**Paper Title:** Heart disease prediction using supervised machine learning algorithms\_ Performance analysis and comparison.

**Publisher:** ELSEVIER

**About:**

* Cardiovascular diseases (CVD) are currently the number one cause of death worldwide and the World Health Organization 2020 estimated this to be around 17.9 million deaths every year.
* Early-stage detection of CVD is an important way of reducing this toll.
* Of the many techniques of improving that this disease detection and diagnosis is, data mining.
* Machine learning and data mining-based approaches to prediction and detection of heart disease would be of great clinical utility, but are highly challenging to develop.
* In most countries there is a lack of cardiovascular expertise and a significant rate of incorrectly diagnosed cases which could be addressed by developing accurate and efficient early-stage heart disease prediction by analytical support of clinical decision-making with digital patient records.
* This study aimed to identify machine learning classifiers with the highest accuracy for such diagnostic purposes.
* Several supervised machine-learning algorithms were applied and compared for performance and accuracy in heart disease prediction.
* Feature importance scores for each feature were estimated for all applied algorithms except MLP and KNN.
* All the features were ranked based on the importance score to find those giving high heart disease predictions.

**Proposed model classification:**

* Using a heart disease dataset collected from Kaggle three-classification based on k-nearest neighbor (KNN), decision tree (DT) and random forests (RF) algorithms the RF method achieved 100% accuracy along with 100% sensitivity and specificity.
* Thus, we found that a relatively simple supervised machine learning algorithm can be used to make heart disease predictions with very high accuracy and excellent potential utility.
* A quick overview of these proposed supervised machine learning formulas for disease detection is given in the corresponding subsection.
  + K-nearest neighbor (KNN): KNN is one of the oldest and easiest classification algorithms or statistical learning techniques. K refers to the number of the nearest neighbours used, which can be defined directly in the object builder or simply calculated using the upper limit provided by the stated value.
    - When an unidentified sample is received, the nearest neighbor algorithm will scan the pattern space for the k training samples next to the unfamiliar sample.
    - From the test instance based on their distance, predictions from several neighbours can be calculated and two distinct methodologies are introduced to transform the distance into a weight.

Advantages:

* It is analytically tractable and very easy to implement.
* The classifier is very effective and performs well in disease prediction especially in HD prediction since it works with a single instance.

In this study, the value of n\_neighbors 2 and leaf\_size 40 were the best fit parameter for the dataset.

* + Random forest: RF is a method for classifying data by ensemble learning based on DT.

It creates a large number of trees and also produces a forest of decision trees, while it is under the training stage. Every tree, a member of the forest, forecast class label for every single instance at the testing period.

* Every tree, a member of the forest, forecast class label for every single instance at the testing period.
* The class label that obtains the largest number of votes is considered as the most appropriate label applied to the test data.
* For every data in the data collection, this cycle is replicated.

The best fit random state value for this study was 123, which gave the best performance for the applied dataset.

* + Decision tree (DT): DT is one of the oldest and most common machine learning algorithms. A DT designs the logic of the decision in such a way that evaluates and matches results for the classification of data items into a structure as like a tree.
    - Evaluation of input variables or features is represented by all internal nodes that contain at least one child node.
    - Depending on the evaluation outcome, the classification techniques branch to the correct child node, where the evaluation and branching process continues before the leaf node is reached.
    - The leaf or terminal nodes refer to the outcomes of the decision.

The maximum depth for this classification algorithm was defined 7 and the classifier by this maximum depth value produced the best result for the applied dataset in this study.

* + AdaboostM1 (ABM1): ABM1 is one kind of ensemble learning based supervised machine learning classifier, which is widely used. It employs an adaptive enhancement approach and produces improved classification results by integrating multiple weak classifiers into a strong classifier.
    - In the initial stage, the same weight is allocated to all the observations.
    - The weights of the observations change with the coefficient of weak classifiers, and the coefficient of the applied classifiers is estimated using the value of the estimation error.
    - . Consequently, the weight of misclassified observations can be raised by the ABM1 algorithm and the weight of correctly identified observations can be reduced.
    - In the subsequent iterations, it will enforce higher weight on the incorrectly classified observations more.
    - Finally, all the weak classifiers developed are combined to form a stronger classifier using a linear combination method to produce accurate classification performance.

The value of n\_estimators was defined as 200, with this classifier providing the best performance in this study.

* + Logistic regression (LR): LR is a strong classifier among supervised machine learning algorithms and is an extension of the general regression modeling as applied to a dataset, represents the probability of occurrence or nonoccurrence of a particular instance.
    - LR identifies the chances of a new observation belonging to a certain class, the result lying between 0 and 1 since it is a probability.
    - Consequently, a threshold is assigned that defines the separation into two classes to implement the LR as binary classification.
    - For instance, a probability value calculated higher than 0.5 is designated as ‘class A’ otherwise ‘class B’

The best fit random state value 1234, and the best fit maximum iteration number 100 were found in this study for the applied dataset.

* + Multilayer perceptron (MLP): MLP is a well-established neural network-based classification algorithm, which consists of three or more types of layers: an input layer, output layer and one or more hidden layers between input and output layers.
    - Every layer contains a number of ‘neurons’ connecting all the layers with each other.
    - The construction of MLP classifiers consists of adequate input variables and specification of the type of network, relevant data pre-processing and partitioning, the configuration of network infrastructure, specification of success parameters, specification of training algorithm (optimization of relation weights), and finally evaluation of model.

The default configuration produced the best result by this classifier in this study.

**Conclusion:** Heart disease is life-threatening, which leads to potentially fatal complications such as heart attacks. Due to its potential for accurate disease prediction rate, the importance of data mining and machine learning techniques could be used to predict its occurrence. Here, we used a heart disease dataset to test the utility of ML approaches to heart disease prediction, finding that three classification algorithms KNN, RF and DT performed extremely well with 100% accuracy. In addition, feature importance scores for each feature was estimated for all the applied algorithms except MLP and KNN. These features were ranked based on feature importance score. The study aimed to find the best ML techniques, among a number of algorithms that are well accepted and easy to implement, finding that, at least for this dataset, they performed well. This is an early stage of using ML approaches but suggests that it could prove to be an excellent adjunct to patient care.