

**VISHWAKARMA INSTITUTE OF TECHNOLOGY, PUNE – 411037**

666, Upper Indiranagar , Bibwewadi, Pune 411037

B.Tech Second Year 2020-21

Semester 2

**Subject :** Data Science **Code** : **MD2201**

**Course Project**

**Topic – Heart disease prediction using machine learning**

**Div.: ME-D Batch 2 – Group 2**

Students Names :

1. D26 Aayush Surawar 11910859

2. D27 Samarth Takbhate 11910739

3. D30 Gopal Tapadiya 11911318

4. D31 Shivam tapre 11910774

5. D43 Umesh Wanare 11911112

**Guided by – Prof. S. Ballal**

**Abstract—** *Machine learning can be used to predict the Heart Disease. We use Cleveland Heart Disease dataset which has some non-linear tendency. We improved Heart Disease Prediction technique by correctly adjusting the Random Forest Machine Learning Model (fetching 85.81% accuracy). Health care data contains hidden information which is useful for making effective decisions. This can help patients in getting a quick diagnosis with a lot less cost.* *Heart plays significant role in living organisms. Diagnosis and prediction of heart related diseases requires more precision, perfection and correctness because a little mistake can cause fatigue problem or death of the person, there are numerous death cases related to heart and their counting is increasing exponentially day by day. To deal with the problem there is essential need of prediction system for awareness about diseases. Machine learning is the branch of Artificial Intelligence(AI), it provides prestigious support in predicting any kind of event which take training from natural events.*

**Keywords —** Data Analysis, Data Visualisation, Machine learning Modelling, Jupyter Notebook.

**Introduction :**

Heart is one of the most extensive and vital organ of human body so the care of heart is essential. Most of diseases are related to heart so the prediction about heart diseases is necessary and for this purpose comparative study needed in this field, today most of patient are died because their diseases are recognized at last stage due to lack of accuracy of instrument so there is need to know about the more efficient algorithms for diseases prediction.

Machine Learning is one of the efficient technology for the testing, which is based on training and testing. It is the branch of Artificial Intelligence(AI) which is one of broad area of learning where machines emulating human abilities, machine learning is a specific branch of AI. On the other hand machines learning systems are trained to learn how to process and make use of data hence the combination of both technology is also called as Machine Intelligence.

As the definition of machine learning, it learns from the natural phenomenon, natural things so in this project we uses the biological parameter as testing data such as cholesterol, Blood pressure, sex, age, etc. and on the basis of these, comparison is done in the terms of accuracy of algorithms such as in this project we have used four algorithms which are decision tree, linear regression, k-neighbour, SVM.

In this project, we calculate the accuracy of different machine learning approaches and on the basis of calculation we conclude that which one is best among them. Section 1 of this paper consist the introduction about the machine learning and heart diseases. Section II described, the machine learning classification. Section III illustrated the related work of researchers. Section IV is about the methodology used for this prediction system. Section V is about the algorithms used in this project. Section VI briefly describes the dataset and their analysis with the result of this project. And the last Section VII concludes the summary of this paper with slight view about future scope of this paper.

**Project Objective :**

**“To build a Heart Disease prediction system to overcome the shortcomings of the prior Heart Disease detection techniques.”**

* Predict whether a patient should be diagnosed with Heart Disease. This is a binary outcome.

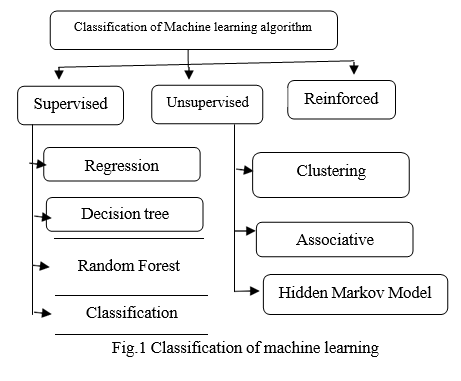
Positive (+) = 1, patient diagnosed with Heart Disease  
Negative (-) = 0, patient not diagnosed with Heart Disease

* Experiment with various Classification Models & see which yields greatest accuracy.
* Examine trends & correlations within our data
* Determine which features are most important to Positive/Negative Heart Disease diagnosis

**Machine learning :**

Machine Learning is one of efficient technology which is based on two terms namely testing and training i.e. system take training directly from data and experience and based on this training test should be applied on different type of need as per the algorithm required.

There are three type of machine learning algorithms:



***Supervised Learning:***

Supervised learning can be define as learning with the proper guide or you can say that learning in the present of teacher .we have a training dataset which act as the teacher for prediction on the given dataset that is for testing a data there are always a training dataset. Supervised learning is based on "train me" concept. Supervised learning have following processes:

* + Classification
  + Random Forest
  + Decision tree
  + Regression

To recognize patterns and measures probability of uninterruptable outcomes, is phenomenon of regression. System have ability to identify numbers, their values and grouping sense of numbers which means width and height, etc. There are following supervised machine learning algorithms:

* + Linear Regression
  + Logistical Regression
  + Support Vector Machines (SVM)
  + Neural Networks
  + Random Forest
  + Gradient Boosted Trees
  + Decision Trees
  + Naive Bayes

***Unsupervised Learning:***

Unsupervised learning can be define as the learning without a guidance which in Unsupervised learning there are no teacher are guiding. In Unsupervised learning when a dataset is given it automatically work on the dataset and find the pattern and relationship between them and according to the created relationships, when new data is given it classify them and store in one of them relation . Unsupervised learning is based on "self sufficient " concept.

For example suppose there are combination fruits mango, banana and apple and when Unsupervised learning is applied it classify them in three different clusters on the basis if there relation with each other and when a new data is given it automatically send it to one of the cluster .

Supervisor learning say there are mango, banana and apple but Unsupervised learning said it as there are three different clusters. Unsupervised algorithms have following process:

* + Dimensionality
  + Clustering

There are following unsupervised machine learning algorithms:

* + t-SNE
  + k-means clustering
  + PCA

***Reinforcement :***

Reinforced learning is the agent ability to interact with the environment and find out the outcome. It is based on "hit and trial" concept. In reinforced learning each agent is awarded with positive and negative points and on the basis of positive points reinforced learning give the dataset output that is on the basis of positive awards it trained and on the basis of this training perform the testing on datasets.

**Methodology :**

1. *Data Collection*

First step for predication system is data collection and deciding about the training and testing dataset. In this project we have used 73% training dataset and 37% dataset used as testing dataset the system.

1. *Attribute Selection*

Attribute of dataset are property of dataset which are used for system and for heart many attributes are like heart bit rate of person, gender of the person, age of the person and many more shown in TABLE.1 for predication system.

1. *Preprocessing of data*

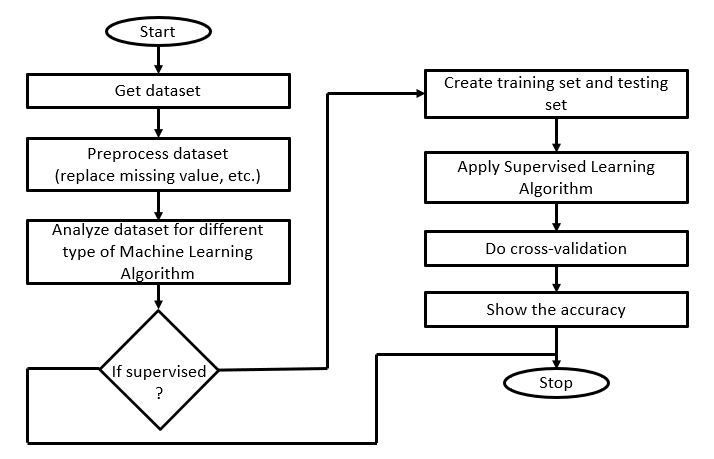
Preprocessing needed for achieving prestigious result from the machine learning algorithms. For example Random forest algorithm does not support null values dataset and for this we have to manage null values from original raw data.

For our project we have to convert some categorized value by dummy value means in the form of “0”and “1” by using following code:

1. *Data Balancing*

Data balancing is essential for accurate result because by data balancing graph we can see that both the target classes are equal. Fig.3 represents the target classes where “0” represents with heart diseases patient and “1” represents no heart diseases pateints.

**Design flow :**



**Machine learning algorithms :**

1. *Linear regression*

It is the supervised learning technique. It is based on the relationship between independent variable and dependent variable as seen in Fig.5 variable “x” and “y” are independent and dependent variable and relation between them is shown by equation of line which is linear in nature that why this approach is called linear regression.

It gives a relation equation to predict a dependent variable value “y” based on a independent variable value “x” as we can see in the Fig.5 so it is concluded that linear regression technique give the linear relationship between x(input) and y(output).

1. *Decision tree*

On the other hand decision tree is the graphical representation of the data and it is also the kind of supervised machine learning algorithms.

For the tree construction we use entropy of the data attributes and on the basis of attribute root and other nodes are drawn.

Entropy= -∑ Pij log Pij (1)

In the above equation of entropy (1) Pij is probability of the node and according to it the entropy of each node is calculated. The node which have highest entropy calculation is selected as the root node and this process is repeated until all the nodes of the tree are calculated or until the tree constructed.

When the number of nodes are imbalanced then tree is create the over fitting problem which is not good for the calculation and this is one of reason why decision tree have less accuracy as compare to linear regression.

1. *Support Vector Machine*

It is one category of machine learning technique which work on the concept of hyperplan means it classify the data by creating hyper plan between them.

Training sample dataset is (Yi, Xi) where i=1,2,3,…….n and Xi is the ith vector, Yi is the target vector. Number of hyper plan decide the type of support vector such as example if a line is used as hyper plan then method is called linear support vector.

Y



X

1. *K-nearest Neighbour*

It work on the basis of distance between the location of data and on the basis of this distinct data are classified with each other. All the other group of data are called neighbor of each other and number of neighbor are decided by the user which play very crucial role in analysis of the dataset.

Y



2

1

3

X

In the above Fig. k=3 shows that there are three neighbor that means three different type of data are there. Each cluster represented in two dimensional space whose coordinates are represented as (Xi,Yi) where Xi is the x-axis, Y represent y- axis and i= 1,2,3,….n.

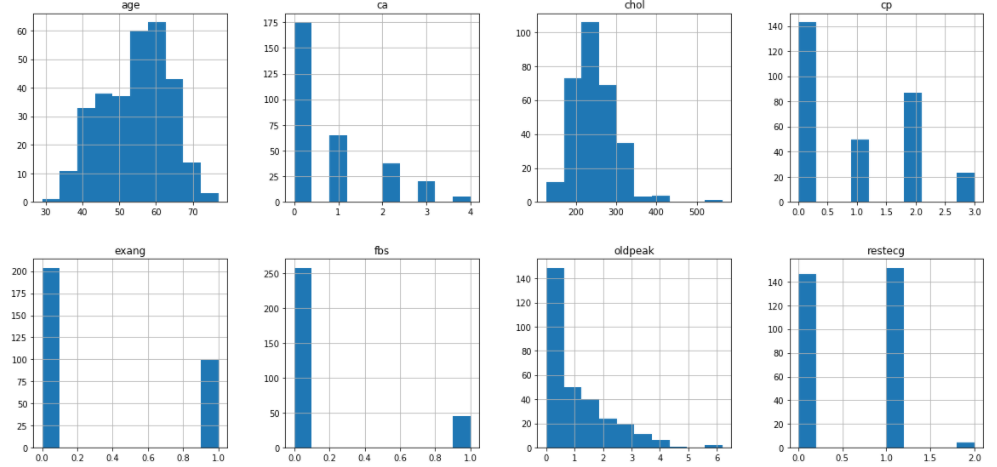
**Results :**

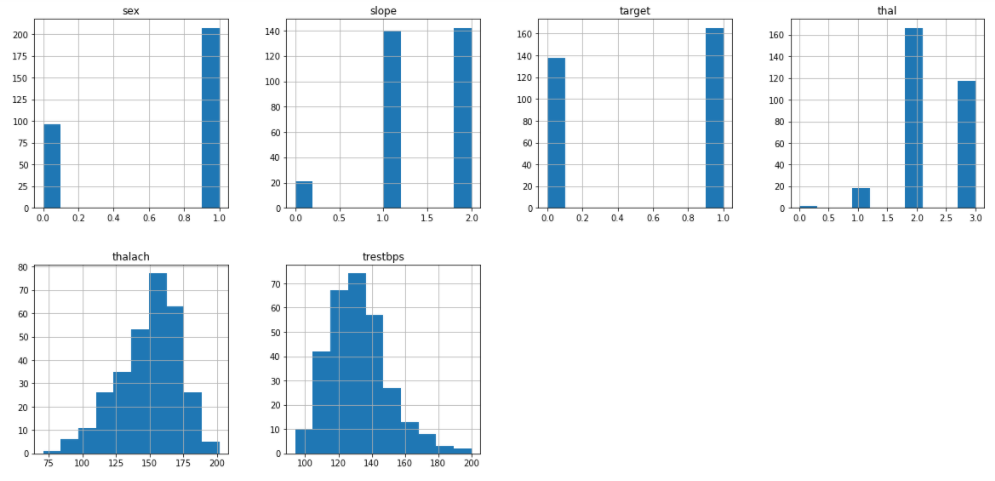
Dataset:

| **Sn0. age** | **sex** | **cp** | **trestbps** | **chol** | **fbs** | **restecg** | **thalach** | **exang** | **oldpeak** | **slope** | **ca** | **thal** | **target** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 0 | 63 | 1 | 3 | 145 | 233 | 1 | 0 | 150 | 0 | 2.3 | 0 | 0 | 1 | 1 |
| 1 | 37 | 1 | 2 | 130 | 250 | 0 | 1 | 187 | 0 | 3.5 | 0 | 0 | 2 | 1 |
| 2 | 41 | 0 | 1 | 130 | 204 | 0 | 0 | 172 | 0 | 1.4 | 2 | 0 | 2 | 1 |
| 3 | 56 | 1 | 1 | 120 | 236 | 0 | 1 | 178 | 0 | 0.8 | 2 | 0 | 2 | 1 |
| 4 | 57 | 0 | 0 | 120 | 354 | 0 | 1 | 163 | 1 | 0.6 | 2 | 0 | 2 | 1 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 298 | 57 | 0 | 0 | 140 | 241 | 0 | 1 | 123 | 1 | 0.2 | 1 | 0 | 3 | 0 |
| 299 | 45 | 1 | 3 | 110 | 264 | 0 | 1 | 132 | 0 | 1.2 | 1 | 0 | 3 | 0 |
| 300 | 68 | 1 | 0 | 144 | 193 | 1 | 1 | 141 | 0 | 3.4 | 1 | 2 | 3 | 0 |
| 301 | 57 | 1 | 0 | 130 | 131 | 0 | 1 | 115 | 1 | 1.2 | 1 | 1 | 3 | 0 |
| 302 | 57 | 0 | 1 | 130 | 236 | 0 | 0 | 174 | 0 | 0.0 | 1 | 1 | 2 | 0 |

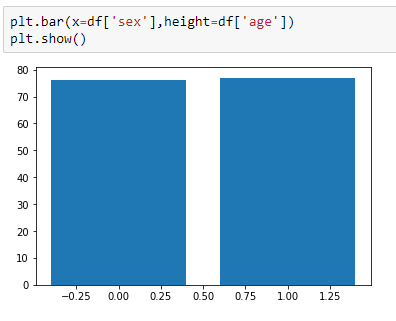
303 rows × 14 columns

**DATA VISUALIZATION :**

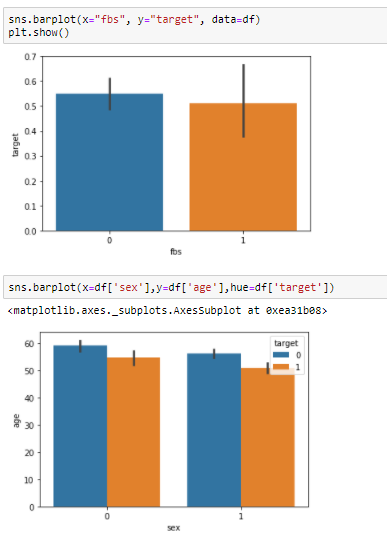




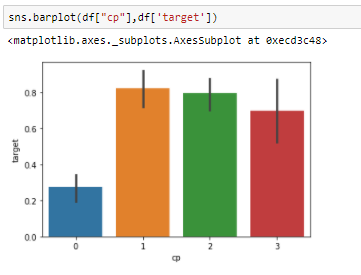
**Histograms of different attributes**



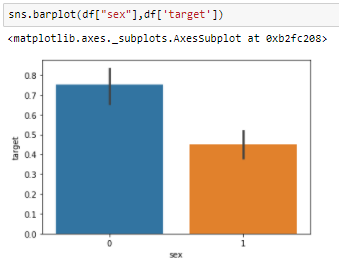
**Barplot of sex Vs age**



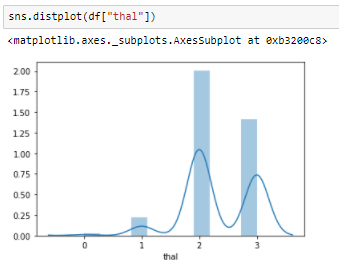
1. **Barplot of fbsVs target 2) Barplot of sex Vs age**



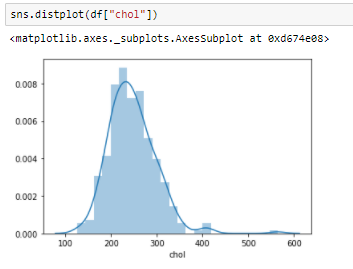
**Barplot of cp Vs target**

****

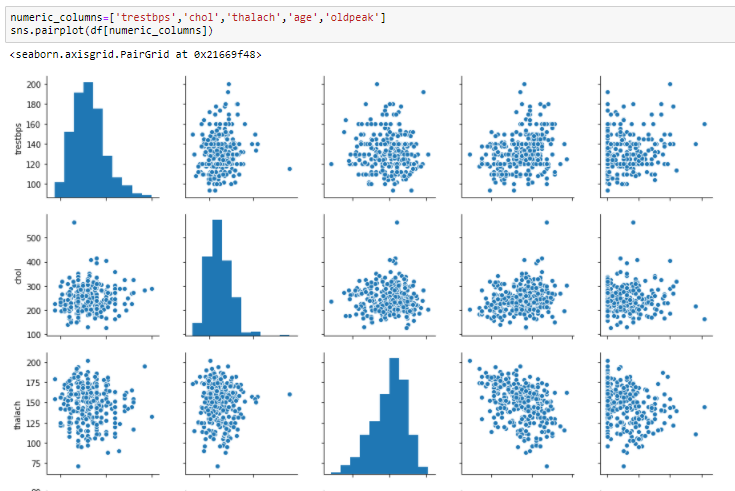
**Barplot of sex Vs target**

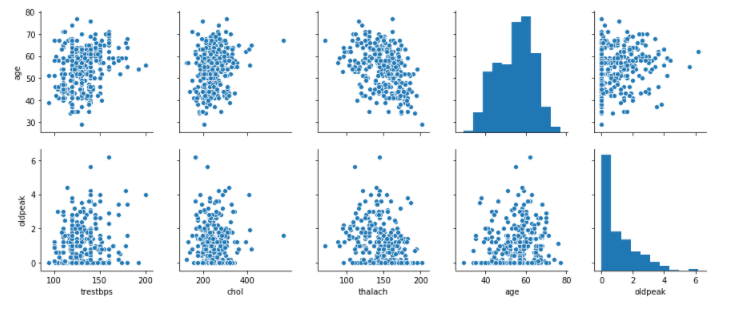
****

**Seaborn distplot of thal**

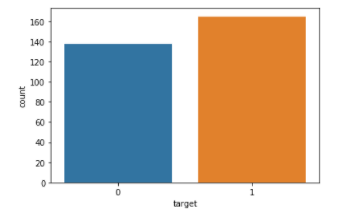


**Seaborn distplot of chol**



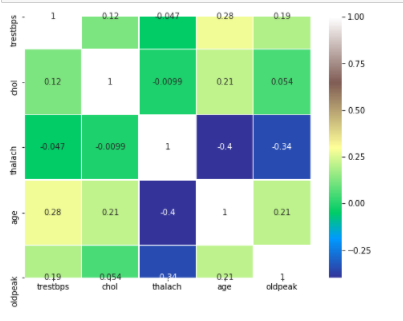


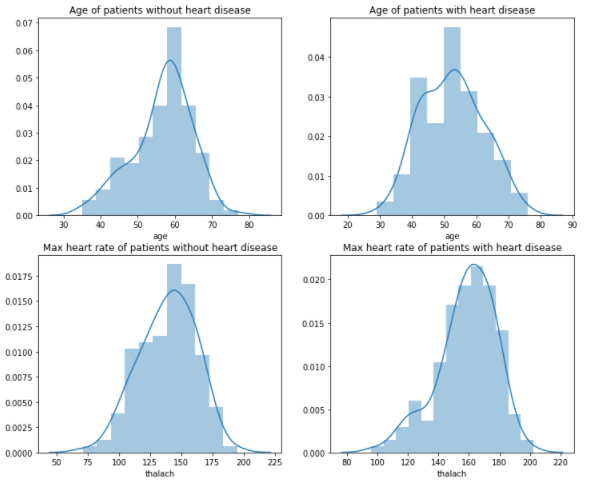
**Pairplot**

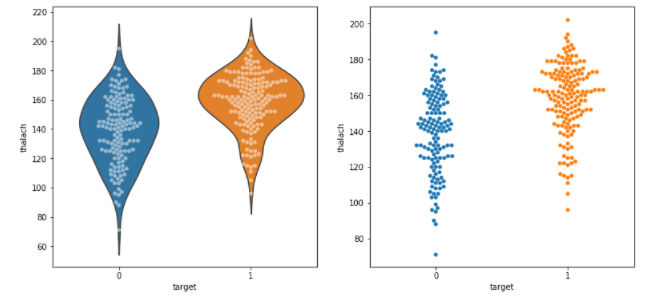


**Histogram of target vs count**

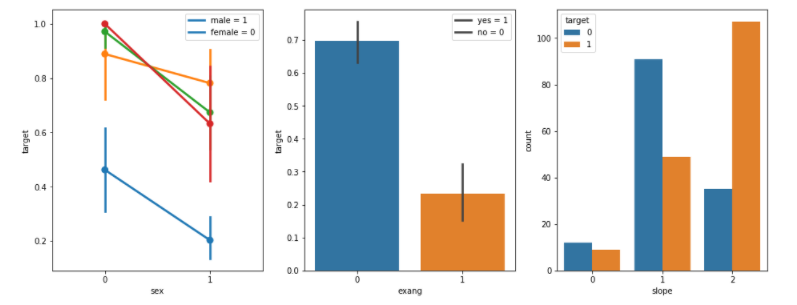
**Co-Relation Heatmap :-**



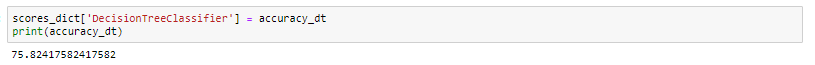


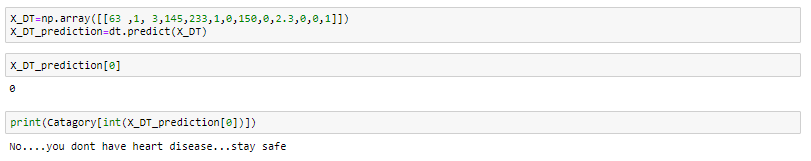


**Voilin and Boxplot**

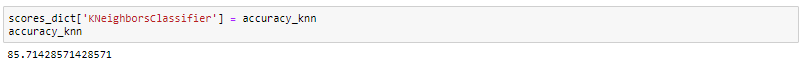


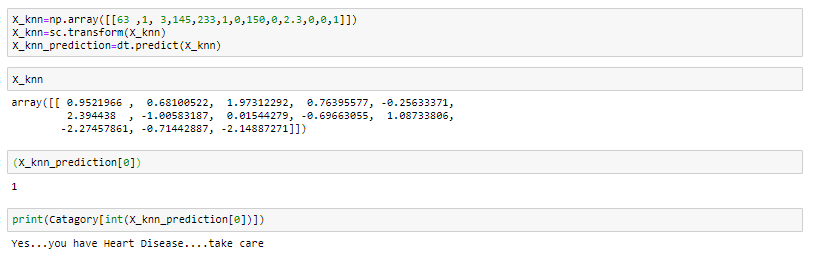
**Prediction using DecisionTree Classifier :-**

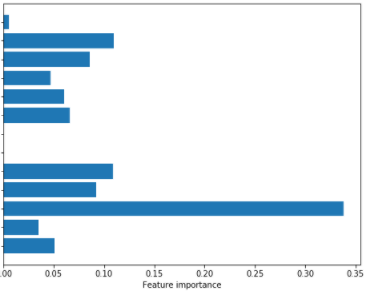




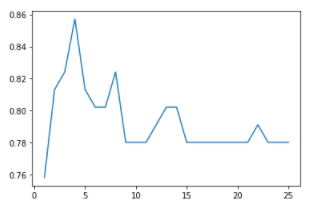
**Prediction using KNN :-**



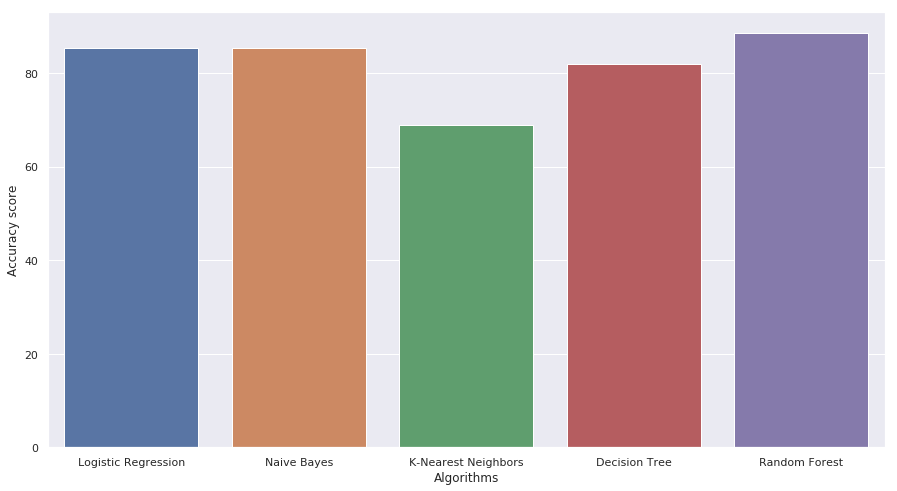




Chest Pain



**N vs Accuracy in KNN**



**Conclusion and future scope :**

Heart is one of the essential and vital organ of human body and prediction about heart diseases is also important concern for the human beings so that the accuracy for algorithm is one of parameter for analysis of performance of algorithms. Accuracy of the algorithms in machine learning depends upon the dataset that used for training and testing purpose. When we perform the analysis of algorithms on the basis of dataset on the basis of confusion matrix, we find **Random forest** is best one.

* Our dataset set had Non-Linear dependency.
* By correctly adjusting the parameters of Random Forest we were able to achieve better accuracy.
* We had lesser amount of dataset so 10-fold cross-validation gave us better result.
* Our solution strategy is not so Robust in nature, every time it needs a few adjustment in parameter.
* The accuracy achieved is satisfactory but can be future improved.
* We can use data warehouses in hospital so that the amount of data increases and a greater accuracy could be achieved.

For the Future Scope more machine learning approach will be used for best analysis of the heart diseases and for earlier prediction of diseases so that the rate of the death cases can be minimized by the awareness about the diseases.

**References:**

1. Santhana Krishnan J and Geetha S, “Prediction of Heart Disease using Machine Learning Algorithms” ICIICT, 2019.
2. Aditi Gavhane, Gouthami Kokkula, Isha Panday, Prof. Kailash Devadkar, “Prediction of Heart Disease using Machine Learning”, Proceedings of the 2nd International conference on Electronics, Communication and Aerospace Technology(ICECA), 2018.
3. Senthil kumar mohan, chandrasegar thirumalai and Gautam Srivastva, “Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques” IEEE Access 2019.
4. Himanshu Sharma and M A Rizvi, “Prediction of Heart Disease using Machine Learning Algorithms: A Survey” International Journal on Recent and Innovation Trends in Computing and Communication Volume: 5 Issue: 8 , IJRITCC August 2017.
5. M. Nikhil Kumar, K. V. S. Koushik, K. Deepak, “Prediction of Heart Diseases Using Data Mining and Machine Learning Algorithms and Tools” International Journal of Scientific Research in Computer Science, Engineering and Information Technology ,IJSRCSEIT 2019.
6. Amandeep Kaur and Jyoti Arora,“Heart Diseases Prediction using Data Mining Techniques: A survey” International Journal of Advanced Research in Computer Science , IJARCS 2015-2019.
7. Pahulpreet Singh Kohli and Shriya Arora, “Application of Machine Learning in Diseases Prediction”, 4th International Conference on Computing Communication And Automation(ICCCA), 2018.
8. M. Akhil, B. L. Deekshatulu, and P. Chandra, “Classification of Heart Disease Using K- Nearest Neighbor and Genetic Algorithm,” Procedia Technol., vol. 10, pp. 85–94, 2013.
9. S. Kumra, R. Saxena, and S. Mehta, “An Extensive Review on Swarm Robotics,” pp. 140–145, 2009.
10. Hazra, A., Mandal, S., Gupta, A. and Mukherjee, “ A Heart Disease Diagnosis and Prediction Using Machine Learning and Data Mining Techniques: A Review” Advances in Computational Sciences and Technology , 2017.
11. Patel, J., Upadhyay, P. and Patel, “Heart Disease Prediction Using Machine learning and Data Mining Technique” Journals of Computer Science & Electronics , 2016.
12. Chavan Patil, A.B. and Sonawane, P.“To Predict Heart Disease Risk and Medications Using Data Mining Techniques with an IoT Based Monitoring System for Post-Operative Heart Disease Patients” International Journal on Emerging Trends in Technology, 2017.
13. V. Kirubha and S. M. Priya, “Survey on Data Mining Algorithms in Disease Prediction,” vol. 38, no. 3, pp. 124–128, 2016.
14. M. A. Jabbar, P. Chandra, and B. L. Deekshatulu, “Prediction of risk score for heart disease using associative classification and hybrid feature subset selection,” Int. Conf. Intell. Syst. Des. Appl. ISDA, pp. 628–634, 2012.
15. https://archive.ics.uci.edu/ml/datasets/Heart+Disease