
299   45    1   3       110   264  ...      0      1.2      1   0     3
300   68    1   0       144   193  ...      0      3.4      1   2     3
301   57    1   0       130   131  ...      1      1.2      1   1     3
302   57    0   1       130   236  ...      0      0.0      1   1     2

[303 rows x 13 columns]
```

Fig.9. sperating X and Y attributes

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Vol 10, Issue 9, September 2023

The code also defines an update_weight function, which is called from the fit method to update the values of W and b. Within the update_weight function, the weights and bias are updated using the provided formulas. The cost function is calculated as the average of the logistic loss over the training data. The partial derivatives of the cost function with respect to W and b are computed and stored in dw and db, respectively.

The predict function is another method within the LogitRegression class. It takes X_test as an input parameter and predicts the output using the sigmoid function. If the predicted value, Y_pred, is greater than or equal to 0.5, the output is set to 1. Otherwise, it is set to 0.

In the main function, the code reads the dataset from a CSV file using pandas and stores it in a variable. The dataset is then split into X (features) and Y (labels). The data is further divided into training and testing sets using the train_test_split function, with a test size of 0.15.

An object of the LogitRegression class is created, with the learning rate set to 0.1 and the number of iterations set to 500,000. The fit method is called to train the logistic regression model using the training data. The predict function is then used to predict the class for the test data.

To evaluate the performance of the model, the code tracks the number of correctly classified and misclassified data points using the variables correctly_classified and mis_classified. If the predicted values (Y_pred) and the actual values (Y_test) are equal, the count of correctly classified data points is incremented. Otherwise, the count of misclassified data points is incremented.

The code outputs the predicted data, the count of correctly classified data points, the count of misclassified data points, the accuracy, and the confusion matrix. The seaborn library is used to visualize the confusion matrix, providing a more visually appealing representation.

VII. RESULTS AND DISCUSSION

In the provided code, two models are implemented and evaluated on the test set. One model is developed from scratch, while the other is imported from the sklearn library. The performance of both models is compared based on their accuracy.

The model developed from scratch demonstrates the ability to create a machine learning model independently, without relying on existing libraries. It achieves an accuracy of 80% on the test set, indicating that it correctly predicts the outcomes for 80% of the test samples. Creating a model from scratch involves manually implementing the necessary algorithms and procedures, which can be a time-consuming and challenging task.

Fig.10. Accuracy of our model from scratch

On the other hand, the second model is imported from the sklearn library, specifically the logistic regression module. Logistic regression is a widely-used algorithm for classification tasks. This model achieves an accuracy of 84.7% on the test set, showcasing the effectiveness of the pre-built implementation provided by sklearn. The sklearn implementation offers optimized algorithms and streamlined processes, allowing for faster model development and evaluation.

Fig.11. Accuracy of our model from sk learn

Comparing the accuracies of both models, we find that they are relatively similar. Both models achieve high accuracy, indicating their effectiveness in predicting outcomes. The slight difference in accuracy between the two models may stem from various factors, such as differences in optimization strategies, specific implementation details, or hyperparameter settings.

The comparable accuracies suggest that the model developed from scratch performs reasonably well in comparison to the sklearn model. It demonstrates the capability of manual implementation, albeit with potential additional effort required. Leveraging the optimized implementation from sklearn provides the convenience of readily available functionalities and efficient algorithms.

This comparison highlights the versatility and performance of logistic regression as a classification algorithm. It also emphasizes the advantages of utilizing pre-existing libraries like sklearn for faster development and evaluation. The choice between developing a model from scratch or using pre-built implementations depends on the specific requirements, resources, and expertise available to the developer.

VIII. CONCLUSION

In conclusion, this project aimed to develop a machine learning technique for effective heart disease prediction. The dataset used in this study consisted of 14 important attributes and 303 patient details. The dataset was pre-processed to ensure its quality and suitability for analysis.The logistic regression model was implemented from scratch and its

accuracy was verified using the logistic regression model provided by the sklearn library. After comparing the accuracies, it was found that our custom logistic regression model exhibited high accuracy in predicting heart disease outcomes. This suggests that the model can be effectively used for heart disease prediction in the medical domain.

Moving forward, there is scope for further research by incorporating other machine learning algorithms such as Naïve Bayes' Classifier, Decision Tree, K-Nearest Neighbor (K-NN), and Random Forest Algorithm. Comparing these algorithms in terms of their accuracy using both custom implementations and sklearn models would provide valuable insights.The results obtained from this project lay the foundation for a comprehensive literature survey paper that will explore and compare various machine learning algorithms for heart disease prediction. By evaluating and benchmarking these algorithms using appropriate evaluation metrics, the paper aims to provide insights into their strengths and weaknesses and guide future research in this domain.

Overall, this study demonstrates the effectiveness of logistic regression in predicting heart disease outcomes. It highlights the potential for further research and exploration of alternative algorithms to enhance the accuracy and reliability of heart disease prediction models in the medical field. By developing accurate prediction models, we can contribute to improving patient outcomes and decision-making in the medical domain.

REFERENCES

[1] Nayab Akhtar et al. “Heart Disease Prediction” In Conference: Heart Disease PredictionAt: Rawalpindi February 2021.

[2] Armin Yazdani et al. “A novel approach for heart disease prediction using strength scores with signifcant predictors” (2021)

[3] Harshit Jindal et al. “Heart disease prediction using machine learning algorithms”. In IOP Conf. Series: Materials Science and Engineering (2021)

[4] Chintan M. Bhatt et al. “Effective Heart Disease Prediction Using Machine Learning Techniques” (2023)

[5] R. Indrakumaria et al. “Heart Disease Prediction using Exploratory Data Analysis”. In International Conference on Smart Sustainable Intelligent Computing and Applications under (ICITETM2020)

[6] Rohit Bharti et al. “Prediction of Heart Disease Using a Combination of Machine Learning and Deep Learning”. In Computational Intelligence and Neuroscience Volume 2021.

[7] Xin Qian et al. “A Cardiovascular Disease Prediction Model Based on Routine Physical Examination Indicators Using Machine Learning Methods: A Cohort Study” In Front. Cardiovasc. Med., 17 June 2022 Sec. Atherosclerosis and Vascular Medicine Volume 9 - 2022 |

[8] Pooja Anbuselvan et al. “Heart Disease Prediction using Machine Learning” In International Journal of Engineering Research & Technology (IJERT) 2020.