**CardioVascularRisk Prediction**

**Tanmay Bohra, Vikram Pratap**

**Amrutha B.S, Mahima Shree, Purnima Rai, Pooja Rana**

**Data science trainees,**

**AlmaBetter, Bangalore**

**Abstract**

Heart Disease prediction is one of the most complicated tasks in the medical field. In the modern era, approximately one person dies per minute due to heart disease. Data science plays a crucial role in processing huge amounts of data in the field of healthcare. As heart disease prediction is a complex task, there is a need to automate the prediction process to avoid risks associated with it and alert the patient well in advance. The proposed work predicts the chances of Heart Disease and classifies the patient's risk level by implementing different algorithms such as Nearest Neighbors, Decision Tree, Logistic Regression and Support Vector Machine. Thus, this paper presents a comparative study by analyzing the performance of different machine learning algorithms. The trial results verify that the Support Vector Machine algorithm has achieved the highest accuracy of 89% compared to other Machine Learning algorithms implemented.

**Keywords**: K Nearest Neighbors, Decision Tree, Logistic Regression, Support Vector Machine Heart Disease Prediction

**1. Problem Statement**

**A** lot of work has been carried out to predict heart disease using this dataset. Different levels of accuracy have been attained using various algorithms which are explained as follows.

We study various different Machine Learning algorithms that can be used for classification of heart disease. Research was carried out to study Decision Tree; KNN and K-Means algorithms that can be used for classification and their accuracy were compared. This research concludes that accuracy obtained by Decision Tree was highest, further it was inferred that it can be made efficient by combination of different techniques and parameter tuning.

We also designed a Machine Learning model comparing five different algorithms. In this research the accuracy of Decision Tree, Logistic Regression and SVM classification algorithms were compared. KNN algorithm had the highest accuracy based on their Accuracy, Precision, Recall and f-measure scores we identified the best classification algorithm which can be used in the heart disease prediction.

**2. INTRODUCTION**

The work proposed in this paper focus mainly on various data mining practices that are employed in heart disease prediction. Human heart is the principal part of the human body. Basically, it regulates blood flow throughout our body. Any irregularity to heart can cause distress in other parts of body. Any sort of disturbance to normal functioning of the heart can be classified as a Heart disease. In today’s contemporary world, heart disease is one of the primary reasons for occurrence of most deaths. Heart disease may occur due to unhealthy lifestyle, smoking, alcohol and high intake of fat which may cause hypertension. According to the World Health Organization more than 10 million die due to Heart diseases every single year around the world. A healthy lifestyle and earliest detection are only ways to prevent heart related diseases.

The main challenge in today's healthcare is provision of best quality services and effective accurate diagnosis. Even if heart diseases are found as the prime source of death in the world in recent years, they are also the ones that can be controlled and managed effectively. The whole accuracy in management of a disease lies on the proper time of detection of that disease. The proposed work makes an attempt to detect these heart diseases at an early stage to avoid disastrous consequences.

Records of large sets of medical data created by medical experts are available for analyzing and extracting valuable knowledge from it. Mostly the medical database consists of discrete information. Hence, decision making using discrete data becomes a complex and tough task. Machine Learning which is a subfield of data mining handles large scale well-formatted dataset efficiently. In the medical field, machine learning can be used for diagnosis, detection and prediction of various diseases. The main goal of this paper is to provide a tool for doctors to detect heart disease at an early stage. This in turn will help to provide effective treatment to patients and avoid severe consequences. Machine Learning plays a very important role to detect the hidden discrete patterns and thereby analyze the given data. After analysis of data ML techniques help in heart disease prediction and early diagnosis. This paper presents performance analysis of various Machine Learning techniques such as K-Nearest Neighbors, Decision Tree, Logistic Regression and Support Vector Machine for predicting risk of heart disease at an early stage.

**3. Cardiovascular Disease Dataset**

The dataset is from an ongoing cardiovascular study on residents of the town of Framingham, Massachusetts. The classification goal is to predict whether the patient has a 10-year risk of future coronary heart disease (CHD). The dataset provides the patients’ information. It includes over 3390 records and 17 attributes.

Variables- Each attribute is a potential risk factor. There are demographic, behavioral, and medical risk factors.

Data Description Demographic:

• Sex: male or female ("M" or "F")

• Age: Age of the patient ;(Continuous - Although the recorded ages have been truncated to whole numbers, the concept of age is continuous)

Behavioral

• is\_smoking: whether or not the patient is a current smoker ("YES" or "NO")

• Cigs Per Day: the number of cigarettes that the person smoked on average in one day.(can be considered continuous as one can have any number of cigarettes, even half a cigarette.) Medical (history)

• BP Meds: whether or not the patient was on blood pressure medication (Nominal)

• Prevalent Stroke: whether or not the patient had previously had a stroke (Nominal)

• Prevalent Hyp: whether or not the patient was hypertensive (Nominal)

• Diabetes: whether or not the patient had diabetes (Nominal) Medical (current) • Tot Chol: total cholesterol level (Continuous)

• Sys BP: systolic blood pressure (Continuous)

• Dia BP: diastolic blood pressure (Continuous)

• BMI: Body Mass Index (Continuous)

• Heart Rate: heart rate (Continuous - In medical research, variables such as heart rate though in fact discrete, yet are considered continuous because of a large number of possible values.)

• Glucose: glucose level (Continuous) Predict variable (desired target)

• 10-year risk of coronary heart disease CHD (binary: “1”, means “Yes”, “0” means “No”) -

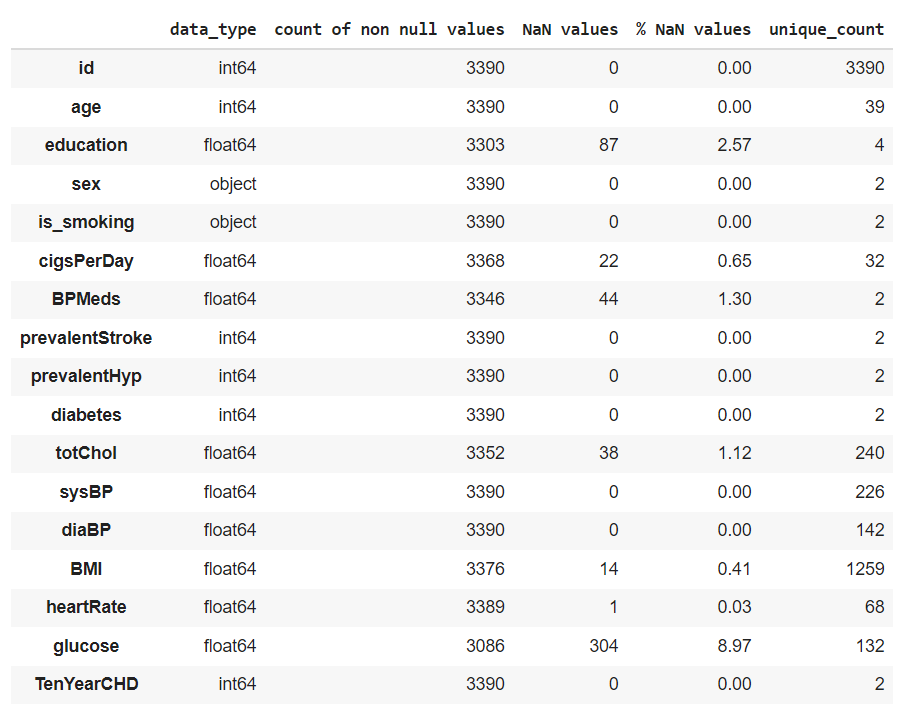
4. **Python**

Most of the info scientist use python due to the good built-in library functions and therefore the decent community. Python now has 70,000 libraries. Python is simplest programming language to select up compared to other language. That’s the most reason data scientists use python more often, for machine learning and data processing data analyst want to use some language which is straightforward to use. That’s one among the most reasons to use python. Specifically, for data scientist the foremost popular data inbuilt open source library is named panda.

Here is the first step to clean the data that will make the results “more” accurate. By finding all unique values of each row the inappropriate values can be identified. Different methods can then be used for removing them or to change those values accordingly to use them to make predictions better. As the proverb goes by saying –

*“The more data we have, the more likely we are to drown in it.” — Nassim Taleb*Not only are we interested in raw data but in the data from which valuable insights can be drawn. To do so, let us take a glimpse at another proverb.

*“More data beats clever algorithms, but better data beats more data.” — Peter Norvig*

After loading the dataset, we can start the exploration but before that, we need to check and see that the dataset is ready for performing several exploration operations or not. To know if there is any missing value or Nan value in the dataset, we can use the isnull () function.

So, we will need to prepare the dataset before performing exploratory data analysis on it.

**5. Data Preparation and Cleaning**

Data preparation is the process of cleaning and transforming raw data prior to processing and analysis. It is an important step prior to processing and often involves reformatting data, making corrections to data, and the combining of data sets to enrich data. Data cleansing or data cleaning is the process of detecting and correcting (or removing) corrupt or inaccurate records from a record set, table, or database and refers to identifying incomplete, incorrect, inaccurate, or irrelevant parts of the data and then replacing, modifying, or deleting the dirty or coarse data.

In the CardioVascular dataset it contains 12% Null or missing values. The column education, totChol, cigsperday, BPMeds, BMI and glucose contain *some* missing values and after the analysis of people having heart disease and people who are not after heart disease we found that we have imbalanced data.

* Imbalanced classification is the problem of classification when there is an unequal distribution of classes in the training dataset.
* The imbalance in the class distribution may vary, but a severe imbalance is more challenging to model and may require specialized techniques.
* Many real-world classification problems have an imbalanced class distribution, such as fraud detection, spam detection, and churn prediction.

Here, this is a problem because typically, the minority class is more important and therefore the problem is more sensitive to classification errors for the minority class than the majority class. So we checked for the NaN values which are there in the minority classes.

Out of 87 NaN values from the Education feature- 13 are from the minority classes.(i.e. 15% are from minority classes)

Out of 44 NaN values from the BPMeds feature- 7 are from the minority classes also for heart rate there is only 1 missing value.

Out of 38 NaN values from the totChol feature- 7 are from the minority classes.(i.e. 18% are from minority classes)

Out of 22 NaN values from the cigsPerday feature- 1 is from the minority classes.(i.e. 16% are from minority classes)

Out of 14 NaN values from the BMI feature- 7 are from the minority classes.(i.e. 50% are from minority classes)

Out of 304 NaN values from the Glucose feature- 39 are from the minority classes.(i.e. 13% are from minority classes)

Out of total 510 NaN values 75 are from minority classes (14% of total NaN value data are having minority classes).So these are useful information to us so we are treating the NaN values as below:

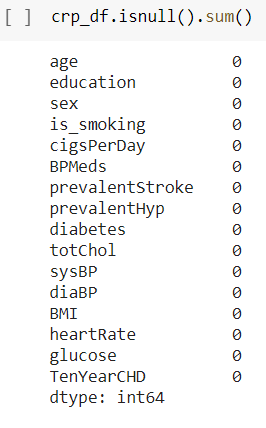
## After that we handle the missing values using KNN imputer

Knn imputer is a popular approach to missing data imputation is to use a model to predict the missing values. This requires a model to be created for each input variable that has missing values. Although any one among a range of different models can be used to predict the missing values, the k-nearest neighbor (KNN) algorithm has proven to be generally effective, often referred to as “*nearest neighbor imputation*.”

For that we identify the categorical and numerical columns to check the data distribution.

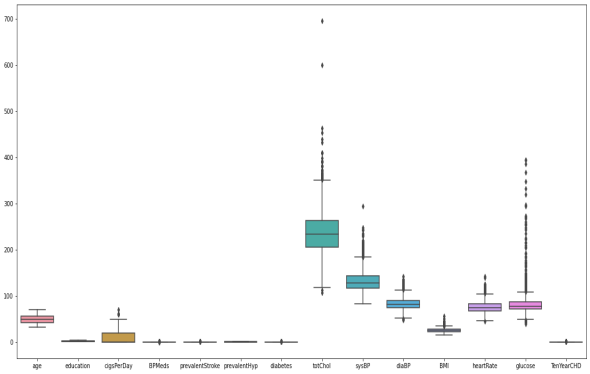
**For numeric variable:** Each missing value is imputed using average of n nearest neighbor values of that particular variable.

**For categorical variables:** Missing values for categorical features are replaced with 'Most Frequent' value for that particular variable.



So this is how we handle our missing values in the attributes.

**5.1 Outlier Detection-**



After that we checked for outliers.

As the image depicts the outliers criteria so here outliers in the features are important since we are concerned with the people who are having high heart risk also outliers are found to improve the prediction rate accurately hence we cannot remove it.

We can clearly see higher values (i.e. 600) in the Totalchol in the box plot which we are replacing with average value for the ease of analysis purpose by this way we treated the outlier in Totalchol.

Changing Categorical Column to Continuous-

* The challenge of working with categorical data when using machine learning and deep learning models.
* How to integer encode and one hot encodes categorical variables for modeling.
* How to learn and embedding distributed representation as part of a neural network for categorical variables.

## **5.2 The Changing Categorical Data into Continuous value**

A categorical variable is a variable whose values take on the value of labels but in machine learning algorithms and deep learning neural networks require that input and output variables are numbers this means that categorical data must be encoded to numbers before we can use it to fit and evaluate a model.

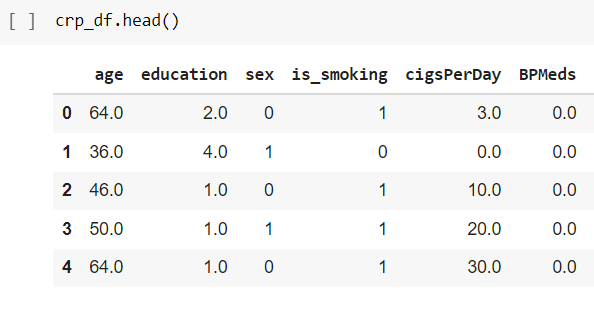
There are many ways to encode categorical variables for modeling, although the three most common are as follows:

1. **Integer Encoding**: Where each unique label is mapped to an integer.
2. **One Hot Encoding**: Where each label is mapped to a binary vector.
3. **Learned Embedding**: Where a distributed representation of the categories is learned.

Here, we use One Hot Encoding to change the categorical column to continuous because one hot encoding is useful for data that has no relationship to each other. Machine learning algorithms treat the order of numbers as an *attribute of significance*. In other words, they will read a higher number as better or more important than a lower number.

While this is helpful for some ordinal situations, some input data does not have any ranking for category values, and this can lead to issues with predictions and poor performance. That’s when one hot encoding saves the day.

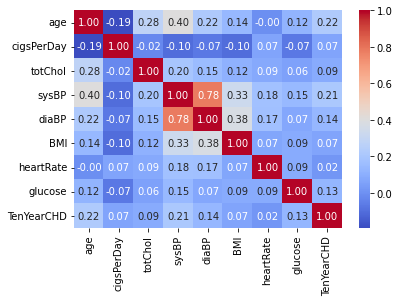
One hot encoding makes our training data more useful and expressive, and it can be rescaled easily. By using numeric values, we more easily determine a probability for our values. In particular, one hot encoding is used for our output values, since it provides more nuanced predictions than single labels.



So this is how we changed our categorical columns namely sex and is\_smoking into continous value.

**5.3 Feature Engineering**

The features in your data will directly influence the predictive models we use and the results we can achieve. We can say that: the better the features that we prepare and choose, the better the results we will achieve. It is true, but it is also misleading. The results we achieve are a factor of the model we choose, the data we have available and the features we prepared. Even framing of the problem and objective measures we’re using to estimate accuracy play a part. We need great features that describe the structures inherent in our data.



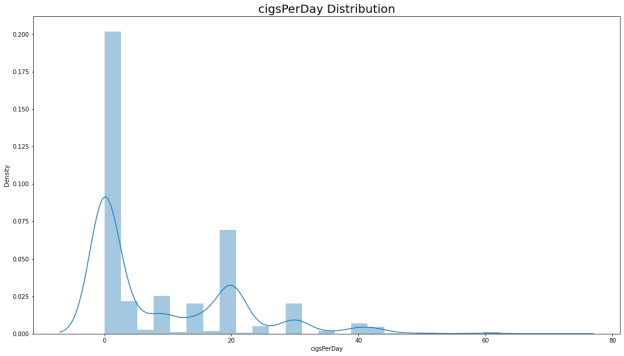
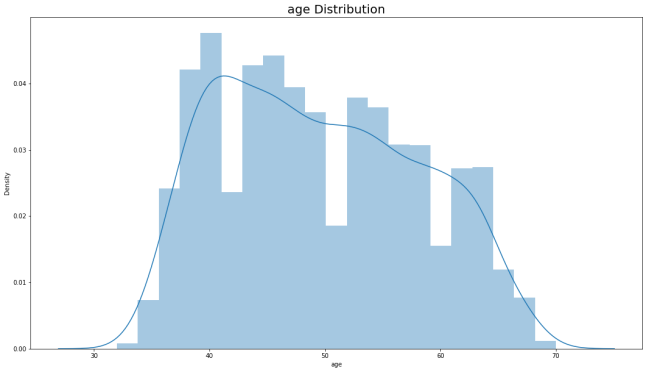
After observing this correlation heatmap we can perform feature engineering with two attribute sysBP and diaBp as they are highly correlated

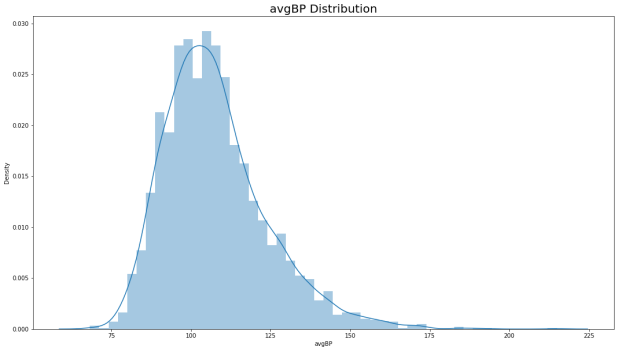
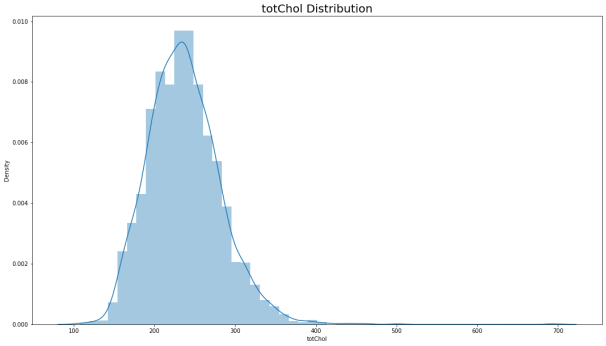
diaBP is Diastolic Pressure sysBP is Systolic Pressure, generally while measuring our BP we look for there Ratio e.g a healthy person has 120/80

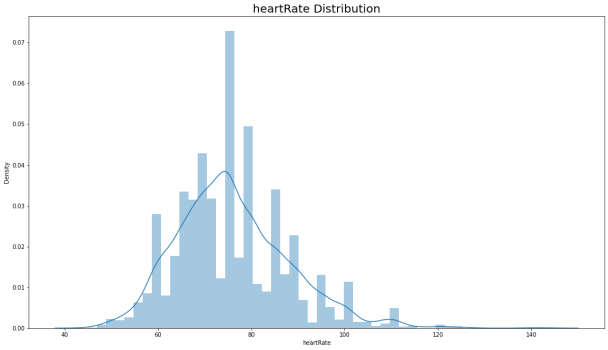
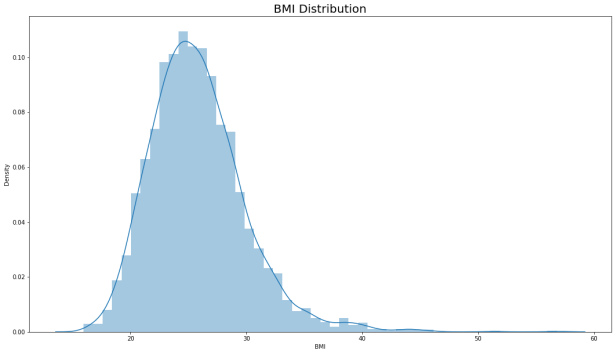
So we derived a new feature **AvgBP = sysBP+diaBP/2** after that we simply dropped diaBP and sysBP.

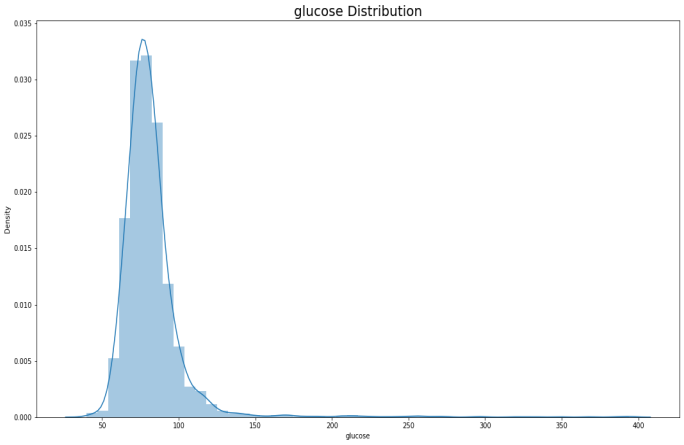
**5.4 Feature Distribution and Selection**

The feature distribution helps in understanding what kind of feature we are dealing with, and what values you can expect this feature to have. We’ll see if the values are centered or scattered. This distribution is important because models learn from data give them incorrect data and they will learn incorrectly. A model is only as good as the data we feed it. If a feature can be seen as a random variable, and enough data is used and the bins are narrow enough, the look of the distribution could be bell-shaped. This is called a [normal distribution](https://en.wikipedia.org/wiki/Normal_distribution) (or standard/Gaussian). The normal distribution is a very important probability distribution that arises in many situations. Generally speaking, when you have a large number of independent samples from a naturally occurring phenomenon, the data will follow a normal distribution.









Among the numerical continuous features:

totChol, avgBP,BMI and heartRate has an uniform distribution.

Glucose shows quite a bit and slight right skewness respectively but since we are concerned about people with cardiovascular issues we consider the right skewed values as an important information so we left it as it is.

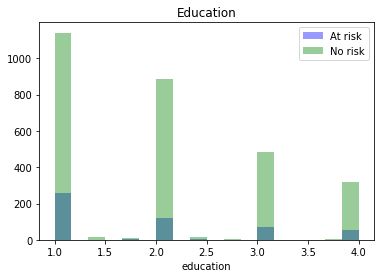
cigsPerDay is not following Gaussian distribution and from the dist plot as well as the kde plot we did not get any inference with respect to target variable.

Among the Categorical(discrete features):

BPMeds,Prevalentstroke,diabetes has relatively very low variance,thus unable to come up with any generalizable conclusion about the correlation between their history and prevalence of cardiovascular risk.

**6. Exploratory Data Analysis**

Analysis the Education feature:



There is no description provided for the values in 'Education', considering them as level of education. '1' being the lowest level of education and '4' being the highest level of education.

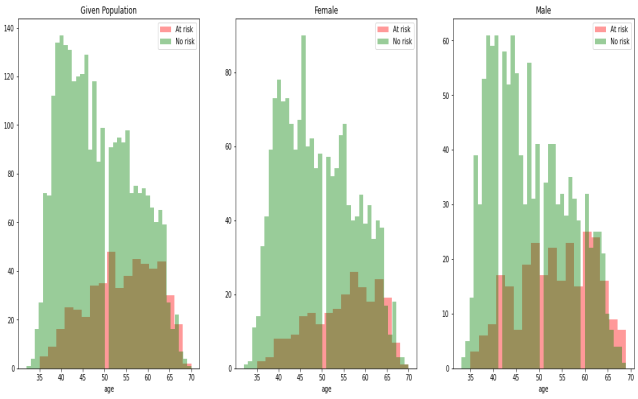
Not much we can interpret from this feature. but according to the data given- we can say people with lower education level are at higher risk of getting heart disease.

**6.1 MultiVariate Analysis**



As we can see from the correlation graph above- the features of low variance- i.e., BPMeds,Prevalentstroke,diabetes has no direct or good linear relationship with the target variable(TenYearCHD).

**6.2 BiVariate Analysis between “Age” and “Sex”**

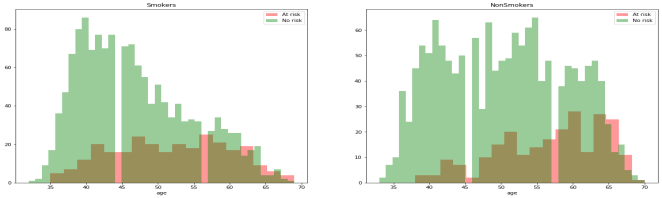


INSIGHTS: From the above graphs we can say that:

Both Women and Men lying in the 50-52 have a high risk of heart disease.

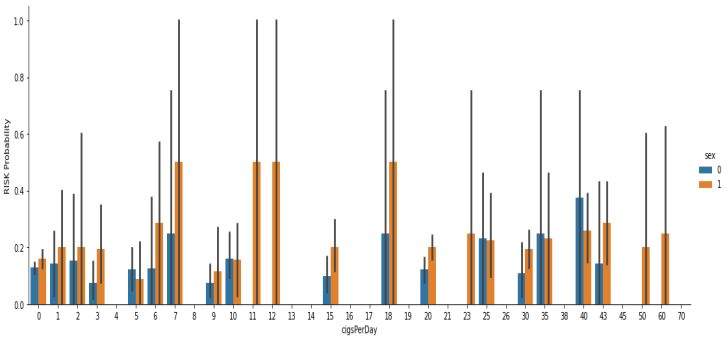
Men lying in the age group 40-42 are at risk.

## **6.3 “Age” and “Smoking”**



We see Risk is High in the age group of (42 to 45),(56,58) and above age 62 despite whether they are Smokers or not.

## **6.4 *“*CigsPerDay”, “Sex” & “RISK”**



From this factor plot there is no exact Inference to say that - less the cigarette consumption means less heart disease and vice versa.

But Male, more than females, are at a higher risk of Heart disease, since the proportion of male smokers v/s females are high and in reality smokers will be having high risk of getting heart disease.

**6.5 Summary of Insights from EDA**

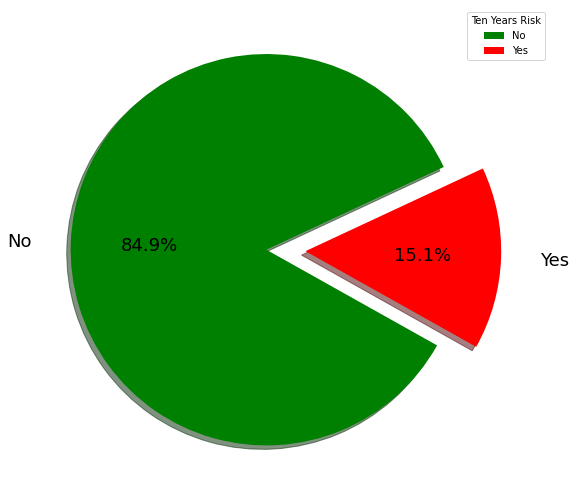
Due to the imbalanced nature of the data set it was difficult to make conclusions but based on what is observed but these are the conclusions that could be drawn:

Slightly more males are suffering from CHD than females. The percentage of people who have CHD is almost equal between smokers and non-smokers.

The percentage of people who have CHD is higher among the diabetic.

Distribution of the ages of the people who had CHD and the number of the sick generally increased with age with the peak being at 63 years old.

**7. Dealing with Imbalanced Dataset**

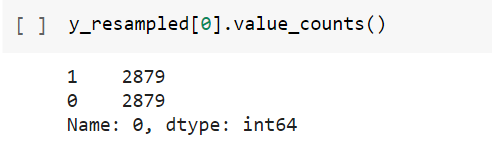
****

From this pie chart we can clearly see that our data is an imbalance target feature- TenYearCHD, the risk rate is relatively rare, only 15% of the people have this disease hence even if our model gives 94% accuracy! It is misleading. All those non-risk cases, we’d have 100% accuracy. Those cases in which a person has a risk, we’d have 0% accuracy classification predictive modeling involves predicting a class label for a given observation.

An imbalanced classification problem is an example of a classification problem where the distribution of examples across the known classes is biased or skewed. The distribution can vary from a slight bias to a severe imbalance where there is one example in the minority class for hundreds, thousands, or millions of examples in the majority class or classes.

Imbalanced classifications pose a challenge for predictive modeling as most of the machine learning algorithms used for classification was designed around the assumption of an equal number of examples for each class. This results in models that have poor predictive performance, specifically for the minority class. This is a problem because typically, the minority class is more important and therefore the problem is more sensitive to classification errors for the minority class than the majority class

There are multiple ways in which the imbalanced dataset can be handled, 'resampling the dataset is one of them. Two ways of resampling are: Undersampling the majority class and oversampling the minority class. In order to reduce overfitting during oversampling, synthetic sampling is used. SMOTE (Synthetic Minority Over Sampling Technique) is one popular algorithm. One approach to addressing imbalanced datasets is to oversample the minority class. The simplest approach involves duplicating examples in the minority class, although these examples don’t add any new information to the model. Instead, new examples can be synthesized from the existing examples. This is a type of [data augmentation](https://machinelearningmastery.com/how-to-configure-image-data-augmentation-when-training-deep-learning-neural-networks/) for the minority class and is referred to as the **Synthetic Minority Oversampling Technique**, or **SMOTE** for short. Instead of using copies of observations to oversample, SMOTE varies attributes of the observations to create new synthetic samples.



This is how our data is balanced and we did not use undersampling because it can cause loss of information and poor generalization to our test set.

**8. Machine Learning**

Machine Learning AI is an upward innovation which permits PCs to gain naturally from past information. AI customs different calculations for building numerical models and making forecasts utilizing old information or data. As of now, it is being utilized for different errands, for example, picture acknowledgment, discourse acknowledgment, email sifting, Facebook auto-labeling, recommender framework, and some more. AI is asserted as a development of man-made reasoning that is to a great extent worried about the development of calculations which grant a PC to gain from the information and past proficiencies all alone. The term AI was first introduced by Arthur Samuel in 1959. We will characterize it in a summed up manner as: "AI empowers a machine to consequently gain from information, improve show from encounters, and anticipate things without being unequivocally modified". A Machine Learning framework gains from authentic information, fabricates the forecast models, and at whatever point it gets new information, predicts the yield for it. The precision of anticipated yield relies on the measure of information, as the huge measure of information assists with building a superior model which predicts the yield all the more precisely. Assume we have a diverse issue, where we need to accomplish a few forecasts, so as opposed to composing a code for it, we simply need to provide information to generic algorithms, with the assistance for these calculations, machine assembles rationale according to information, foresee the yield. AI has changed our perspective about this issue.

**Classification of Machine Learning**

Supervised Learning

Supervised learning is regularly characterized as learning with correct regulator; else you can say that learning inside the presence of educators. The algorithm learns on a labeled dataset with an answer key and does the training and evaluation. Administered learning is anticipated on "train me" idea. Supervised learning has next measures:

• Classification

• Random forest

• Decision tree

• Regression

There are following machine AI algorithms:

• Linear Regression

• Logistical Regression

• Support Vector Machines (SVM)

• Neural Networks

• Random Forest

• Gradient Boosted Trees

• Decision Trees

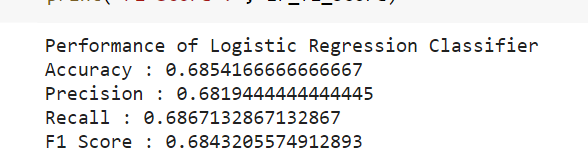
• Naive Bayes

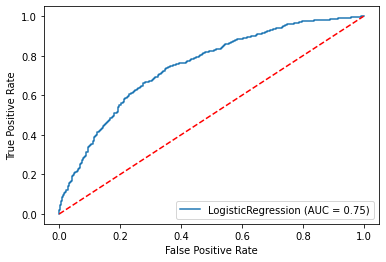
From the above algorithm we trained our model using:

1. Logistic Regression
2. K Nearest Neighbour
3. Decision Tree
4. Support Vector Machine

Train and Test the Model- This stage is to assess the models and upheld the info record. For our motivation of study, we are having the chance to mentor the model utilizing different algorithms.

**8.1 Logistic Regression**-

* Logistic regression is a statistical method for predicting binary classes. The outcome or target variable is dichotomous in nature. Dichotomous means there are only two possible classes. For example, it can be used for cancer detection problems. It computes the probability of an event occurring. It is a special case of linear regression where the target variable is categorical in nature. It uses a log of odds as the dependent variable. Logistic Regression predicts the probability of occurrence of a binary event utilizing a logit function.
* 

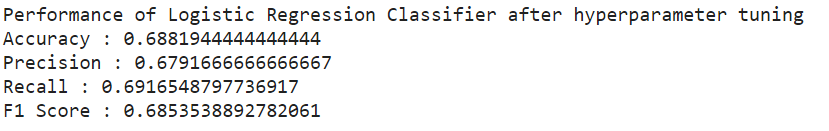


**8.2 Hyperparameter Tuning**

Whenever a machine learning algorithm is implemented on a specific dataset, the performance is judged based on how well it generalizes i.e. how it reacts to new, never-before-seen data. In case the performance of the learning algorithm is not satisfactory or there is room for improvement, certain parameters in the algorithm need to be changed/tuned/tweaked. These parameters are known as ‘hyperparameters’ and the process of varying these hyperparameters to better the learning algorithm’s performance is known as ‘hyperparameter tuning’.

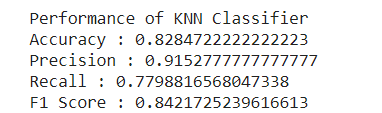
These hyperparameters are not learnt directly through the training of algorithms. These values are fixed before the training of the data begins. They deal with parameters such as learning\_rate, i.e how quickly the model should be able to learn, how complicated the model is, and so on. There can be a wide variety of hyperparameters for every learning algorithm. Selecting the right set of hyperparameters so as to gain good performance is an important aspect of machine learning.

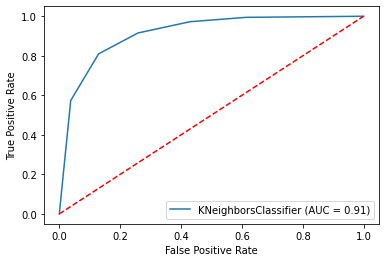
So after hyper parameter tuning the result does not vary for logistic regression much.

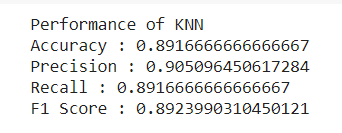


**8.3 K Nearest Neighbor**

K Nearest Neighbor(KNN) is a very simple, easy to understand, versatile and one of the topmost machine learning algorithms. KNN is used in a variety of applications such as finance, healthcare, political science, handwriting detection, image recognition and video recognition. In Credit ratings, financial institutes will predict the credit rating of customers. In loan disbursement, banking institutes will predict whether the loan is safe or risky. In political science, classifying potential voters in two classes will vote or won’t vote. KNN algorithm used for both classification and regression problems. KNN algorithm based on feature similarity approach.



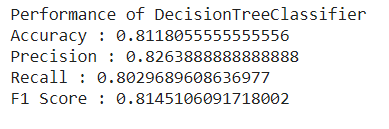


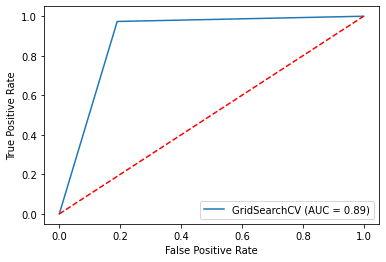


**8.4 Decision Tree**

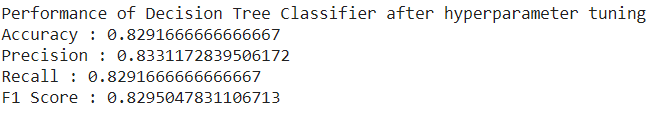
A decision tree is a flowchart-like structure in which each internal node represents a test on a feature (e.g. whether a coin flip comes up heads or tails) , each leaf node represents a class label (decision taken after computing all features) and branches represent conjunctions of features that lead to those class labels. The paths from root to leaf represent classification rules. Decision trees are constructed via an algorithmic approach that identifies ways to split a data set based on different conditions. It is one of the most widely used and practical methods for supervised learning. Decision Trees are a [non-parametric](https://machinelearningmastery.com/parametric-and-nonparametric-machine-learning-algorithms/) supervised learning method used for both classification and regression tasks.

Tree models where the target variable can take a discrete set of values are called classification trees. Decision trees where the target variable can take continuous values (typically real numbers) are called regression trees. Classification And Regression Tree (CART) is the general term for this.



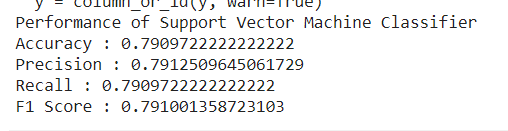


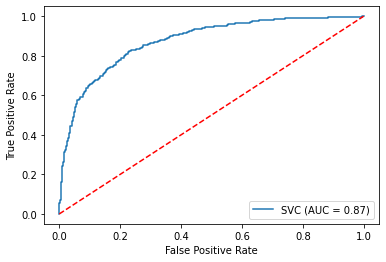
After Hyper Tuning the result is-

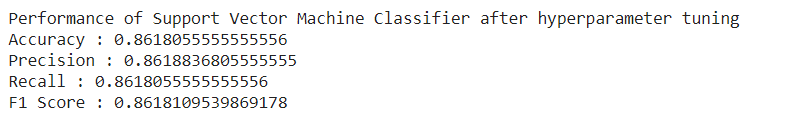


**8.5 Support Vector Machine**

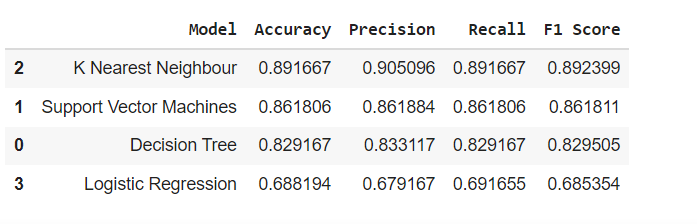
“Support Vector Machine” (SVM) is a supervised [machine learning algorithm](https://courses.analyticsvidhya.com/courses/introduction-to-data-science-2?utm_source=blog&utm_medium=understandingsupportvectormachinearticle) that can be used for both classification and regression challenges. However, it is mostly used in classification problems. In the SVM algorithm, we plot each data item as a point in n-dimensional space (where n is a number of features you have) with the value of each feature being the value of a particular coordinate. Then, we perform classification by finding the hyper-plane that differentiates the two classes very well.







Comparing Performance of all classifiers



Among all the Classifiers used here:

The K Nearest Neighbor classifier has the best performance with highest RECALL and F1 SCORE of 89%.

**Future Scope-**

With the Interest of time we kept certain things for the future work which is as follows:

1. Since the data given to us has fewer records we used an oversampling technique. But In a reality Health data will be imbalanced and will be huge so we can try to work on cost sensitive learning which instead of altering the data records will just give more weightage to the minority and focus on the heart risky patients.
2. Feature Importance with respect to each classifier.

**CONCLUSION**

With the increasing number of deaths due to heart diseases, it has become mandatory to develop a system to predict heart diseases effectively and accurately. The motivation for the study was to find the most efficient ML algorithm for detection of heart diseases. This study compares the accuracy score of Decision Tree, Logistic Regression, Support Vector Machine and K Nearest Neighbors algorithms for predicting heart disease using a given dataset. The result of this study indicates that the K nearest Neighbor algorithm is the most efficient algorithm with an accuracy score of 89% for prediction of heart disease. In future the work since the data given to us has less records we used an oversampling technique but In a reality health data will be imbalanced and will be huge so we can try to work on cost sensitive learning which instead of altering the data records will just give more weightage to the minority and focus on the heart risky patients and feature Importance with respect to each classifier also we can enhanced by developing a web application based on the K Nearest Neighbor algorithm as well as using a larger dataset as compared to the one used in this analysis which will help to provide better results and help health professionals in predicting heart disease effectively and efficiently.

**REFERENCE-**

<https://www.python.org/>

<https://pandas.pydata.org/>

<https://matplotlib.org/>

<https://numpy.org/>

<https://www.w3schools.com/python/pandas/default.asp>

1. Avinash Golande, Pavan Kumar T, Heart Disease Prediction Using Effective Machine Learning Techniques, International Journal of Recent Technology and Engineering, Vol 8, pp.944-950,2019.
2. T.Nagamani, S.Logeswari, B.Gomathy, Heart Disease Prediction using Data Mining with Mapreduce Algorithm, International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-8 Issue-3, January 2019.
3. Fahd Saleh Alotaibi, Implementation of Machine Learning Model to Predict Heart Failure Disease, (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 10, No. 6, 2019.
4. Anjan Nikhil Repaka, Sai Deepak Ravikanti, Ramya G Franklin, Design And Implementation Heart Disease Prediction Using Naives Bayesian, International Conference on Trends in Electronics and Information(ICOEI 2019).
5. Theresa Princy R,J. Thomas,Human heart Disease Prediction System using Data Mining Techniques, International Conference on Circuit Power and Computing Technologies,Bangalore,2016.
6. Nagaraj M Lutimath,Chethan C,Basavaraj S Pol.,Prediction Of Heart Disease using Machine Learning, International journal Of Recent Technology and Engineering,8,(2S10), pp 474-477, 2019.
7. Sayali Ambekar, Rashmi Phalnikar,Disease Risk Prediction by Using Convolutional Neural Network,2018 Fourth International Conference on Computing Communication Control and Automation.
8. C. B. Rjeily, G. Badr, E. Hassani, A. H., and E. Andres, Medical Data Mining for Heart Diseases and the Future of Sequential Mining in the Medical Field, in Machine Learning Paradigms, 2019, pp. 7199.
9. Jafar Alzubi, Anand Nayyar, Akshi Kumar. "Machine Learning from Theory to Algorithms: An Overview", Journal of Physics: Conference Series, 2018
10. Fajr Ibrahem Alarsan., and Mamoon Younes Analysis and classification of heart diseases using heartbeat features and machine learning algorithms,Journal Of Big.