Project Report On

CLASSIFICATION USING K-MEANS CLUSTERING
APPLIED TO
HEART DISEASE PREDICTION

Project submitted

in partial fulfillment of the necessities of the degree of

Master of Computer Application

By

NILADRI DAS

Roll No.: 001910503029

Roll No.: MCA226028

Registration No.: 149890 of 2019-20

Under the supervision of

Dr. Sanjoy Kumar Saha

Department of Computer Science and Engineering

Faculty of Engineering and Technology

Jadavpur University

Kolkata – 700032, India

JADAVPUR UNIVERSITY
FACULTY OF ENGINEERING AND TECHNOLOGY
DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

Certificate of Recommendation

This is to certify that Niladri Das (Registration No.: 149890 of 2019-20, Roll No.: 001910503029, Exam Roll No.: MCA226028) is a student in Master of Computer Application course and the project "**CLASSIFICATION USING K-MEANS CLUSTERING APPLIED TO HEART DISEASE PREDICTION**" is a record of work carried out by him, is accepted in partial fulfillment of the requirement for the degree of Master of Computer Application from the Department of Computer Science and Engineering, Jadavpur University during the academic year 2021-2022. He has been able to follow all the instructions in a calm and responsible way and successfully carried out his research work. All the best wishes to him for his future endeavors.

_____

Dr. Sanjoy Kumar Saha (Project Supervisor)

Assistant Professor, Dept. of Comp. Science & Engineering

Jadavpur University, Kolkata-700032

_____

Prof. Anupam Sinha

Head of the Department, Dept. of Comp. Science & Engineering

Jadavpur University, Kolkata-700032

_____

Prof. Chandan Majumdar

Dean, Faculty Council of Engineering & Technology

Jadavpur University, Kolkata-700032

JADAVPUR UNIVERSITY
FACULTY OF ENGINEERING AND TECHNOLOGY
DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

Certificate of Approval

The foregoing project entitled as "**CLASSIFICATION USING K-MEANS CLUSTERING APPLIED TO HEART DISEASE PREDICTION**" is hereby approved as a creditable study of Master of Computer Applications and presented in a manner satisfactory to warrant its acceptance as a prerequisite to the degree for which it has been submitted. It is understood that by this approval the undersigned do not necessarily endorse or approve any statement made, opinion expressed or conclusion therein but approve this project only for the purpose for which it is submitted.

Signature of Examiner 1	**Signature of Examiner 2**
Date:	**Date:**
Name:	**Name:**

JADAVPUR UNIVERSITY
FACULTY OF ENGINEERING AND TECHNOLOGY
DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

Declaration of Originality and Compliance of Academic Ethics

I, hereby declare that this project contains literature survey and original research work by the undersigned candidate, as part of his Master of Computer Application studies.

All information in this document has been obtained and presented in accordance with academic rule and ethical conduct.

I also declare that, as required by this rules and conduct, I have fully cited and referenced all the materials that are not original to this work.

_____

Signature

Name: Niladri Das

Class Roll No.: 001910503029

Exam Roll No.: MCA226028

Project Title: "CLASSIFICATION USING K-MEANS CLUSTERING APPLIED TO HEART DISEASE PREDICTION"

JADAVPUR UNIVERSITY
FACULTY OF ENGINEERING AND TECHNOLOGY
DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

ACKNOWLEDGEMENTS

I have been interested in Machine Learning from when I first heard about it. Because it is very connected to mathematics and mathematical algorithms which is pulling me to be a part of this subject. Now, at last, I have had the opportunity to research and learn a bit more about it, so I would like to thank the Computer Science and Engineering (CSE) Department of Jadavpur University for allowing me as a candidate for this project. I am especially grateful to my project guide and advisor Dr. Sanjay Kumar Saha for his guidance and attention during the entire project and my project partner Mr. Koushik Majumder. Without his support it would not have been so bearable.

Thank you.

_____

Niladri Das

MCA226028

149890 OF 2019-20

Dr. Sanjoy Kumar Saha

ABSTRACT

Machine Learning is a milestone to reach artificial intelligence and deep learning. It provides various types of powerful algorithms for various types of problems like recognizing patterns, classifying data, segmentation of data and so on. Machine Learning is a child now, but it grows every day. I mean, Machine Learning is a huge platform, growing every day. So, it still remains unknown for the majority of people, and even for most professionals.

This project intends to provide an understandable explanation of what Machine Learning is, what types are there and how Machine Learning can be used. As well as solving a real data classification problem (Classification using K-Means Clustering applied to heart disease prediction), first using K-means algorithm to segment the data, and then classify a new data point according to its nearest cluster or segment, as an introduction to this field.

TABLE OF CONTENTS

1. MACHINE LEARNING(ML)

1.1. INTRODUCTION TO MACHINE LEARNING [9]

As per name and for sake of simplicity, Machine Learning is a technique for learning the machine to work itself.

It might seem that this is a pretty new technology, but, in fact, it isn't. The first ML-related work dates from 72 years ago, in 1950.

Significance of 1950 – Alan Turing creates the "Turing Test". This test determined whether a computer had real intelligence or not. First, experiment on Machine Learning, which influences humans to learn machines like humans.

Arthur Samuel, an American pioneer in the field of computer gaming and artificial intelligence, coined the term "Machine Learning" in 1959 while at IBM.

From those days to now days Machine Learning traveled a long way and filled its bag with various ML algorithms and techniques by different Industries and Scientists. But, there are many things to invent and optimize in Ml, as we know invention is an endless journey. So, the journey will be always on.

1.2. WHAT IS MACHINE LEARNING [9]

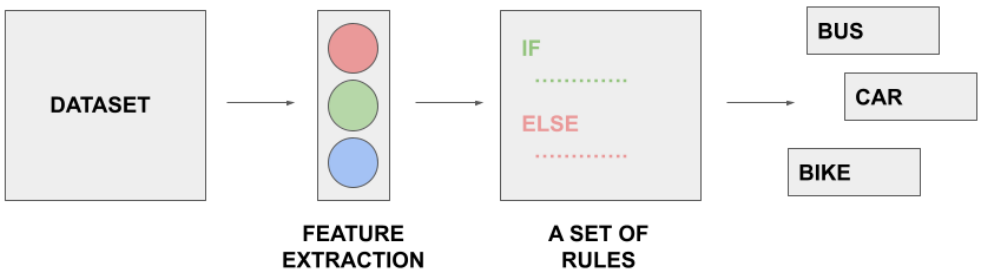
In definition, Machine Learning is the subfield of Computer Science that gives "computers the ability to learn without being explicitly programmed".

A Machine Learning model is, using Machine Learning algorithms to iteratively learn from data and allow Machines to find hidden insights. Simply, Machine Learning models, built by different Machine Learning algorithms, help us in a variety of tasks, such as recognition, summarization, recommendation, segmentation and so on.

The concept of Machine Learning is actually inspired by the human learning process and the main goal of Machine Learning is to build machines like humans.

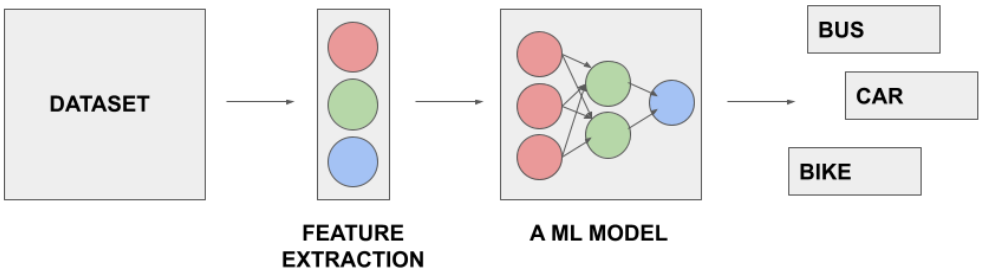
1.3. HOW DOES MACHINE LEARNING WORK AND DIFFERENT FROM NORMAL PROGRAMMING TECHNIQUE [9]

For the sake of simplicity, I explain it by pictorial representation.



IN NORMAL PROGRAMMING

We extract the features from the dataset and specify a set of rules on the features set and it determines what will be the outcome.

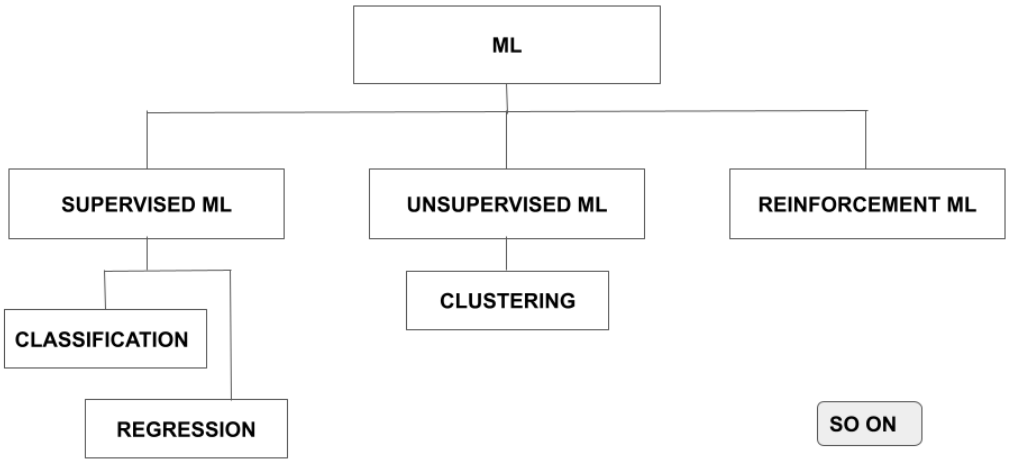


IN MACHINE LEARNING

We extract the features from the dataset. First, build a ML model with respect to the features set. Next, train the model with the dataset. Now, the Model will create a pattern between the features set to find the insight and it determines what will be the outcome.

1.4. TYPES OF MACHINE LEARNING [9]

Types of Machine Learning are bordered by different approaches of Machine Learning.

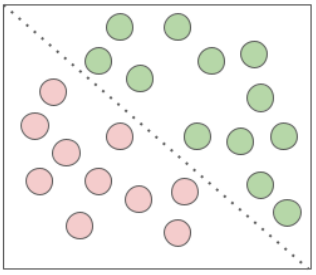


SUPERVISED ML

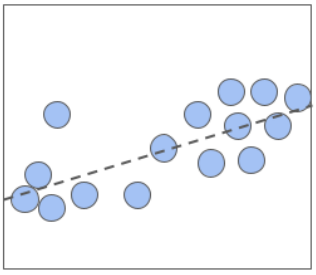
As it sounds, in supervised Machine Learning we need to supervise a ML model. Simply, we teach a model and then with that knowledge it can predict unknown or further instances on labeled data.

Like humans supervising something and predicting something.

Two of the most famous supervised techniques are Classification and Regression.



CLASSIFICATION
Classification is the process of predicting discrete class labels or categories.



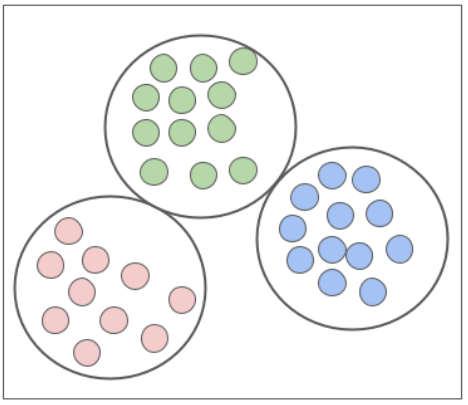
REGRESSION
Regression is the process of predicting continuous values.

UNSUPERVISED ML

As it sounds, in unsupervised Machine Learning we don't need to supervise a ML model.
But, there is a question, what do we need to do?
Simply, we need to let the model work on its own to discover information.

The unsupervised algorithms train on the dataset and draw conclusions or find patterns on unlabeled data.

One of the most famous unsupervised techniques is Clustering.



CLUSTERING
Classification is the process of segmentation the data according to the insights from the data.

1.5. VARIOUS TECHNIQUES FOR VARIOUS ML PROBLEMS / ML APPLICATIONS [9]

REGRESSION/ESTIMATION

This technique is used to predict continuous values.
For example, hotel price prediction, flight/bus fare prediction, and so on.

CLASSIFICATION

This technique is used to predict categorical values.
For example, disease prediction, drug prediction, and so on.

CLUSTERING

This technique is used to create structure on the data.
For example, summarization, customer segmentation, and so on.

RECOMMENDATION SYSTEM

This technique is used to create a recommendation system.
For example, product recommendations, movie/book recommendations, and so on.

SEQUENCE MINING

This technique is used to predict the next item.

ANOMALY DETECTION

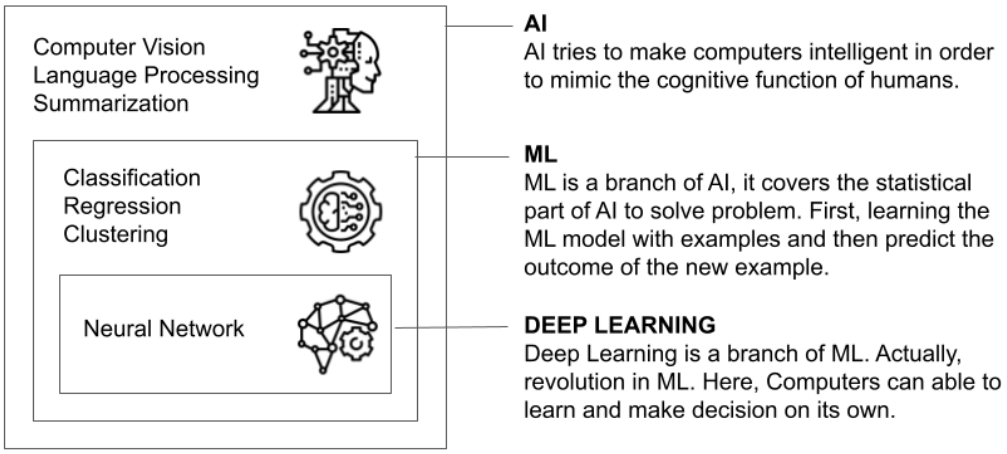
This technique is used to discover abnormal and unusual cases.
For example, credit card fraud detection, crime detection, and so on.

DIMENSION REDUCTION

This technique is used to reduce the size of data in an optimized way.

1.6. DIFFERENCE BETWEEN AI, ML AND DEEP LEARNING [9]

For the sake of simplicity, I explain it by pictorial representation.



2. CLASSIFICATION

2.1. WHAT IS CLASSIFICATION [9]

Classification is a special type of supervised Machine Learning approach, which is used to categorize some unknown items into a discrete set of categories (i.e., categorical values).

Classification first attempts to learn the relationship between a set of feature variables **(independent variable : which variable is not dependent on any other variable)** and a target variable of interest **(dependent variable : which variable is dependent on another variable)** and then determines the class label for an unlabeled test case**.**

2.2. HOW DOES CLASSIFICATION WORK [9]

For the sake of simplicity, I will explain it with an example.

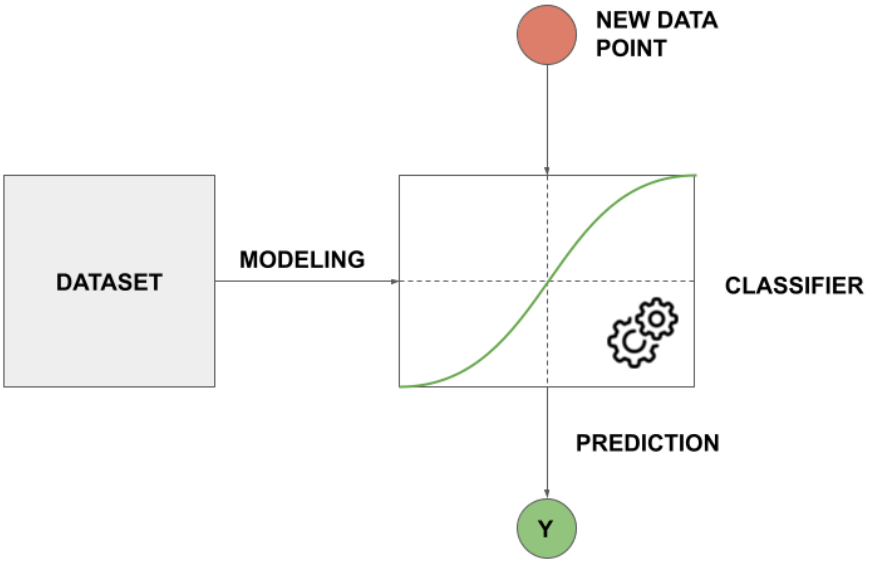
Suppose we have a patient dataset.
In the patient dataset, there are seven independent variables (i.e., **features**) sex, age, education, and so on and one dependent variable (i.e., **target**) TenYearCHD (Ten Year Chance of Heart Disease).

PATIENT DATASET

	INDEPENDENT VARIABLES							DEPENDENT VARIABLE
i	**sex**	**age**	**education**	**cigsPerDay**	**totChol**	**BMI**	**heartRate**	**TenYearCHD**
0	male	39	4.0	0.0	195.0	26.97	80.0	no
1	female	46	2.0	0.0	250.0	28.73	95.0	no
2	male	48	1.0	20.0	245.0	25.34	75.0	no
3	male	61	3.0	30.0	225.0	28.58	65.0	yes
4	male	46	3.0	23.0	285.0	23.1	85.0	no
5	female	40	1.0	13.0	300.0	30.0	95.0	?

Clearly, the relation/pattern between features indicates the possibility of heart disease in the next ten years. We understand, machines do not. So, we need to understand these things to the machine using classification algorithms.

Now, we need to create a classification model which learns the relationship between the set of features and the target variable.



CLASSIFICATION MODEL

Then, the model will give us the prediction of the possibility of heart disease in the next ten years for new patients with the same feature set.

i	sex	age	education	cigsPerDay	totChol	BMI	heartRate	TenYearCHD
0	female	40	1.0	13.0	300.0	30.0	95.0	yes

NOTE: The above example is an example of binary classification. Similarly, we can create models for multiclass classification.

2.3. VARIOUS CLASSIFICATION ALGORITHMS [9]

There are many classification algorithms for building both binary and multiclass classifiers or classification models.

Some of those are,

I. **K Nearest Neighbors (KNN)**
II. **Decision Tree (ID3, C4.5, C5.0)**
III. **Logistic Regression (LR)**
IV. **Support Vector Machine (SVM)**
V. **Native Bayes**
VI. **Neural Network**
, and so on.

All of the above classification algorithms are used for different real world works.

2.4. APPLICATIONS OF CLASSIFICATION [9]

There are many uses of classifications in the real world.

We can use classification to categorize a customer. Like, banks need to categorize customers for loan approval, telecommunication service providers need to know about whether a customer switches to another provider or not, advertisement companies need to know whether a customer responds or not, and so on.

There is a wide range of use of classification in the medical sector. Like, whether a patient is affected by a disease or not. Also, we can use classification to find perfect drugs for a patient.

2.5. MODEL EVALUATION TECHNIQUES FOR CLASSIFICATION [9]

The goal of classification is to build a model to accurately predict an unknown case or class or category.

But, how can we do it?
It is by going through a model evolution technique.

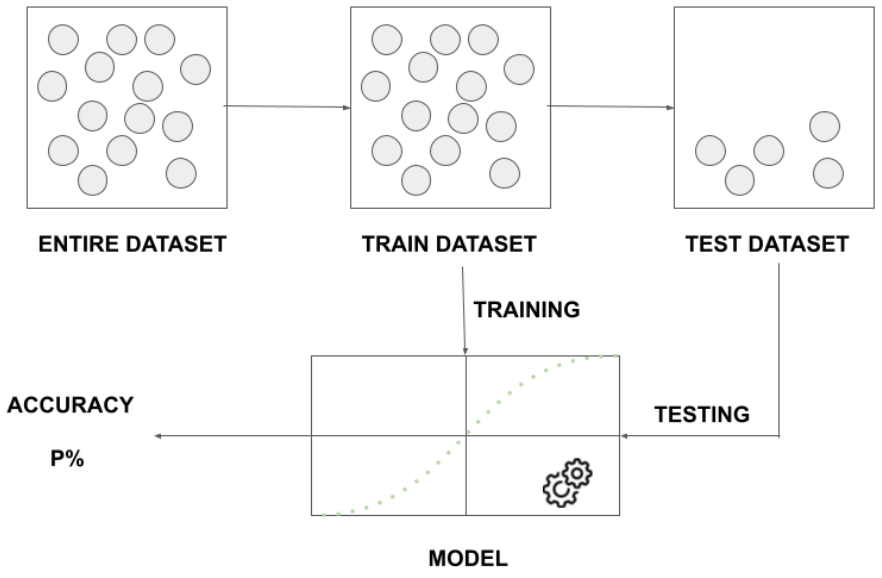
There are few model evaluation techniques for classification used for different purposes.
Those are,

I. Train and Test on the same dataset (**good**)
II. Train Test Split (**better**)
III. K-Fold Cross Validation (**best**)

To know which approach is the best. First we need to know about two terms:

Training Accuracy
Training Accuracy is the percentage of current predictions that the model makes when using the test data.
High training accuracy is not necessarily a good thing. It causes **overfitting** (the model is overly trained to the dataset which may capture noise and produce a non-generalized model).

Out-of-sample Accuracy
Out-of-sample Accuracy is the percentage of correct predictions that the model has not trained on.
Low Out-of-sample Accuracy is not a good thing. We need to high it for better accuracy.



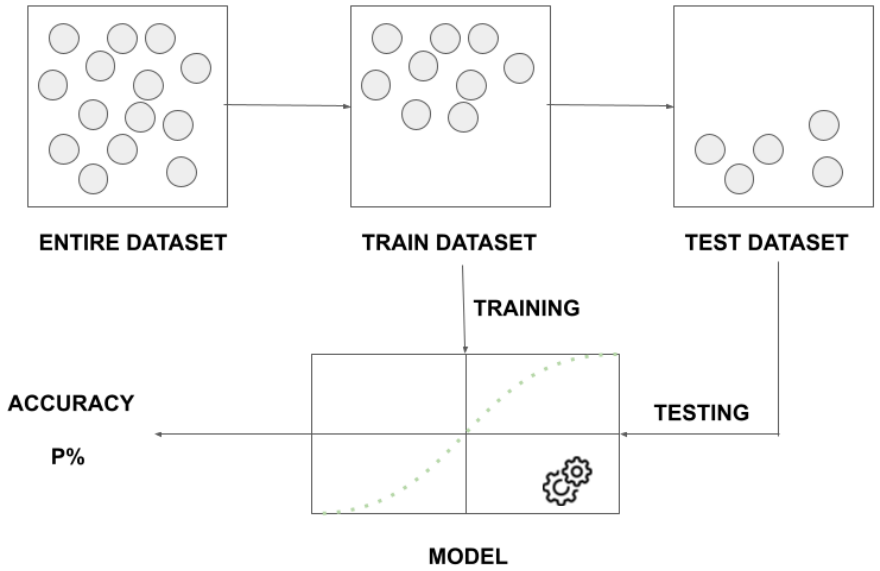
TRAIN AND TEST ON THE SAME DATASET

The name of the approach says it all.

We need to first train the model with the entire dataset and then test it using a portion of the entire dataset.

I. High “Training Accuracy”

II. Low “Out-of-sample Accuracy”

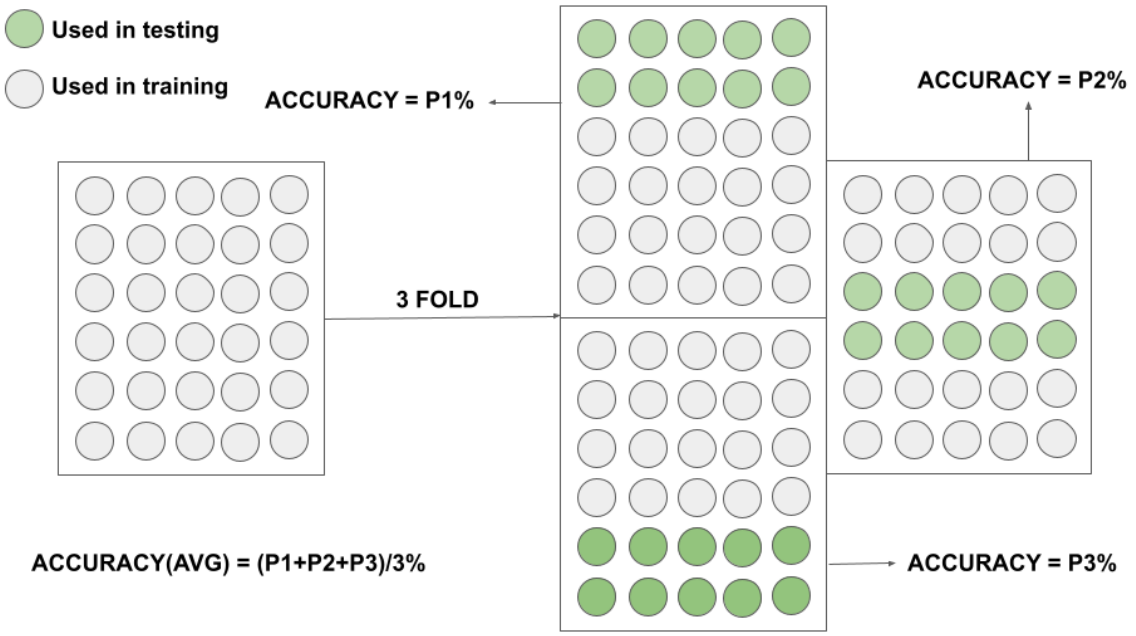


TRAIN TEST SPLIT

Here, we need to first split the dataset into two datasets, one is the Train dataset and another is the Test dataset. Then train the model with the train dataset and next test it using the test dataset.

I. More accurate “Out-of-sample Accuracy”

II. Highly dependent on which data sets the model is trained and tested



K FOLD CROSS VALIDATION

K Fold cross validation is the simplest form that performs multiple train-test splits using the same dataset, where each split is different. In each split Test dataset contains 1/K th portion of the dataset and the Train dataset contains (K-1)/K th portion of the dataset.

For each split, train the model with the train dataset and next test it using the test dataset.

The result of the average accuracy according to the k splits is the final accuracy of the model.

More preferable and correct Model Evaluation Technique in Classification.

2.6. EVALUATION METRICS FOR CLASSIFICATION [9]

As I say, the goal of classification is to build a model to accurately predict an unknown case or class or category.

But, how can we find how accurate our classification model is?
It can be sorted out using Evaluation Metrics for classification.
Evaluation Metrics actually explain the performance of a model.

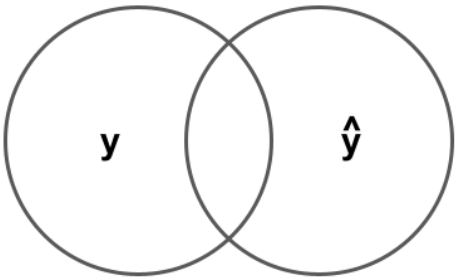
There are different model evaluation metrics for classification. Some of them are,

I. Jaccard Index
II. F1 Score
III. Log Loss

JACCARD INDEX

One of the simplest accuracy measurements evaluation metric is Jaccard Index also known as Jaccard similarity coefficient.

Jaccard Index = (number of objects in common) / (total number of objects)



Y : Actual Labels
Ŷ : Predicted Labels

$$J(y,\hat{y}) = \frac{|\, y \cap \hat{y} \,|}{|\, y \cup \hat{y} \,|}$$

$J(y,\hat{y}) = 0.0$
If no matches

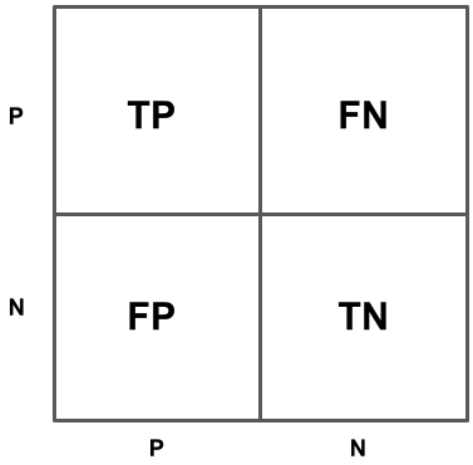
$J(y,\hat{y}) = 1.0$
If all matches

F1 SCORE

One of the most used Evaluation Metric for accuracy measurement in Classification Problem.

F1 Score = 2 x (Precision x Recall) / (Precision + Recall)

If, the F1 Score is high. Then, the accuracy will be high.
If, the F1 Score is low. Then, the accuracy will be low.



CONFUSION MATRIX

Precision : measure of accuracy

Recall : trueness rate

F1 Score : harmonic average of Precision and Recall

$$\text{Precision} = \frac{TP}{TP + FP} \quad \frac{TN}{TN + FN}$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad \frac{TN}{TN + FP}$$

FOR P FOR N

$$\text{F1 Score} = \frac{2 \times (\text{Precision} \times \text{Recall})}{(\text{Precision} + \text{Recall})}$$

NOTE: The above confusion matrix and F1 Scores are for Binary Classification. It is also used for Multiclass Classification.

LOG LOSS

Sometimes the output of a classifier is the probability of a class label instead of the label itself. Like, in the Logistic Regression algorithm.

In those situations, we can't use previous Evaluation Metrics.

Here, we need to use Log Loss.
Logarithmic Loss measures the performance of a classifier where the predicted output is a probability.

$$H_p(q) = -\frac{1}{N} \sum_{i=1}^{N} y_i \cdot log(p(y_i)) + (1 - y_i) \cdot log(1 - p(y_i))$$

Classifiers with lower Log Loss have better accuracy.

3. CLUSTERING

3.1. WHAT IS CLUSTERING [9]

Clustering is a special type of unsupervised Machine Learning approach, it is used to find clusters/segments in a dataset by reading the pattern in between the features set of the dataset.

Now, the question is, what is a cluster?
A cluster is a group of data points or objects in a dataset that are similar to other data points or objects in the group, and dissimilar to data points in other clusters.

So, Clustering is used to create mutually exclusive groups in a dataset in an unsupervised way, based on similarity of the features set of the dataset.

3.2. HOW DOES CLUSTERING WORK [9]

For the sake of simplicity, I will explain it with an example.

Suppose we have a patient dataset.
We have to find which patient is in which category according to their age and weight.

PATIENT DATASET

i	sex	age	education	totChol	BMI	thyroid
0	male	39	4.0	195.0	26.97	no
1	female	46	2.0	250.0	28.73	no
2	male	48	1.0	245.0	25.34	no
3	male	61	3.0	225.0	28.58	yes
4	male	46	3.0	285.0	23.1	no
5	female	40	1.0	300.0	25.0	yes

Clearly, by seeing the relation/pattern between the age and BMI features we can say which patient is in which type. We understand, machines do not. So, we need to understand these things to the machine using clustering algorithms.

Now, we need to create a clustering model which learns the relationship between the set of features and segment the dataset.

CLUSTERING MODEL

Then, the model will give us the segmented data.

i	sex	age	education	totChol	BMI	thyroid	typeOfPatient
0	male	39	4.0	195.0	26.97	no	Young and overweight
1	female	46	2.0	250.0	28.73	no	Middle aged and overweight
2	male	48	1.0	245.0	25.34	no	Middle aged and overweight
3	male	61	3.0	225.0	28.58	yes	Old and overweight
4	male	46	3.0	285.0	23.1	no	Middle aged and middleweight
5	female	40	1.0	300.0	25.0	yes	Middle aged and overweight

NOTE: The above example is an example of a clustering model on two-dimensional space. Similarly, we can create clustering models on multi-dimensional space.

3.3. VARIOUS CLUSTERING ALGORITHMS [9]

There are many clustering algorithms for building clustering models. According to the methodology of the algorithms, clustering algorithms are divided into three types.

Those are,

I. **Partition Based Clustering**
   These algorithms are relatively efficient and are used for medium and large sized datasets.
   Main drawback of these algorithms is finding the best partitions.
   E.g., K-Means, K-Median, Fuzzy C-Means.

II. **Hierarchical Clustering**
   The main methodology of these types of algorithms is producing trees of Clusters.
   These algorithms are very intuitive and are generally good for use with small datasets.
   E.g., Agglomerative, Divisive Algorithms.

III. **Density Based Clustering**
   Produces arbitrary shaped clusters.
   These are especially good algorithms when dealing with special clusters or when there is noise in the dataset.
   E.g., DBSCAN.

All of the above clustering algorithms are used for different purposes of real world work.

3.4. APPLICATIONS OF CLUSTERING [9]

There are many uses of Clustering in the real world. One of the main uses of clustering is segmentation of data. In marketing we can identify buying patterns of customers, recommending products to the new customers, etc. In banking or insurance sector fraud detection in credit card use, identifying types of customers (loyal/churned), so on. In the medical sector, characterizing patient behaviors, steps of a disease, etc. And many more.

3.5. K-MEANS CLUSTERING ALGORITHM [9]

At the start of all, mind it all the data points are actually considered as n dimensional vectors, where n = number of features in the data points.

K-Means algorithm is a partition based clustering algorithm.

Main Objectives of K-Means are,

I. It divides the dataset into K non-overlapping clusters without any cluster internal structure or labels.
II. K-means is used to form clusters in such a way that similar samples go into a cluster and dissimilar samples fall into different clusters.
III. K-means tries to minimize the **intra cluster distances** and maximize the **inter cluster distances**.
So, the distance of the sample data points from each other is used to find the shape of the cluster.

Intra cluster distance : It is the distance between the data points belonging to the same cluster.
inter cluster distance : It is the distance between the data points belonging to the different clusters.

Now, the question comes,
How can we calculate the dissimilarity or distance of two data points?

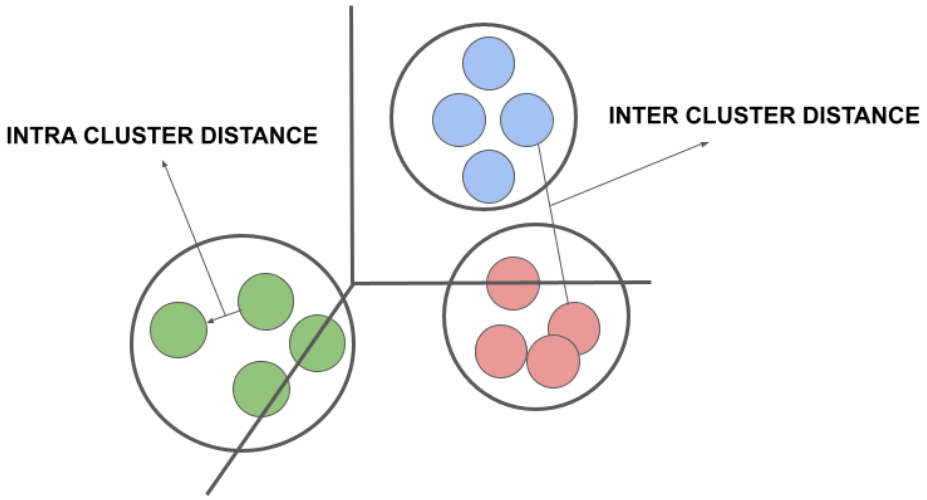
There are many metrics to find the distance between two points, like Euclidean Distance, Cosine Similarity, Average Distance and so on.

But, the most popular distance metric is Euclidean Distance.

The formula of the euclidean distance in n dimensional space from one data point to another data point is.

d(p,q)=d(q,p)=$\sqrt{(p1 - q1)^2 + (p2 - q2)^2 + \ldots + (pn - qn)^2}$

Where,
p ≡ (p1, p2, …., pn)
q ≡ (q1, q2, …., qn)



In the K-Means algorithm K actually indicates the number of clusters.

Before jumping to the steps of the K-Means algorithm we must need to know about one more thing.
The Centroid of a cluster. Each cluster must have a centroid.
It is hypothetically the center point of the cluster. In K-Means the centroid of a cluster is defined as,

$$\left(\sum_{i=1}^{m} p_{1i} / \sum_{i=1}^{m} i,\ \sum_{i=1}^{m} p_{2i} / \sum_{i=1}^{m} i,\ \ldots\ldots,\ \sum_{i=1}^{m} p_{ni} / \sum_{i=1}^{m} i\right)$$

Where,
n = number of features (i.e., number of dimensions of the vector)
m = number of data points (i.e., number of vectors)

STEPS OF K-MEANS CLUSTERING ALGORITHM

I. **Initialize K value and randomly placed K centroids, one for each cluster.**

   But, here is a question.
   How do we know which K value is the best K value for our dataset?
   And the answer is using the Elbow Method. I will describe it after describing the steps of K-Means Algorithm.

II. **Calculate distance of each data point from each centroid.**

III. **Assign each data point to its closest centroid and create a cluster.**

IV. **Calculate the position of the new K centroids.**

V. **Repeat the steps from II to IV, until the centroids no longer move. Please note that, whenever a centroid moves, the distance from the recent old centroid to the recent new centroid is measured.**

Yes, K-means is an iterative algorithm and we have to repeat steps II to IV until the algorithm converges.

These are steps to the K-Means algorithm.

Some observations about the K-Means algorithm :

a. **It is a heuristic algorithm, there is no guarantee that it will converge to the global optimum and the result may depend on the initial clusters. The algorithm is guaranteed to converge to a local optimum and the result is not necessarily the best possible outcome.**
b. **It is relatively efficient for medium and large size datasets.**
c. **It produces sphere-like clusters because the clusters are shaped around the centroids.**
d. **Its drawback is that we should pre specify the number of clusters, and this is not an easy task.**

Before jumping to the Elbow Method. Let's discuss the K-Means model accuracy. That means how accurate our model is.

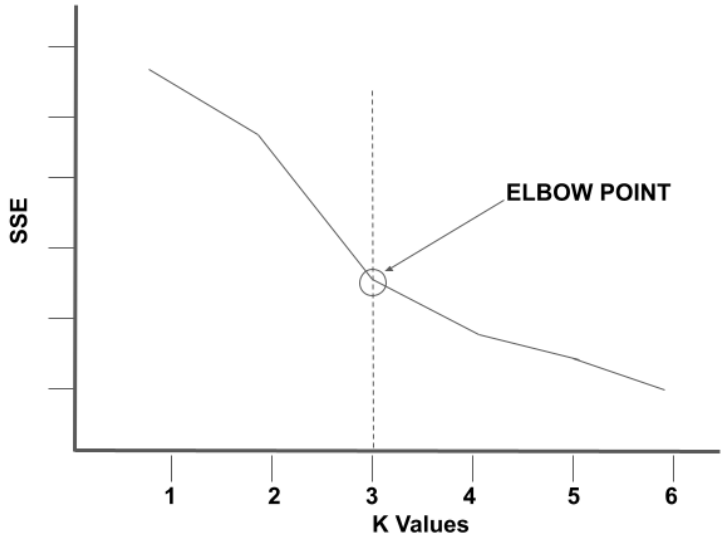
To know that, there is a metric to find error in our model.
Which is known as Sum of the Squared Difference(SSE) between each point and its centroid.

$$\textbf{SSE} = \sum_{i=1,j=1}^{i=n,j=k} (x_i - c_j)^2$$

Where,
x_i **indicates to each data points**
k indicates to number of clusters
n indicates to number of data points
c_j **indicates to cluster centroid of j th cluster**

Low SSE implies the clusters shaped in such a way that the total distance of all the members of a cluster from its centroid be minimized.

ELBOW METHOD



The correct choice of K is often ambiguous because it is very dependent on the distribution of the points in a dataset.

One of the best methods and commonly used methods for finding K value is to run a clustering model across the different values of k and looking at SSE for minimum error for clustering model.

Then, looking at the change of SSE with respect to K values.
But, the problem is the increasing K value will always reduce the error i.e., SSE.

So, choose the elbow point where the rate of decrease sharply shifts. It is the right K value for our clustering model.

This method is known as the Elbow Method.

4. DESCRIPTION OF SOME OF OTHER METHODOLOGIES AND ALGORITHMS USED IN THIS PROJECT

I. Chi-Square Test

When the input is categorical and output is also categorical we have to use the Chi-Square test for Feature selection for best outcomes.

A Chi-Square test is performed on two distributions to measure the degree of similarity of their relative variances. It presupposes that the given distributions are independent in their null hypothesis. So, this can be used to select the optimal features for a given dataset by determining which features are most reliant on the output class label.

The χ^2 value is calculated for each feature in the dataset and then arranged in descending order based on the χ^2 value. The greater the χ^2 value, the more dependent the response is on the feature and the greater the importance of the feature in deciding the output. It indicates the hypothesis of independence is incorrect. [1]

Allow m attribute values for the feature in question and k class labels for the result. Then the following expression gives the value of χ^2.

$$\chi^2 = \sum_{i=1}^{m} \sum_{j=1}^{k} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

where O_{ij} = Observed Frequency and E_{ij} = Expected Frequency.

A contingency table with m rows and k columns is built for each feature. Each cell (i,j) represents the number of rows with the attribute feature i and the class label k. Observed frequency is denoted by each cell in this table. To compute the expected frequency for each cell, first compute the

proportion of the feature value in the overall dataset, then multiply it by the total number of the current class label. [1]

Steps to perform Chi-Square test:

Step 1: Define Hypothesis

Step 2: Create a Contingency Table

Step 3: Find expected values

Step 4: Calculate Chi-Square statistic

Step 5: Accept or Reject Null Hypothesis [1]

II. ANOVA Test

ANOVA stands for Analysis of Variance. It is a statistical approach for comparing the means of two or more groups that are significantly different from one another. When the input is numerical and output is categorical we have to use ANOVA test for Feature selection. It assumes hypothesis as

H0: The means of all groups are the same.

H1: At least one mean in each group differs.

So we will compare between-group variability to within-group variability in ANOVA.

Types:

(1) One-Way ANOVA

An ANOVA test that has only one independent variable [2]

Steps to perform this:

Step 1: Define Hypothesis

Step 2: Calculate the sum of squares

Step 3: Calculate Degrees of Freedom

Step 4: Calculate F-Value

Step5: Accept or reject hypothesis

(2) Two-Way ANOVA

An ANOVA test that has two independent variable

(3)n-Way ANOVA

An ANOVA test that has more than two independent variable

Equations:

We can write the formula of F-score for One-way ANOVA test as following

$$F = \frac{SS_B/(k-1)}{SS_W/(n-k)}$$

Where, $SS_B = \sum_i n_i(\overline{y}_i - \overline{y})^2$ and $SS_W = \sum_{ij}(\overline{y}_{ij} - \overline{y})^2$

Where ,

y_i = sample mean in the i^{th} group

n_i = number of observations in the i^{th} group

y = total mean of dataset

k = total number of groups

$y_{ij} = j^{th}$ observation in the out of k group

N = overall sample size [3]

III. SMOTE

SMOTE stands for Synthetic Minority Oversampling Technique. If a dataset contains too few examples of the minor class then we can use this method to oversample the dataset. SMOTE works by selecting instances in the feature space that are close together, drawing a line between the examples, and drawing a new sample at a position along that line. [4]

Steps in SMOTE:

Step 1: Minority group Set A is completed, and the k-closest neighbors of x are obtained by calculating the Euclidean distance between x and each example in set A.

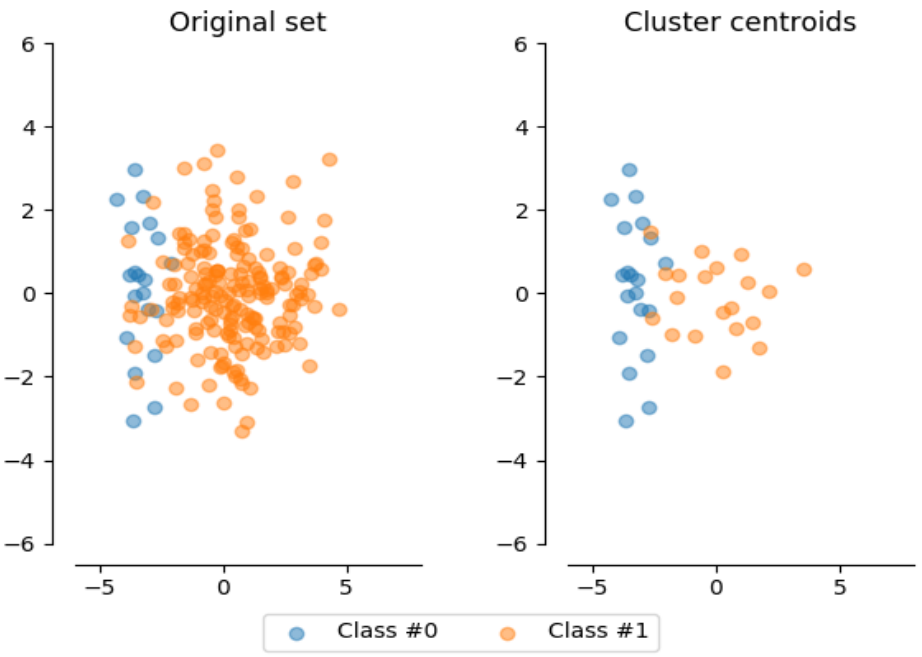
Step 2: The unbalanced extent determines the testing rate N. For each, N models (x1, x2,... xn) are arbitrarily selected from their k-closest neighbors and used to construct the set.

Step 3: The following equation is used to generate another model for each model (k= 1, 2, 3,...N). rand(0, 1) is used to address an irregular number between 0 and 1.[4]

IV. Cluster Centroids

We can use this method to undersample the dataset if we have an imbalanced dataset. This methodology generates a new set based on centroids using clustering methods, which results in undersampling. The algorithm creates a new set based on the cluster centroid of a KMeans algorithm. A method for undersampling the majority class by replacing the cluster centroid of a KMeans algorithm with a cluster of majority samples. Instead of the original samples, the newly produced set is synthesized

using the centroids of the K-means approach. It mainly changes the majority classes. [5]

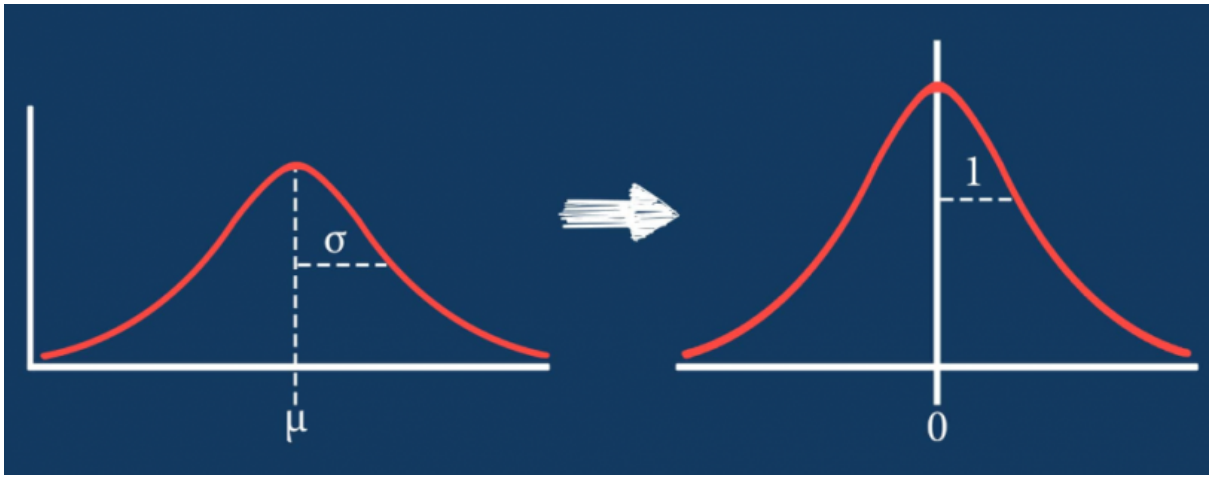


V. Feature Scaling

We often encounter different types of variables in the same dataset. A significant issue is that the range of the variables may differ a lot. Using the original scale may put more weight on the variables with a large range. In order to deal with this problem, we need to apply the technique of feature rescaling to independent variables or features of data in the step of data pre-processing. The terms **normalization** and **standardization** are sometimes used interchangeably, but they usually refer to different things. [6]

The goal of applying Feature Scaling is to make sure features are on almost the same scale so that each feature is equally important and makes it easier to process by most ML algorithms.

Standardisation (Z-score Normalization)	Max-Min Normalization
$x_{\text{stand}} = \frac{x - \text{mean}(x)}{\text{standard deviation } (x)}$	$x_{\text{norm}} = \frac{x - \min(x)}{\max(x) - \min(x)}$



This technique to re-scale features value with the distribution value between 0 and 1 is useful for the optimization algorithms, such as gradient descent, that are used within machine learning algorithms that weight inputs (e.g., regression and neural networks). Rescaling is also used for algorithms that use distance measurements, for example, K-Nearest-Neighbors (KNN). [6]

5. Heart Disease Prediction

In this chapter, the problem, the procedure and the algorithms will be explained along with the results obtained in each case.

5.1. DESCRIPTION OF THE PROBLEM

I. WHAT IS THE PROBLEM?

Before coming to the solution of the project, I will just briefly describe what the problem is.
The problem that I am trying to solve is actually a heart disease prediction.
I have a dataset which contains 10 years of historical data about patients which are either affected by heart disease or not affected by heart disease in the next 10 years.
Now the task is to create a perfect classification model which is a good fit at this dataset and predict the chance of heart disease in the next 10 years for a patient.

II. HOW IS THE PROBLEM GOING TO SOLVE?

Now, the question is, how do I solve the problem.
The problem is based on predicting the chance of heart disease of a patient in the next 10 years. So, It is actually a classification problem.
But, here I don't use any classification algorithm to solve this problem.
To classify the problem here I use clustering.

Methodology behind my solution

First, I am going to split the dataset into a train dataset and a test dataset.

Afterwards, I use the K-Means clustering algorithm for segmenting the training dataset and mark the clusters as diseased clusters which have more diseased clusters and similarly for the non-diseased clusters .

Then, I take each point of the test dataset to find the minimum distance from the clusters. If the point is near to a diseased cluster then the point is referred to as a diseased point and if the point is near to a non diseased cluster then the point is referred to as a non diseased point.

The above process is going through multiple stages.
After analyzing all the results of all stages the final model of classification which is best fit for the dataset will be decided.

Finally, we are at a solution.

III. EVALUATION MODEL AND EVALUATION METRICS USED FOR RESULT SUBMISSION

I use the **train-test split** as the preferred Evaluation Model for this project.

I use the **Euclidean Distance** as a Distance Metric.

Submissions are evaluated using the **F1-Score** and **Jaccard Index**.

*** I described all the evaluation metrics and evaluation models in depth previously.**

5.2. DATASET ANALYSIS

The dataset [7] has different “features”, each one provides information for each record.
The dataset contains,
Number of Records : 4238
Number of Features : 16

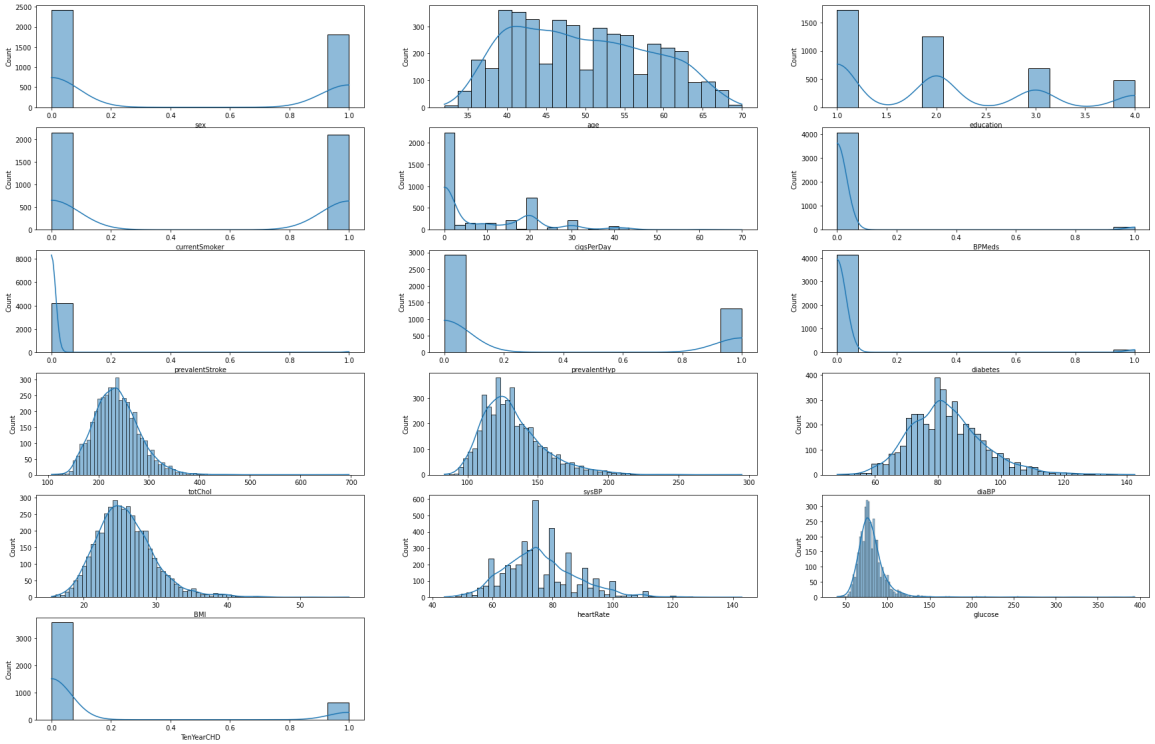
3594 of them are indicating to non diseased records and 644 of them are indicating to diseased records.

EACH RECORD CONTAINS THE FOLLOWING INFORMATIONS

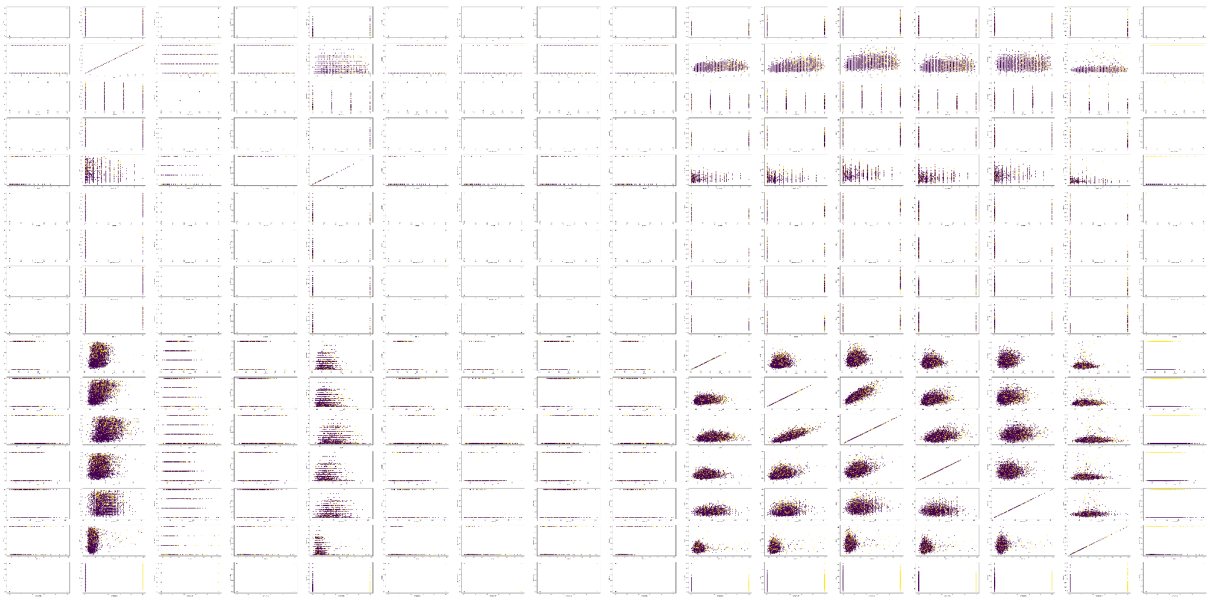
NAME OF THE FEATURE	NON-NULL COUNT	DATA TYPE	DESCRIPTION	VALUE TYPE (1 => TRUE, 0 => FALSE)
sex	4238	INT64	GENDER	CATEGORICAL (0/1)
age	4238	INT64	AGE	NUMERICAL
education	4133	FLOAT64	EDUCATION	CATEGORICAL (1.0/2.0/3.0 /4.0)
currentSmoker	4238	INT64	CURRENT SMOKER STATUS	CATEGORICAL (0/1)
cigsPerDay	4209	FLOAT64	CIGARETTES PER DAY CONSUMED BY THE PATIENT	NUMERICAL
BPMeds	4185	FLOAT64	BLOOD PRESSURE MEDICINE STATUS	CATEGORICAL (0.0/1.0)
prevalentStoke	4238	INT64	PREVALENT STOKE STATUS	CATEGORICAL (0/1)
prevalentHyp	4238	INT64	PREVALENT HYPERTENSION STATUS	CATEGORICAL (0/1)
diabetes	4238	INT64	PREVALENT DIABETES STATUS	CATEGORICAL (0/1)
totChol	4188	FLOAT64	TOTAL CHOLESTEROL LEVEL	NUMERICAL
sysBP	4238	FLOAT64	SYSTOLIC BLOOD PRESSURE	NUMERICAL
diaBP	4238	FLOAT64	DIASTOLIC BLOOD PRESSURE	NUMERICAL
BMI	4219	FLOAT64	BODY MASS INDEX	NUMERICAL
heartRate	4237	FLOAT64	AVERAGE HEART RATE	NUMERICAL
glucose	3850	FLOAT64	GLUCOSE LEVEL	NUMERICAL
TenYearCHD	4238	INT64	THE CHANCE OF HEART DISEASE IN	CATEGORICAL (0/1)

NAME OF THE FEATURE	NON-NULL COUNT	DATA TYPE	DESCRIPTION	VALUE TYPE (1 => TRUE, 0 => FALSE)
			THE NEXT 10 YEARS	

THE ABOVE INFORMATION PROJECTED INTO THE NEXT PICTURE.



THE POSITIONS OF THE DATA POINTS ACCORDING TO EACH FEATURES PROJECTED INTO THE NEXT PICTURE.



MISSING VALUES IN THE DATASET

feature	number of missing values
sex	0
age	0
education	105
currentSmoker	0
cigsPerDay	29
BPMeds	53
prevalentStroke	0
prevalentHyp	0
diabetes	0
totChol	50
sysBP	0
diaBP	0
BMI	19
heartRate	1
glucose	388
TenYearCHD	0

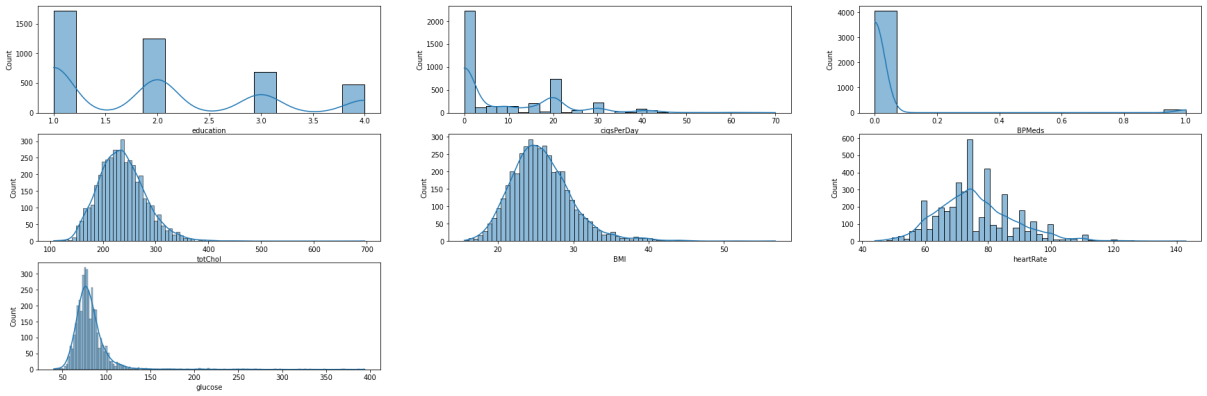
ALL ABOUT THE DATASET.

5.3. DATASET TREATMENT

I. MISSING VALUES FILLING

Here, I use the **Imputation method** for missing values filling.

*** I described all the used missing values filling methods in depth previously.**



According to the distribution of the data

feature	Central Tendency for filling the missing values
education	MODE
cigsPerDay	MEAN
BPMeds	MODE
totChol	MEDIAN
BMI	MEDIAN
heartRate	MEAN
glucose	MEDIAN

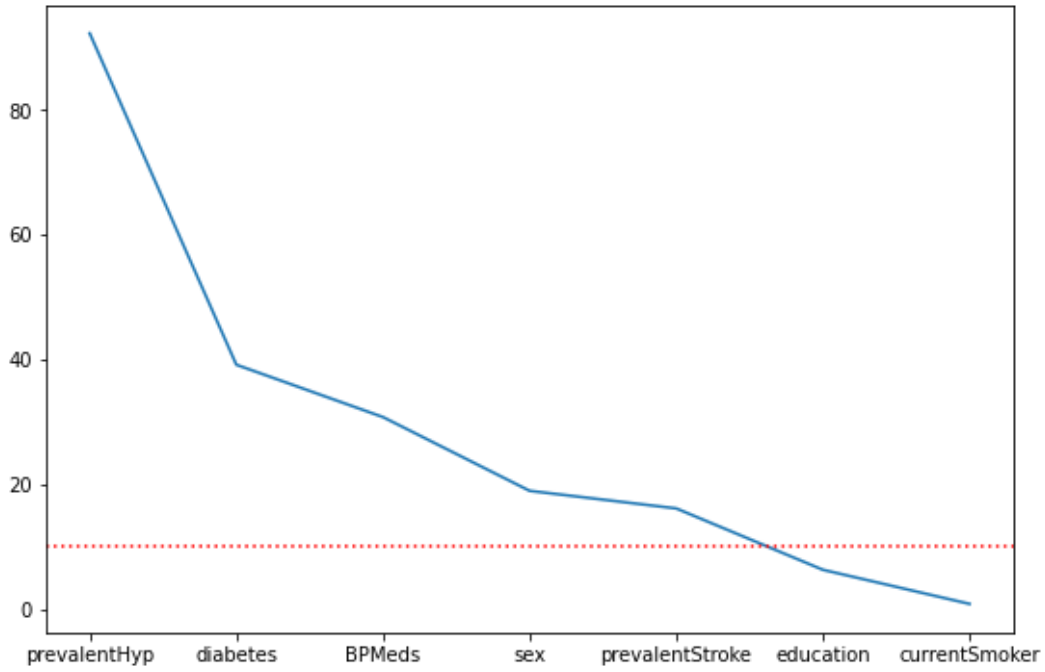
All the missing values are filled now.

II. FEATURES SELECTION

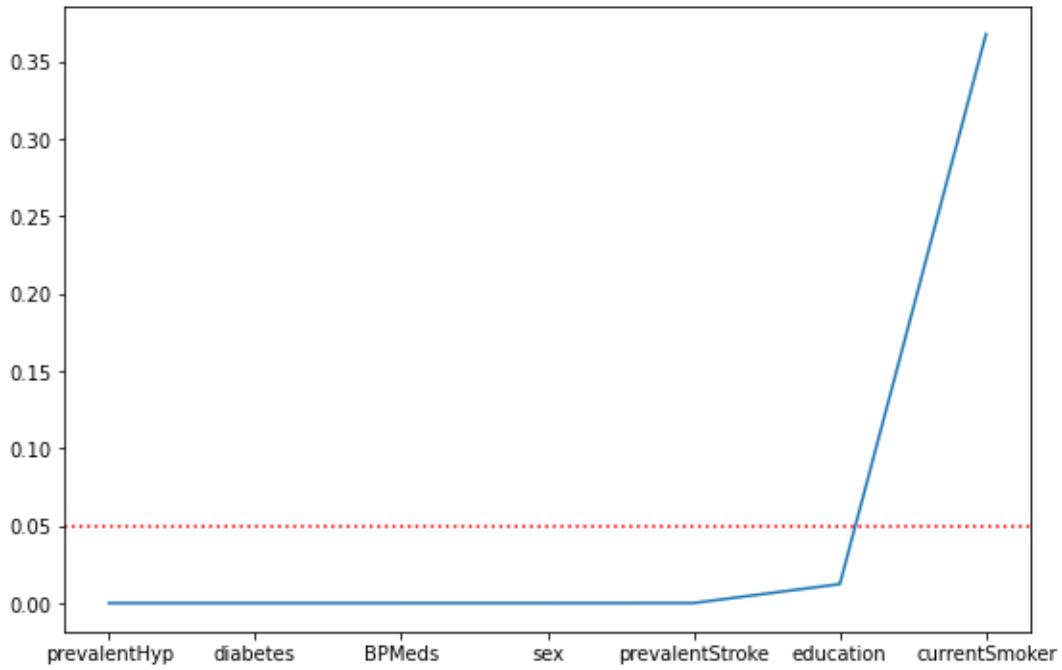
Here, I use the **Chi-Square** method for selecting categorical features and **Anova test** method for selecting numerical features features.

*** I described all the used feature selection methods in depth previously.**

CATEGORICAL FEATURE SELECTION USING CHI-SQUARE METHOD



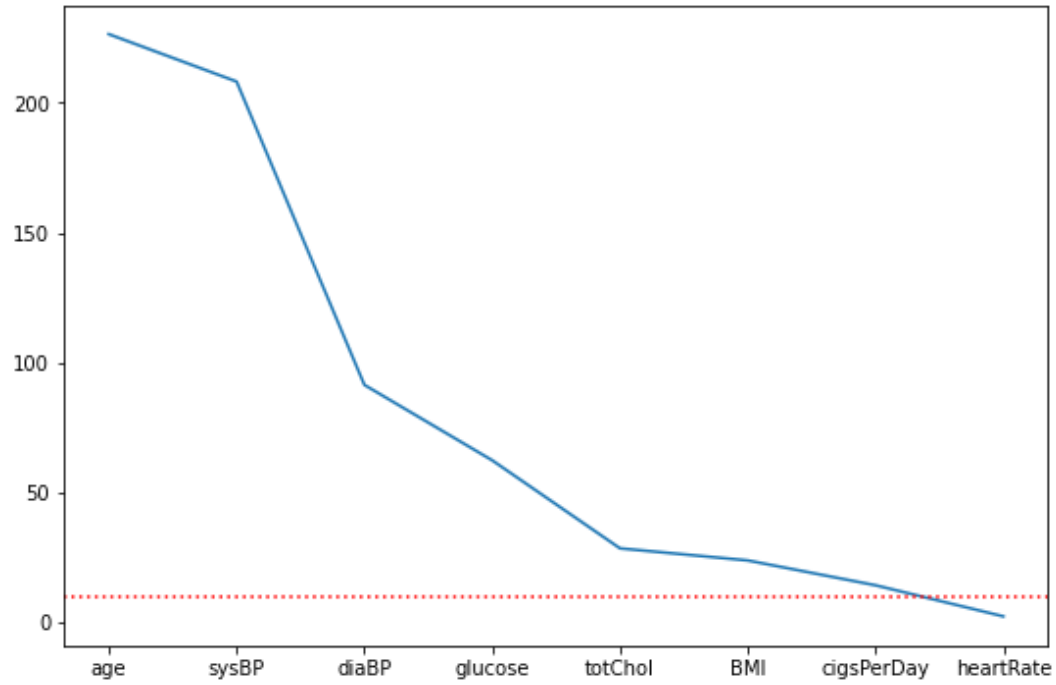
F-SCORE<5.0



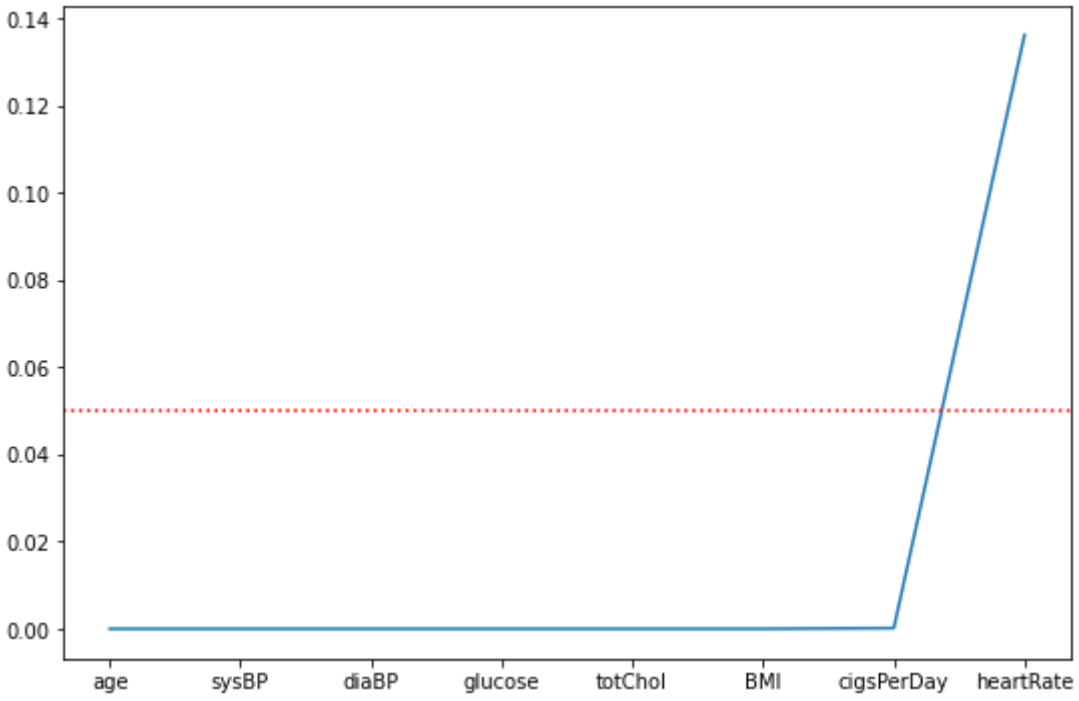
P-VALUE>0.05

SO, ACCORDING TO THE CHI-SQUARE FEATURE SELECTION METHOD "currentSmoker" FEATURE IS THE LEAST CATEGORICAL FEATURE.

CATEGORICAL FEATURE SELECTION USING CHI-SQUARE METHOD



F-SCORE<5.0



P-VALUE>0.05

SO, ACCORDING TO THE CHI-SQUARE FEATURE SELECTION METHOD “heartRate” FEATURE IS THE LEAST CATEGORICAL FEATURE.

FEATURE SELECTION IS COMPLETED.

5.4. TREATED DATASET ANALYSIS

The dataset contains,
Number of Records : 4238
Number of Features : 14

3594 of them are indicating to non diseased records and 644 of them are indicating to diseased records.

EACH RECORD CONTAINS THE FOLLOWING INFORMATIONS

NAME OF THE FEATURE	NON-NULL COUNT	DATA TYPE	DESCRIPTION	VALUE TYPE (1 => TRUE, 0 => FALSE)
sex	4238	INT64	GENDER	CATEGORICAL (0/1)
age	4238	INT64	AGE	NUMERICAL
education	4238	FLOAT64	EDUCATION	CATEGORICAL (1.0/2.0/3.0 /4.0)
cigsPerDay	4238	FLOAT64	CIGARETTES PER DAY CONSUMED BY THE PATIENT	NUMERICAL
BPMeds	4238	FLOAT64	BLOOD PRESSURE MEDICINE STATUS	CATEGORICAL (0.0/1.0)
prevalentStoke	4238	INT64	PREVALENT STOKE STATUS	CATEGORICAL (0/1)
prevalentHyp	4238	INT64	PREVALENT HYPERTENSION STATUS	CATEGORICAL (0/1)
diabetes	4238	INT64	PREVALENT DIABETES STATUS	CATEGORICAL (0/1)
totChol	4238	FLOAT64	TOTAL CHOLESTEROL LEVEL	NUMERICAL
sysBP	4238	FLOAT64	SYSTOLIC BLOOD PRESSURE	NUMERICAL
diaBP	4238	FLOAT64	DIASTOLIC BLOOD PRESSURE	NUMERICAL
BMI	4238	FLOAT64	BODY MASS INDEX	NUMERICAL

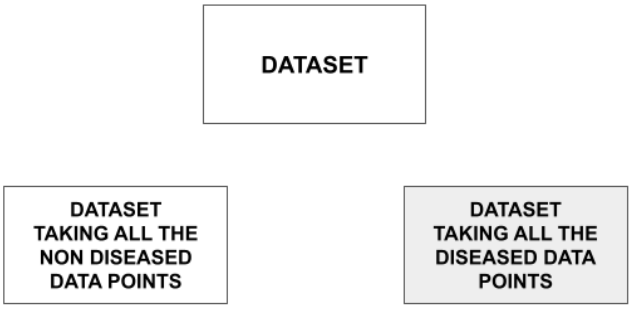
NAME OF THE FEATURE	NON-NULL COUNT	DATA TYPE	DESCRIPTION	VALUE TYPE (1 => TRUE, 0 => FALSE)
glucose	4238	FLOAT64	GLUCOSE LEVEL	NUMERICAL
TenYearCHD	4238	INT64	THE CHANCE OF HEART DISEASE IN THE NEXT 10 YEARS	CATEGORICAL (0/1)

5.5. CLASSIFICATION MODEL CREATION USING CLUSTERING AND RESULTS

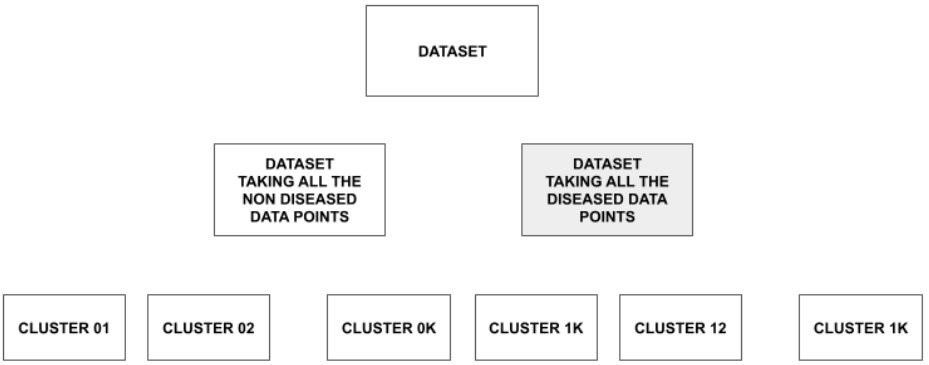
My approach to solve this problem is making a two layer tree structure clustering algorithm for segmenting the dataset as diseased segment and non-diseased segment to classify whether a data point is diseased or non-diseased.

The main methodology behind this approach is,

First , divide the dataset into two clusters. One cluster is taking all the non-diseased data points from the dataset and another cluster is taking all the diseased data points from the dataset.



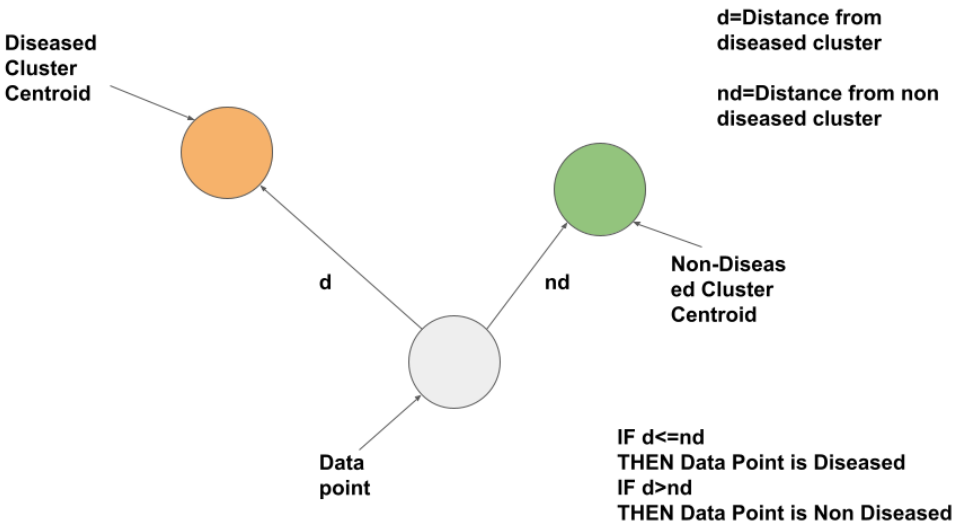
Then , apply the K-Means Clustering algorithm on both of the dataset.

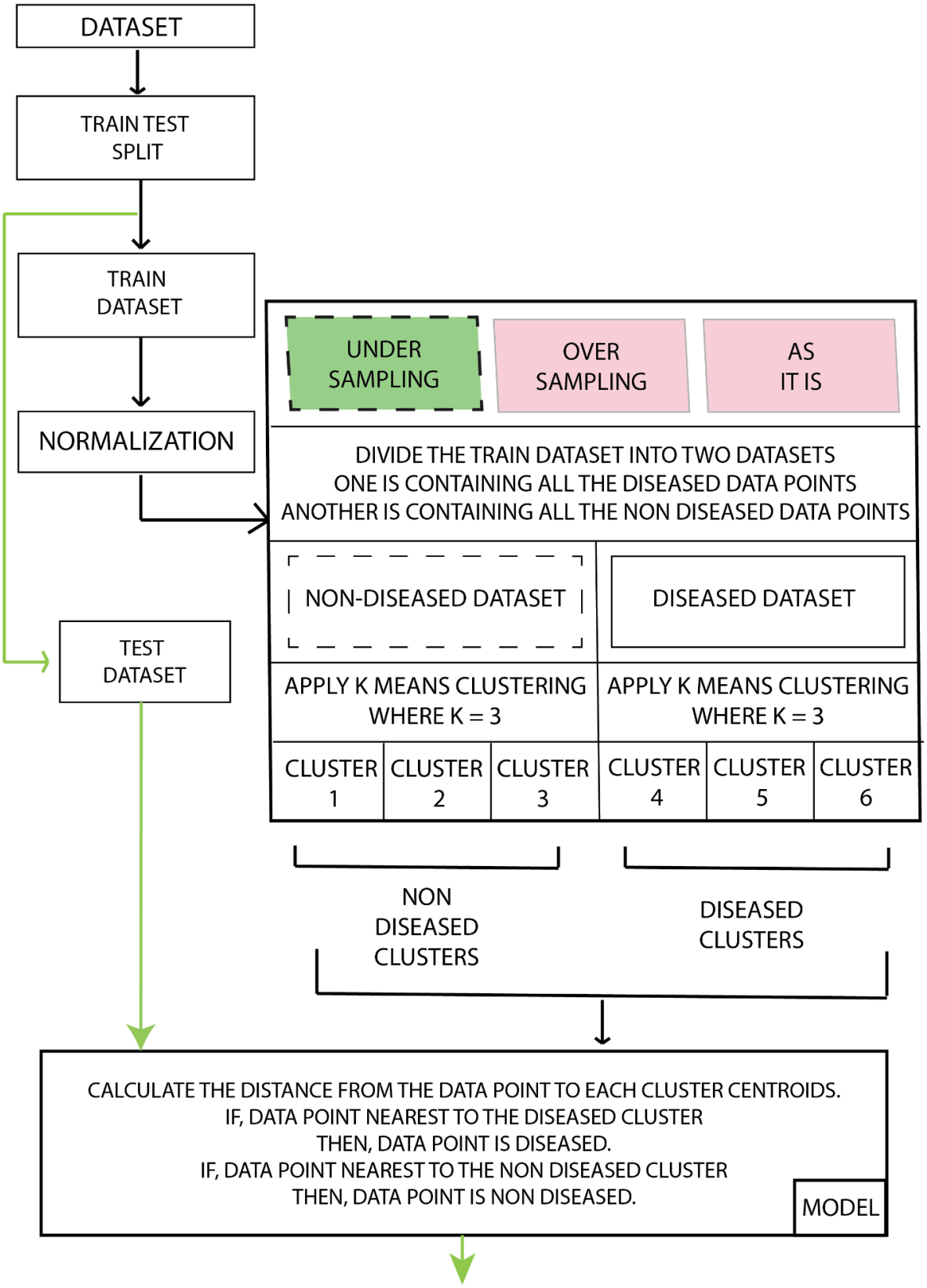


Here,
CLUSTER 01, CLUSTER 02, … , CLUSTER 0K are non diseased clusters.
CLUSTER 11, CLUSTER 12, … , CLUSTER 1K are non diseased clusters.

Find the best K value using the Elbow Method and assign it to K separately.

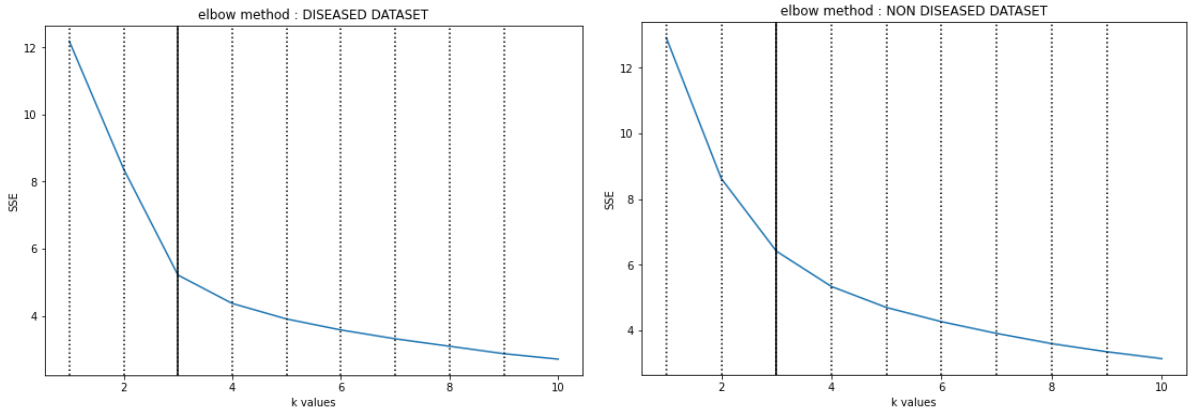
Finally, I need to calculate the Euclidean Distance from the data point to each cluster and take the nearest cluster to the point. If the nearest cluster of the data point belongs to the diseased cluster then the point is diseased and If the nearest cluster of the data point belongs to the non-diseased cluster then the point is non-diseased.




DATASET
TRAIN TEST SPLIT
TRAIN DATASET
NORMALIZATION
TEST DATASET
UNDER SAMPLING
OVER SAMPLING
AS IT IS
DIVIDE THE TRAIN DATASET INTO TWO DATASETS
ONE IS CONTAINING ALL THE DISEASED DATA POINTS
ANOTHER IS CONTAINING ALL THE NON DISEASED DATA POINTS
NON-DISEASED DATASET
DISEASED DATASET
APPLY K MEANS CLUSTERING WHERE K = 3
APPLY K MEANS CLUSTERING WHERE K = 3
CLUSTER 1
CLUSTER 2
CLUSTER 3
CLUSTER 4
CLUSTER 5
CLUSTER 6
NON DISEASED CLUSTERS
DISEASED CLUSTERS
CALCULATE THE DISTANCE FROM THE DATA POINT TO EACH CLUSTER CENTROIDS.
IF, DATA POINT NEAREST TO THE DISEASED CLUSTER
THEN, DATA POINT IS DISEASED.
IF, DATA POINT NEAREST TO THE NON DISEASED CLUSTER
THEN, DATA POINT IS NON DISEASED.
MODEL

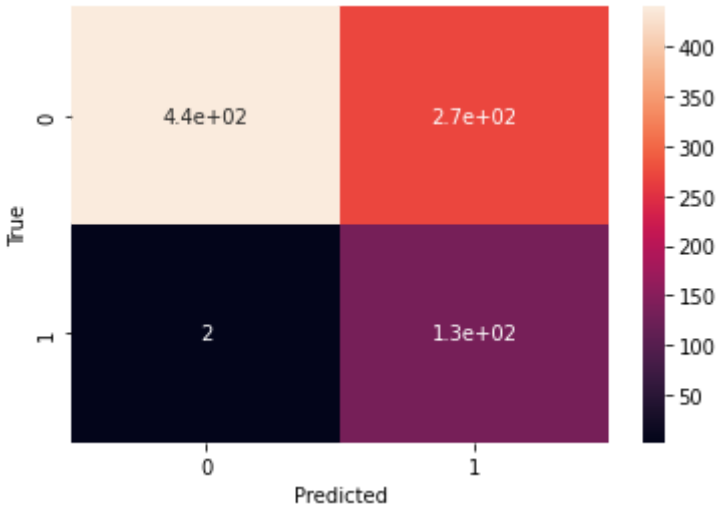
APPROACH 1

BEST K FOR BOTH NON DISEASED AND DISEASED DATASET



F1 SCORE

	PRECISION	RECALL	F1-SCORE	SUPPORT
0 (NON-DISEASED)	1.00	0.62	0.76	712
1 (DISEASED)	0.33	0.99	0.49	136
ACCURACY			**0.68**	848
MACRO AVG	0.66	0.80	0.63	848
WEIGHTED AVG	0.89	0.68	0.72	848

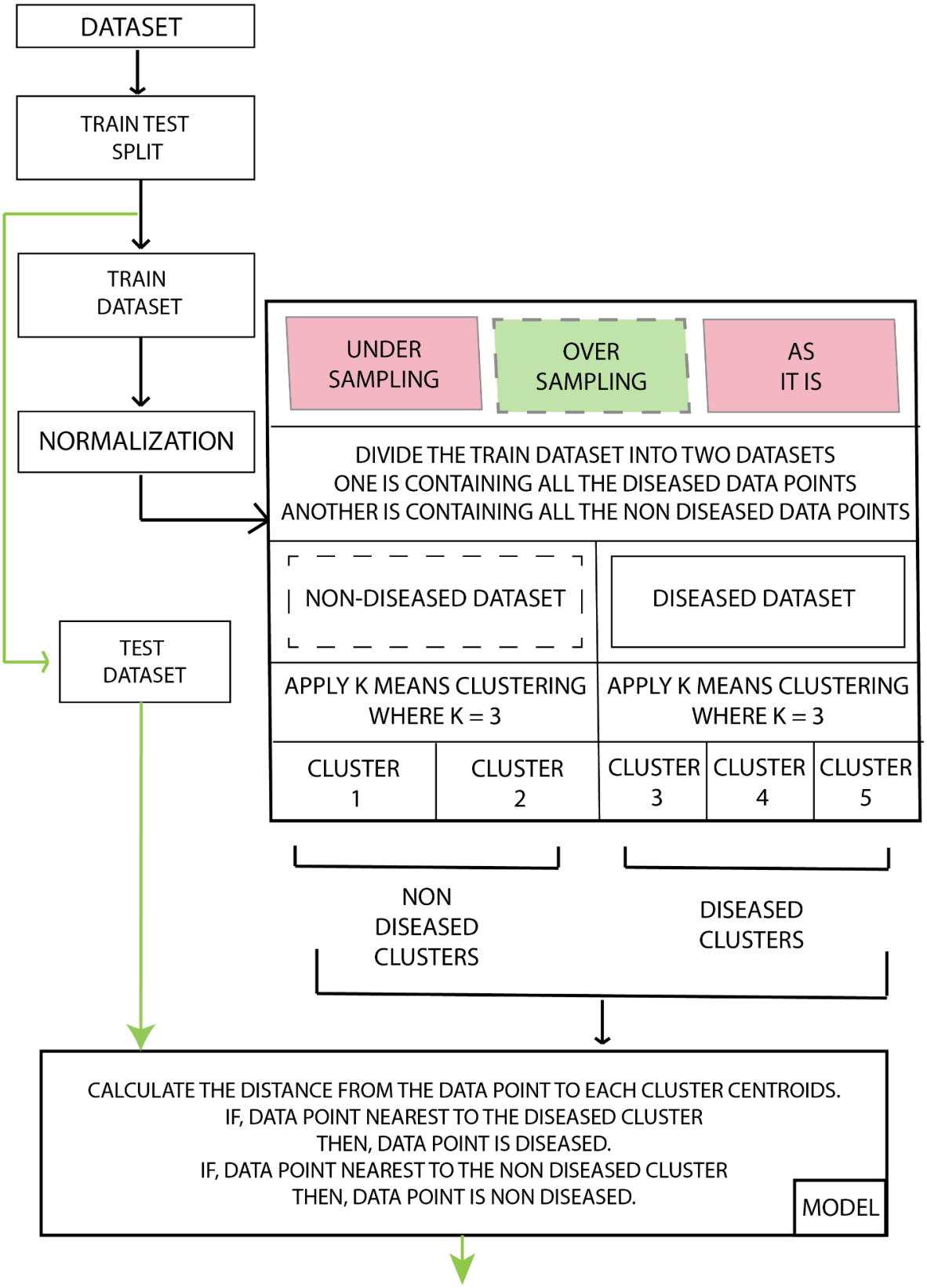


CONFUSION MATRIX

JACCARD SCORE

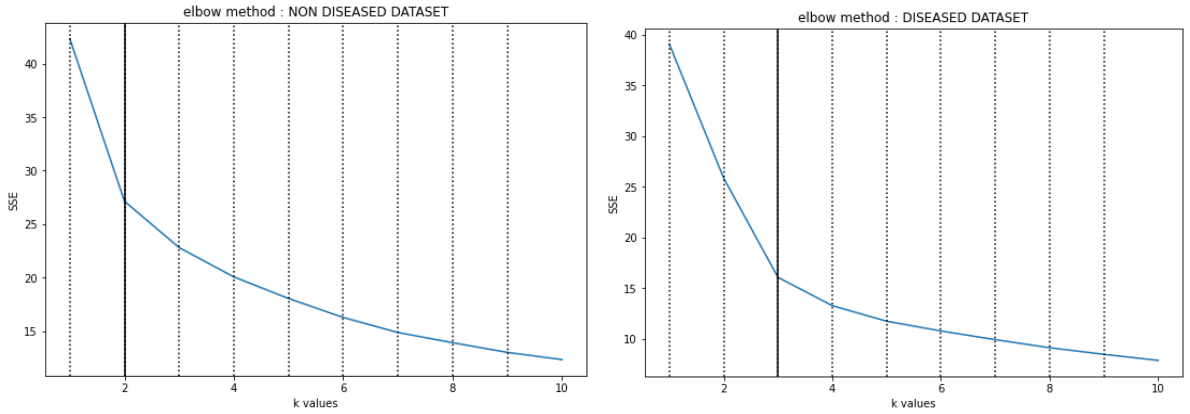
	JACCARD SCORE
0 (NON-DISEASED)	0.62

1 (DISEASED)	0.33


DATASET
TRAIN TEST SPLIT
TRAIN DATASET
NORMALIZATION
TEST DATASET
UNDER SAMPLING
OVER SAMPLING
AS IT IS
DIVIDE THE TRAIN DATASET INTO TWO DATASETS
ONE IS CONTAINING ALL THE DISEASED DATA POINTS
ANOTHER IS CONTAINING ALL THE NON DISEASED DATA POINTS
NON-DISEASED DATASET
DISEASED DATASET
APPLY K MEANS CLUSTERING WHERE K = 3
APPLY K MEANS CLUSTERING WHERE K = 3
CLUSTER 1
CLUSTER 2
CLUSTER 3
CLUSTER 4
CLUSTER 5
NON DISEASED CLUSTERS
DISEASED CLUSTERS
CALCULATE THE DISTANCE FROM THE DATA POINT TO EACH CLUSTER CENTROIDS.
IF, DATA POINT NEAREST TO THE DISEASED CLUSTER
THEN, DATA POINT IS DISEASED.
IF, DATA POINT NEAREST TO THE NON DISEASED CLUSTER
THEN, DATA POINT IS NON DISEASED.
MODEL

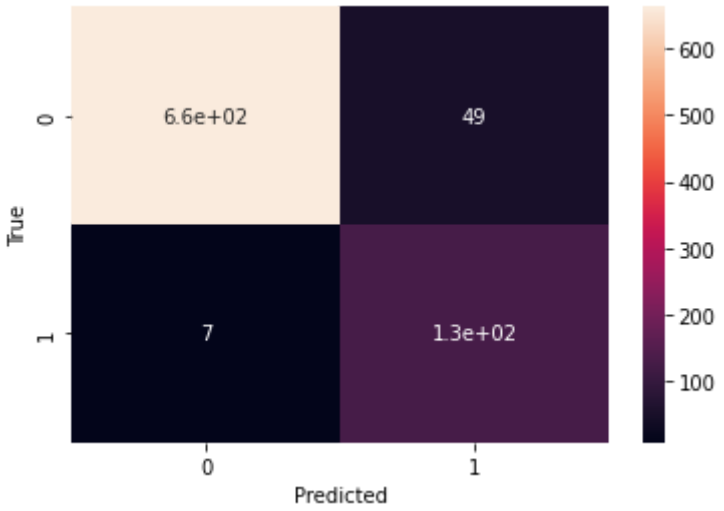
APPROACH 2

BEST K FOR BOTH NON DISEASED AND DISEASED DATASET



F1 SCORE

	PRECISION	RECALL	F1-SCORE	SUPPORT
0 (NON-DISEASED)	0.99	0.93	0.96	712
1 (DISEASED)	0.72	0.95	0.82	136
ACCURACY			**0.93**	848
MACRO AVG	0.86	0.94	0.89	848
WEIGHTED AVG	0.95	0.93	0.94	848

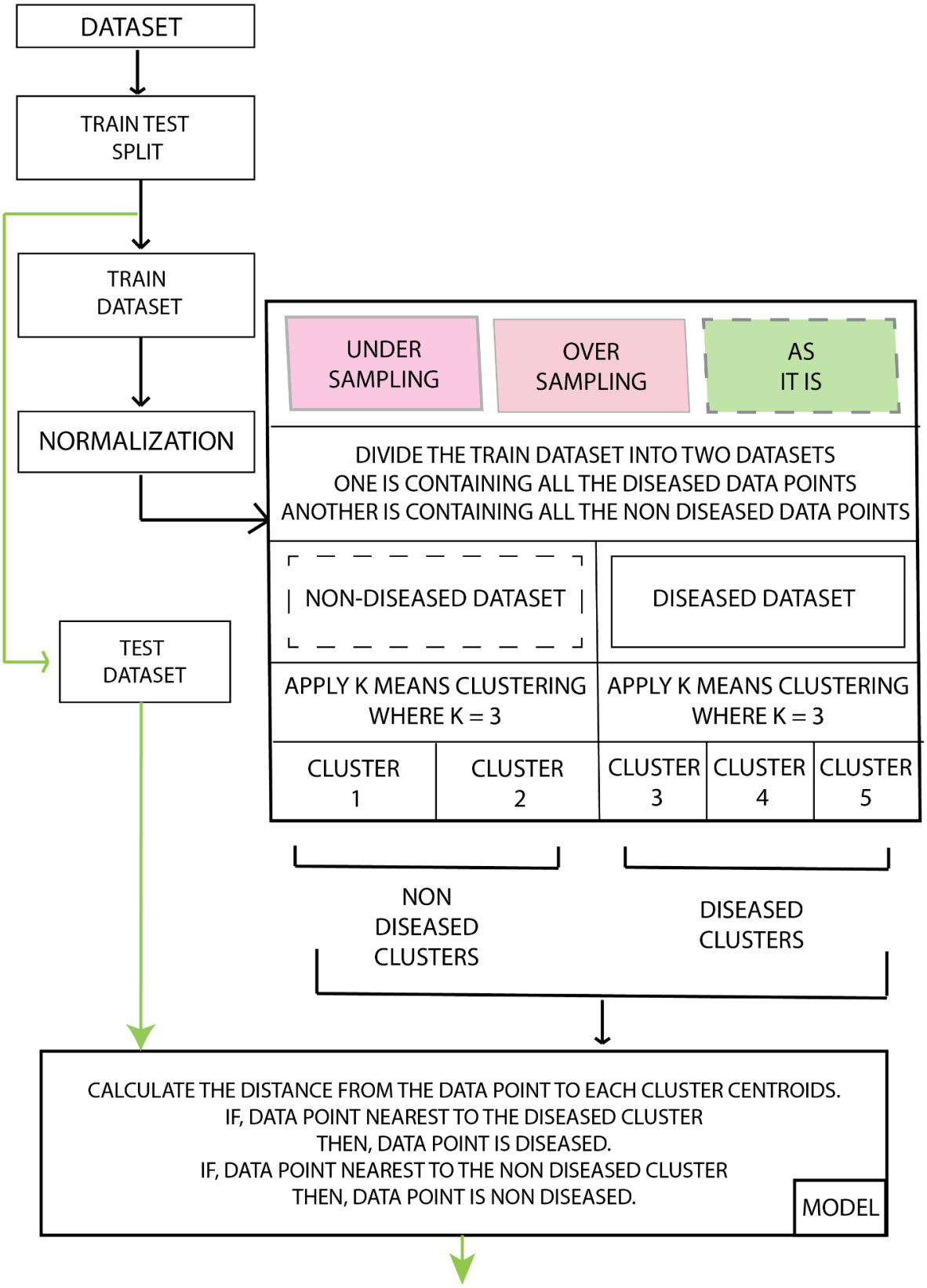


CONFUSION MATRIX

JACCARD SCORE

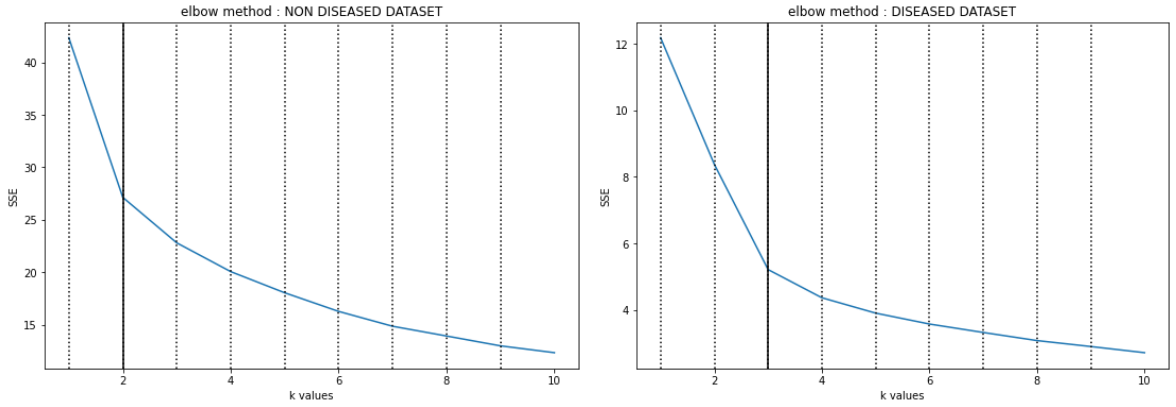
	JACCARD SCORE
0 (NON-DISEASED)	0.93

1 (DISEASED)	0.70


DATASET
TRAIN TEST SPLIT
TRAIN DATASET
NORMALIZATION
TEST DATASET
UNDER SAMPLING
OVER SAMPLING
AS IT IS
DIVIDE THE TRAIN DATASET INTO TWO DATASETS
ONE IS CONTAINING ALL THE DISEASED DATA POINTS
ANOTHER IS CONTAINING ALL THE NON DISEASED DATA POINTS
NON-DISEASED DATASET
DISEASED DATASET
APPLY K MEANS CLUSTERING WHERE K = 3
APPLY K MEANS CLUSTERING WHERE K = 3
CLUSTER 1
CLUSTER 2
CLUSTER 3
CLUSTER 4
CLUSTER 5
NON DISEASED CLUSTERS
DISEASED CLUSTERS
CALCULATE THE DISTANCE FROM THE DATA POINT TO EACH CLUSTER CENTROIDS.
IF, DATA POINT NEAREST TO THE DISEASED CLUSTER
THEN, DATA POINT IS DISEASED.
IF, DATA POINT NEAREST TO THE NON DISEASED CLUSTER
THEN, DATA POINT IS NON DISEASED.
MODEL

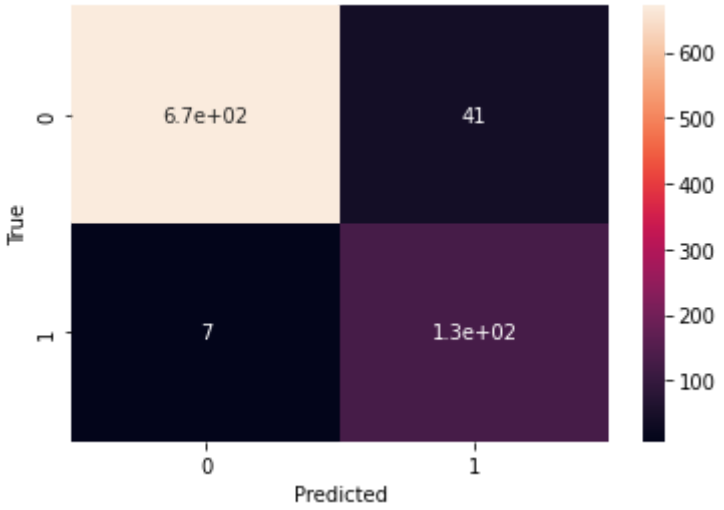
APPROACH 3

BEST K FOR BOTH NON DISEASED AND DISEASED DATASET



F1 SCORE

	PRECISION	RECALL	F1-SCORE	SUPPORT
0 (NON-DISEASED)	0.99	0.94	0.97	712
1 (DISEASED)	0.76	0.95	0.84	136
ACCURACY			**0.94**	848
MACRO AVG	0.87	0.95	0.90	848
WEIGHTED AVG	0.95	0.94	0.95	848



CONFUSION MATRIX

JACCARD SCORE

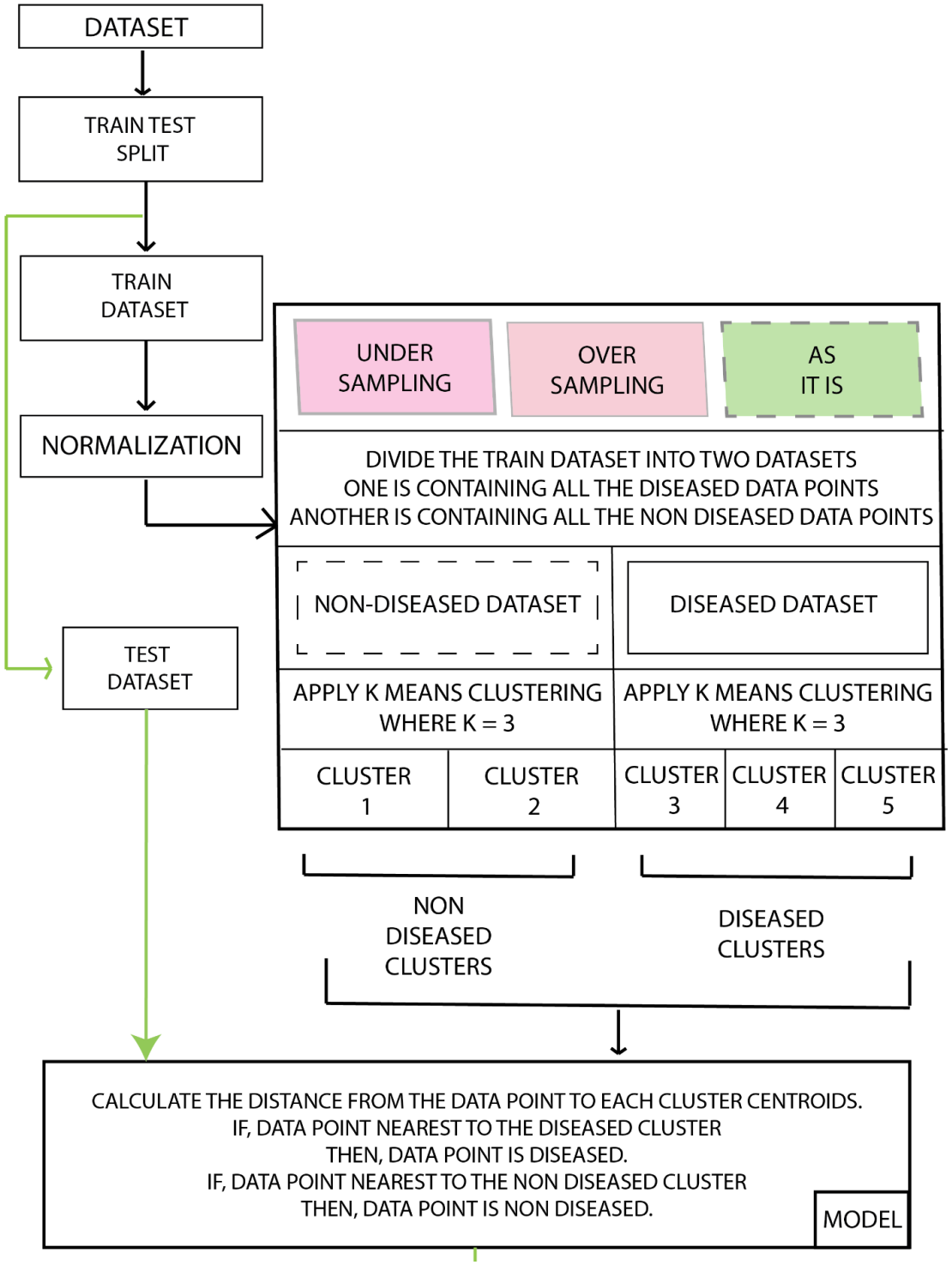
	JACCARD SCORE
0 (NON-DISEASED)	0.94

1 (DISEASED)	0.73

RESULT ANALYSIS

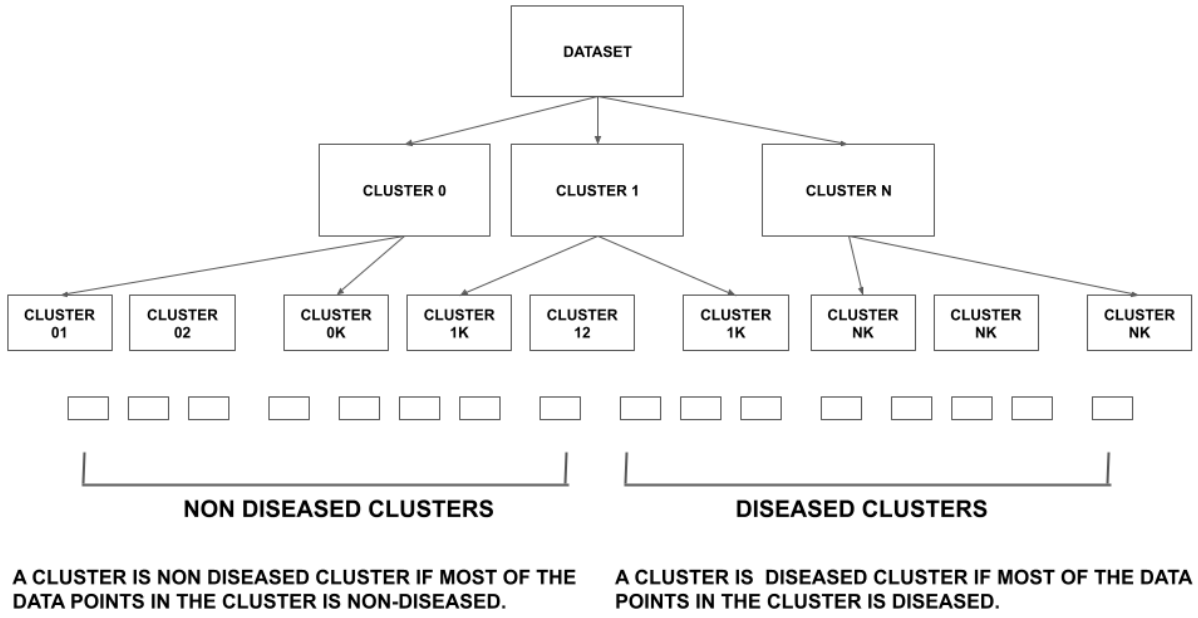
By analyzing all the above approaches and those results, we can say the best fitted model for this dataset according to the Evaluation Metrics is:

APPROACH 3



CONCLUSIONS AND FUTURE DEVELOPMENT

By analyzing all the approaches and the results, this is a valuable step to building a tree-like clustering structure for segmenting the data with minimum noise. Tree-like Clustering means clustering over the clusters in a hierarchical manner.



Using this methodology we can create the K-Means algorithm more powerful. So, we can find more reliable and trustable data segments.

BIBLIOGRAPHY

1. https://towardsdatascience.com/chi-square-test-for-feature-selection-in-machine-learning-206b1f0b8223
2. https://towardsdatascience.com/anova-for-feature-selection-in-machine-learning-d9305e228476
3. https://www.javatpoint.com/anova-test-in-python
4. https://www.geeksforgeeks.org/ml-handling-imbalanced-data-with-smote-and-near-miss-algorithm-in-python/
5. https://hersanyagci.medium.com/under-sampling-methods-for-imbalanced-data-clustercentroids-randomundersampler-nearmiss-eae0eadcc145
6. https://www.linkedin.com/pulse/standardization-machine-learning-sachin-vinay/
7. https://www.kaggle.com/datasets/dileep070/heart-disease-prediction-using-logistic-regression
8. https://en.wikipedia.org/wiki/ML
9. https://www.coursera.org/learn/machine-learning-with-python/home/