Heart Disease Prediction using Supervised Machine Learning Algorithms

G. Anuhya(AP19110010462), K. Varshini Reddy(AP19110010403) , M. Tanushq(AP19110010416), Sripada Srikar(AP19110010474), Souvik Mandal(AP19110010485)

Abstract: Cardiovascular disease (CVD) is a challenging health problem in today's world. It is reported that four out of every five CVD deaths are caused by heart-attacks. CVD claims the lives of 17.9 million people each year, accounting for almost 32% of all deaths worldwide. We need to put in place a system that can detect heart-attack symptoms early and hence can be prevented. It is infeasible for a common person to examine expensive tests like the ECG at a regular interval. It is vital to recognise and treat such disorders as soon as possible by developing a system that is convenient and reliable in estimating the likelihood of a heart disease. To address this challenge, machine learning (ML) can be used to anticipate diseases in a variety of fields, including healthcare. The purpose of this research is to apply various ML algorithms (e.g., logistic regression, Random Forest, etc.) to find the most important predictors of heart disease and to forecast total risks as a binary classification task. The main objective is to determine whether the patient has a 10-year risk of heart disease or not. To implement and validate our model, the data collection, preparation, classification, and results are analysed using the Python language in Google Collab environment. We can predict the risk of a heart attack in a person based on information such as age, blood pressure, artery thickness, and so on. Our method can be used tackle the possibility of heart disease.

Keywords: CVD, ECG: Electrocardiography, ML: Machine Learning.

INTRODUCTION

The heart is essential to the proper functioning of the human body. According to the World Health Organization, heart illnesses cause 12 million deaths worldwide each year. Cardiovascular disease account for half of all deaths.

Health is wealth. In today's society, it is more vital than ever to live a healthy lifestyle, but this is not achievable for everyone due to a variety of circumstances such as poverty, stress, unhealthy diets, and so on. According to the World Health Organization, 50 million deaths occur each year, with heart disease accounting for 12 million deaths. Majority of all deaths are caused by cardiovascular disease.

This occurs because people will not attend to regular check-ups because they cannot afford it, and because diagnostic labs and skilled doctors are very few in most countries. In this scenario, machine learning comes into play, as we can forecast the early risk of cardiovascular disease using machine learning algorithms.

The risk of death from heart disease can be reduced if it is detected early.

Machine learning algorithms are capable of dealing with enormous data sets, pre-processing the data and identifying patterns. Data sets are divided into testing and training data, with the training set being trained using various machine learning algorithms to predict the target variable. To obtain the appropriate accuracy, the model's accuracy is compared to the accuracy of other models.

Literature Survey

Many studies have been conducted to evaluate the classification accuracy of various machine learning algorithms using various datasets.

Singh and Choudary et al. [1] created a robust classifier using the Ada-Boost Algorithm. It creates a model using training data and then creates a second model to fix the inaccuracy in the first model. Ada-boost, a binary classification technique, was the first effective boosting algorithm. Their model was overfitting when they used the decision tree approach on 13 attributes from their data set. For the optimization of the output generated by the Decision tree, they used the Ada-Boost algorithm. They achieved an accuracy of 89% using the Ada-Boost algorithm.

Sharma, Yadav, and Gupta et al.[2] have looked into a variety of applications that demonstrated the importance of machine learning approaches in a variety of areas. They have used a benchmark dataset of UCI Heart disease prediction for this research work, which consists of 14 different parameters related to heart disease. They used a variety of machine learning techniques, such as random forest and decision trees, to achieve accuracy of 99 and 88 percent, respectively. They discovered that SVM, with a 98 percent accuracy, and random forest, with a 99 percent accuracy, produced the greatest results when compared to all other ML algorithms.

Archana and Kumar et al[3] used a variety of machine learning algorithms, which include linear regression, decision tree, support vector machine, and k-nearest-neighbour, to train the

UCI machine learning data set, which contains 303 samples and 14 input measures, with 73 percent of the data used for training and 37 percent for testing. When using a decision tree, they found that the number of nodes is unbalanced, resulting in overfitting and lower accuracy. After experimenting with various algorithms, they discovered that Knn has the best accuracy with 87%.

Bertsimas et al. [4] proposed a novel methodology to extract ECG-related features and predict the type of ECG recorded. Our models lever-age a collection of almost 40 thousand ECGs labelled by expert cardiologists across different hospitals and countries and are able to detect 7 types of signals: Normal, AF, Tachycardia, Bradycardia, Arrhythmia, Other or Noisy. We exploit the XGBoost algorithm, a leading machine learning method, to train models achieving sample F1 Scores in the range 0.93 – 0.99.

Methodology

I. Data Collection:

The system processing begins with data collection, for which we used the Cleveland repository dataset from Kaggle, which has been well validated by a number of researchers. There are 1025 instances and 14 attributes in this collection. In this research, we used 80 percent as the training dataset and 20 percent as the testing dataset.

II. Attribute Selection:

Attribute of dataset is a dataset property that is utilised for system and for our work many attributes like chol of person, gender of person, age of person and many more displayed in TABLE.1 for prediction system.

S.No	Attribute	Description	Type
1	Age	In years	Numeric
2	Sex	Male, Female	Nominal
3	Cp	Angina-1, abnang-2, notang-3, asympt-4	Nominal
4	trestbps	Resting Blood pressure in mm hg	Numeric
5	Chol	Serum choklestereol in mg/dl	Numeric
6	fbs	Fasting blood sugar	Nominal
7	restecg	Resting cardiographic results	Nominal
8	thalach	Maximum heart rate observed	Numereical
9	exang	Exercise with angina has occurred	Nominal
10	oldpeak	ST depression induced through exercise	Numerical
11	slope	Slope of ST segment	Nominal
12	ca	Heart status	Numerical
13	thal	Number of major vessels ranging from 0-3 color by fluoroscopy	Nominal
14	target	Output class	Nominal

III. Data pre-processing:

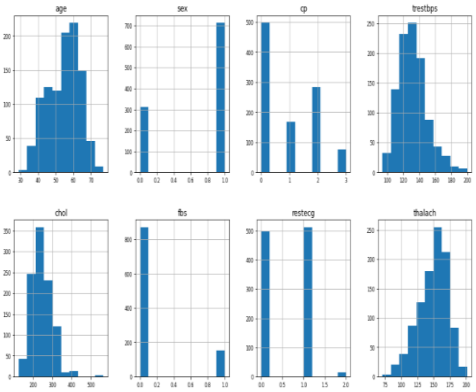
Some values in datasets may be missing, resulting in imperfect accuracy. To overcome this, we must either remove or replace the missing data. Dropping the key attribute with the most missing values may cause the model's overall performance to be biased. Instead of dropping, we can replace the missing values with the attribute's mean/median/mode.

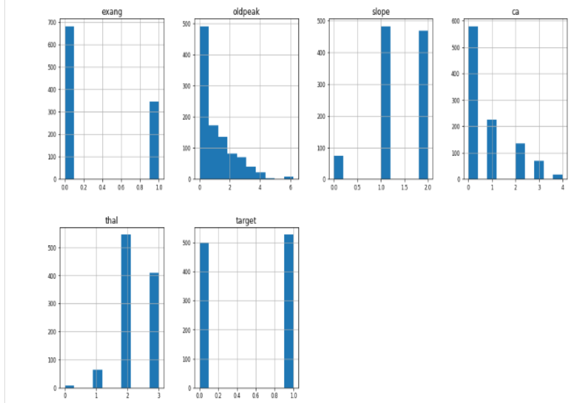
There are no missing values in our dataset. As a result, we did not perform data cleaning.

IV. Data Visualization:

Histogram of attributes:

The attribute histogram displays the range of dataset attributes as well as the code used to build it.



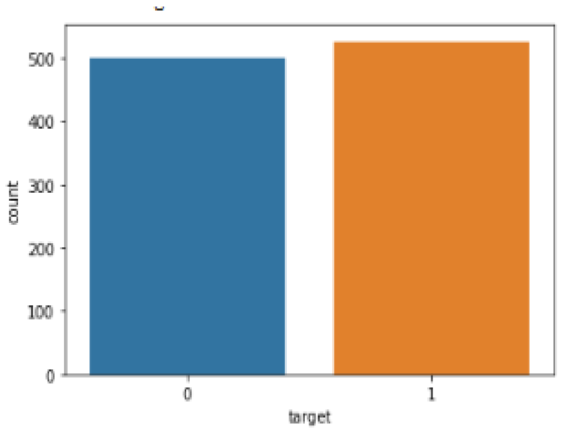


Target attribute:

Distribution of target attribute

0 – represents no heart disease

1 – represents heart disease



V. Building Model:

The model is built in Jupiter notebook. The software handles common data mining tasks like data preparation, clustering, classification, regression, visualisation, and feature selection with ease. It provides a simple environment for loading data in the form of files, URLs, or databases.

Three accuracy measures have been considered for comparison of the four models, they are as follows:

- Accuracy:
  The percentage of all correct predictions divided by the total number of instances is used to calculate a classifier's accuracy.
  Accuracy = [Number of True Positives + True Negatives]/ [Total Instances]

- Precision:
  It is the average likelihood of retrieving relevant data.
  Precision = Number of true positives/Number of true positives + False positives.

- Recall:
  It is the average likelihood of complete retrieval.
  Recall= True positives/True positives + False negative.

Proposed Models

A. Logistic Regression:

Logistic regression is a Machine Learning algorithm, which comes under the Supervised Learning technique. It's a method for predicting a categorical dependent variable from a set of independent variables. The dependant variable in logistic regression is always binary (with two categories). Our model is predicting whether or not a person will get heart disease (0-Healthy,1-sick).

Logistic Regression involves fitting an equation of the form to the data [5]:

Y = ß0 + ß1x1 + ß2x2 + … + ßnxn – >eq. 1

B. SVM:

Support Vector Machine is a type of supervised learning technique that may be used to solve both classification and regression tasks. SVM is used when data has exactly two classes. SVM algorithm helps to find the best line or decision boundary; this best boundary or region is called a hyperplane. SVM algorithm finds the closest point of the lines from both classes. SVM is based on mathematical functions and is used to model complex and real-world problems. SVM performs well on data sets that have many attributes [1].

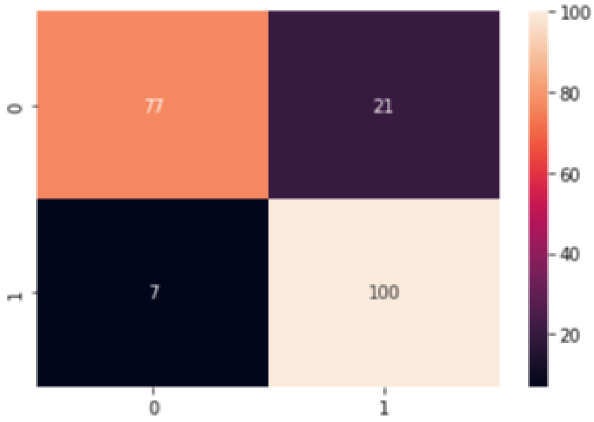
C. **Knn**:

The K-nearest neighbours' algorithm is a method for supervised classification. It categorises objects that are dependent on their nearest neighbours. The data are clustered based on their similarity, and K-NN can be used to fill in the missing values in the data. Various prediction approaches are applied to the data set when the missing values are filled. Nearest neighbour classification is used mainly when all the attributes are continues.

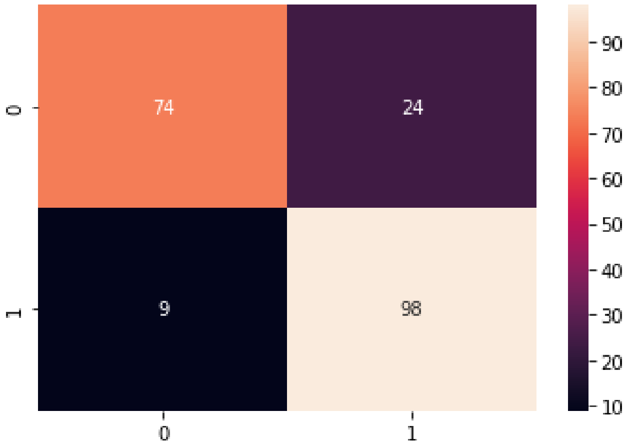
Results

At the end of our trial, the findings demonstrate that Logistic regression and SVM outperformed knn. The logistic regression model produced an accuracy of 86.33, which is 7% higher than knn. Similarly, the model constructed using SVM has an accuracy of 79.5 %, which is 5% higher than knn. Unfortunately, we did not find knn acceptable for our data.

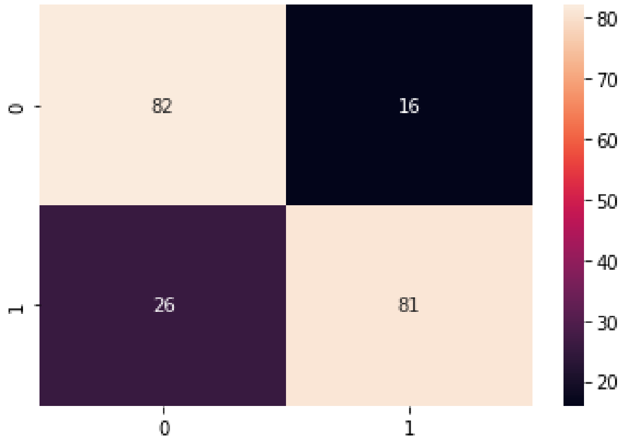
Confusion Matrix of Logistic Regression:



Confusion Matrix of SVM:



Confusion Matrix of Knn:



Performance measure of the models

Model	Accuracy	Precision	Recall
LogisticRegression	86.34	82.64	93.45
SVM	83.9	80.32	91.58
Knn	79.1	83.5	75.70

Conclusion

In our research, we used various machine learning algorithms to predict whether or not a person would develop heart disease. We utilised a dataset from Kaggle that had 1025 instances and 14 attributes. To test the accuracy, we used three different algorithms. By the end of the implementation phase, we observed that logistic regression has the highest accuracy level of our dataset, which is 86.34 percent, and knn has the lowest accuracy, which is 79.1 percent. Other algorithms may work better in other cases and with various datasets, however for our condition, we discovered this result.

References

1. Choudhary, G., & Singh, S. N. (2020, October). Prediction of heart disease using machine learning algorithms. In *2020 International Conference on Smart Technologies in Computing, Electrical and Electronics (ICSTCEE)* (pp. 197-202). IEEE.

2. Sharma, V., Yadav, S., & Gupta, M. (2020, December). Heart disease prediction using machine learning techniques. In *2020 2nd International Conference on Advances in Computing, Communication Control and Networking (ICACCCN)* (pp. 177-181). IEEE.

3. Singh, A., & Kumar, R. (2020, February). Heart disease prediction using machine learning algorithms. In *2020 international conference on electrical and electronics engineering (ICE3)* (pp. 452-457). IEEE.

4. Bertsimas, D., Mingardi, L., & Stellato, B. (2021). Machine learning for real-time heart disease prediction. *IEEE Journal of Biomedical and Health Informatics*, *25*(9), 3627-3637

5. Mythili, T., Mukherji, D., Padalia, N., & Naidu, A. (2013). A heart disease prediction model using SVM-decision trees-logistic regression (SDL). International Journal of Computer Applications, 68(16).

6. https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset