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 as follows:

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 a total of 6 machine learning algorithms we originally planned to apply
5. We also performed Principal component analysis to reduce the number of features from 13 to 7

II. Literature Survey

Using classical machine learning algorithms with the help of analysis and results of the 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 the 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 diastolic bp is greater than systolic bp, data with duplicate instances
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.

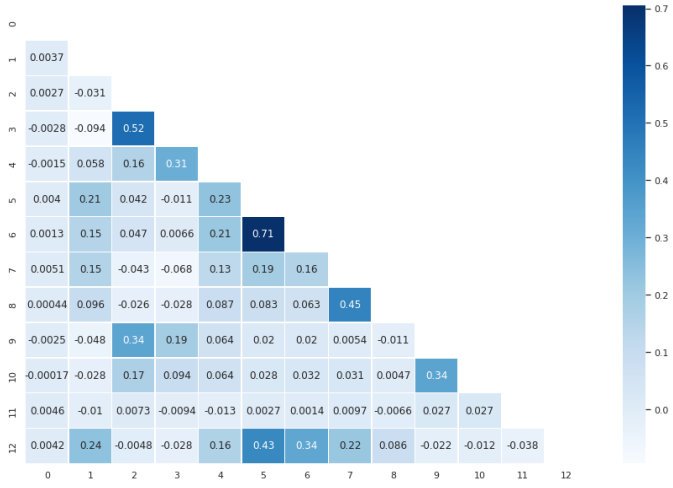


Fig. 1 A lower triangular correlation matrix represented in the form of a heatmap

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 the 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 data

2) *Logistic Regression*: As for Logistic Regression, no special provisions were made. The normalized dataset initially created is 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-values 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.

3) Naive Bayes: The ideal assumption behind using the Naive Bayes algorithm was that all features are independent from each other irrespective of their classes. Bayes' theorem relates the conditional and marginal probabilities of two random events. In simple words, if a patient is observed for certain symptoms then by using Bayes theorem we can compute the probability of a proposed diagnosis given the observation. By using the Naive Bayes algorithm we were able to achieve 70% accuracy on testing dataset points. The advantage of using the Naive Bayes algorithm is that it requires only a small amount of training data to make an estimation of the different parameters mean, variance of the variables.

4) Decision Tree: Decision tree is a supervised form of learning where certain parameters are divided into various data. The main agenda is to split the data into 2 or more homogenous sets which are done based on the most significant attributes to make as many distinct groups as possible. It has mainly 2 categories which are leaves and decision nodes and the algorithm is used for both continuous and categorical dependent variables. The decision tree imitates the human thinking ability when we are making decisions and compares the value of attributes to root attributes along with sub nodes and then it moves further.

5) Support Vector Machine (SVM): In SVM the coordinates of the point will be the support vector for classifying the data into the various groups and in our case it is whether the patient is affected with coronary heart disease or not. The point is plotted based on the number of the features taken, based on the dimensions. SVM is known for its discriminative power for small sample sizes but a larger number of features are involved in high dimensional space. When the SVM algorithm is applied on the testing dataset we obtained 72.67% accuracy on the testing dataset which is the highest amongst other machine learning algorithms.

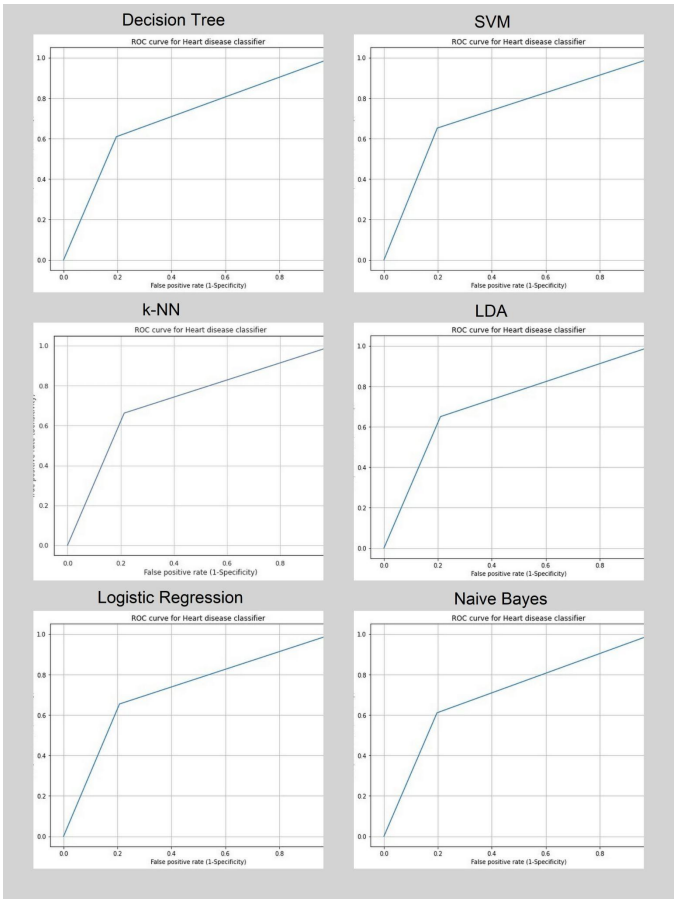
6) Linear Discriminant Analysis (LDA) : LDA is a dimensionality reduction technique for supervised classification problems. LDA identifies a projection vector which helps to maximize the scatter among the classes and minimizes the within-class scatter matrix in feature space. The main goal of using the LDA in coronary heart prediction is to project the higher dimensional features into a lower dimensional feature to avoid the curse of dimensionality and also to reduce the cost of dimensions and resources. Using LDA we got an accuracy of 72.14%.

7) Principal Component Analysis (PCA) : PCA is an unsupervised machine learning algorithm used for the dimensionality reduction method which simplifies the complexity of high dimensional data while returning the particular pattern and trend into lower dimensions. Main purpose is to transfer data into fewer dimensions which act like summaries of the features. In order to observe trends, clusters and outliers of a multivariate data table as a smaller set of features, PCA is very useful.

IV. Results

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

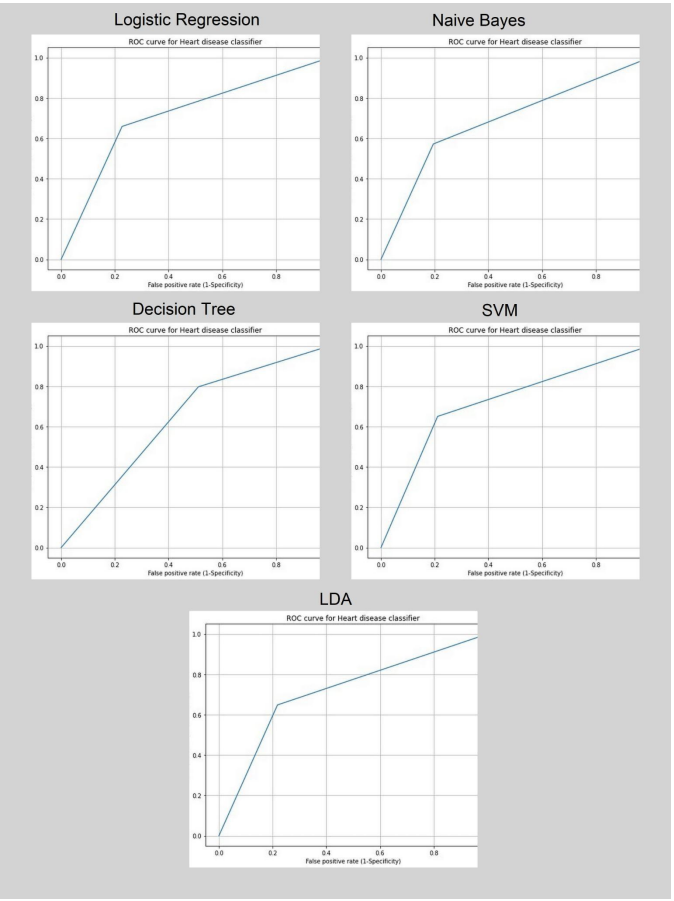
Classifier	Test Accuracy	Precision	Recall	F1 score
KNN	0.7227	0.73	0.72	0.72
Logistic Regression	0.7201	0.72	0.72	0.72
Naive Bayes	0.70863	0.71	0.71	0.71
Decision Tree	0.70639	0.71	0.71	0.70
SVM	0.72679	0.73	0.73	0.73
LDA	0.72143	0.72	0.72	0.72



From the results stated above, it can be observed that SVM performs the best among all other machine learning algorithms with the accuracy of 72.67% while Decision Tree performs worst with the lowest overall accuracy of 70.63%. These accuracies are measured when the PCA is not applied over the dataset.

The following results are of the classifiers after Principal Component Analysis:

Classifier	Test Accuracy	Precision	Recall	F1 score
Logistic Regression	0.7159	0.72	0.72	0.71
Naive Bayes	0.6883	0.70	0.69	0.68
Decision Tree	0.6430	0.66	0.64	0.63
SVM	0.7196	0.72	0.72	0.72
LDA	0.7150	0.72	0.72	0.71



From the results stated below, it can be observed that SVM performs the best among all other machine learning algorithms with the accuracy of 71.96% while Decision Tree performs worst with the lowest overall accuracy of 64.30%.These accuracies are measured after the PCA is applied over the dataset.

V. SVM Hyperparameters

1) Gamma : It controls the distance of influence of a single training point. Its low value indicates a large similarity radius which results in more points being grouped together. For high values, the points

need to be very close to each other in order to be considered in the same group.

2) C : The C parameter tells the SVM optimizer how much you want to avoid misclassifying each training example. If c is small, the penalty for misclassified points is low so a decision boundary with a large margin is chosen. If c is large, SVM tries to minimize the number of misclassified examples due to the high penalty which results in a decision boundary with a smaller margin.

C	Gamma	Accuracy
1	0.001	0.499
1	0.1	0.609
10	0.1	0.684
10	1	0.694
50	1	0.694

We tried to modify the values of gamma and c to see if the accuracy of prediction increased. With their values 10 and 1 respectively we achieved an accuracy of 69.45%. However, it could not beat the previous accuracy of SVM which was 72.679%. It was not a useful technique for the given dataset but could well be handy in some other case

VI. 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. Traditional algorithms worked reasonably well while Support Vector Machine worked the best out of all. These algorithms show similar accuracy with SVM showing better results as compared to the rest. When explored in depth, SVM with kernel tricks, the accuracy again dropped further down, indicating that the default kernel i.e., the linear kernel is the best fit for the data.

VII. GitHub Link

https://github.com/just-a-rookie-2001/CSE523-Machine-Learning-2022-Bug-Smashers

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.

[6] Salhi, D., Tari, A. and Kechadi, M., 2022. Using Machine Learning for Heart Disease Prediction.

[7] Canadian Journal of Medicine, 2021. Analysis and Prediction of Heart Disease Using Machine Learning and Data Mining Techniques.

[8] F Pedregosa. G. Varoquaux, A. Gramfort, V. Michel. B. Thirion, O. Grisel, M. Blondel. P. Prettenhofer, R. Weiss, V. Dubourg, J. Vander plas, A. Passos. D. Cournapeau, M. Brucher, M. Perrot, and E. Duch esnay. "Scikit-learn: Machine learning in Python." Journal of Machine Learning Research, vol. 12, pp. 2825-2830, 2011.

[9] A. Géron, Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems. O'Reilly Media, 2019.