Statistical and AI Techniques in Data Mining

Subject Code: MTH552A
Semester II

Project

"Heart Disease Prediction using KNN, Decision Tree and Logistic regression"

Arghadeep Basu – 211274
Sahil Mallick-211365
Sankhadeep Mitra – 211369

Under the Supervision
Of
Dr. Amit Mitra

Department of Mathematics and Statistics
Indian Institute of Technology, Kanpur

Abstract

This data is about prediction of heart disease among patients based on their age, physical and mental health, past record of diseases and unhealthy habits like smoking. The dataset contains 17 explanatory variables and 1 target variable which is binary, i.e., it tells whether the person may be a victim of heart disease or not. After preliminary exploratory works on the data like plotting the distribution of the different variables, we perform feature scaling and split the data on train and test datasets and fit different classification models like K-Nearest-Neighbour Method, Decision Tree and Logistic Regression. Then we test the different fitted models on the test dataset and hence analyse and compare their performances.

ACKNOWLEDGEMENT

Our journey of accomplishing the project really involves many ones to whom we are highly obliged. We would like to express our heartfelt gratitude to Dr. Amit Mitra, Department of Mathematics and Statistics, IIT Kanpur, for assigning this project to us.

It has been a great learning experience and has also provided us with a practical insight of the theoretical knowledge gathered during the course.

We would also like to thank our seniors for their extensive support throughout the session. Their constant encouragement has enabled us to complete the project within the stipulated time period.

Context

Heart diseases are one of the most common and dangerous diseases among the present-day world population. It is on the rise, claiming millions of lives every year and is not showing any signs to decrease anytime soon. The key factors for heart diseases are considered to be high blood pressure, diabetes, high cholesterol, smoking and drinking habits and several other similar factors.

Hence it becomes important to diagnose a patient's condition and and hence be able to predict beforehand whether he/she can be a probable heart disease patient or not. In this way it will become easy for the doctors and the hospitals and they can start the medication and treatment of the patient before his/her case gets worse.

Statistical data mining and machine learning techniques can play an important role in the prediction of heart disease in the way that one can look into past records of patients who had heart disease and hence build statistical models which are able to learn the pattern of possibility of heart disease in those patients. Afterwards, one can test those models on new datasets and try to predict the possibility of disease in them too.

In this project we aim to use models like K-Nearest-Neighbours method, Decision tree and Logistic Regression and hence compare them based on their performances.

Dataset Description

Originally, the dataset come from the CDC and is a major part of the Behavioural Risk Factor Surveillance System (BRFSS), which conducts annual telephone surveys to gather data on the health status of U.S. residents. As the CDC describes: "Established in 1984 with 15 states, BRFSS now collects data in all 50 states as well as the District of Columbia and three U.S. territories. BRFSS completes more than 400,000 adult interviews each year, making it the largest continuously conducted health survey system in the world." In this dataset, there are different factors (questions) that directly or indirectly influence heart disease. The data has 319795 data points and 19 variables.

Variable description

Name of Variable	Description
HeartDisease	Respondents that have ever reported having coronary heart disease (CHD) or myocardial infarction (MI).(yes /no)
BMI	Body Mass Index (BMI)
Smoking	Have you smoked at least 100 cigarettes in your entire life? (The answer Yes or No.
AlcoholDrinking	Heavy drinkers (adult men having more than 14 drinks per week and adult women having more than 7 drinks per week) (yes/no)
Stroke	Whether the person have had a stroke before(yes/no)
PhysicalHealth	Physical health includes physical illness and injury. For how many days during the past 30 days was the respondent's physical health not good? (0-30 days).
MentalHealth	For how many days during the past 30 days was the respondent's mental health not good? (0-30 days)
DiffWalking	Does the respondent have serious difficulty walking or climbing stairs? (yes/no)

Sex	Respondent male or female
AgeCategory	Fourteen-level age category.
Race	Imputed race/ethnicity value
Diabetic	Whether the respondent has ever been diagnosed with Diabetes?(yes/no)
PhysicalActivity	Adults who reported doing physical activity or exercise during the past 30 days other than their regular job.(yes/no)
GenHealth	Would you say that in general your health is? (5 categories, 'poor', 'fair', 'good', 'very good', 'excellent')
SleepTime	On average, how many hours of sleep do the respondent get in a 24-hour period?
Asthma	Whether the respondent has ever been diagnosed with asthma?(yes/no)
KidneyDisease	Not including kidney stones, bladder infection or incontinence, was the respondent ever diagnosed with kidney disease?(yes/no)
SkinCancer	Whether the respondent has ever been diagnosed with skin cancer?(yes/no)

Objectives

1. Pre-process the data and take care of missing values (if any)
2. Perform exploratory data analysis on the data and draw the distributional plots corresponding to the different variables
3. Perform feature scaling
4. Split the data into train and test datasets and fit
   a. K-Nearest-Neighbours model
   b. Decision Tree model
   c. Logistic Regression model
5. Analyse and compare the performance of the models on the test datasets

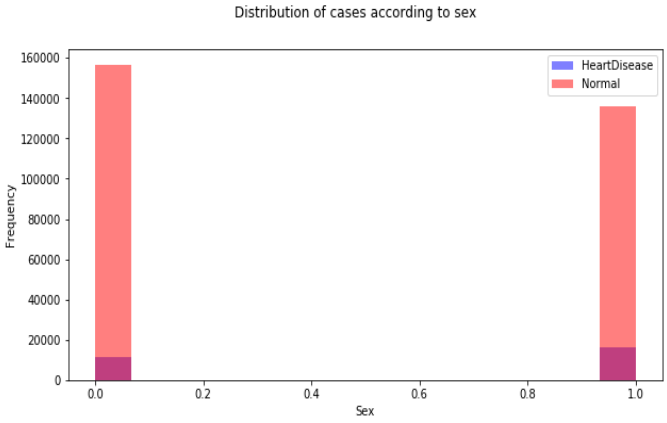
Methodology

The sample data looks like

	HeartDisease	BMI	Smoking	AlcoholDrinking	Stroke	PhysicalHealth	MentalHealth	DiffWalking	Sex	AgeCategory	Race	Diabetic	PhysicalActivity	GenHealth	SleepTime	Asthma	KidneyDisease	SkinCancer
0	No	16.60	Yes	No	No	3	30	No	Female	55-59	White	Yes	Yes	Very good	5	Yes	No	Yes
1	No	20.34	No	No	Yes	0	0	No	Female	80 or older	White	No	Yes	Very good	7	No	No	No
2	No	26.58	Yes	No	No	20	30	No	Male	65-69	White	Yes	Yes	Fair	8	Yes	No	No
3	No	24.21	No	No	No	0	0	No	Female	75-79	White	No	No	Good	6	No	No	Yes
4	No	23.71	No	No	No	28	0	Yes	Female	40-44	White	No	Yes	Very good	8	No	No	No
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
319790	Yes	27.41	Yes	No	No	7	0	Yes	Male	60-64	Hispanic	Yes	No	Fair	6	Yes	No	No
319791	No	29.84	Yes	No	No	0	0	No	Male	35-39	Hispanic	No	Yes	Very good	5	Yes	No	No
319792	No	24.24	No	No	No	0	0	No	Female	45-49	Hispanic	No	Yes	Good	6	No	No	No
319793	No	32.81	No	No	No	0	0	No	Female	25-29	Hispanic	No	No	Good	12	No	No	No
319794	No	46.56	No	No	No	0	0	No	Female	80 or older	Hispanic	No	Yes	Good	8	No	No	No

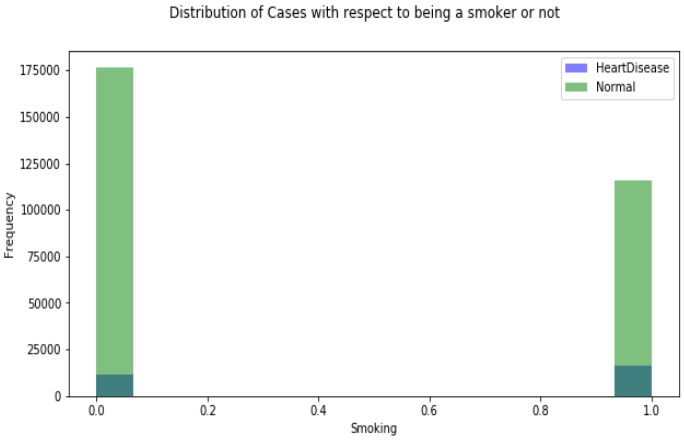
Exploratory Data Analysis

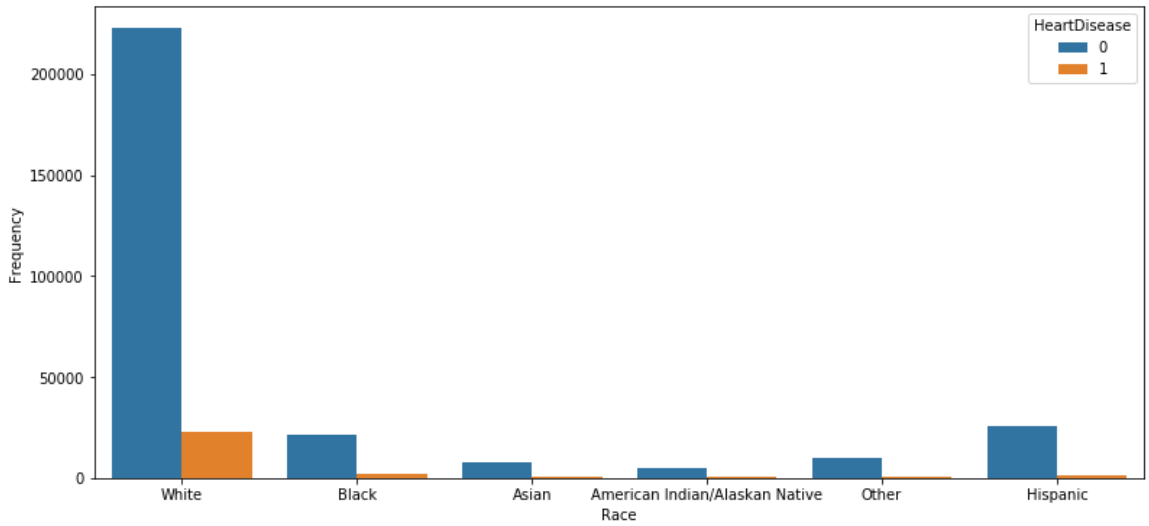
Here we study the distributions of study variables according to the presence of heart disease.



In this figure distribution of cases of heart disease according to sex (Male='1', Female='0'). This diagram tells us that proportion of males affected by heart disease is more than that of females.

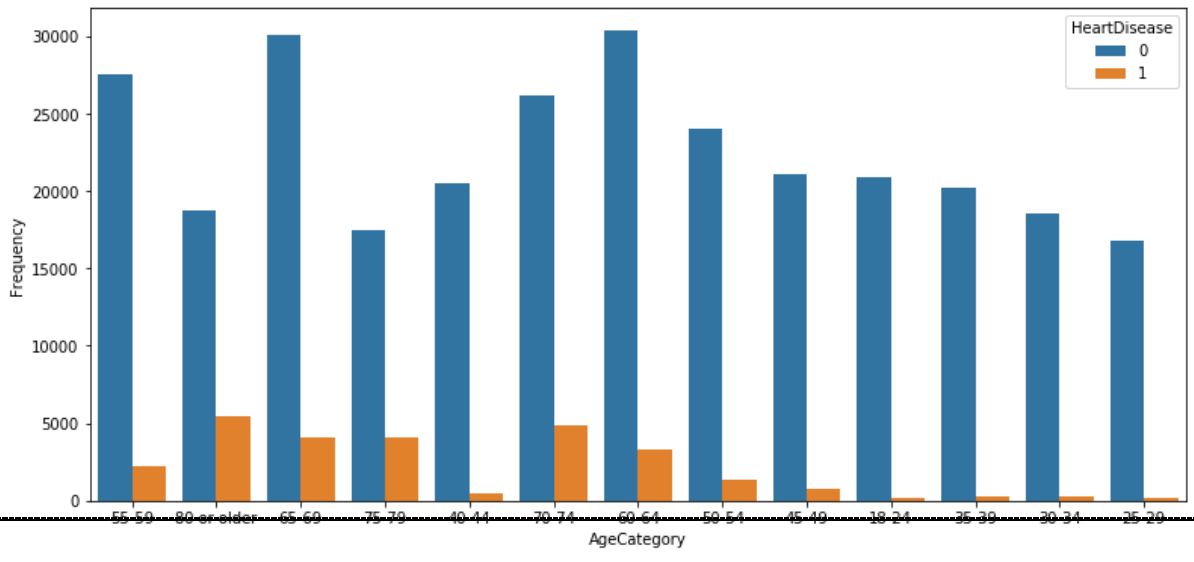
This figure shows the distribution of heart disease cases with respect to being a smoker or not. It is natural that a smoker is more vulnerable to heart disease than a non-smoker, which the figure depicts.



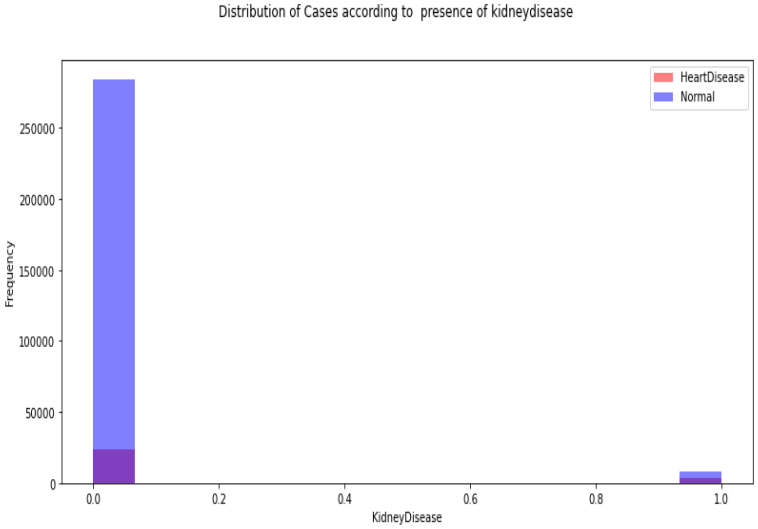


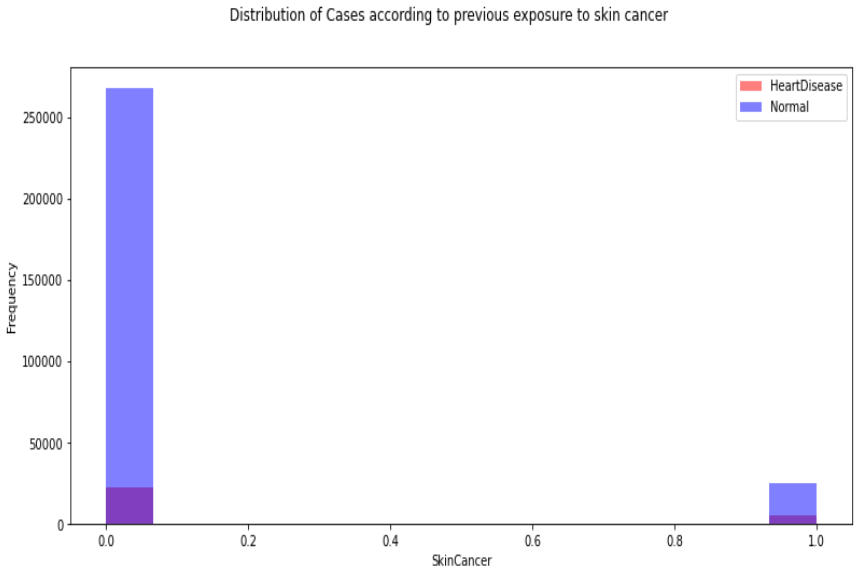
The above diagram shows how the cases of heart disease are distributed among the human races. As our data is US based, count of people suffering from heart disease is maximum for white people.

The following diagram plots the cases of heart disease against the age category. A first look at the plot tells that the elderly people (more than 70 years of age) are more prone to heart disease, followed by the age group 60-70 years.

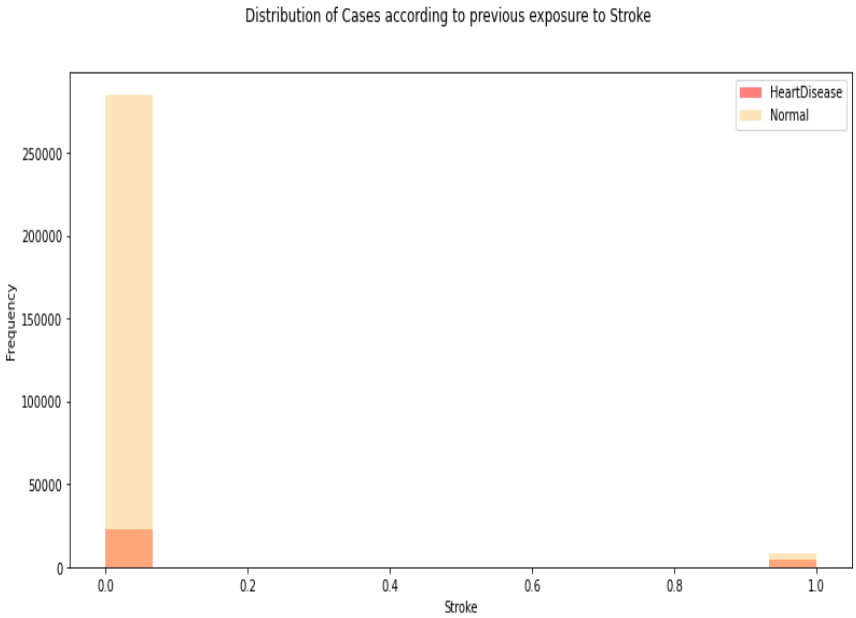


The diagram shows the distribution of cases according to presence of kidney disease. Heart disease is more common to the patients who are diagnosed with kidney disease.



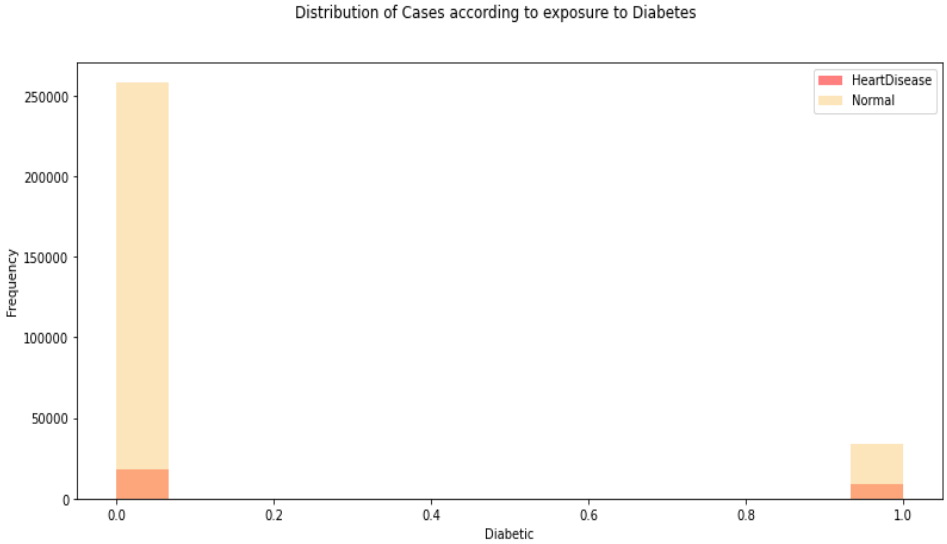


The diagram shows the distribution of cases of heart disease w.r.t. previous exposure to skin cancer. Previous exposure to skin cancer does not signify the presence of heart disease.

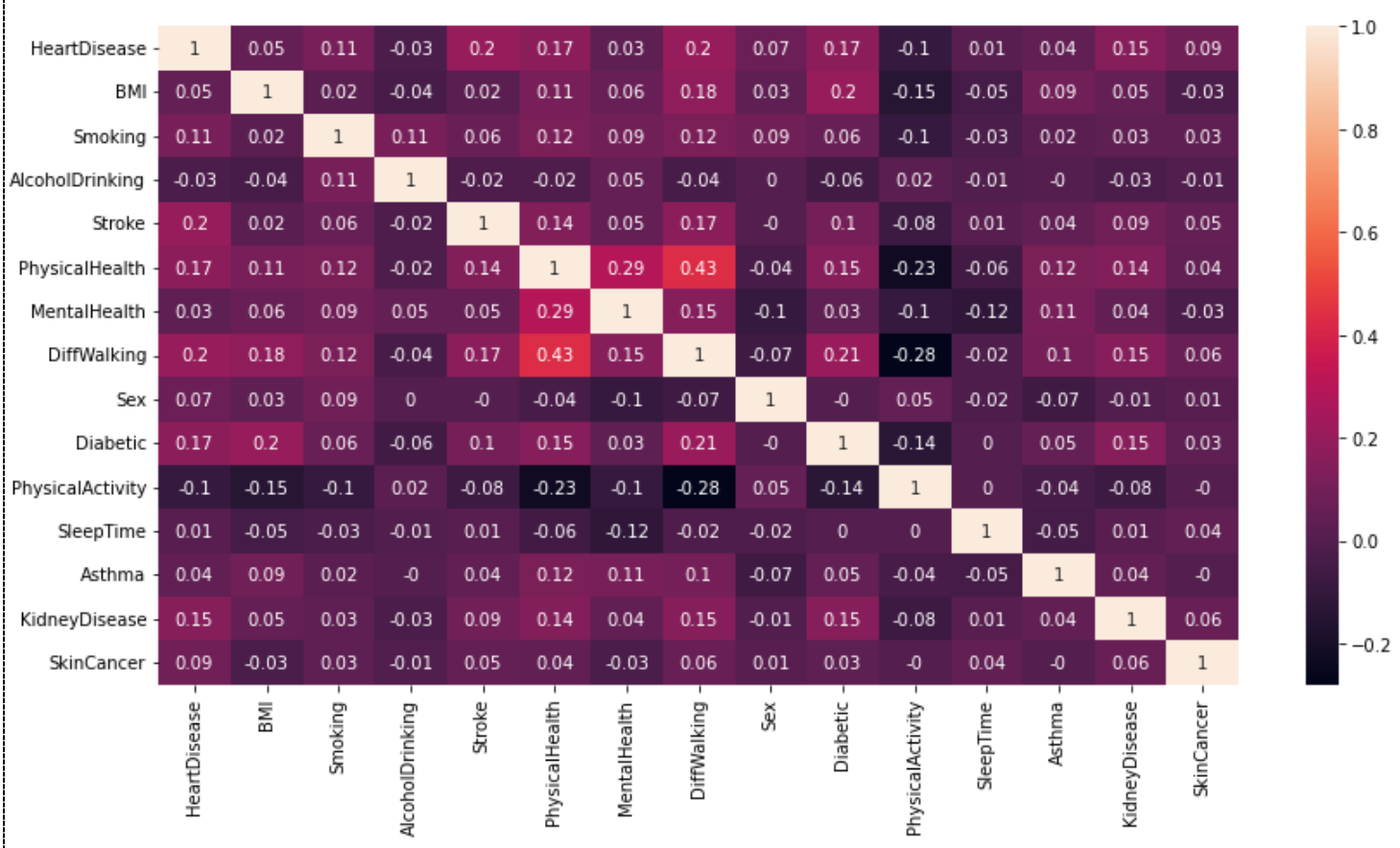


The diagram shows that a person having previous exposure to stroke has higher chance of heart disease.

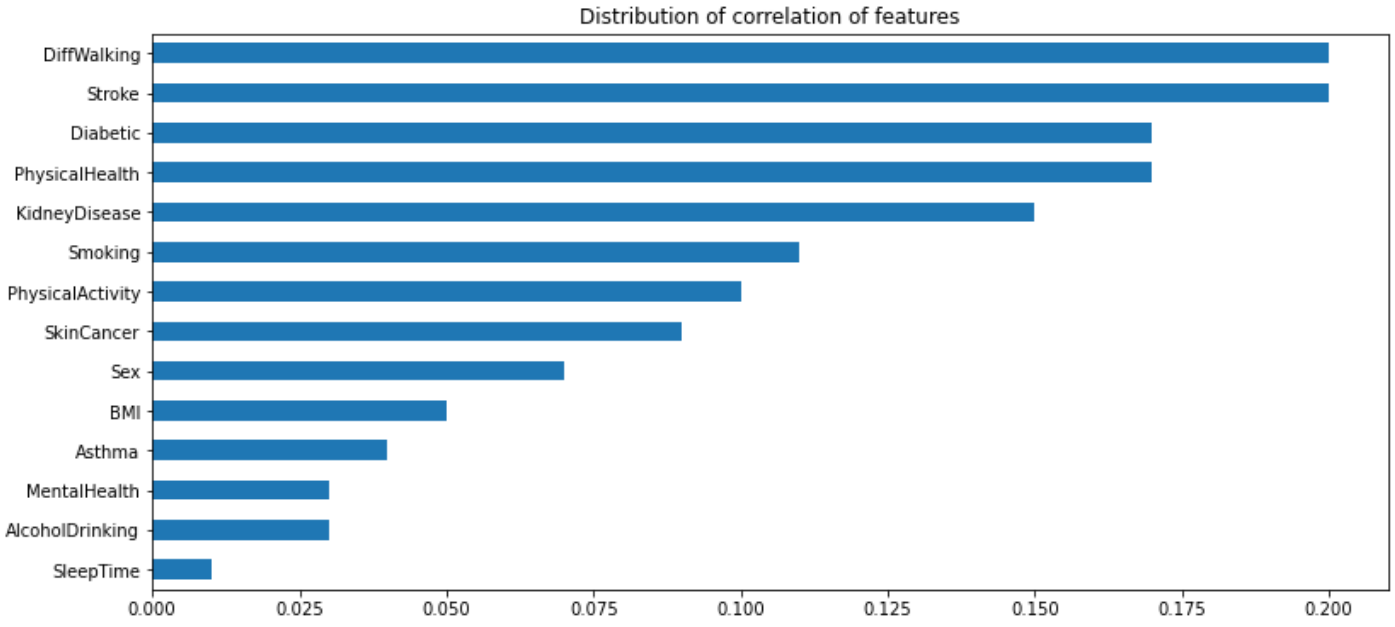
This diagram shows that diabetes patients are more prone to have heart disease.



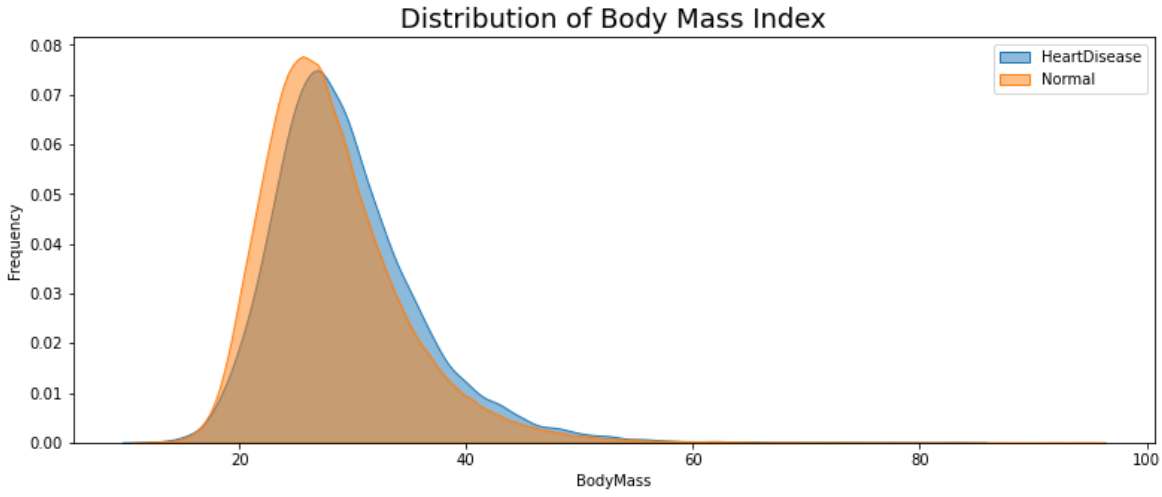
Correlation Heatmap



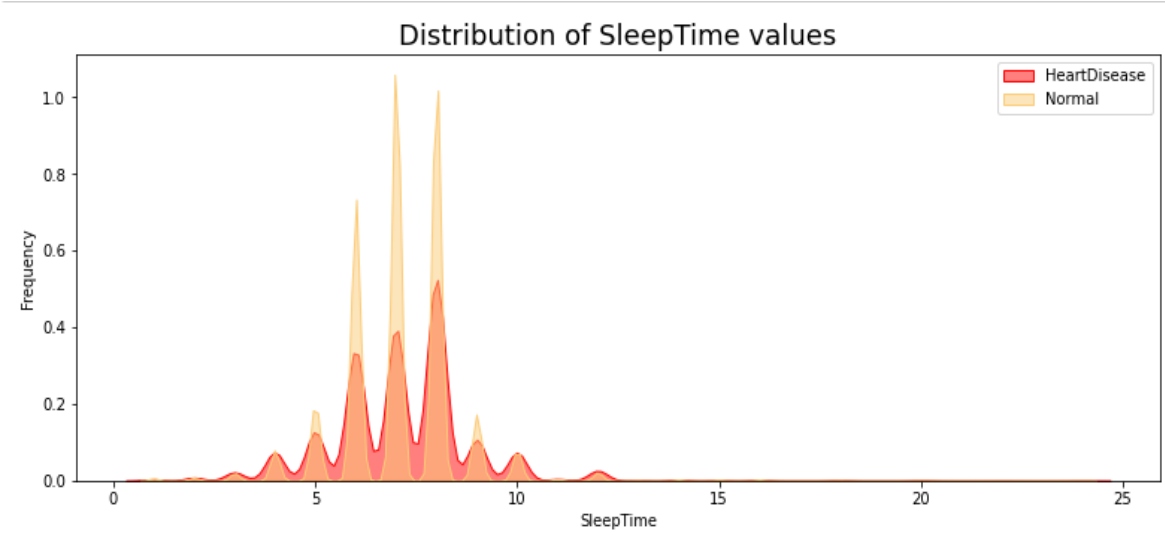
Form the correlation heat map; it is evident that there is no significant association between any of the variables.



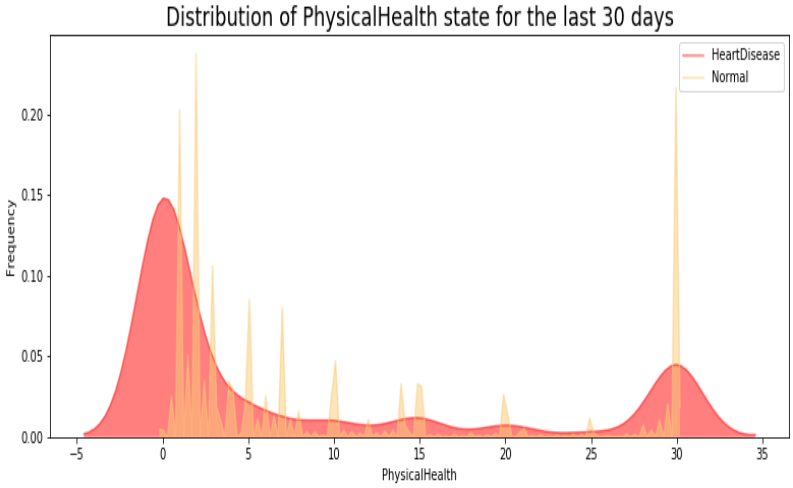
The above diagram depicts the order of association of different explanatory variables with the response variable. It is observed that persons having difficulty in walking (*Diffwalking*) and ever been diagnosed with stroke and/or diabetes have highest chance/occurance of heart disease.



The above diagram shows the distribution of body mass index of heart disease patients and non-heart disease patients. It is observed that patients having heart disease have higher mean BMI.

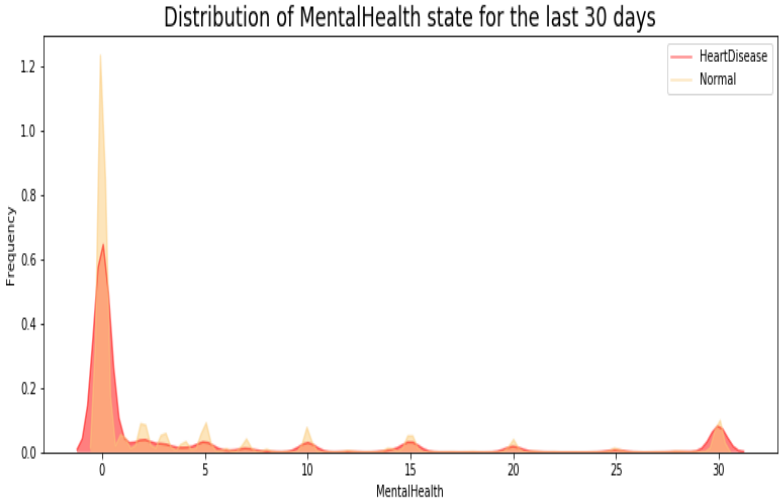


From the above figure , we see the distributions of sleep-times of heart disease patients and non-heart disease patients. Patients having sleep-times that are too low or high have higher chance/occurance of heart disease.



Here, we can see the distributions of the number of days in the previous month both heart disease and non-heart disease patient had reported bad physical health .

Here, we can see the distributions of the number of days in the previous month both heart disease and non-heart disease patient had reported bad mental health .



Feature Scaling

Now we perform feature scaling during the data pre-processing to handle highly varying magnitudes or values or units. If feature scaling is not done, then any classification or machine learning algorithm tends to weight greater values higher and consider smaller values as the lower values, regardless of unit of the values.

Here we use standardization technique on feature values so that it has mean=0 and variance =1

$$Z = \frac{X_i - X_{mean}}{Standard\ Deviation}$$

Dummy Variables

A dummy independent variable (also called a dummy explanatory variable) which for some observation has a value of 0 will cause that variable's coefficient to have no role in influencing the dependent variable, while when the dummy takes on a value 1 its coefficient acts to alter the intercept.

Let there be k categorical features in the model $(v_1, v_2, \dots, v_k)$ where $\boldsymbol{v_j\ has\ c_j\ categories}, \boldsymbol{j = 1(1)k}$

Then we have to introduce $\sum_{j=1}^{k}(c_j - 1)$ dummy variables in total.

The dummy variables for v_j are defined as:

$$z_{ik} = \begin{cases} 1, & if\ \text{the}\ i^{th}\ observation\ belongs\ to\ \text{the}\ k^{th}\ category \\ 0, & otherwise \end{cases}$$

k= 1(1) $(c_j - 1)$

Here we define dummy variables for the following categories:

a. **Race ---------- 6 categories ---------- 5 dummy variables**
b. **GenHealth---- 5 categories ---------- 4 dummy variables**
c. **AgeCategory----13 categories---------12 dummy variables**

Hence total number of dummy variables: 21

Splitting of Dataset

We divide the whole dataset into two parts, **train data and test data** randomly. We take **80% of the total observations as train dataset** on which we initially fit the model. Remaining **20% of the total observations is used as test dataset** where we check the adequacy for the different models. The number of observations in 2 parts is **255836 and 63959** respectively.

Model: Logistic Regression

Our response variable 'Heart disease', is a categorical variable with two categories, 0 indicates absence of heart disease and 1 which indicates presence of heart disease. Here we cannot use linear regression models.

Suppose the model has the form

$$\boldsymbol{Y_i = \underline{x_i}' \underline{\beta} + \underline{\epsilon}}$$

Where $\underline{x_i}' = [1, x_{i1}, x_{i2}, \cdots, x_{ik}]$ and $\underline{\beta}' = [\beta 0\ ,\ \beta 1\ ,\ \beta 2\ ,\ .\ .\ .\ ,\ \beta k\ ]$ and the response variable takes the value either 0 or 1.

We will assume that the response variable y_i is a Bernoulli random variable with probability distribution as follows:

$$P(Y_i = 1) = \pi_i$$
$$P(Y_i = 0) = 1 - \pi_i\ , 0 < \pi_i\ < 1, \forall\ i = 1(1)p$$

The expected value of the response variable is

$$E(Y_i|x_i) = \underline{x_i}'\ \underline{\beta} = \ \pi_i$$

Note that if the response is binary then the error terms ϵ_i can only take two values ,

$$\epsilon_i = 1 - \underline{x_i}'\ \underline{\beta}\ when\ Y_i = 1$$

And, $\epsilon_i = -\underline{x_i}'\ \underline{\beta}\ when\ Y_i = 0$

So here ϵ_i's does not follow normality.

Let us consider the generalised linear model (GLM):

$$Y_i = \ f\left(\underline{\beta}|\underline{x_i}\right) + \ \varepsilon_i, \forall\ \mathbf{i} = 1(1)\mathbf{n}$$

$$where\ \underline{x_i} = (x_{i1}, x_{i2}, \ldots, x_{ik})'$$

f need not be a linear function of parameters and is a function of linear predictor $\underline{x_i}'\ \underline{\beta}$.

We get different GLMs for different functional form of f.

Let $\eta_i = \ \underline{x_i}'\underline{\beta}$

Logistic Regression Model is a particular GLM where

$$f\left(\underline{\beta}|\underline{x_i}\right) = \frac{e^{\underline{x_i}'\underline{\beta}}}{1+ e^{\underline{x_i}'\underline{\beta}}} = \frac{e^{\eta_i}}{1+ e^{\eta_i}} = \pi_i$$

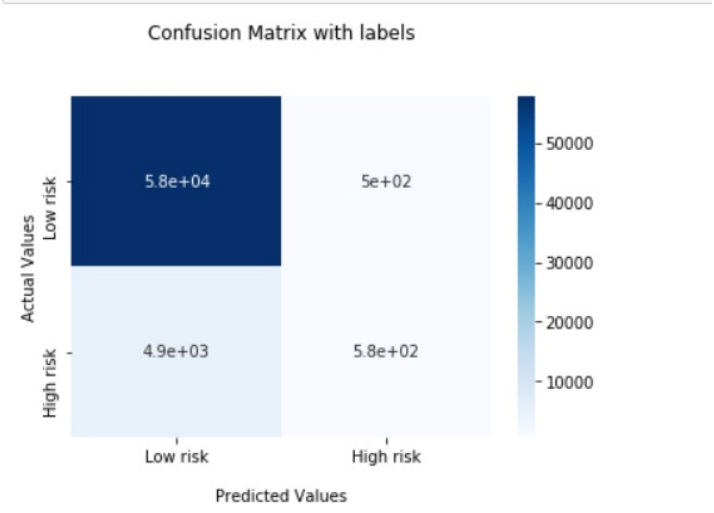
$\Rightarrow \eta_i = ln\left(\frac{\pi_i}{1-\pi_i}\right)$ is **the logit link function** which is linear in parameters.

Training the dataset

```
glm(formula = HeartDisease ~ ., family = binomial, data = training)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1189  -0.4122  -0.2437  -0.1285   3.6145

Coefficients:
                                    Estimate Std. Error z value Pr(>|z|)
(Intercept)                       -1.119015   0.060053 -18.634  < 2e-16 ***
BMI                                0.057799   0.008097   7.139 9.42e-13 ***
Smoking                            0.355487   0.016062  22.132  < 2e-16 ***
AlcoholDrinking                   -0.231466   0.037335  -6.200 5.66e-10 ***
Stroke                             1.069972   0.025247  42.381  < 2e-16 ***
PhysicalHealth                     0.022918   0.007672   2.987  0.00281 **
MentalHealth                       0.038622   0.007852   4.919 8.71e-07 ***
DiffWalking                        0.214854   0.020291  10.589  < 2e-16 ***
Sex                                0.707344   0.016220  43.610  < 2e-16 ***
Diabetic                           0.455398   0.018384  24.772  < 2e-16 ***
PhysicalActivity                   0.007579   0.017922   0.423  0.67239
SleepTime                         -0.034794   0.006952  -5.005 5.59e-07 ***
Asthma                             0.267885   0.021531  12.442  < 2e-16 ***
KidneyDisease                      0.562700   0.027228  20.666  < 2e-16 ***
SkinCancer                         0.094124   0.021847   4.308 1.64e-05 ***
AgeCategory_18.24                 -3.245805   0.101929 -31.844  < 2e-16 ***
AgeCategory_25.29                 -3.099128   0.100316 -30.894  < 2e-16 ***
AgeCategory_30.34                 -2.739750   0.079061 -34.653  < 2e-16 ***
AgeCategory_35.39                 -2.663779   0.071071 -37.480  < 2e-16 ***
AgeCategory_40.44                 -2.277863   0.058114 -39.197  < 2e-16 ***
AgeCategory_45.49                 -1.934641   0.048782 -39.659  < 2e-16 ***
AgeCategory_50.54                 -1.473008   0.038658 -38.104  < 2e-16 ***
AgeCategory_55.59                 -1.260779   0.033529 -37.602  < 2e-16 ***
AgeCategory_60.64                 -0.988779   0.029817 -33.161  < 2e-16 ***
AgeCategory_65.69                 -0.734856   0.028149 -26.106  < 2e-16 ***
AgeCategory_70.74                 -0.458459   0.027075 -16.933  < 2e-16 ***
AgeCategory_75.79                 -0.269081   0.028431  -9.464  < 2e-16 ***
GenHealth_Excellent               -1.904631   0.045849 -41.542  < 2e-16 ***
GenHealth_Fair                    -0.382434   0.031747 -12.046  < 2e-16 ***
GenHealth_Good                    -0.845703   0.034268 -24.679  < 2e-16 ***
GenHealth_Very.good               -1.418041   0.037433 -37.882  < 2e-16 ***
Race_American.Indian.Alaskan.Native  0.040324   0.071609   0.563  0.57336
Race_Asian                        -0.549294   0.088380  -6.215 5.13e-10 ***
Race_Black                        -0.306790   0.053486  -5.736 9.70e-09 ***
Race_Hispanic                     -0.172988   0.054655  -3.165  0.00155 **
Race_White                        -0.027334   0.044568  -0.613  0.53967
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 149676  on 255606  degrees of freedom
Residual deviance: 116204  on 255571  degrees of freedom
AIC: 116276

Number of Fisher Scoring iterations: 7
```

Checking Model Adequacy

Confusion Matrix



For this data

Accuracy score: 0.92

Precision: 0.54

Recall: 0.11

F1 score: 0.18

```
Confusion Matrix:
 [[58010   503]
 [ 4861   585]]
```

Model: K-Nearest Neighbor

It is a non-parametric classifier approach.

Let there be a learning sample L ,

L = $\{(x_i , y_j), i = 1(1)n, j=1(1)c\}$, where n is the number of observations

The feature vectors x_i's in the learning sample are classified into c possible classes y_j s ,j=1(1)c.

NN classifier uses observations in L closest to the given observation, closest in the feature vector space.

Let $N_k(X)$ denote the neighborhood of X defined by the k closest points x_i's in the Learning/training set.

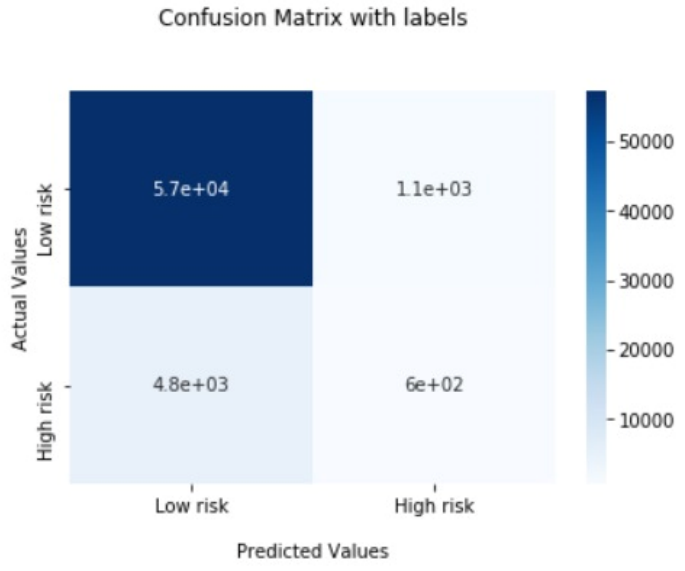
Now we obtain a "majority voting rule" for the k nearest neighbours, i.e. , points within $N_k(X)$; voting with respect to their class membership. The class which is the winner(having maximum counts for cases inside $N_k(X)$) is the assigned class.

In our case it is a 2-class classification problem

The two classes are "Presence of heart disease ", denoted by 1 and "Absence of heart disease" , denoted by 0.

We take 5 nearest neighbors for our model.

Checking Model Adequacy



For this data

Accuracy score: 0.91

Precision: 0.36

Recall: 0.11

F1 score: 0.17

```
Confusion Matrix:
 [[57397 1116]
 [ 4825   621]]
```

Model: Decision Tree

Classification/Decision tree is an example of a multistage decision process.

It is a non-parametric approach and does not require any distributional assumption.

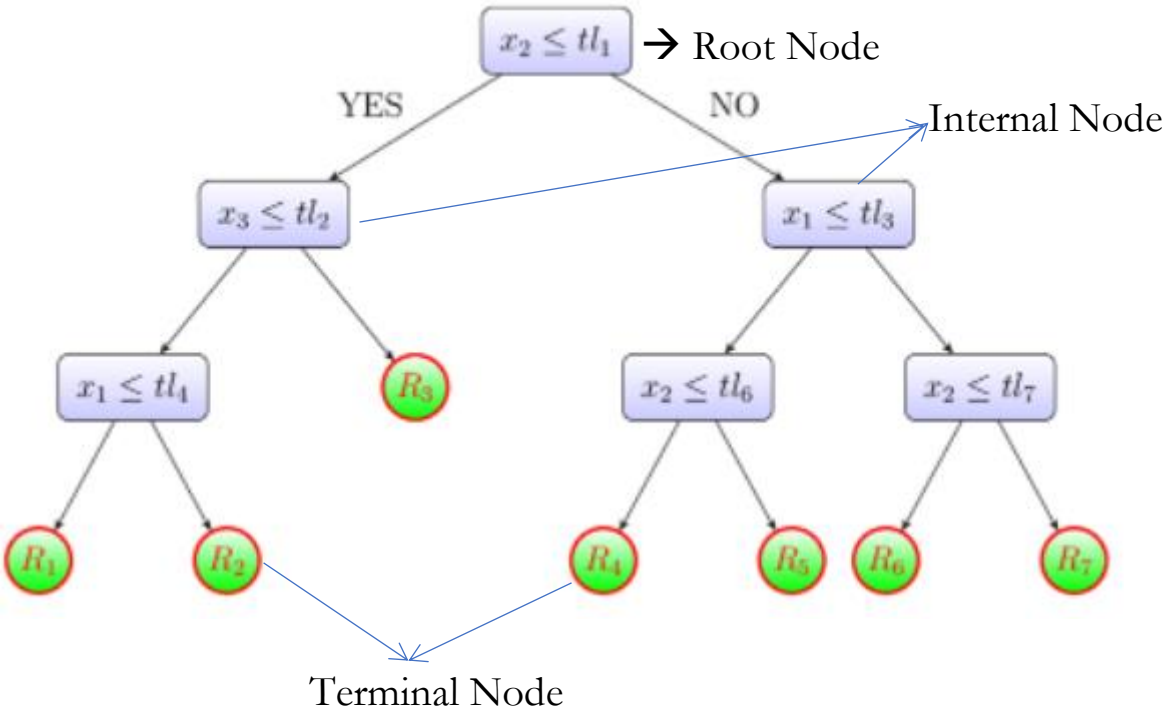
Rather than using a complete set of features jointly to make a output decision, different subsets of features are used at different levels of the tree.

A tree has its starting point called the root node.

There are internal nodes connected by branches, from each node there emerges two branches for two types of decisions (Yes/No)

And finally there are the terminal nodes which classify a feature vector into a class label.

A typical tree looks like:



Constuction of classification tree would revolve around

i. How to choose the variable associated with any internal node (including root node)

ii. How to find the threshold associated with the variable at any internal node.
iii. How to assign class label at the terminal nodes
iv. When to declare a node as terminal node.

For a vector X from a p dimensional feature space, let us consider a split on X_k at a split level 'l' , such that the split rule is $\{ X : X_k < l\}$

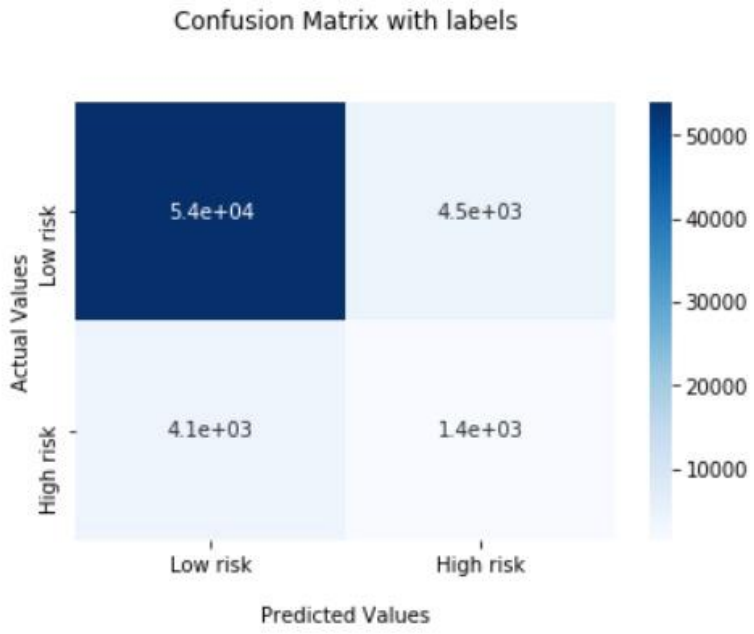
Starting from the root node , at each step we have to find optimum k, l , at each node such that the goodness of split measure (i.e . the change in impurity/misclassification due to split is maximised.

k = 1(1)p where p is the dimension of the feature space

l is varied over a grid of possible values of the dimension of the feature vector

By repeating the above procedure we stop splitting if the reduction in impurity due to split is less than a threshold.

Checking Model Adequacy



For this data

Accuracy score: 0.87

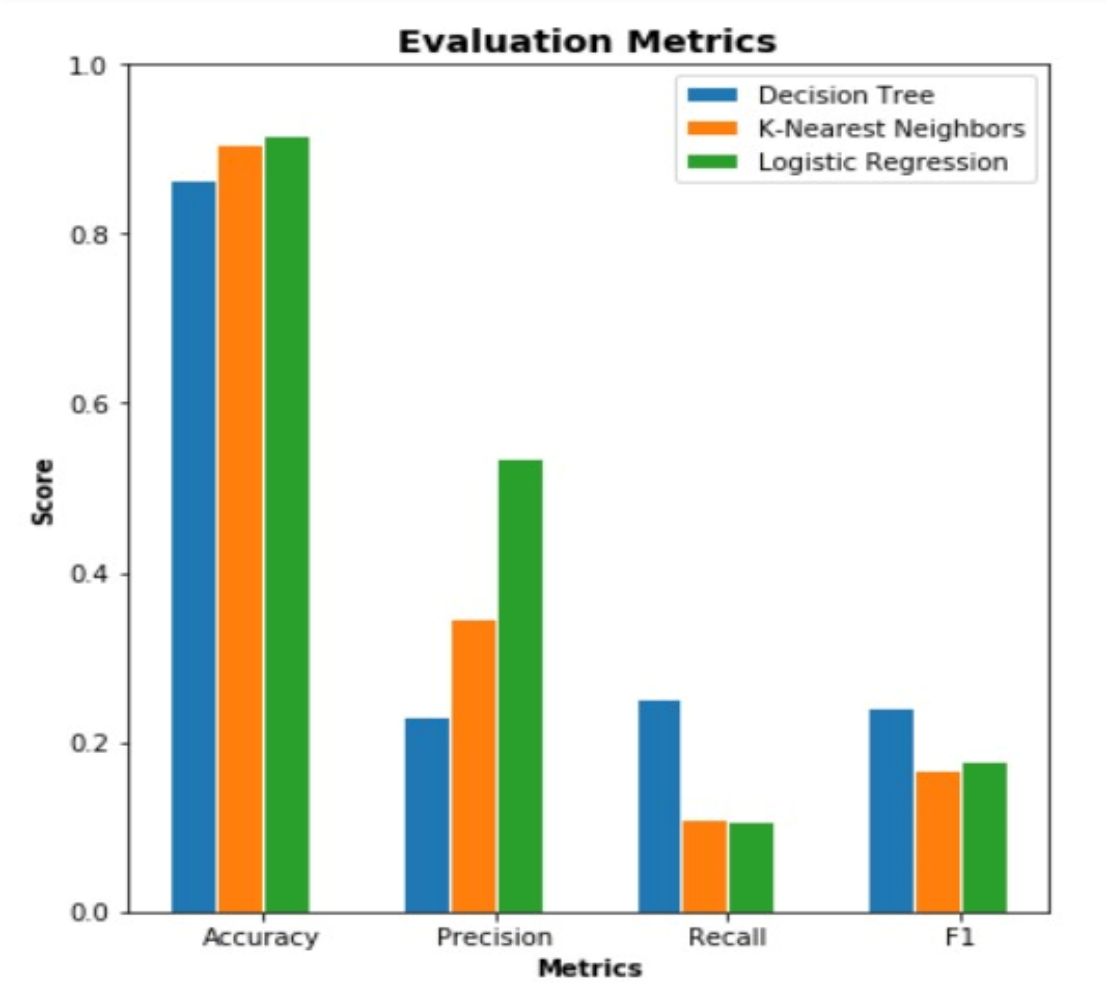
Precision: 0.23

Recall: 0.25

F1 score: 0.24

```
Confusion Matrix:
 [[53969  4544]
 [ 4052  1394]]
```

Comparing Evaluation Metrics for the Three Models

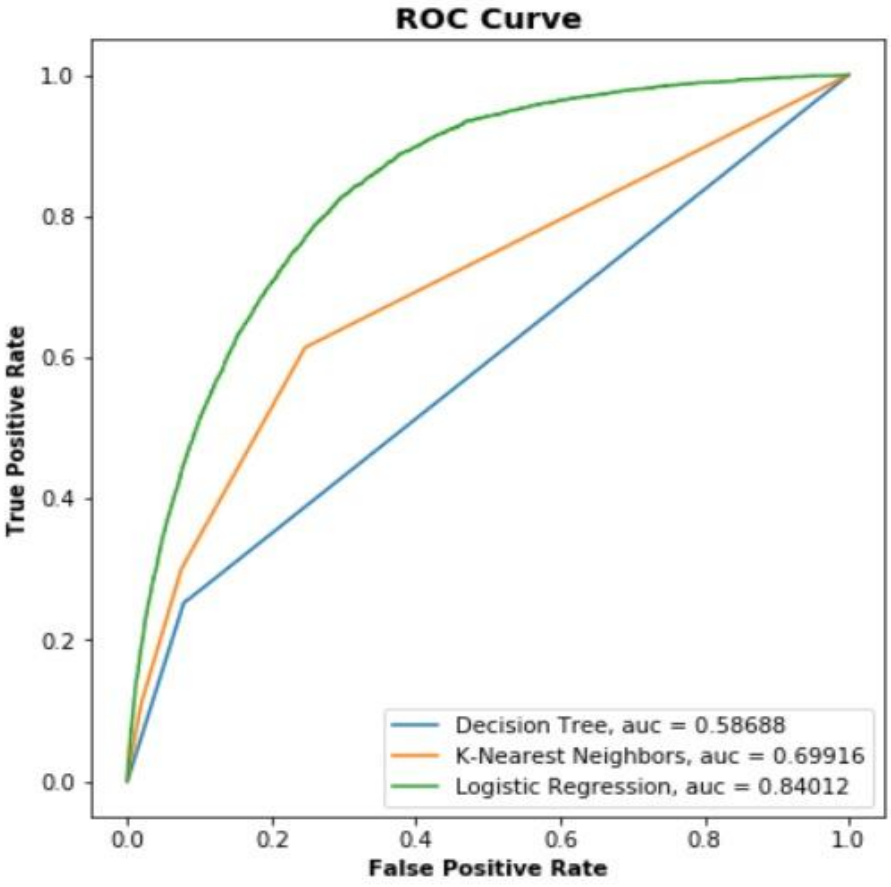


Here it is clearly evident that logistic regression model has higher accuracy and precision score whereas decision tree model has higher recall and F1 score.

ROC Curve

An ROC curve (receiver operating characteristic curve) is a graph showing the performance of a classification model at all classification thresholds. This curve plots two parameters: True Positive Rate, False Positive Rate. The ROC curve is produced by calculating and plotting the true positive rate against the false positive rate for a single classifier at a variety of thresholds.

For example, in logistic regression, the threshold would be the predicted probability of an observation belonging to the positive class.



Hence for our data we can clearly see that logistic regression model has the highest AUC, hence in our case logistic regression model is the best classifier.

Conclusion and Interpretation

- In this case, we can clearly see that the logistic classifier model is a somewhat better classifier in terms of both the evaluation metrics and the ROC curve.
- But, if we look closely, we see that the recall score for decision tree is higher than that of the other two methods. Now, recall score, which is the sensitivity or true positive rate from the confusion matrix is an indicator of how well the model can predict the 1's which are actually 1's, i.e., in our case presence of heart disease.
- Again, precision score, which is the fraction of actual 1's (presence of heart disease) among all the predicted 1's, is high for the logistic classifier.
- But, in the F1 score, which is the harmonic mean of precision and recall and a measure of precision and robustness of the model, is slightly higher for decision tree than the others.

❖ Hence, we can say although the logistic classifier is the best classifier, the decision tree model actually performs better in identifying the presence of heart disease.

References

- An **I**ntroduction to **S**tatistical **L**earning with Applications in **R**- *Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani*

- https://hello.iitk.ac.in/mth552a22/#/home -course homepage of MTH552A

- Machine Learning using Python – *Manaranjan Pradhan, U Dinesh Kumar*