Coronary Heart Disease Prediction Using Machine Learning Algorithms

Moksh Doshi[#1], Jaimik Patel[#2], Nandish Patel[#3], Jenil Bagadiya[#4]

School of Engineering and Applied Science, Ahmedabad University

[#1]AU1940028,[#2]AU1940120,[#3]AU1940130,[#4]AU1940164

***Abstract*—The ability to predict whether or not a patient will have a heart disease in the future using current body observations is a great asset. This article applies some of the machine learning algorithms to the data in order to predict the outcome. It also details the literature related to the algorithms along with a comparison based on standard performance metrics.**

***Keywords*— machine learning, heart, disease, prediction, classification**

I. Introduction

Range of conditions that affect a person's heart are termed Coronary Heart Diseases (CHDs). In modern times, CHDs are one of the major causes of death worldwide with estimates of around 17.9 million deaths in 2019 according to WHO report[2]. This is further magnified by modern lifestyle choices such as stress, smoking, hypertension, abnormal blood pressure, etc.

We aim to answer the question whether it is possible to predict a coronary heart disease using Machine Learning algorithms. Along with that, we learn about the significance of features involved, and decide if they need to be considered or not.

Our currently accomplished tasks are following:

1. We found a suitable dataset which contains the required features and has enough data to build suitable models.
2. We performed Exploratory Data Analysis (EDA).
3. We did feature analysis.
4. We applied 2 out of the total machine learning algorithms we wish to apply in future.

II. Literature Survey

Using classical machine learning algorithms with help of analysis and results of Heart Disease dataset 94.2% accuracy can be achieved. Using the SVM (Supervised Machine Learning) algorithm we can predict the future outcome whether a patient will have heart disease or not using the different data of the patients like age, blood pressure level, sugar level and whether the patient is diabetic or not. The cause behind 17.9 million deaths is because of heart disease.

We can use different machine learning approaches to make a heart disease prediction model like K-Nearest Neighbour, Random Forest, Decision Trees, Naive Bayes. According to the article [1] if we can apply some modifications to the machine learning technique we can achieve a good amount of accuracy in predicting the disease. This research paper gives us a holistic view of various algorithms which could be used to predict whether or not a person will be diagnosed with heart failure in the future.

The other research paper [3] which was used for the literature survey tells us about the pre-selection and post selection of the attributes and also the Naive Bayes algorithm which is based on Bayes theorem for making the independent assumptions. Another ID3 algorithm was used to build the decision trees. And also the KNN clustering algorithm technique was used for clustering of the datasets.

III. Implementation

A. Introduction to Dataset

The dataset was obtained from Kaggle[4]. It consists of 11 different features and 1 target variable. The features are of different types and are obtained from distinct sources. For example systolic & diastolic blood pressure are obtained from physical examination and are of float type while others such as smoking & alcohol intake are binary categorical features obtained from patient responses i.e. subjective. The dataset has presence of factual information in the form of objective features such as height, weight, age, gender, etc. The target variable is a binary categorical variable which depicts the presence or absence of CHD in a specific patient.

B. Problem Statement

Our goal is to predict whether the patient will have coronary heart disease or not (classification) in the future using various different machine learning models.

C. *Pre-Processing Steps*

The data is processed in the following manner to better fit the machine learning models and result in a better outcome.

1) The useless data such as 'id', etc. is dropped
2) The data is cleaned of NaN/Null values i.e. rows with such data are dropped
3) Next, the interquartile range of each and every variable in considered for the next step i.e. 2nd and 3rd quartiles while the 1st and 4th quartiles are dropped
4) Erroneous data is also cleaned. For example where where diastolic bp is greater than systolic bp
5) Extra features such as BMI, Mean Arterial Pressure are engineered from base features such as height,weight and systolic and diastolic blood pressure
6) Finally, the data is normalized using min-max scaling

D. *Exploratory Data Analysis*

Univariate and bivariate data analysis were performed. Various graphical outputs such as box plots, histograms, bar graphs, pairplots and many more were obtained which helped us gain a holistic view of the data distribution. Correlation matrix in particular helped in gaining a preliminary idea of which features affect the outcome the most.

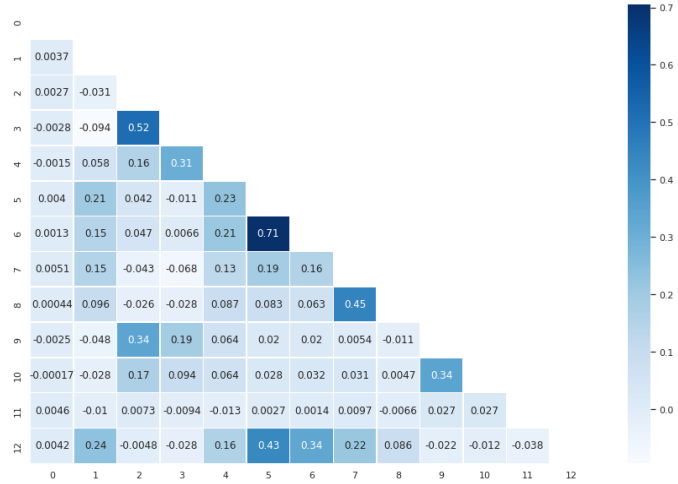

Fig. 1 A lower triangular correlation matrix represented in the form of a heatmap

E. *Algorithms*

As the target variable of the dataset is a binary categorical variable, classification algorithms must be utilized to predict the class i.e. output. Till date we have implemented two of the vast plethora of classification algorithms, namely K-Nearest Neighbours and Logistic Regression. To measure the model performance along with accuracy many other metrics were used such as precision, F1-score, recall, support, etc.

1) *K-Nearest Neighbours (KNN):* The preprocessed data was fed into KNN model to search for the appropriate hyperparameter. The model was measured on the value of K ranging from 1 all the way to 10001 with odd values of K accounting for even numbers of samples (odd K serves as a tiebreaker). The ideal value of K was found out to be 101 neighbours. The generalization gap i.e. the difference between training and testing scores of the dataset was found to be minimal with that specific value of K. Thus, it can be concluded that the algorithm could accurately model the true data as it could the empirical dat

2) *Logistic Regression*: As for Logistic Regression, no special provisions were made. The normalized dataset initially created was passed to the library function as is.

Both of the algorithms were run multiple times with different subsets of the original datasets. For the first iteration the original normalized dataset was passed as arguments to the model. For the second iteration the original dataset with some engineered features (BMI was created from height and weight) was passed as arguments to the models. The features were checked for significance levels of 5% i.e. α=0.05. Features with p-value less than the threshold were kept while the rest were dropped. For the third and final iteration, only objective and examined features (such as height, weight, systolic and diastolic blood pressure, glucose, etc) were passed as arguments to the model.

IV. Results

A popular machine learning library named Scikit-Learn was used to code the machine learning pipeline.

Classifier	Train Accuracy	Test Accuracy	Precision	Recall	F1 score
KNN	0.7299	0.7182	0.72	0.72	0.72
KNN (feature engineering)	0.7293	0.7227	0.73	0.72	0.72
KNN (only objective features)	0.7309	0.7191	0.72	0.72	0.72
Logistic Regression	0.7250	0.7151	0.72	0.72	0.71
Logistic Regression (feature engineering)	0.7251	0.7151	0.72	0.72	0.71
Logistic Regression (only objective features)	0.7223	0.7201	0.72	0.72	0.72

From the results stated above, it can be observed that Logistic Regression performs the best when only objective features are considered with the accuracy of 72.01% while KNN performs best in the case of engineered features with the highest overall accuracy of 72.27%. Taking the mean of all KNN and Logistic Regression, the KNN mean is higher than Logistic mean, hence it can be concluded that KNN is better fit for this dataset.

V. Conclusion

It can be concluded that the possibility of a patient having Coronary Heart Disease (CHD) can be fairly accurately modelled using machine learning algorithms. With the help of more efficient data processing methods along with feature engineering the performance of current models can be improved further. We also plan to implement more advanced machine learning models to find the best fit.

References

[1] R. Bharti, A. Khamparia, M. Shabaz, G. Dhiman, S. Pande, and P. Singh, “Prediction of Heart Disease Using a Combination of Machine Learning and Deep Learning,” Computational Intelligence and Neuroscience, vol. 2021. Hindawi Limited, pp. 1–11, Jul. 01, 2021. doi: 10.1155/2021/8387680.

[2] “Cardiovascular diseases,” World Health Organization. [Online]. Available: https://www.who.int/health-topics/cardiovascular-diseases. [Accessed: 19-Mar-2022].

[3] N. Rajesh, M. T, S. Hafeez, and H. Krishna, “Prediction of Heart Disease Using Machine Learning Algorithms,” International Journal of Engineering & Technology, vol. 7, no. 2.32. Science Publishing Corporation, p. 363, May 31, 2018. doi: 10.14419/ijet.v7i2.32.15714

[4] S. Ulianova, “Cardiovascular disease dataset,” Kaggle, 20-Jan-2019. [Online]. Available: https://www.kaggle.com/datasets/sulianova/cardiovascular-disease-dataset. [Accessed: 19-Mar-2022].

[5] D. Shah, S. Patel, and S. K. Bharti, “Heart disease prediction using Machine Learning Techniques,” SN Computer Science, vol. 1, no. 6, 2020.