MACHINE LEARNING ALGORITHMS FOR OPTIMISING HEART DISEASE PREDICTION THROUGH HYPERPARAMETER TUNING

Puneet Misra

Assistant Professor
Department of Computer Science,
University of Lucknow, Lucknow,India

Abstract:
The eighth goal of the Global Action Plan for Noncommunicable Diseases states that at least 50% of eligible people should receive medication and counselling (including blood sugar control) to prevent heart attacks and strokes. Using holistic cardiovascular risk methods to prevent heart attacks and strokes is more cost-effective than treatment decisions based solely on individual risk factor thresholds, and should be part of the universal health coverage core benefit plan. At the individual level, in order to prevent the first heart attack and stroke, individual health interventions should target individuals with high overall cardiovascular risk or individuals with individual risk factors above traditional thresholds (such as hypertension and hypercholesterolemia). However, the delayed recognition and diagnosis of celiac disease can cause permanent damage to the heart. Heart failure can be life-threatening, but early treatment for heart disease can help prevent complications. This can cause heart disease, which can lead to complications and problems.Cardiovascular disease (CVD), despite significant advances in diagnosis and treatment, continues to be the leading cause of morbidity and mortality worldwide. To improve and optimize cardiovascular disease outcomes, AI can fundamentally change the way we approach cardiology, especially in imaging, offering us new tools for interpreting data and making clinical decisions. Artificial intelligence techniques such as machine learning and deep learning can also improve medical knowledge by increasing the volume and complexity of data, providing clinically relevant information.

***Index Terms* - Classification, Diagnosis, Heart Disease, Machine Learning, Supervised Algorithm, K-nearest neighbor, Random Forest, Naïve Bayes, Logistic Regression, Hyperparameter tuning.**

I. INTRODUCTION

Cardiovascular disease (CVD) has become the leading cause of death in India. Cardiovascular diseases account for a quarter of all deaths. Coronary artery disease and stroke are the predominant causes and are responsible for> 80% of deaths from cardiovascular disease. Cardiovascular diseases, especially coronary heart disease (CHD), are epidemic in India. According to the Global Burden of Disease, nearly a quarter (24.8 per cent) of all deaths in India is due to CVDs. Even an analysis of the medical certification of cause of death (MCCD) reports points to an increase in the proportion of deaths due to CVD. It went from 20.4 per cent in 1990 to 27.1 per cent in 2004[2]. The first nationally representative study monitoring cardiovascular mortality in India in 15 years found that the chances of dying from coronary artery disease (CAD) - heart problems caused by narrowing of the arteries of the heart - in a population aged 30 to 69 decreased. up. A 2018 study led by Prabhat Jha, director of the Center for Global Health Research at St. Michaels Hospital in Toronto, Canada, and a professor at the University of Toronto, found that deaths from heart attacks rose rapidly in rural India and exceeded those in urban areas. geographic areas from 2000 to 2015. By 2020, coronary artery disease (CAD) is projected to become the most common cause of death worldwide, including India.1 CAD is high among people of Indian descent currently living overseas. 2-6 Gupta and Gupta7 reviewed the research determining the prevalence of coronary artery disease in India. They found that the prevalence of coronary heart disease in urban areas increased from 1% in the 1960s to 9% in the 1990s. In rural areas, this figure increased from 2% in the 1970s to 4% in the 1990s. The incidence is lowest in the northern region (20/100,000), highest in the eastern and central regions (50/100,000 and 63/100,000, respectively), and highest in the southern region (135/100,000). Kerala, Punjab and Tamil Nadu have the highest prevalence rates, followed by Andhra Pradesh, Himachal Pradesh, Maharashtra, Goa and West Bengal [2].

New measures and tools are needed to monitor improvements in cardiovascular health and cardiovascular care over the next decade. As health-technology innovations continue to transform medicine, the ethical implications of machine learning must also be considered.Machine learning in healthcare, also known as deep medicine, is a general term used to describe the use of artificial intelligence to mimic human cognition in a variety of areas, from healthcare to education. AI can fundamentally change the way we approach cardiology, especially in imaging, offering us new tools for interpreting data and making clinical decisions.

Machine learning (ML) is a branch of artificial intelligence (AI) that is increasingly being used in the field of cardiovascular medicine. Basically, this is how computers parse data and decide or classify activities with or without human control [4]. The conceptual framework of machine learning is based on models that take inputs (such as images or text) and predict outcomes (such as favorable, unfavorable, or neutral) through a combination of mathematical optimization and statistical analysis.

Several machine learning algorithms have been applied to daily activities. For example, a general machine learning algorithm labeled SVM can recognize nonlinear patterns for use in facial recognition, handwriting interpretation, or fraudulent credit card transactions1,2. So-called enhancement algorithms used for prediction and classification have been used to identify and process spam. Another algorithm called random forest (RF) can facilitate decision making by averaging multiple nodes.

II. Supervised Machine Learning Classification Algorithms

Classification has been an age-old problem. Early in the 4th century BC, Aristotle tried to group organisms into two classes depending on whether they are beneficial or harmful to a human. He also introduced the concept of classifying all forms of life for organizing the rich diversity in living organisms. Classification is defined as the process of finding a set of models (or functions) that describe and distinguish data classes and concepts, with the goal being to use the model to predict the classes of objects whose class labels are unknown. Thus, classification is a supervised learning problem where the task is to predict the value of a discrete output variable given a set of training examples and a test sample where each training example is a pair consisting of the input object and the desired class. Generally, data classification is a two-step process. In the first step, a classifier is built describing a pre-determined set of data classes or concepts. This is the learning step (or training phase), where a classification algorithm builds the classifier by analyzing or learning from a training set. In the second step, the model is used for classification.

Supervised Learning is an approach to machine learning defined using labeled datasets to train data classification algorithms and predict outcomes. A tagged dataset has an output tag that matches the input for the machine to figure out what to look for in the invisible data. The tagged data is used to train the classifier so that the algorithm works well with unlabeled (not yet tagged) data. Supervised learning models can be a viable solution to eliminate manual classification work and to predict the future based on labeled data. Supervised learning is useful for classification and regression tasks, such as determining the category a news article belongs to, or predicting sales for a specific date in the future. Supervised classification is one of the most common tasks performed by intelligent systems.

Supervised machine learning techniques include linear and logistic regression, multiclass classification, decision trees, and helper vector machines. Some of the more common supervised learning algorithms include support vector machines (SVMs), logistic regression, naive bayes, neural networks, K-nearest Neighborhood (KNN), and Random Forest.

III. Algorithms and Techniques Used

1) *K-NN CLASSIFICATION ALGRORITHM:*The K-NN is supervised learning algorithm. The K-Nearest Neighbor is one of the introductory supervised classifier algorithms i.e., supervised classification leaning algorithm. K-nn address the pattern recognition problems and the best choices for addressing some of the classification related tasks. The simple version of the K-nearest neighbor classifier algorithms is to predict the target label by finding the nearest neighbor class. The closest class will be identified using the distance measures like Euclidean distance. It is a learning method based on instances that does not require a learning phase. The working principle behind KNN is it presumes that alike data points lie in same surroundings. It reduces the burden of building a model, adapting a number of parameters, or building furthermore assumptions. It catches the idea of proximity based on mathematical formula called as Euclidean distance, calculation of distance between two points in a plane. Suppose the two points in a plane are A (x0, y0) and B (x1, y1) then the Euclidean distance between them is calculated as follows.

$$\sqrt{(x0 - x1)^2 + (y0 - y1)^2}$$

An object to be classified is allotted to the respective class, which represents the greater number of its nearest neighbors. If *k* takes the value as 1, then the data point is classified into the category that contains only one nearest neighbor.

2) *NAÏVE BAYES's ALGORITHM*: Naive Bayes algorithm is a Bayes theorem-based classification algorithm that can be used for both exploratory and predictive modeling. The naive Bayesian classifier is part of a family of very simple probabilistic classifiers based on Bayes' theorem. The Naive Bayesian Classifier is one of the simple and efficient classification algorithms that helps you create fast machine learning models that allow you to make fast predictions. When used to classify text, a naive Bayesian classifier often achieves a higher success rate than other algorithms due to its ability to perform well on multiclass problems while remaining independent. Bayes' Theorem is a simple mathematical formula used to calculate conditional probabilities. Conditional probability is a measure of the likelihood of an event occurring given that another event has occurred (by hypothesis, presumption, statement, or evidence). Using Bayes' theorem, we can find the probability that A will happen, given that B has happened. This is the same as predicting Y when only the X variables in the test data are known. If we assume that X follows a certain distribution, you can enter the probability density function of that distribution to compute the likelihood probability. Bayes' Rule provides a formula for calculating the probability of an exit (Y) given an entry (X). In real-life problems, unlike the hypothetical hypothesis of one input function, we have several variables X.

3) *RANDOM FOREST CLASSIFIER ALGORITHM:*Random forest is a supervised classification machine learning algorithm. It builds decision trees on different samples and take their majority vote for classification and average in case of regression. A random forest algorithm randomly selects observations and characteristics to construct various decision trees, and then averages the results. However, when multiple decision trees form a coherent whole in the random forest algorithm, they predict more accurate results, especially when the individual trees are not related to each other. So, in our random forest, we get trees that not only train on different datasets (thanks to packaging), but also use different functions to make decisions. The most important feature of random forest is that it can handle the data set containing continuous variables as in case of regression and categorical variable in case of classification.They also produce forecasts with high accuracy, stability and ease of interpretation. As such, every data scientist should learn these algorithms and use them in their machine learning projects. Tree Algorithms are popular machine learning methods used to solve supervised learning problems.

4) *LOGISTIC REGRESSION:*Logistic regression is another powerful supervised ML algorithm used for binary classification problems (when target is categorical).Logistic regression uses the same basic formula as linear regression, but regresses for the likelihood of a categorical outcome. It can have any of an infinite number of possible values. In logistic regression, the result (dependent variable) has only a limited number of possible values. Dependent variable Linear regression is used when the response variable is continuous. Logistic regression is an important machine learning algorithm because it can determine probabilities and classify new data using continuous and discrete datasets. Logistic regression can be used to rank observations using a variety of data types and can easily determine the most efficient variables to use for ranking. Logistic regression is a supervised learning classification algorithm used to predict the probability of a target variable. This is where logistic regression comes in. Logistic regression is a popular statistical model used for binary classification, that is, for predictions like this or that, yes or no, A or B, etc.

IV. HYPERPARAMETER TUNING

The purpose of this research work is to explore various hyperparameter optimization [1] strategies for machine learning models. The model can have many hyperparameters, and finding the best combination of parameters can be seen as a search problem. Unlike model parameters, hyperparameters are set by machine learning engineers before training. The number of trees in the random forest is a hyperparameter, and the weight in the neural network is the model parameter obtained during training. We like to think of hyperparameters as model settings that need to be adjusted so that the model can best solve machine learning problems. Each machine learning model can set different hyperparameters. We want to be absolutely clear, hyperparameters are not model parameters and cannot be trained directly from the data. Given a set of input functions (hyperparameters), tuning hyperparameters optimizes the model for the selected metric. Without automated technologies such as AI Platform Training hyperparameter optimization, you need to manually change hyperparameters over many trainings runs to achieve optimal values. Each hyperparameter you choose to optimize can increase the number of tests required to successfully optimize[5]. Once you know the set of hyperparameters that work best, you can define a new model, set values for each hyperparameter, and then adapt the model to all available data. We can then select the optimal hyperparameter values based on this posterior expectation as our next candidate model. This process is repeated with a different set of values for the same hyperparameters until optimal accuracy is obtained or the model reaches optimal error. However, if you use test data for this evaluation, you will end up "fitting" the model architecture to the test data, losing the ability to actually measure model performance on invisible data. The best approach is to objectively search for different values of the model's hyperparameters and select the subset that leads to the model that performs best on the given dataset. For specific learning algorithms, one can compute the gradient over hyperparameters and then optimize the hyperparameters using gradient descent.

V. Experimentation

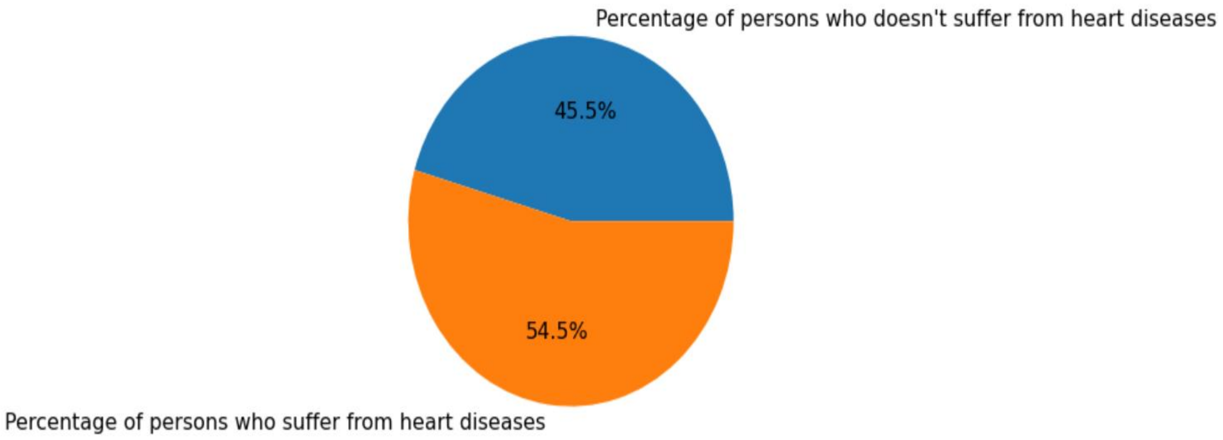
Predictive models built using machine learning (ML) algorithms can assist clinicians in early CAD detection and improve outcomes. There has been a lot of research done and various machine learning models for classification and prediction have been used to diagnose heart disease. This study aims to predict the likelihood of heart disease as the likely cause of computer-assisted heart disease prediction, which is useful in the medical field for physicians and patients. In this paper, several machine learning are applied to compare results and analyze the UCI Machine Learning Heart Disease dataset.

We compared four supervised machine learning algorithms [3] using the UC Irvine Cleveland dataset to predict disease outcomes. the Cleveland database the "goal" field refers to the presence of heart disease in the patient. It is integer valued from 0 (no presence) to 4. Experiments with the Cleveland database have concentrated on simply attempting to distinguish presence (values 1,2,3,4) from absence (value 0). The evaluation of performance of learning methods required the database to be divided into two parts: the training dataset that represents the initial base for which the class of different clinical cases are known, and the testing set for predictive analysis. We have used Cross validation technique, which is used to divide the database randomly between training data set and testing dataset. We have compared the accuracy scores of all the respective models. Since the result we expected to more optimized results and efficiency we used the hyperparameter tuning technique to optimize and tune our models.

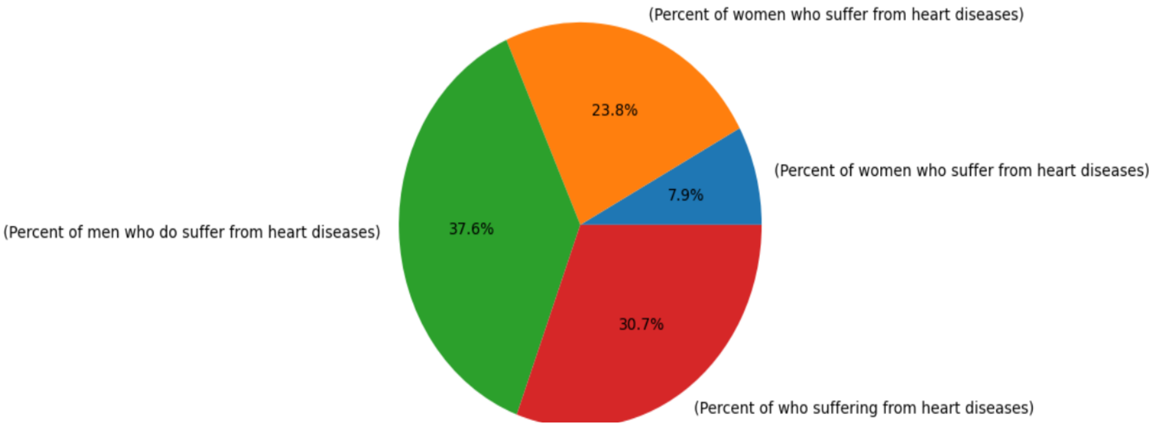
IV. Results and Discussions:

Through our dataset analysis we predicted:

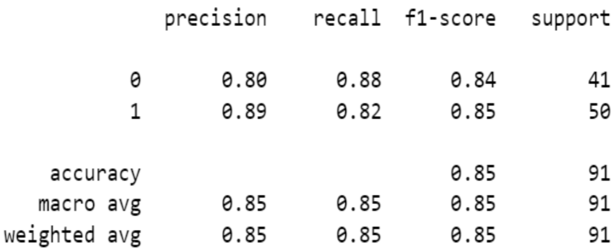
The percentage of people who are either suffering from heart disease or not.



Based on gender attribute

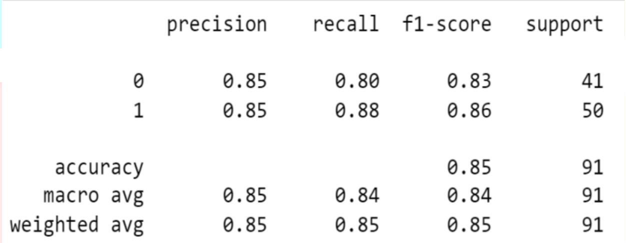


Classification Report, Accuracy Score of K-nn algorithm is 84.61%:



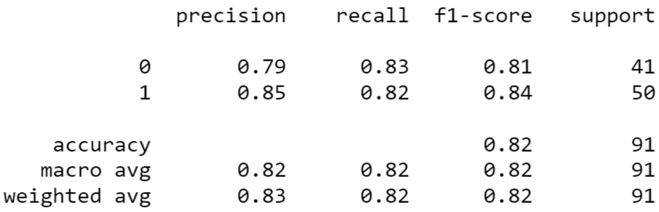
```
              precision    recall  f1-score   support

           0       0.80      0.88      0.84        41
           1       0.89      0.82      0.85        50

    accuracy                           0.85        91
   macro avg       0.85      0.85      0.85        91
weighted avg       0.85      0.85      0.85        91
```

Classification Report, Accuracy Score of Naïve Bayes' algorithm is 84.61%:



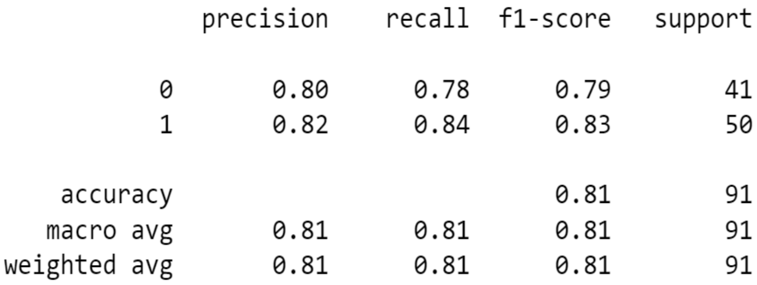
```
              precision    recall  f1-score   support

           0       0.85      0.80      0.83        41
           1       0.85      0.88      0.86        50

    accuracy                           0.85        91
   macro avg       0.85      0.84      0.84        91
weighted avg       0.85      0.85      0.85        91
```

Classification Report, Accuracy Score of Random Forest algorithm is 82.41%:



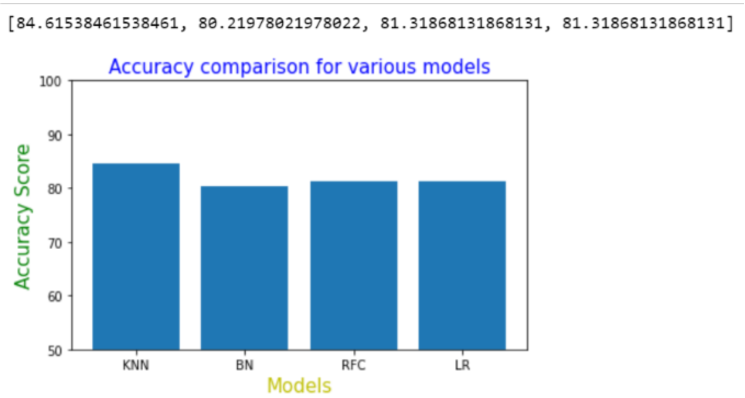
```
              precision    recall  f1-score   support

           0       0.79      0.83      0.81        41
           1       0.85      0.82      0.84        50

    accuracy                           0.82        91
   macro avg       0.82      0.82      0.82        91
weighted avg       0.83      0.82      0.82        91
```

Classification Report, Accuracy Score of Logistic Regression algorithm is 81.31%:



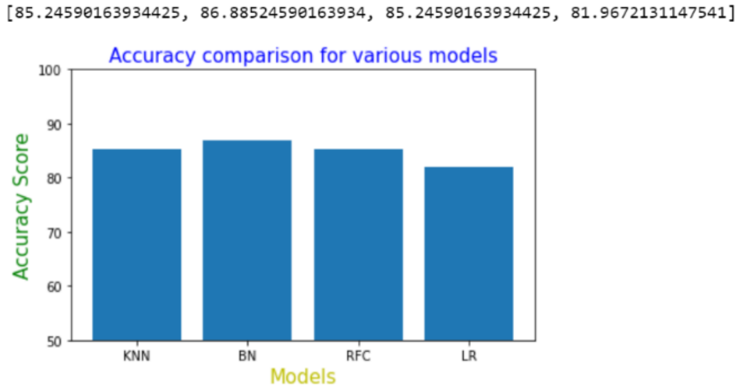
```
              precision    recall  f1-score   support

           0       0.80      0.78      0.79        41
           1       0.82      0.84      0.83        50

    accuracy                           0.81        91
   macro avg       0.81      0.81      0.81        91
weighted avg       0.81      0.81      0.81        91
```

Accuracy comparison of all the used models.



```
[84.61538461538461, 80.21978021978022, 81.31868131868131, 81.31868131868131]
```

To optimize the accuracy score of the model used we use hyperparameter tuning which helps in optimizing the performance and speed the model.



```
[85.24590163934425, 86.88524590163934, 85.24590163934425, 81.9672131147541]
```

After applying hyperparameter tuning the accuracy score changes. The performance of naïve bayes enhances the most from 80.21% to 86.88%. and random forest from 81.31% to 85.24%.

V. Conclusion

The following points lead to the conclusion that there is a huge scope for intelligent models in predicting cardiovascular disease or related heart diseases. Each above used machine learning algorithm have performed extremely well in some cases but not so good also in some cases. Efficient and accurate prediction with a lesser number of attributes and tests is our goal. In this study, we considered only 14 essential attributes. We applied four supervised classification learning techniques, K-nearest neighbor, Naive Bayes, Random Forest, and Logistic Regression and then applied Hyperparameter Tuning on the models to optimize for better performance and speed. Systems based on machine learning algorithms and techniques have been very accurate in predicting heart related diseases but still there is a lot scope of research to be done on how to handle high dimensional data and overfitting. A lot of research can also be done on the correct ensemble of algorithms to use for a particular type of data and optimize their accuracy score performance and speed.

References

[1] J. Bergstra, Y. Bengio Random search for hyper-parameter optimization Journal of Machine Learning Research, 13 (1) (2012), pp. 281-305

[2] Agoston E Eiben and Selmar K Smit. Parameter tuning for configuring and analyzing evolutionary algorithms. Swarm and Evolutionary Computation, 1(1):19–31, 2011.

[3] Frank Hutter, Holger H Hoos, and Kevin Leyton-Brown. Identifying key algorithm parameters and instance features using forward selection. In International Conference on Learning and Intelligent Optimization, pages 364–381. Springer, 2013.