Cardiovascular Disease Prediction

Rahul Vallapureddy, Vijaya Lakshmi Challagulla, Kalpana Danaboyina

Abstract **- Recent advances in machine learning have significantly enhanced our capacity to predict and identify health crises, disease populations, and disease stages. This technology may be very beneficial for the early detection and prevention of cardiac disease, which has become the leading cause of death worldwide. This is one of the areas where it may be especially useful. Predicting cardiac disease is a complex, involved procedure that requires the use of sophisticated data analysis techniques. The early detection of heart disease has the potential to reduce the number of fatalities caused by cardiovascular ailments. We employed a variety of machine learning strategies for the early diagnosis of cardiac disease in order to surmount this obstacle. Among these techniques were Decision Tree, Random Forest, Support Vector Machine, and Gradient Boosting. We were able to determine a person's risk of developing heart disease using these algorithms, which accounted for a variety of variables, including age, gender, cholesterol levels, blood pressure, blood glucose levels, and maximal heart rate. In this research, we utilized two distinct datasets. To determine which method produced the most accurate results, we compiled and organized two datasets before separating them into test and training sets. Support Vector Machine achieved the highest accuracy rating of 92% for the first dataset, while Random Forest achieved the highest accuracy rating of 89% for the second dataset, as determined by our research. We were able to enhance the Gradient Boosting classifier's accuracy to 85% by combining the two datasets. These results highlight the importance of using advanced data analysis techniques to address complex health challenges, demonstrate the potential benefits of machine learning algorithms in predicting and preventing heart disease, and identify which clinical features are more strongly associated with predicting cardiovascular disease.**

I. INTRODUCTION

The heart is regarded as the most important organ in the human body. Globally, heart disease is currently the primary cause of death. Each year there are 17,9 million fatalities worldwide. The term "heart disease" encompasses a variety of cardiac conditions. Four out of every five cardiovascular deaths are caused by heart attacks and strokes, and premature mortality accounts for one-third of all cardiovascular deaths in those younger than 70 years old. We can reduce our likelihood of developing heart disease by modifying our diet and by taking medication. The three most important risk factors for coronary heart disease are cigarette smoking, elevated blood pressure, and excessive cholesterol levels. More than half of the population, or 47% of all Americans, have at least one of these risk factors, according to CDC estimates [1]. Several other factors, such as diabetes, obesity, being overweight, having poor eating habits, being physically inactive, and excessive alcohol consumption, may also increase the risk of developing heart disease. Premature deaths are preventable if the individuals at greatest risk for developing cardiovascular diseases are identified and appropriate preventative measures are taken [2] for these individuals. We are able to reduce the number of cardiovascular disease-related deaths by reducing the prevalence of risk factors, developing standards of care, enhancing the healthcare system's capacity to provide care for people with cardiovascular diseases, and monitoring disease patterns and trends to inform national and international interventions.

Machine learning is a technique that can be used to manipulate data and acquire previously unknown or previously known but potentially useful information about it. The discipline of machine learning has an extraordinarily broad and diverse spectrum of applications, the scope of which is continually expanding. Machine learning employs a wide variety of classifiers derived from supervised learning, unsupervised learning, and ensemble learning to make predictions and assess the reliability of the presented dataset. In the present day, the healthcare industry generates immense quantities of complex data pertaining to patients, medical equipment, maladies, hospital resources, electronic patient records, and a variety of other topics. There are many obstacles to performing an excellent analysis, including time-dependent performance, a lack of experience, inaccurate results, and difficulty in updating knowledge due to the numerous characteristics present in heart datasets, which include both important and irrelevant and redundant characteristics. All of these factors can make it challenging to conduct an exceptional analysis. In the healthcare industry, machine learning-based technologies have the potential to reduce costs while simultaneously increasing efficiency and efficacy. It is possible to autonomously predict the cardiovascular health of a patient, which has the potential to significantly improve medical diagnosis and treatment. It is possible to accurately predict whether or not an individual has cardiovascular disease based on their clinical history.

Our study can use medical history to predict heart problems. Heart disease signs may be identified with it. The research uses machine learning algorithms to identify whether a patient's medical factors such as gender, age, chest discomfort, fasting blood sugar level, etc., indicate cardiovascular heart disease. Cost-effective methods are utilized to assess heart disease risk.

This study was inspired by Machine Learning research in cardiovascular heart disease diagnostics. Nabaousia Lauridi et al. [3] proposed a machine-learning-based cardiovascular disease diagnosis method. They tried mean value, KNN, MICE, RF, and filling-in-the-blanks algorithms to categorize cardiovascular disease patients. Stacking surpassed other machine learning methods in accuracy, F score, precision, and sensitivity.

Ahmed Ismail et al. [4] proposed a real-time health issue prediction system based on big medical data processing on the cloud, where medical parameters are sent to Apache Spark to extract attributes and apply machine learning algorithms to predict healthcare risks and send alerts and recommendations to users and healthcare providers. Compared to literature, their heart disease prediction algorithm has a 90.6% predictability. Using usable characteristics, they used SVM to improve prediction.

Jaishri Wankhede et al. [5] used feature selection to predict cardiac disease. Their proposed method of combining decision function-based chaotic salp swarm algorithm and IESFO algorithm outperformed other classifiers like support vector

machine, KNN, random forest, decision tree, gradient boosting, and logistic regression with 98.7 percent accuracy for the CVD dataset and 98 percent for the UCI dataset.

II. DATA

In our research we used dataset from Kaggle to forecast heart disease.

A. *Dataset:*

The Dataset contains 70,000 records of patient data. It contains 11 attributes with the target variable describing the presence (represented as 1) or absence (represented as 0) of heart disease. Input features are of three types. Objective information which contains the descriptive information of the patients. Examination information containing the results of medical examination. Subjective information containing the information given by the patients. The input features include age, gender, height, weight, systolic blood pressure, diastolic blood pressure, cholesterol, glucose, smoking, alcohol activity and physical activity.

Table 1Attributes of the Datasets

Sl.No	Observation	Description
1	Age	Patients age in years
2	Height	Height of the person in centimetres
3	Weight	Weight of the person in kilograms
4	Gender	Male and female
5	ap_hi	Systolic blood pressure
6	ap_lo	Diastolic blood pressure
7	cholesterol	1: normal, 2: above normal, 3: well above normal
8	Glucose	1: normal, 2: above normal, 3: well above normal
9	Smoking	Whether patient smokes or not
10	Alcohol activity	Consumption of alcohol
11	Physical activity	Presence of physical activity or no
14	output	Patient has heart disease or not

B. *Data Preprocessig:*

Python's panda's library is used to open the CSV file, and the dataFrame function of head () is used to output the first five rows of data. This is done after making certain that all requirements have been satisfied and that all required modules have been imported. Seventy thousand records relating to cardiovascular disease are included in this file, along with thirteen characteristics that connect to each record. After that, the structure of the whole dataset will be printed.

We have used the 'change' function to rename the columns of the panda data frame to 'ap_hi: systolic' and 'ap_lo:Diastolic,' 'gluc: Glucose,' 'alco:Alcohal,' 'active:Physical_activity,' 'cardio:CV_disease,' and then printed all the characteristics. To provide details such as each column's non-null values and datatype, the code makes use of the info methods of the pandas' data frame. The result displays the number of rows, columns, datatypes, and non-null values found through information checking for missing values and data handling requirements.

We have converted days to years using the pandas age format. The computation divides the age column by 365 days to year, rounding the closest integer results with the round () function and converting datatype integer values with the as type () function. Following the execution of the code, a new column will be added to the data frame.

The plot was made in Python using the pie () function of the seaborn package. This would result in a pie chart displaying the percentage of values in the DataFrame df's 'CV_disease' column. It aids in distinguishing between cardiovascular and non-cardiovascular illness.

Outliers may have a significant influence on the mean, standard deviation, and variance of a dataset. To improve data analysis, remove outliers. Reducing noise: Outliers in a dataset may not be representative of the population. Outlier removal enhances data representation. Outliers may wreak havoc on statistical and machine learning models. Outliers should be removed to improve model accuracy and insights.

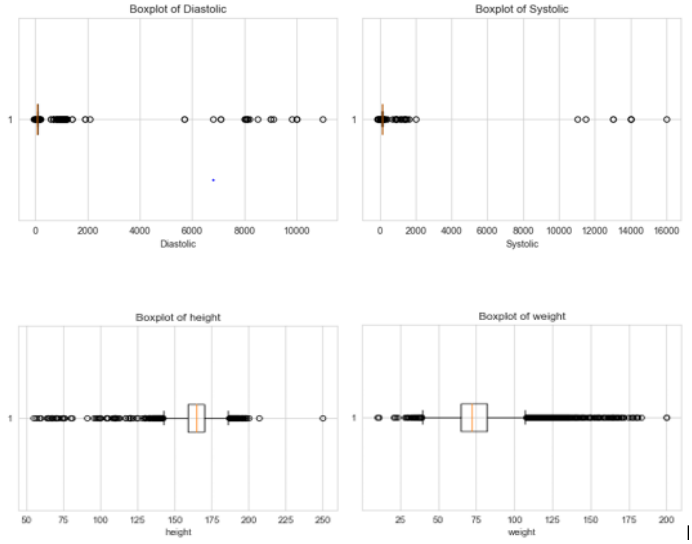

Figure 1Box plot of attributes having outliers.

Using the matplotlib. pyplot module and imported as plt, create a histogram to show all information in each column of a data frame with the chosen plot size and bins. The Show () function displays the findings, while the Histogram function visualizes the distribution of a variable, including the shape, center, and spread of the data. They may also demonstrate tendencies like as outliers, skewness, and multimodality. By creating a histogram for each column in a dataset, we can rapidly observe how the data is distributed and identify any issues, such as skewed data or outliers.

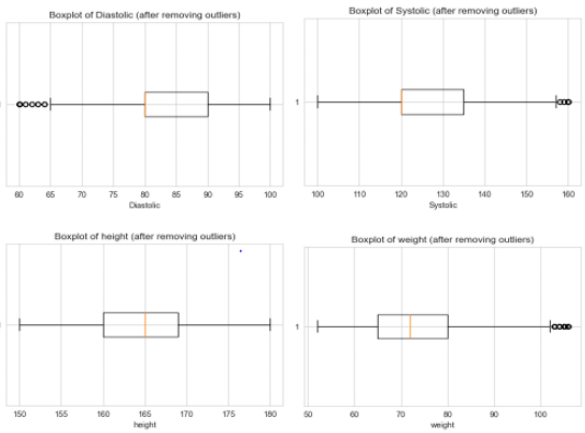

Figure 2Box plot of attributes after removing outliers.

Calculating Body mass index using height and the weight values which are given in the dataset and creating a new column in the dataset. The body mass index (BMI) is a standard measurement for comparing a person's weight to their height. It is used to determine if a person is underweight, at a healthy weight, overweight, or obese. BMI is a measure of general health and the likelihood of developing specific diseases. By computing BMI for everyone in a dataset, we can assess the distribution of BMI values and identify patterns or trends.

$$BMI = \frac{weight\ \left(\frac{kg}{lb}\right)}{height2\ \left(\frac{m2}{in2}\right)}$$

An individual data set is created for each patient using their systolic and diastolic blood pressure values. To approximate the average pressure in the arteries during a cardiac cycle, the formula for MAP is (2/3 * diastolic) + (1/3 * systolic). Using the MAP values, we can create a new column in the DataFrame called "MAP" to store the results. The systolic reading is at the top of the column and the diastolic number at the bottom. Since it includes both the systolic and diastolic values, the mean arterial pressure (MAP) is a decent surrogate for the average pressure in the arteries during a cardiac cycle. It's feasible to measure blood pressure accurately. MAP is widely used in clinical settings for a variety of purposes, including blood pressure monitoring and the calculation of cardiac output and systemic vascular resistance. By determining each subject's MAP, we may examine the overall distribution, identify trends and outliers.

$$MAP = \frac{2\ Diastollic\ Blood\ Pressure + Sistolic\ Blood\ Pressure}{3}$$

The method first generates a histogram for a specified column age in a dataset using the Seaborn package, and then sorts the histogram's produced data according to CV_disease. Within the dataset, there is a variable referred to as "column variable" that holds a list of the column names that will be used for the visualization.

We created the histogram plot by using the Sean born library, and we grouped the data by the "CV_disease" variable. This allowed us to build separate histograms for age, height, weight, gender, systolic and diastolic blood pressure, alcohol intake, and body mass index (BMI). The graph that was developed may be used to make age comparisons between people in the sample who suffer from cardiovascular disease and those who do not have the condition. By constructing histograms for each column, we can investigate the distribution of each column variable, look for trends and patterns, and detect patterns and trends.

When investigating the distribution of continuous data, histograms are a very helpful tool. In this scenario, the plot may assist in determining whether there are any significant variations in the height distribution of those who suffer from cardiovascular disease and those who do not. Demonstrate, by way of illustration, that those who suffer from cardiovascular disease are more likely to be of a shorter or taller stature than those who do not have the condition. Using this information, possible risk factors for cardiovascular disease might be identified.

We display the distribution of people who have cardiovascular disease by each column in every plot. This represents both people who have the illness and those who do not have the condition, and it allows us to detect variations in the disease that are caused by different characteristics of those who have it.

To illustrate the pairwise correlations between the numerical characteristics in the dataset data, this piece of code generates a heatmap by utilizing the heatmap () function that is provided by Seaborn. A grid is used to create the heatmap, and inside each cell of the grid is a representation of the correlation coefficient between two characteristics. The correlation coefficient is a statistical metric that reflects the degree and direction of a linear connection between two variables. It does this by comparing the values of the two variables.

The fig size option allows the figure's size to be customized in inches. Because True is the value that is assigned to the cannot parameter, the process of adding the numerical values of the correlation coefficients in each cell is carried out. The c map option is responsible for customizing the color map that is used when displaying the correlation coefficients.

The heatmap that was produced therefore may assist in determining which characteristics are significantly connected with one another. A greater connection (either positive or negative) is represented by darker colors, while a lesser correlation or no correlation at all is represented by lighter hues. A positive correlation between two variables indicates that those variables tend to change in the same direction at the same time, while a negative correlation between two variables indicates that those variables tend to change in the opposite direction at the same time.

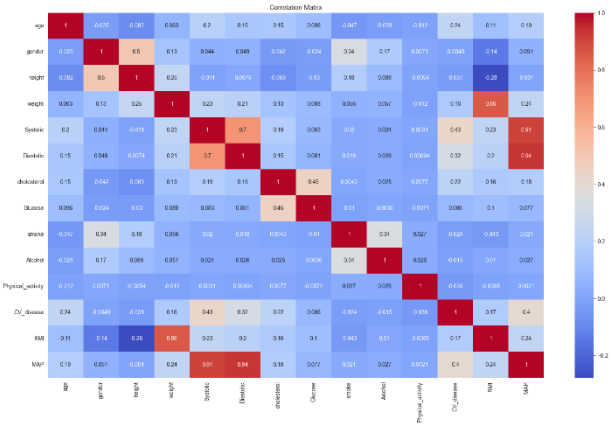

Figure 3Correlation heatmap

This information may be valuable for feature selection or feature engineering since it can help discover redundant or strongly correlated characteristics that may not be effective for developing a predictive model. Additionally, this information can help identify features that may not be relevant for building a predictive model. In addition to this, it may assist in the identification of probable correlations between characteristics, which may be of use in comprehending the underlying processes of the issue that is being investigated.

In order to carry out one-hot encoding on the categorical variables included in the dataset data1, this code makes use of the get_dummies() function provided by Pandas. The process of transforming categorical information into numerical variables that may be included into machine learning models is referred to as one-hot encoding.

The observation belongs to that category if its value is 1, else it is 0. The get_dummies() method generates binary columns for each category in the categorical variables. Binary columns replaced category columns. The columns argument in this piece of code allows the user to choose the categorical variables that are to be one-hot encoded. This process has resulted in the creation of new columns to the dataset def, one for each distinct category included inside the variables.

III. METHODOLOGY

The data set is split into two parts-one part is features, labeled as variable X- and second-part target variable Y. Splitting the data into training and testing set using the train_test_split function from the Sklearn model selection library will evaluate the performance of machine learning models. Random state parameters ensure the data is split in the same way every time. Moreover, text size parameters are used for testing data. To improve the performance of machine learning models and to scale the input structures we use the standardscaler from the preprocessing library. To ensure the same scaling parameters are upper on both test and train we used fit transform method on the training data to fit the standard scalar and then transform the data using scaling parameters from the training data. Standardization helps to increase the accuracy of machine learning models as well as data splitting will help to improve the performance of new model or unseen data that ensures reliability in the meantime of predicting outcomes.

A. Logistic Regression

Logistic regression is a popular supervised Machine learning algorithm Used for predicting the presence or absence of a heart disease. It works by estimating the probability of variable according to various factors Such as age, gender, blood Pressure, Cholesterol levels and other medical indicators. This method is useful when the target is two possible outcomes whether technician has a heart disease or not. Moreover. This method is a popular choice for healthcare professionals who don't have extensive knowledge of machine learning models, as it is a simple algorithm. Logistic regression can be upgraded with the new data, allowing for ongoing monitoring of heart disease medication. It also provides valuable insights into the importance and various predictive variables in determining the likelihood of heart disease. In this. Prediction, we used logistic regression algorithms that has tuning hyperparameters such as C, penalty, solver to train the data using cross validation. When the model was trained followed by making predictions on the test data to generate a classification to evaluate the performance, this allowed to optimize the model accurately to predict the target variable by avoiding overfitting on the training data.

B. Decision Tree Classifier

Decision tree is a supervised learning algorithm that have a set of rules to protect the output the primary objective of this algorithm is to keep suffering the feature space by giving the rules until all the data points are specified correctly and then by separating the data at the root node and splitting it using the future and the best parameters are prizes and if you stop criterion is met by reaching the maximum depth of the tree when we put this plate would not improve the classification accuracy. To evaluate the performance of this classifier was with the data into test and training in that respect to values. This parameter specifies the number of cross validations to use while training the model to help reduce overfitting. Moreover, we use scoring parameter to specify the evaluation metric to use for selecting the best hyperparameters.

C. Naïve Bayes

It is based around the base theorem that states the probability of hypothesis belonging with class given some observed evidence which is proportional to the product of probability of the hypothesis and the likelihood of the evidence given by the hypothesis. New Balance has several variants including gussian NB and multinomial and depending on the nature of the futures it used when the features are continuous and following gaussian distribution. Need bias is a simple but powerful algorithm which is widely used classification task especially in text analysis and natural language processing. In this model we imported required libraries from the glassware for calculating the accuracy score and generating the classification report. Next create an instance of ambiguous fair and train it on the training set using a bit function and reduce the training classifier to make predictions on the test set using a product function for using the function and we generated the classification report using function which provides precision recall F1 score and supporting metrics for each class.

IV. RESULTS

We analyze the results by evaluating several metrics including precision, recall and accuracy. These metrics show the performance of models in predicting the target variable.

False positive (FP)- If a model predicts a positive outcome incorrectly.
False Negative (FN)- if a model predicts a negative outcome incorrectly.
True positive (TP)- if a model predicts a positive outcome correctly.
True Negative (TN)- If a model predicts a negative outcome correctly.

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

$$F1\ Score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

Based on those results. it appears the logistic regression has the highest accuracy of 73 percent compared to the other two. Even the precision score is also highest for the decision tree algorithm. Whereas Recall score is highest for Naïve bayes and the F1 score is highest for Decision tree with 74 percent.

Table 2Observations

Name of model	Accuracy	Precision	Recall	F1 score
Decision Tree	0.73	0.72	0.76	0.74
LR	0.72	0.70	0.76	0.73
NB	0.59	0.56	0.87	0.68

V. Conclusion

Body mass index, Cholesterol, Mean arterial pressure are the most related clinical features for cardiovascular diseases. Hyperparameter tuning helped to get improved accuracy which means the classifiers can predict the presence or absence of heart disease accurately by taking clinical features into consideration. We may investigate deep learning algorithms to see better results.

In this research we used three different machine learning algorithms such as Decision tree, Logistic regression, Naïve bayes for prediction of heart disease. We used dataset which contains 70,000 records of patient data which is available in Kaggle. Among three algorithms Decision Tree and LR performed similarly, with Decision Tree performing slightly better, while NB performed significantly worse than the other two models in terms of accuracy and precision. However, NB had the highest recall, indicating that it was better at identifying positive cases but had more false positives than the other two models. The F1 score, which is a harmonic mean of precision and recall, also shows that Decision Tree had the best overall performance, followed by LR and then NB.

CONTRIBUTIONS

Contributions to this project – performed equally by everyone. Data extraction and data preprocessing is done by Vijayalakshmi, data visualization is done by Rahul and algorithms where done by Kalpana. Power point and report were done equally by everyone.

References

[1] Centers for Disease Control and Prevention. (2023, April 14). Heart disease. Centers for Disease Control and Prevention. Retrieved April 21, 2023, from https://www.cdc.gov/heartdisease/index.htm

[2] World Health Organization. (n.d.). Cardiovascular diseases. World Health Organization. Retrieved April 21, 2023, from https://www.who.int/health-topics/cardiovascular-diseases#tab=tab_1

[3] Louridi, N., Douzi, S., & El Ouahidi, B. (2021). Machine learning-based identification of patients with a cardiovascular defect. Journal of Big Data, 8, 1-15.

[4] Ismail, A., Abdlerazek, S., & El-Henawy, I. M. (2020). Big data analytics in heart diseases prediction. Journal of Theoretical and Applied Information Technology, 98(11), 15-19.

[5] Wankhede, J., Kumar, M., & Sambandam, P. (2020). Efficient heart disease prediction-based on optimal feature selection using DFCSS and classification by improved Elman-SFO. IET systems biology, 14(6), 380-390.