Technical Seminar on

HEART DISEASE PREDICTION USING MACHINE LEARNING TECHNIQUES

BY
MADHUKESH BHAT M (1BM18CS051)

ABSTRACT

Heart disease, alternatively known as cardiovascular disease, encases various conditions that impact the heart and is the primary basis of death worldwide over the span of the past few decades.

This research presents various attributes related to heart disease, and the model on basis of supervised learning algorithms as Naïve Bayes, decision tree, K-nearest neighbor, and random forest algorithm. It uses the existing dataset from the Cleveland database of UCI repository of heart disease patients.

The mode implemented aims to envision the probability of developing heart disease in the patients. It can also be helpful to the medical practitioners at their clinic as decision support system. The results portray that the highest accuracy score is achieved with Naïve Bayes.

INTRODUCTION

OVERVIEW

Machine Learning is a branch of AI research and has become a very popular aspect of data science. The Machine Learning algorithms are designed to perform a large number of tasks such as prediction, classification, decision making etc.

Various data mining techniques such as regression, clustering, association rule and classification techniques like Naïve Bayes, Decision Tree, Random Forest, Support Vector Machine and K-nearest neighbor are used to classify various heart disease attributes in predicting heart disease. In this research, a dataset from the UCI repository have been taken. The classification model is developed using classification algorithms for prediction of heart disease. A discussion of algorithms used for heart disease prediction and comparison among the existing systems is made here. Further research and advancement possibilities is also mentioned in here.

MOTIVATION

Over the last decade, heart disease or cardiovascular remains the primary basis of death worldwide. An estimate by the World Health Organization, that over 17.9 million deaths occur every year worldwide because of cardiovascular disease, and of these deaths, 80% are because of coronary artery disease and cerebral stroke. The vast number of deaths is common amongst low and middle-income countries.

The efficient and accurate and early medical diagnosis of heart disease plays a crucial role in taking preventive measures to prevent death. This leads to the proposed system which determines the best Machine Learning algorithm to predict the occurrence of heart disease in patients beforehand.

OBJECTIVE

Heart disease affects millions of people, and it remains the chief cause of death in the world. Medical diagnosis should be proficient, reliable, and aided with computer techniques to reduce the effective cost for diagnostic tests.
The objective of this research is to use different classification techniques to predict heart disease and determine the best of them.

LITERATURE SURVEY

1. Heart Disease Prediction using Machine Learning Techniques:
   The work of Devansh Shah, et. al. presents various attributes related to heart disease, and the model on basis of supervised learning algorithms as Naïve Bayes and K-nearest neighbor. Of these algorithms, K-Nearest Neighbor classifier turned out to be the best one with an accuracy of 78.94% on testing set with K = 2 neighbors.

2. Prediction of Heart Disease Using a Combination of Machine Learning and Deep Learning:
   In this paper, Rohit Bharthi et al. proposed three methods in which comparative analysis was done and promising results were achieved. The conclusion which was found is that machine learning algorithms performed better in this analysis. For the 13 features which were in the dataset, K-Neighbors classifier performed better in the ML approach when data preprocessing is applied.

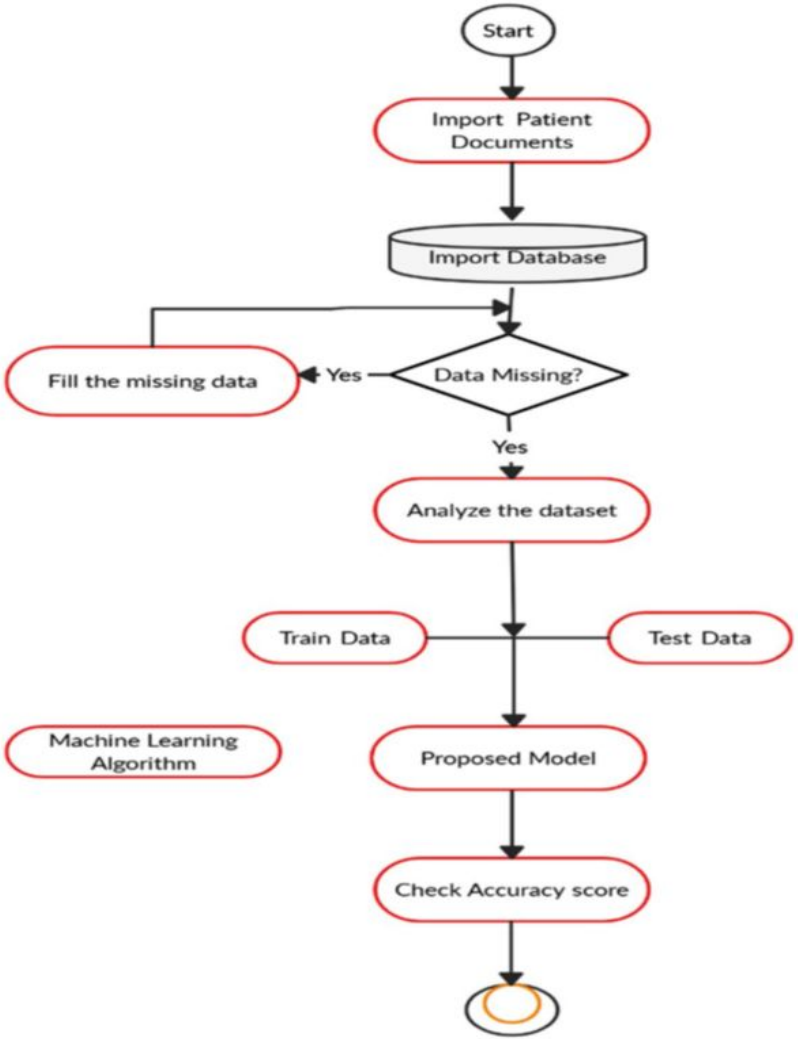
3. Heart Disease Prediction using Machine Learning:
   In the research of Vijeta Sharma et. al., Machine Learning algorithms such as Random Forest, Support Vector Machine (SVM), Naive Bayes and Decision tree have been used for the development of model. In the research they have also tried to find the correlations between the different attributes available in the dataset with the help of standard Machine Learning methods and then using them efficiently in the prediction of chances of Heart disease. Result showed that compared to other ML techniques, Random Forest gives more accuracy in less time for the prediction.

4. Developing a Hyperparameter Tuning Based Machine Learning Approach of Heart Disease Prediction
   In this paper, the traditional and proposed system was implemented by Emrana Kabir Hashi et. al., to predict Cleveland heart disease dataset. Without hyperparameters tuning, the LR, KNN, SVM, DT, and RF classifiers provide an accuracy rate of 88.52%, 90.16%, 88.52%, 81.97%, and 85.25% respectively. However, with the hyperparameters tuning approach, the LR, KNN, SVM, DT, and RF classifiers in the refined set takes the accuracy rate 90.16%, 91.80%, 90.16%, 86.89%, and 85.25% respectively. Hence, it is concluded that the proposed model is more efficient.

METHODOLOGY/TECHNIQUES USED

DATA PREPROCESSING

The real-life information contains large numbers with missing and noisy data. These data are pre-processed to overcome such issues and make predictions vigorously.



ALGORITHMS USED

Naive Bayes Classifier

- Naïve Bayes' Classifier is a supervised algorithm.
- It is a simple classification technique using Bayes theorem.
- It assumes strong (Naive) independence among attributes.
- The predictors are neither related to each other nor have correlation to one another.
- All the attributes independently contribute to the probability to maximize it.
- It is able to work with Naïve Bayes model and does not use Bayesian methods.
- Many complex real-world situations use Naive Bayes classifiers.

Decision Tree Classifier

- Decision tree is used for creating tree-like structures.
- Decision tree is simple and widely used to handle medical dataset. It is easy to implement and analyse the data in tree-shaped graph.
- The decision tree model makes analysis based on three nodes:
  - Root node: main node, based on this all other nodes functions.
  - Interior node: handles various attributes.
  - Leaf node: represent the result of each test.
- This algorithm splits the data into two or more analogous sets based on the most important indicators.
- The entropy of each attribute is calculated and then the data are divided, with predictors having maximum information gain or minimum entropy.

K-Nearest Neighbors Classifier

- The K-nearest neighbors algorithm is a supervised classification algorithm method.
- It classifies objects dependant on nearest neighbor.
- It is a type of instance-based learning.
- The calculation of distance of an attribute from its neighbors is measured using Euclidean distance.
- It uses a group of named points and uses them on how to mark another point.
- The data are clustered based on similarity amongst them, and is possible to fill the missing values of data using K-NN.
- Once the missing values are filled, various prediction techniques apply to the data set.
- It is possible to gain better accuracy by utilizing various combinations of these algorithms.

Random Forest Algorithm

- Random forest algorithm is a supervised classification algorithmic technique.
- In this algorithm, several trees create a forest.
- Each individual tree in random forest lets out a class expectation and the class with most votes turns into a model’s forecast.
- In the random forest classifier, the more number of trees give higher accuracy.

Support Vector Machine

- Support Vector Machine is a classification technique of Machine learning, which is used to analyze data and discover patterns in classification and regression analysis.
- SVM is typically considered when data is characterized as two class problem.
- In this strategy, data is characterized by finding the best hyperplane that isolates all data points of one class to the other class.
- The higher separation or edge between the two classes is, the better is the model considered.

DESCRIPTION OF TOOL SELECTED

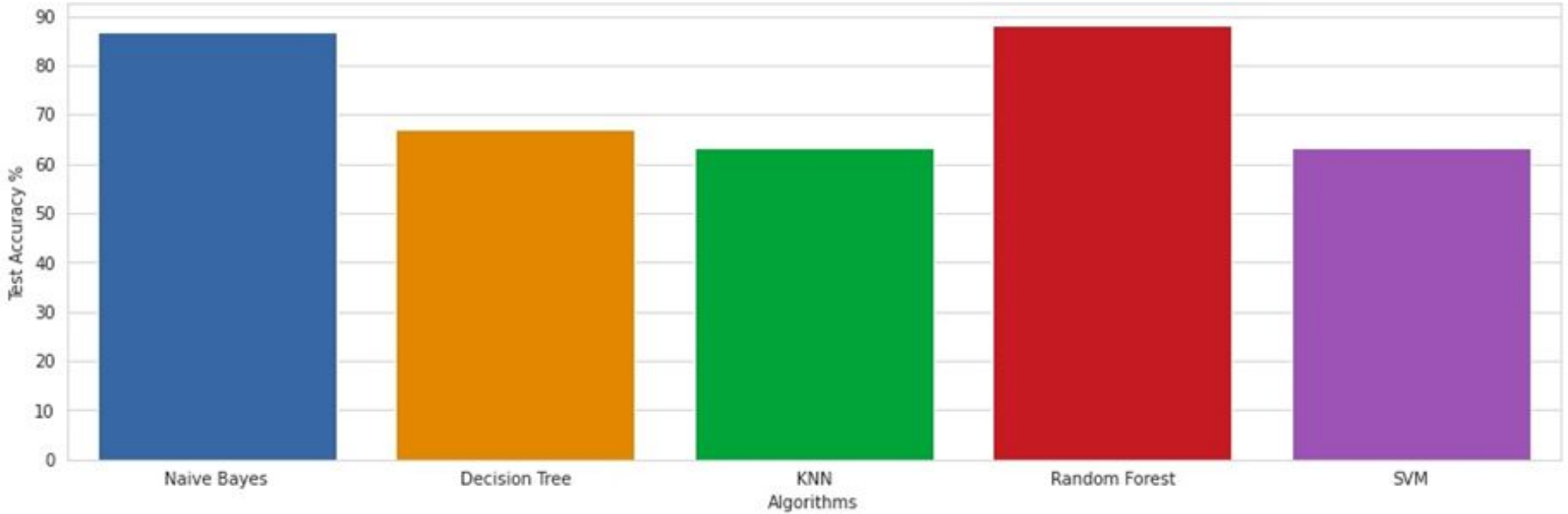
- Scikit-learn: Scikit-learn is the most useful library for machine learning in Python. The sklearn library contains a lot of efficient tools for machine learning and statistical modeling including classification, regression, clustering and dimensionality reduction.
- Numpy: NumPy is the package for scientific computing in Python. NumPy arrays eases advanced mathematical and other types of operations on large numbers of data.
- Pandas: Pandas is a Python package providing fast, flexible, and expressive data structures designed to make working with relational or labeled data both easy and intuitive. It aims to be the fundamental high-level building block for doing practical, real-world data analysis in Python.
- Matplotlib: Matplotlib is a cross-platform, data visualization and graphical plotting library for Python and its numerical extension NumPy.
- Jupyter Notebook: The Jupyter Notebook is an open-source web application that allows data scientists to create and share documents that integrate live code, equations, computational output, visualizations, and other multimedia resources, along with explanatory text in a single document.

DETAILED DESCRIPTION OF MODULES IMPLEMENTED

DATA PREPROCESSING AND MODEL BUILDING

Dataset was classified and split into a training set (75%) and a test set (25%). Pre-processing of the data is done and supervised classification techniques mentioned above are applied to get accuracy score. The accuracy score results of different classification techniques were noted using for training and test data sets. Percentage accuracy scores are depicted in the below figure for different algorithms.

	model	train_accuracy	test_accuracy
0	Naive Bayes	81.06	86.84
1	Decision Tree	100.00	67.11
2	KNN	77.97	63.16
3	Random Forest	100.00	88.16
4	SVM	66.08	63.16



Bar Chart showing the comparison of all the algorithms implemented

K FOLD VALIDATION

Cross-validation is a resampling procedure used to evaluate machine learning models on a limited data sample.

The procedure has a single parameter called k that refers to the number of groups that a given data sample is to be split into. As such, the procedure is often called k-fold cross-validation. When a specific value for k is chosen, it may be used in place of k in the reference to the model, such as k=10 becoming 10-fold cross-validation.

Cross-validation is primarily used in applied machine learning to estimate the skill of a machine learning model on unseen data. That is, to use a limited sample in order to estimate how the model is expected to perform in general when used to make predictions on data not used during the training of the model.

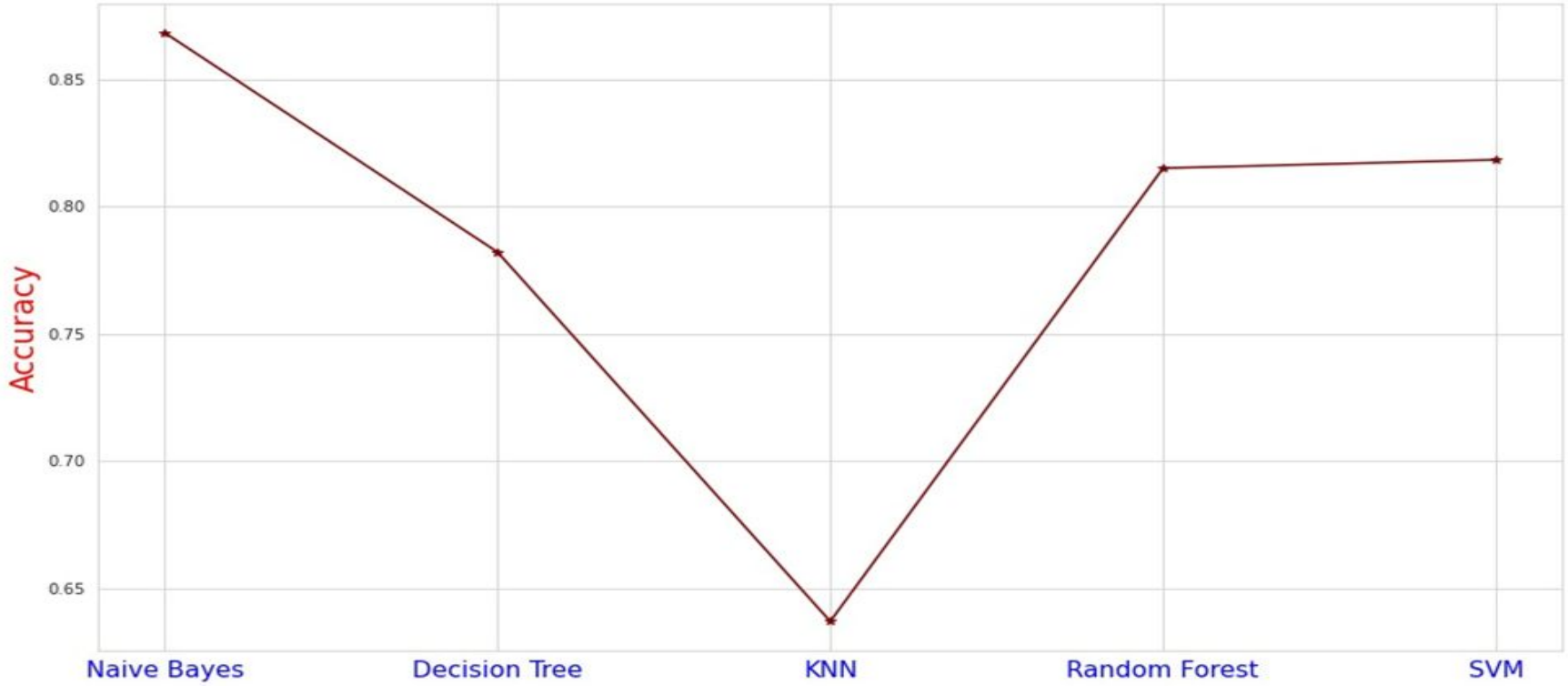
HYPERPARAMETER TUNING

Hyperparameter tuning works by running multiple trials in a single training job. Each trial is a complete execution of your training application with values for our chosen hyperparameters, set within limits you specify. We can keep track of the results of each trial and makes adjustments for subsequent trials. When the job is finished, we can get a summary of all the trials along with the most effective configuration of values according to the criteria we specified.

FINALISED MODELS' ACCURACY

Percentage accuracy scores after performing these two techniques are depicted in the below figure for different algorithms.

	model	best_score	best_parameter
0	Naive Bayes	0.868421	None
1	Decision Tree	0.782178	{'criterion': 'entropy', 'max_features': 'sqrt'}
2	KNN	0.636964	{'algorithm': 'auto', 'n_neighbors': 5}
3	Random Forest	0.815182	{'n_estimators': 50, 'random_state': 100}
4	SVM	0.818482	{'C': 50, 'kernel': 'linear'}



Comparison by plot for various algorithms implemented

NEW LEARNINGS FROM THE TOPIC

During this research, I learnt about

- ➔ The importance of machine learning domain in the medical field.
- ➔ The benefits of predicting heart disease before hand.
- ➔ Few data preprocessing methods used on the dataset.
- ➔ Five machine learning models implemented to predict the heart disease.
- ➔ Cross validation and Hyperparameter tuning techniques used to improve the accuracy of algorithms implemented.

REFERENCES

1. Devansh Shah, Samir Patel, and Santosh Kumar Bharti, “Heart Disease Prediction using Machine Learning Techniques”, Springer Nature Singapore Pte Ltd 2020, SN Computer Science (2020) 1:345
2. Rohit Bharthi, Adithya Kamparia et. al., “Prediction of Heart Disease Using a Combination of Machine Learning and Deep Learning”, Hindawi, Computational Intelligence and Neuroscience, Volume 2021
3. Vijeta Sharma, Shrinkala Yadav, and Manjari Guptha, “Heart Disease Prediction using Machine Learning Techniques”, 2020 2nd International Conference on Advances in Computing, Communication Control and Networking (ICACCCN)
4. Emrana Kabir Hashi and Md. Shahid Uz Zaman, “Developing a Hyperparameter Tuning Based Machine Learning Approach of Heart Disease Prediction”, Journal of Applied Science & Process Engineering, Vol. 7, No. 2, 2020

THANK YOU!