

PROJECT REPORT ON **HEART DISEASE PREDICTION**

(Using Various Data Mining and Machine Learning Algorithm)

For the partial fulfillment of

Data Mining

Semester VI



Submitted by:

Ishika Bhardwaj (8209)

Akkati Chethan (8186)

Submitted to:

Dr.. Manju Sardana

Department of Computer Science

Hansraj College

Department of Computer Science

Hansraj College

University of Delhi

April - 2022

CERTIFICATE

It is hereby certified that the Project Report of Data Mining entitled “**Heart Disease Prediction**” is submitted by Ishika Bhardwaj and Akkati Chethan, the students of B.Sc. (H) Computer Science, Hansraj College (University of Delhi). It has been found satisfactory and hereby approved for submission.

Dr. Manju Sardana

(Assistant Professor Department of CS, HRC)

Date:

DECLARATION

We, the students of B.Sc (H) Computer Science, Hansraj College, University of Delhi declare that the work entitled "**Heart Disease Prediction**" has been successfully completed under the guidance of our Professor Dr. Manju Sardana, Computer Science Department, Hansraj College, University of Delhi. This dissertation work is submitted in partial fulfillment of the requirements for the of Data Mining in Computer Science during the academic year 2021-2022. Further the matter embodied in the project report has not been submitted previously by anybody for the award of any degree or diploma to any university.

Place:

Date:

Ishika Bhardwaj (8209)

Akkati Chethan (8186)

ACKNOWLEDGEMENT

We wish to express our sincere gratitude to our Professor Dr. Manju Sardana, for providing his invaluable guidance, comments and suggestions throughout the course of the project.

We also wish to express our gratitude to the officials and other staff members of various college fractions who rendered their help in requirement gathering in the Project.

Lastly, we would like to thank our parents and friends for all their moral support they have given us during the completion of this work.

TABLE OF CONTENTS

| | |
|--|----|
| 1. Introduction..... | 7 |
| 2. Approach..... | 9 |
| a. Dataset..... | 9 |
| b. Pre-processing..... | 10 |
| c. Data clean..... | 11 |
| d. Visualization..... | 11 |
| e. Training & Testing..... | 16 |
| 3. Classification Techniques used and results..... | 17 |
| a. K-Nearest Neighbor..... | 17 |
| b. Naive Bayes Classifier..... | 18 |
| c. Decision Tree Classifier..... | 19 |
| 4. Modules..... | 21 |
| 5. Results..... | 23 |
| 6. Conclusion..... | 24 |
| 7. Limitations..... | 24 |
| 8. Future Scope..... | 25 |
| 9. References..... | 25 |
| 10. Appendix..... | 25 |

HEART DISEASE PREDICTION

Abstract—The world has seen an unprecedented and exponential increase in cases of heart disease worldwide every day. In the paper, the early prognosis of heart disease through careful treatment and the implementation of a healthy lifestyle through other studies will help prevent many cardiovascular diseases. This project discusses a statistical model of heart disease that, based on basic parameters of the patients' health history, will help medical examiners and cardiac practitioners forecast heart disease. To build this prediction model, three different Classifier Models are used, namely, K-Nearest Neighbors Classifier, Naive Bayes Classifier and Decision Tree Classifier. Different important clinical features of a patient, critical for deciding a patient's heart disease, are taken in the first section and, secondly, different Classifiers are defined on the given dataset and their accuracy calculated.

I. INTRODUCTION

In terms of time, accuracy, and cost, medical dictation has always remained a high maintenance field. Human beings are susceptible to mistakes and can make errors. At an exponential pace, cases of cardiovascular diseases are that, and that is very troubling. In order to point out any prediction based on many factors, the human mind cannot process too much estimation and can therefore provide incorrect feedback several times, leading to vital risk to the patient. Data Mining has proved itself to be very effective in forecasting diverse scenarios for numerous fields. With data mining and deep learning paired with each other, many models have been designed to forecast specific scenarios for us to operate on them. Similarly, the Dataset for Heart Disease research was used to train a model using three distinct classifier algorithms to forecast heart disease with the highest accuracy. Our models are trained on 14 Dataset parameters, namely the K-Nearest Neighbor Classifier, Naive Bayes Algorithm, and Decision tree Classifier, which are very simple for early detection of heart disease, helping patients to sustain a healthier lifestyle along with taking sufficient precautionary steps to prevent future heart disease.

In the prediction of heart disease, computational intelligence has an important function. The relationships between patient characteristics such as heart rate, obesity and other conditions can be defined by terms used in computer intelligence.

To predict illness, it utilizes vast databases and past medical history. Today, depending on the risk factors and clinical history of the patient, ML algorithms and techniques are used to indicate heart diseases. These variables are taken as a parameter, such as heart rate, age, blood pressure (BP), obesity, sex, and more, and algorithms are applied to compare the characteristics and predict

heart disease, namely KNN, Decision trees, logistic regression, and random forest classifier.

In order to explain the nature of our data, the data is analyzed in the first stage with different Exploratory Data Analysis (EDA) techniques, followed by the application of some standardization to correct the data in the event of some empty data cell errors. Again, the data is analyzed in order to learn various types and samples of data under EDA techniques. In order to train and test models, the data is then severed and split into two parts.

The project is subdivided into the following section:

1. Loading necessary libraries.
2. Loading Dataset from a CSV file or from a Table.
3. Summarization of Data to understand Dataset (Descriptive Statistics).
4. Visualization of Data to understand Dataset (Plots, Graphs etc.).
5. Data pre-processing and Data transformation.
6. Applying different learning algorithms on the training dataset.
7. Evaluating the performance of the fitted model using evaluation metrics like confusion matrix, precision recall curves.

II. APPROACH

The Proposed methodology's implementation begins with downloading publicly available dataset. Then data cleaning and normalization is done as a step of preprocessing of data. After cleaning and normalizing, the data is visualized. The data visualization helps to observe the trends and the relationship of attributes in the dataset.

The dataset is then splitted into test and train data. The model is trained by applying algorithms on the training dataset, and then the model is tested. At last, the comparison of all classification algorithms implemented in this model is done.

A. Dataset

Dataset download link:

<https://www.kaggle.com/datasets/rishidamarla/heart-disease-prediction>

The dataset's open-source dataset registry is used in analysis. It has numerous disease-related databases. It is the dataset existence category data category. With 303 instances and 75 properties, it is a multivariate dataset. The dataset includes both knowledge that is helpful and attributes that are not useful. So, in pre-processing the useful data is selected and data cleaning is done to remove the null values.

B. Pre-Processing

As at this preprocessing level, this is the important step; meaningful data is derived from the dataset of heart disease. This phase is compulsory because the raw data is not reliable and unfinished, so pre-processing is performed for more steps to render ready raw data. In this approach, during pre-processing, 14 attributes are extracted to understand the nature of patients' health better. The extricated 14 attributes include BP, sex, heart rate, chest, and others. The attribute's values are normalized and converted into numerical form.

Data contains:

- Age - age in years
- Sex - (1 = male; 0 = female)
- Chest pain type - chest pain type
- BP - resting blood pressure (in mm Hg on admission to the hospital)
- Cholesterol - serum cholesterol in mg/dl
- FBS - (fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)
- EKG results - resting electrocardiographic results
- Max HR - maximum heart rate achieved
- Exercise angina - exercise induced angina (1 = yes; 0 = no)
- ST depression - ST depression induced by exercise relative to rest
- Slope of ST - the slope of the peak exercise ST segment
- Number of vessels fluro - number of major vessels (0-3) colored by fluoroscopy
- Thallium - 3 = normal; 6 = fixed defect; 7 = reversible defect
- Heart Disease - have disease or not (1=yes, 0=no)

C. Data Clean

The quality of data plays an essential role, and the most carefully depicted thing to be. For this research, data cleaning has improved the quality of our dataset. Data cleaning is necessary as it removes unnecessary or irrelevant attributes of data from the dataset. This step of the model will make the dataset more precise and exact. In this part of the approach, the Null (NaN) values are removed from the dataset to make it more useful as these values decrease the productivity of the algorithm. At the data cleaning stage, the dataset is also normalized to not have any ambiguity after cleaning

D. Visualization

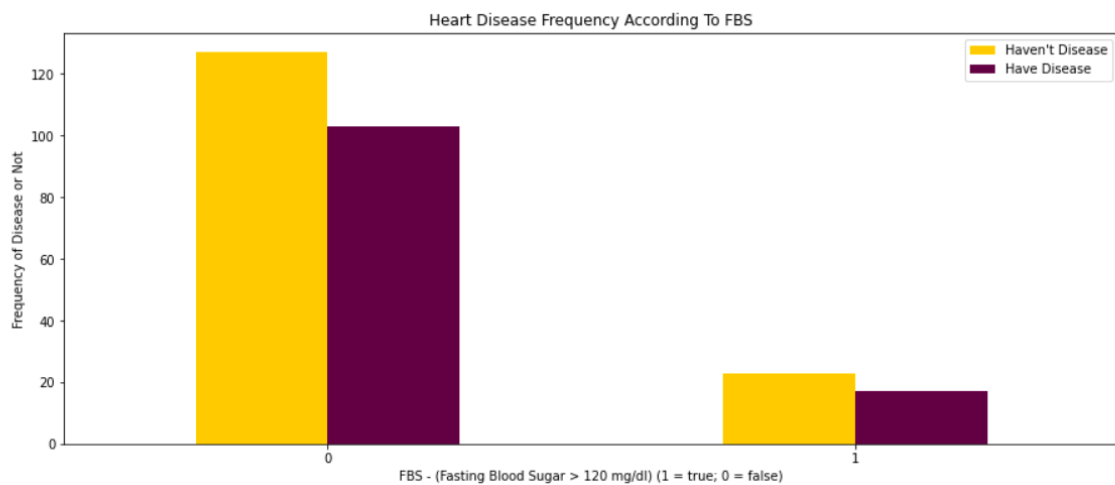
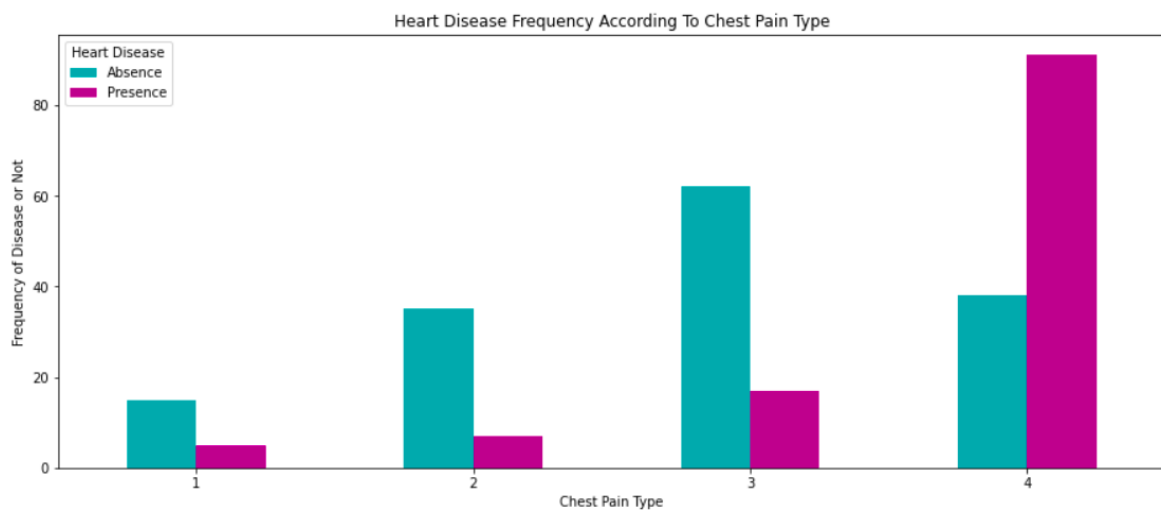
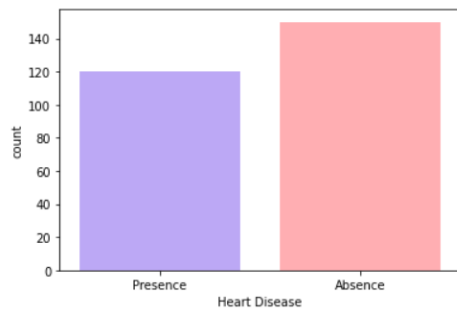
The dataset is in tabular form and it is hard to observe and understand the data in this or any other form. So the data is visualized graphically. It helps in knowing the trend of the data. Data visualization in this approach is a graphical representation of the data. In this analysis, using bar charts and scatter plots, the cleaned data acquired by pre-processing is visualized. It illustrates the actions of data attributes. It makes it easy to grasp the attribute's complicated relationship by graphical representation.

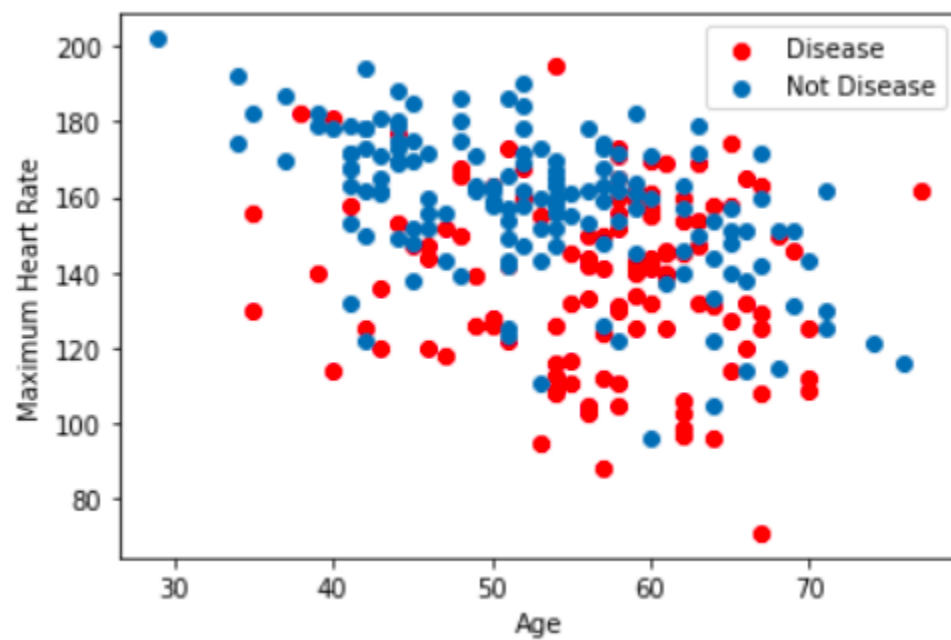
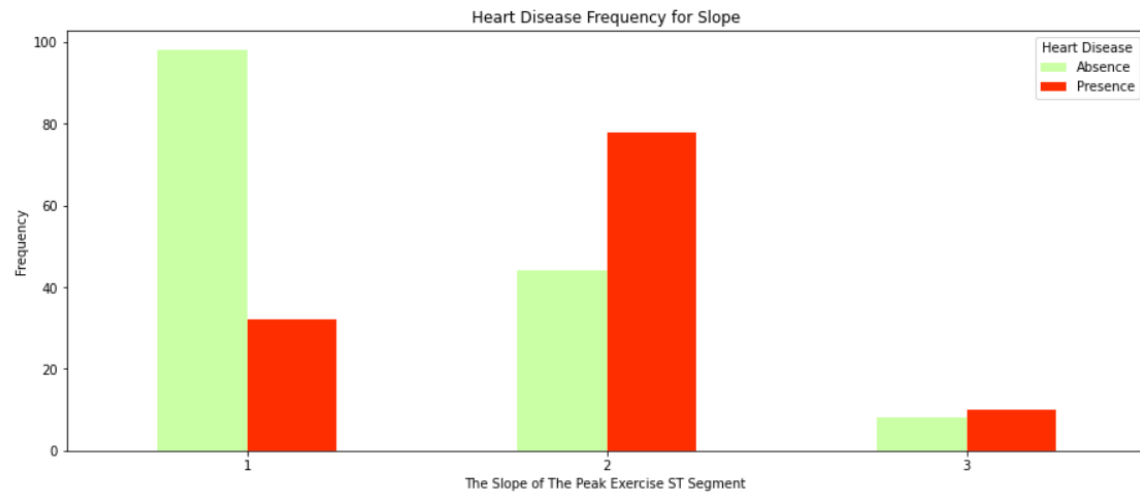
As mentioned above, this visualization plays a crucial role in data exploration. The various parameters of the dataset plotted depend on the age of patients.

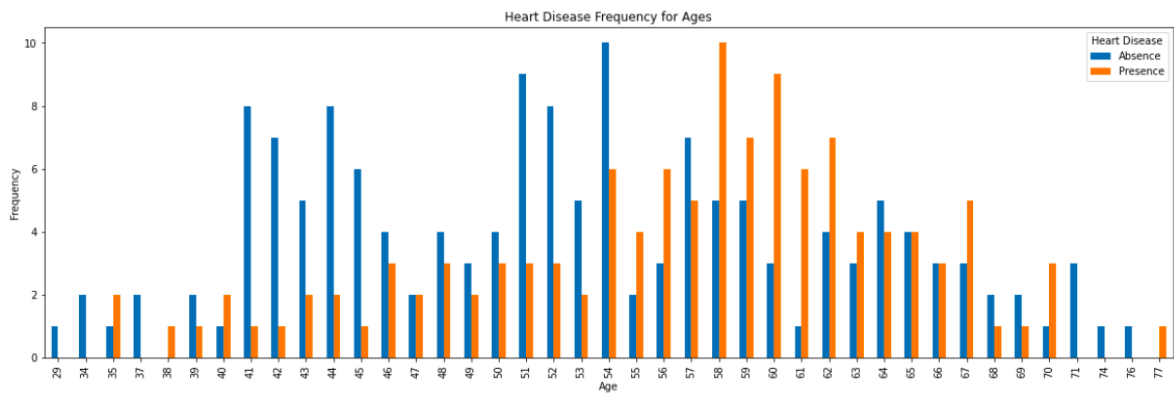
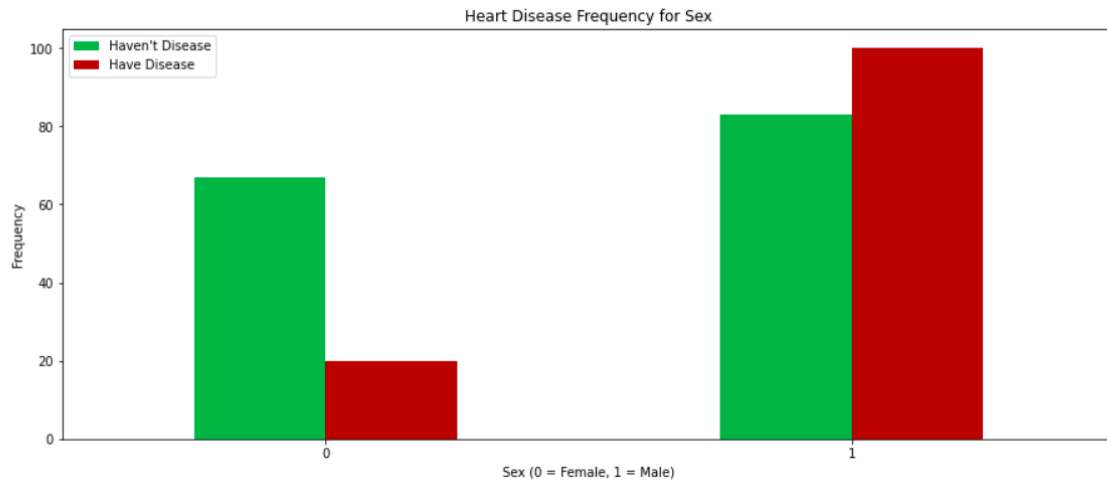
```
In [12]: df["Heart Disease"].value_counts()
```

```
Out[12]: Absence    150  
         Presence    120  
         Name: Heart Disease, dtype: int64
```

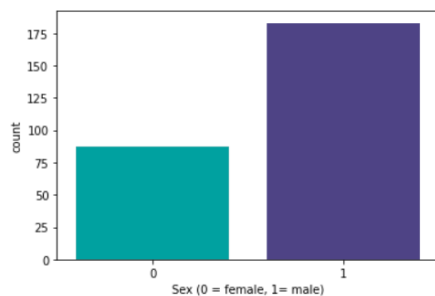
```
In [13]: sns.countplot(x="Heart Disease", data=df, palette="bwr")  
plt.show()
```







```
In [15]: sns.countplot(x='Sex', data=df, palette="mako_r")
plt.xlabel("Sex (0 = female, 1= male)")
plt.show()
```



```
In [16]: countFemale = len(df[df.Sex == 0])
countMale = len(df[df.Sex == 1])
print("Percentage of Female Patients: {:.2f}%".format((countFemale / (len(df.Sex))*100)))
print("Percentage of Male Patients: {:.2f}%".format((countMale / (len(df.Sex))*100)))
```

Percentage of Female Patients: 32.22%
Percentage of Male Patients: 67.78%

Using this Pairplot method from the Seaborn and Matplotlib library of Python, it can be inferred the cardiac disorder, aka. Cardiovascular disorders rely primarily on the patient's age.

Correlation Among the features of each currency : As shown earlier the features of each currency are highly correlated and here is another proof in support of our prediction.

| | Age | Sex | Chest pain type | BP | Cholesterol | FBS over 120 | EKG results | Max HR | Exercise angina | ST depression | Slope of ST | Number of vessels fluro | Thallium |
|-------------------------|-----------|-----------|-----------------|-----------|-------------|--------------|-------------|-----------|-----------------|---------------|-------------|-------------------------|-----------|
| Age | 1.000000 | -0.094401 | 0.096920 | 0.273053 | 0.220056 | 0.123458 | 0.128171 | -0.402215 | 0.098297 | 0.194234 | 0.159774 | 0.356081 | 0.106100 |
| Sex | -0.094401 | 1.000000 | 0.034636 | -0.062693 | -0.201647 | 0.042140 | 0.039253 | -0.076101 | 0.180022 | 0.097412 | 0.050545 | 0.086830 | 0.391046 |
| Chest pain type | 0.096920 | 0.034636 | 1.000000 | -0.043196 | 0.090465 | -0.098537 | 0.074325 | -0.317682 | 0.353160 | 0.167244 | 0.136900 | 0.225890 | 0.262659 |
| BP | 0.273053 | -0.062693 | -0.043196 | 1.000000 | 0.173019 | 0.155681 | 0.116157 | -0.039136 | 0.082793 | 0.222800 | 0.142472 | 0.085697 | 0.132045 |
| Cholesterol | 0.220056 | -0.201647 | 0.090465 | 0.173019 | 1.000000 | 0.025186 | 0.167652 | -0.018739 | 0.078243 | 0.027709 | -0.005755 | 0.126541 | 0.028836 |
| FBS over 120 | 0.123458 | 0.042140 | -0.098537 | 0.155681 | 0.025186 | 1.000000 | 0.053499 | 0.022494 | -0.004107 | -0.025538 | 0.044076 | 0.123774 | 0.049237 |
| EKG results | 0.128171 | 0.039253 | 0.074325 | 0.116157 | 0.167652 | 0.053499 | 1.000000 | -0.074628 | 0.095098 | 0.120034 | 0.160614 | 0.114368 | 0.007337 |
| Max HR | -0.402215 | -0.076101 | -0.317682 | -0.039136 | -0.018739 | 0.022494 | -0.074628 | 1.000000 | -0.380719 | -0.349045 | -0.386847 | -0.265333 | -0.253397 |
| Exercise angina | 0.098297 | 0.180022 | 0.353160 | 0.082793 | 0.078243 | -0.004107 | 0.095098 | -0.380719 | 1.000000 | 0.274672 | 0.255908 | 0.153347 | 0.321449 |
| ST depression | 0.194234 | 0.097412 | 0.167244 | 0.222800 | 0.027709 | -0.025538 | 0.120034 | -0.349045 | 0.274672 | 1.000000 | 0.609712 | 0.255005 | 0.324333 |
| Slope of ST | 0.159774 | 0.050545 | 0.136900 | 0.142472 | -0.005755 | 0.044076 | 0.160614 | -0.386847 | 0.255908 | 0.609712 | 1.000000 | 0.109498 | 0.283678 |
| Number of vessels fluro | 0.356081 | 0.086830 | 0.225890 | 0.085697 | 0.126541 | 0.123774 | 0.114368 | -0.265333 | 0.153347 | 0.255005 | 0.109498 | 1.000000 | 0.255648 |
| Thallium | 0.106100 | 0.391046 | 0.262659 | 0.132045 | 0.028836 | 0.049237 | 0.007337 | -0.253397 | 0.321449 | 0.324333 | 0.283678 | 0.255648 | 1.000000 |

E. Training & Testing

Any data mining algorithm relies on data for input and output. The input data is called a training dataset, and data is checked for the knowledge to verify model functioning. The Heart dataset is split into databases of train and test. Through applying algorithms on the training dataset to learn, the models are generated and then evaluated using the testing dataset. Random signals and inputs are used in the testing process to verify whether or not the model is operating correctly.

Algorithms play a vital role in the construction of machine learning and data science models entirely. These algorithms can include supervised algorithms, semi supervised algorithms, unsupervised algorithms, and reinforcement algorithms. These types are divided into different algorithms, and the following three algorithms used in this project, namely, K-Nearest Neighbor, Naive Bayes Classifier and Decision Tree algorithm all are Supervised Algorithms.

III. Classification Techniques Used and Results

1. K-Nearest Neighbor

K-nearest neighbors (KNN) algorithm is a type of supervised ML algorithm which can be used for both classification as well as regression predictive problems.

However, it is mainly used for classification of predictive problems in industry.

K-nearest neighbors (KNN) algorithm uses 'feature similarity' to predict the values of new data points which further means that the new data point will be assigned a value based on how closely it matches the points in the training set.

KNN extracts from the dataset the data points and estimates the closest output. It provides very high predictive precision. This technique is used because it fits well with pattern recognition, because there are several features in the heart disease dataset. KNN extracts logic and knowledge based on the Euclidean distance Samples function $d(x_i, x_j)$ along with the majority of KNN.

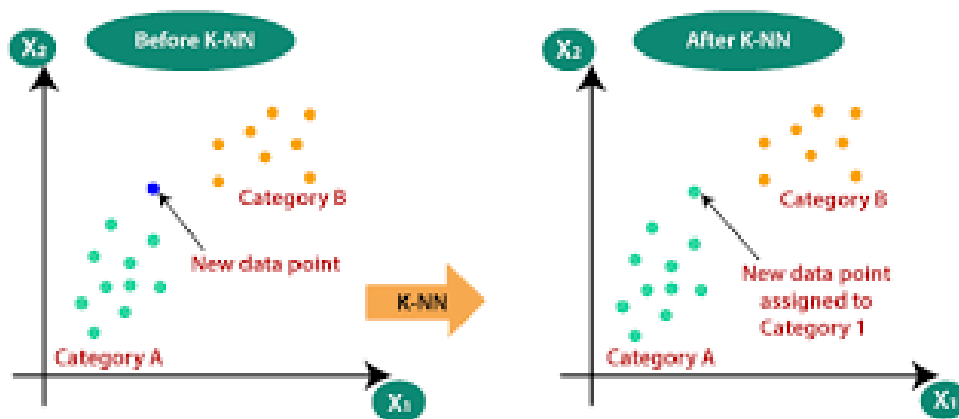


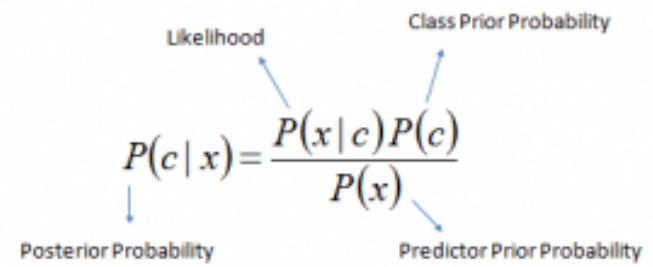
Fig: KNN Classifier

In the project, KNN technique produced an accuracy of 79.63%.

2. Naive Bayes Classifier

Naive Bayes is a kind of classifier which uses the Bayes Theorem. It predicts membership probabilities for each class such as the probability that a given record or data point belongs to a particular class. The class with the highest probability is considered as the most likely class. This is also known as **Maximum A Posteriori (MAP)**.

It is a classification technique based on Bayes' Theorem with an assumption of independence among predictors. In simple terms, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature. For example, a fruit may be considered to be an apple if it is red, round, and about 3 inches in diameter. Even if these features depend on each other or upon the existence of the other features, all of these properties independently contribute to the probability that this fruit is an apple and that is why it is known as 'Naive'. Naive Bayes model is easy to build and particularly useful for very large data sets. Along with simplicity, Naive Bayes is known to outperform even highly sophisticated classification methods. Bayes theorem provides a way of calculating posterior probability $P(c|x)$ from $P(c)$, $P(x)$ and $P(x|c)$. Look at the equation below:



The diagram shows the Bayes Theorem equation $P(c|x) = \frac{P(x|c)P(c)}{P(x)}$ with four labels and arrows: 'Likelihood' points to $P(x|c)$, 'Class Prior Probability' points to $P(c)$, 'Posterior Probability' points to $P(c|x)$, and 'Predictor Prior Probability' points to $P(x)$.

$$P(c|x) = \frac{P(x|c)P(c)}{P(x)}$$
$$P(c|X) = P(x_1|c) \times P(x_2|c) \times \dots \times P(x_n|c) \times P(c)$$

Fig: Bayes Theorem

- $P(c|x)$ is the posterior probability of class (c, target) given predictor (x, attributes).
- $P(c)$ is the prior probability of class.
- $P(x|c)$ is the likelihood which is the probability of predictor given class.
- $P(x)$ is the prior probability of predictor.

In the project, Naive Bayes technique produced an accuracy of 68.52%.

3. Decision Tree Classifier

Decision Tree is a Supervised learning technique that can be used for both classification and Regression problems, but mostly it is preferred for solving Classification problems. It is a tree-structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome. In a Decision tree, there are two nodes, which are the Decision Node and Leaf Node. Decision nodes are used to make any decision and have multiple branches, whereas Leaf nodes are the output of those decisions and do not contain any further branches. The decisions or the test are performed on the basis of features of the given dataset.

It is a graphical representation for getting all the possible solutions to a problem/decision based on given data. It is called a decision tree because, similar to a tree, it starts with the root node, which expands on further branches and constructs a tree-like structure. A decision tree simply asks a question, and based on the answer (Yes/No), it further splits the tree into subtrees.

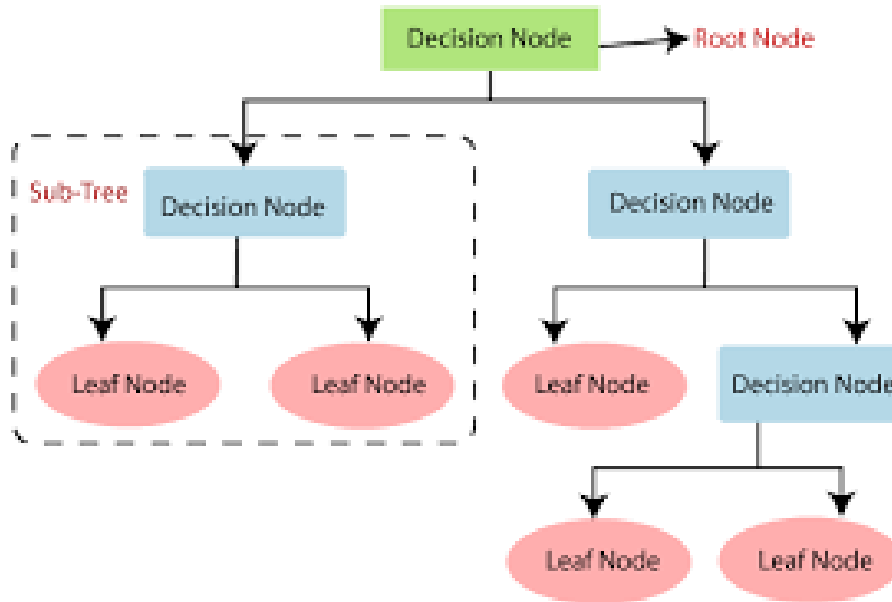
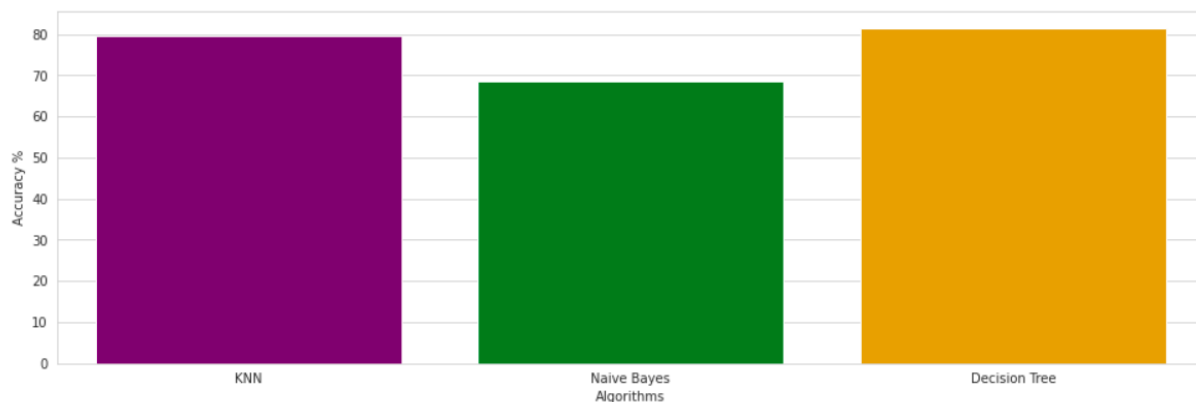


Fig: Decision Tree working

In the project, Decision Tree classification technique produced an accuracy of 81.48%.

Therefore, for the Heart Disease Prediction dataset, Decision Tree Classifier has performed the best with an accuracy of 81.48%.



IV. MODULES

1. **Pandas** : Pandas allows importing data from various file formats such as comma-separated values, JSON, SQL, Microsoft Excel. Pandas allows various data manipulation operations such as merging, reshaping, selecting, as well as data cleaning, and data wrangling features.
2. **NumPy** : NumPy is the fundamental package for scientific computing in Python. It is a Python library that provides a multidimensional array object, various derived objects and an assortment of routines for fast operations on arrays, including mathematical, logical, shape manipulation, sorting, selecting, I/O, discrete Fourier transforms, basic linear algebra, basic statistical operations, random simulation and much more.
3. **Matplotlib** : Matplotlib is a plotting library for the Python programming language and its numerical mathematics extension NumPy. It provides an object-oriented API for embedding plots into applications using general-purpose GUI toolkits like Tkinter, wxPython, Qt, or GTK.
4. **Naïve Bayes Classifier** : Naive Bayes is a supervised learning algorithm used for classification tasks. Hence, it is also called Naive Bayes Classifier. As other supervised learning algorithms, naive bayes use features to make a prediction on a target variable. The key difference is that naive bayes assume that features are independent of each other and there is no correlation between features. However, this is not the case in real life. This naive assumption of features being uncorrelated is the reason why this algorithm is called “naive”.

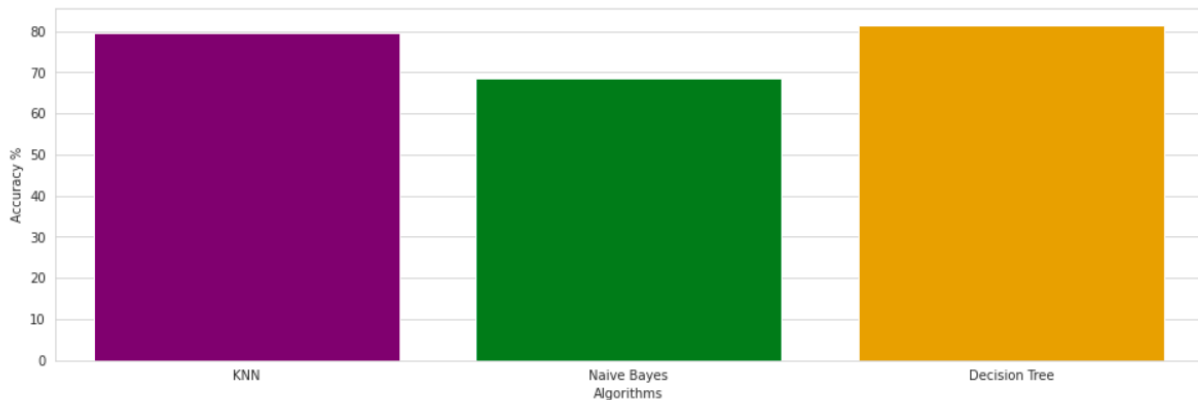
5. **KNN Model** : The k-nearest neighbors (KNN) algorithm is a simple, supervised machine learning algorithm that can be used to solve both classification and regression problems. It is one of the simplest machine learning algorithms and is used in a wide array of institutions. KNN is a non-parametric, lazy learning algorithm.
6. **Decision Tree**: Decision Tree is a Supervised Machine Learning Algorithm that uses a set of rules to make decisions, similarly to how humans make decisions. Decision trees can perform both classification and regression tasks, so you'll see authors refer to them as the CART algorithm: Classification and Regression Tree. This is an umbrella term, applicable to all tree-based algorithms, not just decision trees. The intuition behind Decision Trees is that you use the dataset features to create yes/no questions and continually split the dataset until you isolate all data points belonging to each class.
7. **Seaborn** : Seaborn is a library for making statistical graphics in Python. It builds on top of matplotlib and integrates closely with pandas data structures. Seaborn helps you explore and understand your data. Its plotting functions. Operate on dataframes and arrays containing whole datasets and internally perform the necessary semantic mapping and statistical aggregation to produce informative plots. Its dataset-oriented, declarative API lets you focus on what the different elements of your plots mean, rather than on the details of how to draw them.
8. **Scikit-Learn** : Scikit-learn (Sklearn) is the most useful and robust library for machine learning in Python. It provides a selection of efficient tools for machine learning and statistical modeling including classification, regression, clustering and dimensionality reduction via a consistent

interface in Python. This library, which is largely written in Python, is built upon NumPy, SciPy and Matplotlib.

V. RESULTS

Datasets are applied to KNN, Naive Bayes and Decision Tree classifiers.

Concerning the algorithm and the proportion of test data, the results differ. When the proportion of test data is 0.2 percent, the maximum accuracy achieved is 81.48% percent by Decision Tree Classifier.



As mentioned, different models constitute different results, accuracy levels, and error rates. Throughout time, the Heart Disease dataset has been used due to its easy availability and ease of usage, which can be seen since many approaches are being used.

VI. CONCLUSION

A very detailed, useful, and highly preferable Learning based model in this project that helps medical practitioners diagnose heart diseases at an early stage to enable patients to take precautionary measures in a rectification window. Having three separate classifiers used in a model, based on the findings shown above, it can be inferred that the proportion of test and training data plays an enormous role in a classification model.

VII. LIMITATIONS

1. Certain limitations of our project are that one should know the dataset and all the components in full details as it is the major part.
2. Secondly, one should have the knowledge of python and various models of machine learning to apply them.
3. The knowledge of working of these models and the output or results given by these models should be known.
4. The scope of the project is limited to the models and the relative dataset on which these models are applied.
5. Moreover we are not providing any user interface so you have to search and load your own dataset.

VIII. FUTURE SCOPE

By adding more attributes in the dataset, the model described in the paper can further be elongated to diversify. For better accuracy, you can add the number of attributes used for prediction. Moreover, the dataset's size can be increased too; this will also help get preferable accuracy.

Algorithms such as Help Vector Machine, Linear Regression, and others may be beneficial in order to get a greater degree of precision for the results. Centered on diseases and algorithms, a comparative study of the performance of this model would be facilitated.

IX. REFERENCES

- <https://www.geeksforgeeks.org/k-nearest-neighbor-algorithm-in-python/>
- <https://towardsai.net/p/latest/capital-assets-pricing-modelcapm%E2%80%8BAusing-python>
- <https://www.geeksforgeeks.org/k-nearest-neighbors-with-python-ml/>
- https://www.geeksforgeeks.org/ml-sklearn-linear_model-linearregressionin-python/
- <https://medium.com/@yrnigam/how-to-write-a-data-science-report181bd49d8f4d>

X. APPENDIX

- Jupyter Notebook
- Dataset: Heart_Disease_Prediction.csv