# COMPARATIVE ANALYSIS OF MACHINE LEARNING ALGORITHMS IN HEART DISEASE PREDICTION

[1] I.SAPTHAMI, [2] YAARRADODLA NISHWANTH THARAK REDDY, [3]POLUBOINA KALPANA, [4]MALLAVARAPU SWAROOPA RANI, [5]KALAJULAPATI BALA SAI SUKESH

[1]Assistant professor, Dept. of CSE, Visvodaya Engineering College, Kavali, AP, India.

[2,3,4,5] Student, Dept. of CSE, Visvodaya Engineering College, Kavali, AP, India.

***Abstract*** – As per the recent study by WHO, heart related diseases are increasing. 17.9 million people die every year. A better approach to health care is to prevent a disease with early intervention rather than taking a treatment after it is diagnosed. Data in the health care sector is growing beyond dealing capacity of the health care organizations and is expected to increase significantly in the coming years. This data is effectively used for analysis and prediction. The main aim of this project is to make a comparative study of machine learning algorithms such as Support Vector Machine, Extra Tree Classification and K – Nearest Neighbor Classification in predicting heart disease. In this project we are using real time heart disease dataset collected from kaggle website and the data set is preprocessed as per the requirements to perform comparative study of the specified machine learning algorithms. Confusion matrix, F1-score and accuracy have been considered as metrics for evaluating the performance of algorithms in the context of predicting heart disease.

*Index terms* – heart disease prediction, SVM, ETC, KNN.

## I. INTRODUCTION

In today’s chaotic world we all have very busy life, tough schedule and competitive activities for growing up and to achieve success in our life but we are neglecting our health issues because of this, we encounter many diseases that are threat to our lives. People don’t pay attention to symptoms and this negligence cause death. Many diseases if they will not be treated properly they cause death. Heart disease is one of the chronic illnesses that produce different signals from early stage but we fail to recognize these signals which lead to long term illness or loss of life.

Heart disease mostly occurs in man then female according to the report of WHO world health organization’s statistics 24% death in world that are not communicable happens because of heart illness.

We have proposed to used Random Forest algorithm with correlation based selected features on a given data set. An improve results with highest accuracy compare with

Hoeffding tree method. We compare results with the following models Na¨ıve Byes, Logistic regression, Gradient Boost,Support Vector Classification and also compare the result with a paper named as A Classification for Patients with Heart Disease

## II. BACKGROUND WORK

### *A. Machine Learning based Medical Information Analysis, Estimations and Approximations over Present Health Research Domain*

The concept of Machine Learning is widely started using in medical domain based on the prediction and analysis got succeed and it is continuing using in Medical domain these days due to many advantages things such as flexibility, can take quick decisions, portability, reliability, user friendly etc. so this helps us to improve the health care quality by machine learning. Machine Learning is modern and highly sophisticated technological applications became a huge trend in the industry. Machine Learning is Omni present and is widely used in various applications. It is playing a vital role in many fields like Medical science, finance, forecasting and in security.

### *B. Machine Learning and Big Data Implementation on Health Care data*

Healthcare is the most prominent field suitable for the applications of machine learning and big data on health care data. So the implementations of health care with big data and machine learning is increased with the client health requirements. The electronic health record applications are being increased in this current situation, which is needed to be focused on utilizing the data generated by those applications. There is a large volume of data in health care that is related to different health care domains especially neuro and cardiac. These data need a special focus and the architectures currently focusing on these domains has to implement the latest technologies to predict some patterns. The implementation of different health care architecture is focussed, which uses live data gathered from different sources over the globe. In this article, machine learning approaches and the big data framework are combined to design a prediction model and data handling techniques. Machine learning is used to discover patterns from medical data sources and provide excellent capabilities to predict diseases. In this paper, they have reviewed various machine learning algorithms used for developing efficient decision support for healthcare applications. This paper helps in reducing the research gap for building efficient decision support system for medical applications.

### *C. Predicting Heart Disease at Early Stages using Machine Learning: A Survey*

Predicting and detection of heart disease has always been a critical and challenging task for health care practitioners. Predicting heart disease at the early stages will be useful to the people around the world so that they will take necessary actions before getting severe. Machine Learning shows effective results in making decisions and prediction from the broad set of data produced by the healthcare industry.Some of the supervised machine learning techniques used in this prediction of heart disease are artificial neural network (ANN), decision tree (DT), random forest (RF), support vector machine (SVM), naïve Bayes) (NB).

### *D. Improving Health – Care systems by disease prediction*

Predicting the likelihood of a person developing diabetes and to get the good accuracy in prediction by comparing machine learning algorithms. In this paper they have compared different algorithms and plotted comparison graphs. Finally chosen the best algorithms for prediction of the disease in the early stage. This paper also focuses on a comparison of 5 prediction algorithms which are Artificial Neural Networks, Logistic Regression, Decision Tree and Random Forest algorithm. According to the results, neural networks produce the highest accuracy after model evaluation.

## III. PROPOSED WORK

### *A. System Architecture*

Machine Learning algorithms are totally subject to data since it is the most vital perspective that makes model training possible. On the other hand, if we won't be able to make sense out of that data, before feeding it to ML algorithms, a machine will be useless. In straightforward words, we generally need to take care of the right data for example the data in the right scale, group, and containing important features, for the problem we need a machine to solve.

This makes data preparation the most important step in the ML process. Data preparation may be defined as the procedure that makes our dataset more appropriate to work within the ML process.

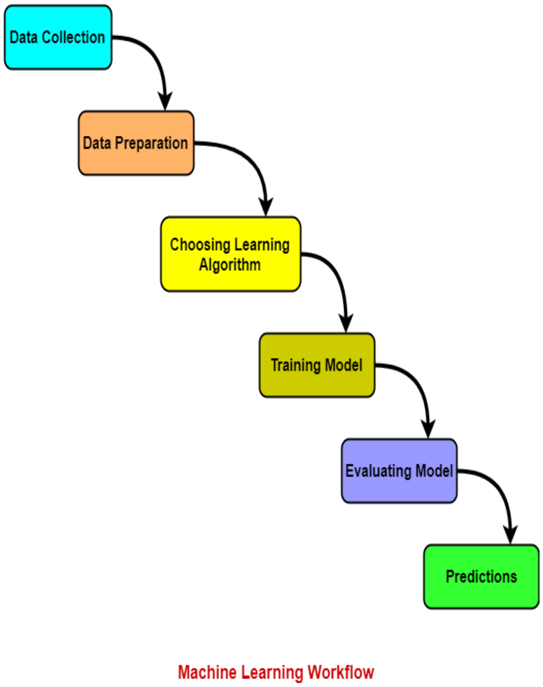


**Fig 1. Machine Learning Workflow**

A dataset can be viewed as a collection of data objects, which are often also called as records, points, vectors, patterns, events, cases, samples, observations, or entities.

Data objects are described by a number of features that capture the basic characteristics of an object, such as the time at which an event occurred, etc... Features are often called as variables, characteristics, fields, attributes, or dimensions.

A feature is an individual measurable property or characteristic of a phenomenon being observed. Features can be categorical (Nominal, Ordinal), Numerical (Interval, Ratio).

### *B. Dataset*

Machine learning heavily depends on data and dataset makes machine learning training feasible. In this project dataset has been collected from kaggle website. A dataset is used to train the model for performing various actions, to work automatically. The training dataset is a dataset in which machine learning algorithms have been trained and the dataset we use to validate the accuracy of our model is called testing dataset.

### *C. Data Preprocessing*

Data preprocessing is a process of preparing the raw data and making it suitable for a machine learning model. It is the first and crucial step while creating a machine learning model.

When creating a machine learning project, it is not always a case that we come across the clean and formatted data. And while doing any operation with data, it is mandatory to clean it and put in a formatted way. So for this, we use data preprocessing task.

**Need of Data Preprocessing**

When creating a machine learning project, it is not always a case that we come across the clean and formatted data. And while doing any operation with data, it is mandatory to clean it and put in a formatted way. So for this, we use data preprocessing task.

Steps Involved in Data Preprocessing:

1. **Data quality assessment**
2. **Feature Sampling**
3. **Feature Scaling**

### *D. Train / Validation / Test Split*

After feature scaling is done, our dataset is ready for the exciting machine learning algorithms. But before we start deciding the algorithm which should be used, it is always advised to split the dataset into 2 or sometimes 3 parts. Machine Learning algorithms, or any algorithm has to be first trained on the data distribution available and then validated and tested, before it can be deployed to deal with real-world data.

**Training data**: This is the part on which your machine learning algorithms are actually trained to build a model. The model tries to learnthe dataset and its various characteristics.

**Validation data**: This is the part of the dataset which is used to validate our various model fits. In simpler words, we use validation data to choose and improve our model parameters. The model does not learn the validation set but uses it to get to a better state of parameters.

**Test data**: This part of the dataset is used to test our model hypothesis. It is left untouched and unseen until the model and parameters are decided, and only after that the model is applied on the test data to get an accurate measure of how it would perform when deployed on real-worlddata.

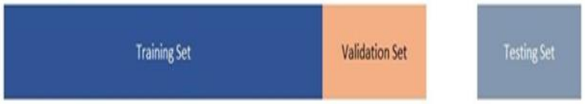


**Fig. 2. Data Splitting**

### *E. Algorithms*

We are using supervised machine learning algorithms for disease prediction and analyzation. The following are the algorithms used in this project. They are:

1. K – Nearest Neighbor Classification
2. Support Vector Machine
3. Extra Tree Classification

### *i) K – NEAREST NEIGHBOR CLASSIFICATION (KNN)*

- K-Nearest Neighbor is one of the simplest Machine Learning algorithms based on Supervised Learning technique.
- K-NN algorithm stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suite category by using K- NN algorithm.
- K-NN is a non-parametric algorithm, which means it does not make any assumption on underlying data.

**Working of KNN**

The K-NN working can be explained on the basis of the below algorithm.

**Step-1:** Select the number K of the neighbors.

**Step-2:** Calculate the Euclidean distance of **K number of neighbors.**

**Step-3:** Take the K nearest neighbors as per the calculated Euclidean distance.

**Step-4:** Among these k neighbors, count the number of the data points in each category.

**Step-5:** Assign the new data points to that category for which the number of the neighbor is maximum.

**Step-6:** Our model is ready.

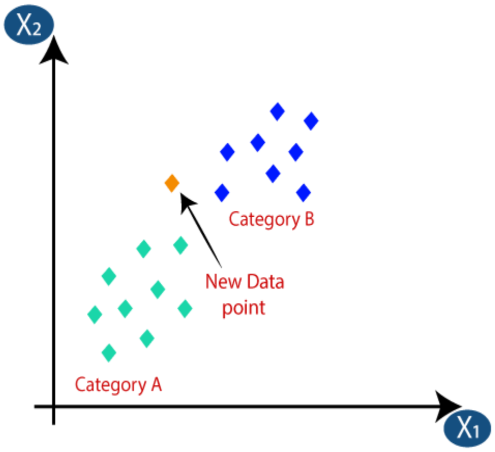


**Fig. 3: KNN-1**

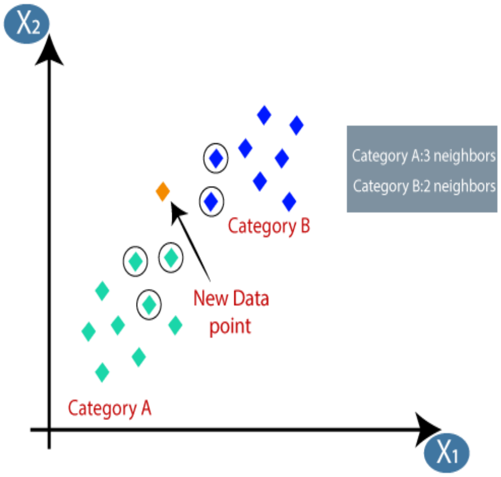


**Fig. 4: KNN-2**

**Selecting K Value**

- There are no pre-defined statistical methods to find the most favorable value of K.
- Initialize a random K value and start computing.
- Choosing a small value of K leads to unstable decision boundaries.
- The substantial K value is better for classification as it leads to smoothening the decision boundaries.
- Derive a plot between error rate and K denoting values in a defined range. Then choose the K value as having a minimum error rate.

### ii) SUPPORT VECTOR MACHINE (SVM)

- Support Vector Machine or SVM is one of the most popular Supervised Learning algorithms, which is used for Classification as well as Regression problems. However, primarily, it is used for Classification problems in Machine Learning.
- The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane.
- SVM chooses the extreme points/vectors that help in creating the hyperplane. These extreme cases are called as support vectors, and hence algorithm is termed as Support Vector Machine.

***Types of SVM***

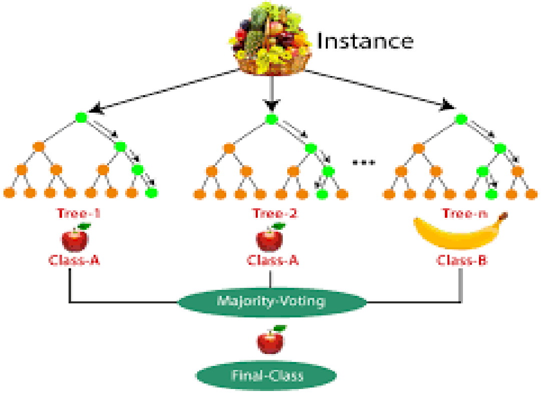
***SVM can be of two types***

- **Linear SVM:** Linear SVM is used for linearly separable data, which means if a dataset can be classified into two classes by using a single straight line, then such data is termed as linearly separable data, and classifier is used called as Linear SVM classifier.
- **Non-linear SVM:** Non-Linear SVM is used for non-linearly separated data, which means if a dataset cannot be classified by using a straight line, then such data is termed as non-linear data and classifier used is called as Non-linear SVM classifier.

### iii) EXTRA TREE CLASSIFICATION

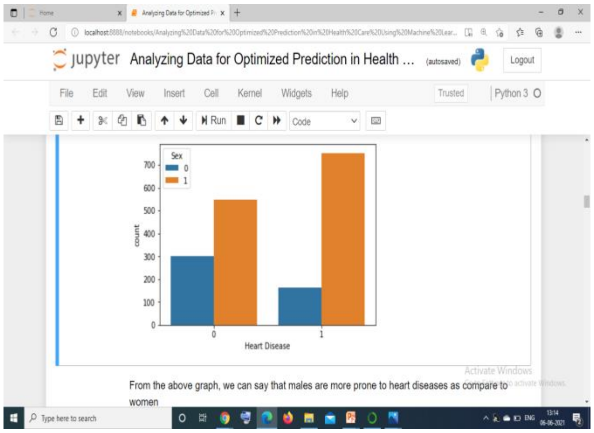
- **Extremely Randomized Trees**, or Extra Trees for short, is an ensemble machine learning algorithm.
- Specifically, it is an ensemble of decision trees and is related to other ensembles of decision trees algorithms such as bootstrap aggregation (bagging) and random forest.
- The Extra Trees algorithm works by creating a large number of unpruned decision trees from the training dataset. Predictions are made by averaging the prediction of the decision trees in the case of regression or using majority voting in the case of classification.
- Unlike bagging and random forest that develop each decision tree from a bootstrap sample of the training dataset, the Extra Trees algorithm fits each decision tree on the whole training dataset.
- Like random forest, the Extra Trees algorithm will randomly sample the features at each split point of a decision tree. Unlike random forest, which uses a greedy algorithm to select an optimal split point, the Extra Trees algorithm selects a split point at random.
- There are three main hyper parameters to tune in the algorithm; they are the number of decision trees in the ensemble, the number of input features to randomly select and consider for each split point, and the minimum number of samples required in a node to create a new split point.
- The random selection of split points makes the decision trees in the ensemble less correlated, although this increases the variance of the algorithm. This increase in variance can be

countered by increasing the number of trees used in the ensemble.



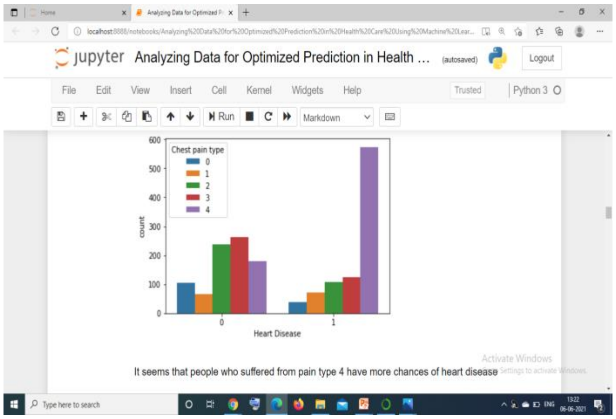
### Fig. 5: Extra Tree Classification

## IV. RESULTS


Home
Analyzing Data for Optimized P
localhost:8888/notebooks/Analyzing%20Data%20for%20Optimized%20Prediction%20in%20Health%20Care%20Using%20Machine%20Lear...
jupyter
Analyzing Data for Optimized Prediction in Health ...
(autosaved)
Logout
File
Edit
View
Insert
Cell
Kernel
Widgets
Help
Trusted
Python 3
Run
Code
Sex
0
1
count
700
600
500
400
300
200
100
0
0
1
Heart Disease
Activate Windows
Type here to search
ENG
13:14
06-06-2021

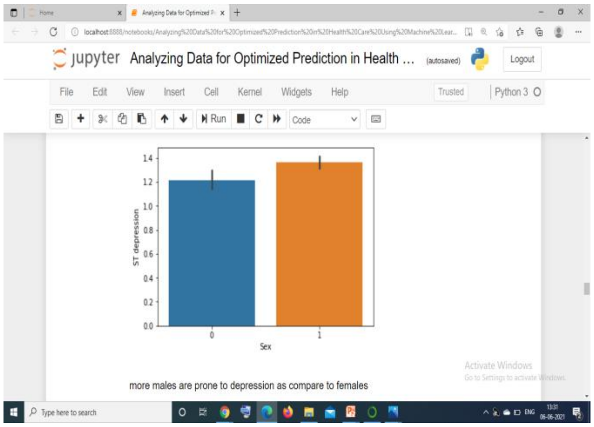
From the above graph, we can say that males are more prone to heart diseases as compare to women

### Fig 6: count of heart disease patients



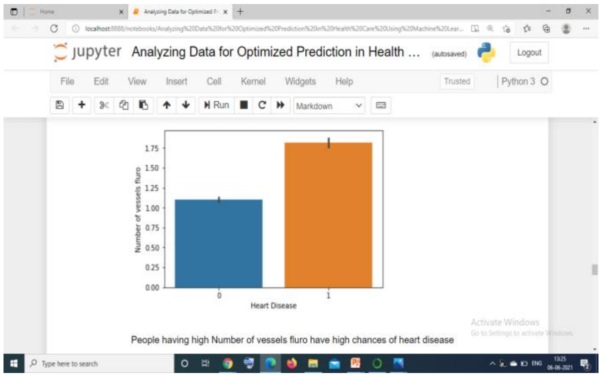
It seems that people who suffered from pain type 4 have more chances of heart disease

### Fig 7: chest pain type Vs heart disease



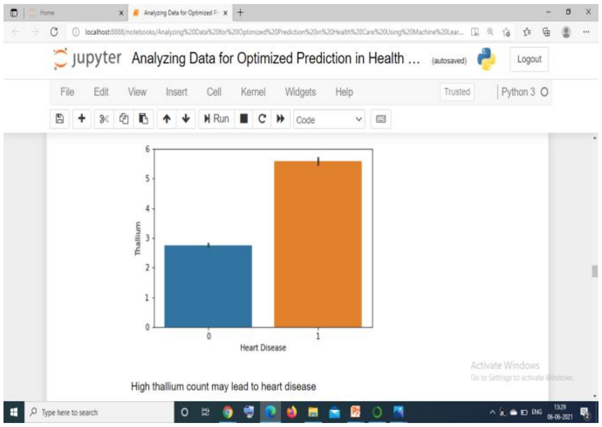
more males are prone to depression as compare to females

### Fig 8: Sex Vs Depression



People having high Number of vessels fluro have high chances of heart disease

### Fig 9: heart disease Vs number of vessels fluro



High thallium count may lead to heart disease

### Fig 10: heart disease Vs thallium

## V. CONCLUSION

Machine learning algorithms are vastly used in the prediction of diseases. The results obtained from the comparative study of supervised machine learning models like K – Nearest Neighbor

Classification, Support Vector Machine and Extra Tree Classification show that K- Nearest Neighbor is the best model for predicting heart disease.

## REFERENCES

1. Methaila, Aditya, Prince Kansal, Himanshu Arya, and Pankaj Kumar. "Early heart disease prediction using data mining techniques." Computer Science & Information Technology Journal (2014): 53-59.
2. L. Ali, A. Rahman, A. Khan, M. Zhou, A. Javeed, and J. A. Khan, “An Automated Diagnostic System for Heart Disease Prediction Based on $\chi 2$ Statistical Model and Optimally Configured Deep Neural Network,” IEEE Access, vol. 7, pp. 34938–34945, 2019.
3. Purushottam, K. Saxena, and R. Sharma, “Efficient Heart Disease Prediction System,” Procedia Comput. Sci., vol. 85, pp. 962–969, 2016.
4. M. Kavitha;G. Gnaneswar;R. Dinesh;Y. Rohith Sai;R. Sai Suraj”Heart Disease prediction using Hybrid Machine Learning Model ” .
5. Alperen Endogen, Selda Güney from Başkent Üniversitesi, Ankara, Türkiye,” Heart Disease Prediction by Using Machine Learning Algorithm
6. https://youtu.be/38SUUaMX5Rg{--youtube}
7. B. Dhomse Kanchan;M. Mahale Kishore(2016)”study of machine learning algorithms for special disease prediction using principle of component analysis”
8. Vijeta Sharma; Shrinkhala Yadav ,Manjari Gupta ”Heart disease prediction using
9. Machine learning techniques” (2020): ICACCCN….IEEE