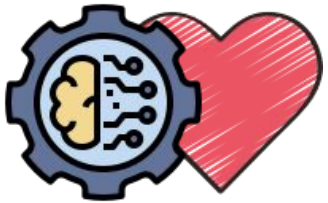
CLASSIFICATION USING K-MEANS CLUSTERING APPLIED TO HEART DISEASE PREDICTION



By Niladri Das

Under the supervision of Dr. Sanjoy Kumar Saha

INTRODUCTION

MACHINE LEARNING

CLASSIFICATION

CLUSTERING

EVALUATION MODELS FOR CLASSIFICATION

EVALUATION METRICS FOR CLASSIFICATION

FEATURE SELECTION METHODS

OVERSAMPLING METHODS

UNDERSAMPLING METHODS

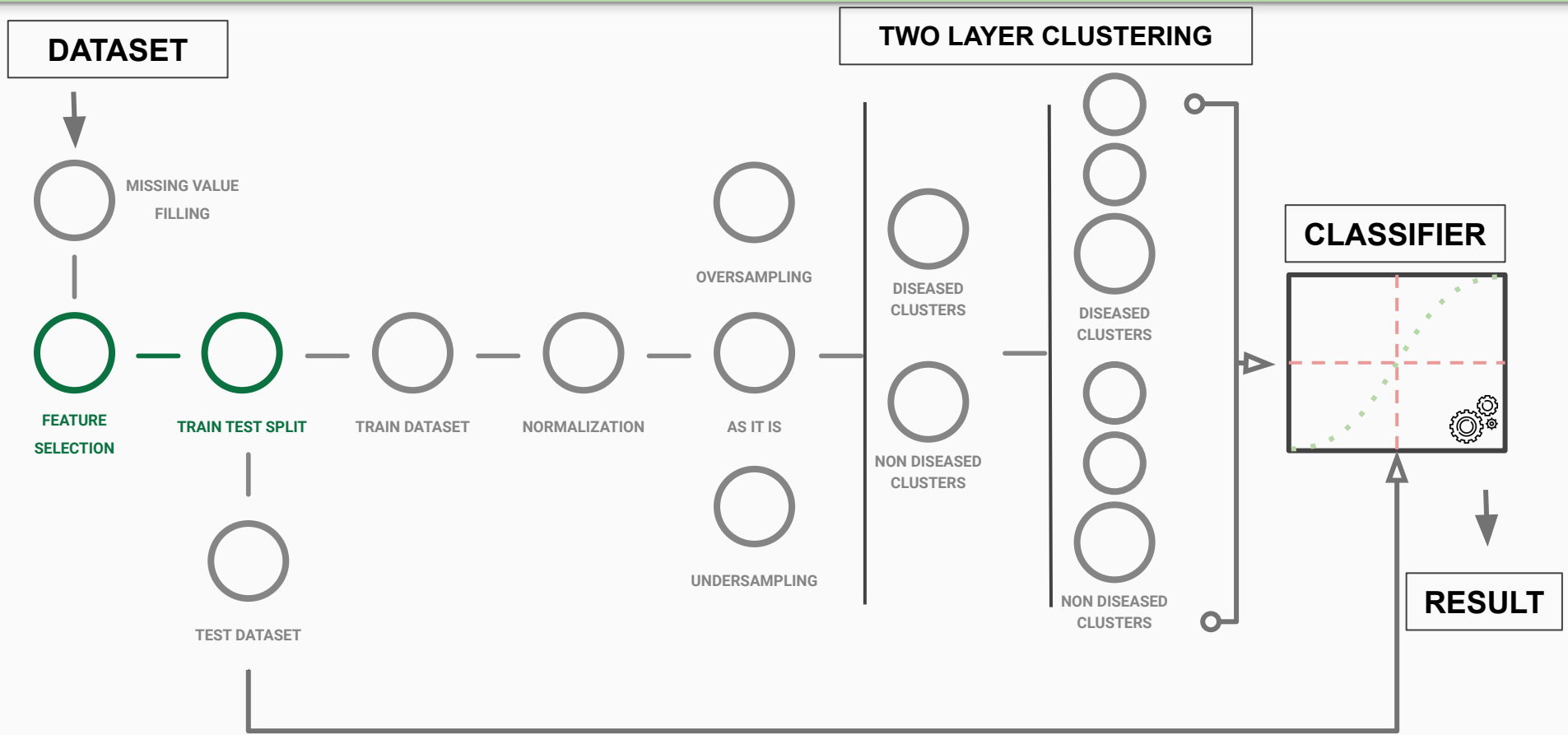
FEATURE SCALING METHODS

PROPOSED WORK

INTRODUCTION

As per name this is a project on the subtopics of Machine Learning.
So, in this presentation first I will discuss about the Machine Learning and its subtopics like classification and clustering.
Actually, this is a classification problem solved by segmenting the dataset using clustering algorithms.
Then, I will discuss about the Evaluation models and the Evaluation metrics.
Next, I will decide which Evaluation model and Evaluation metrics best fit for my project work.
After that, I will discuss about some special techniques used to preprocess the dataset in this project.
Finally, I jump to the workflow of the project and discuss about its results.

WORKFLOW OF THE PROJECT



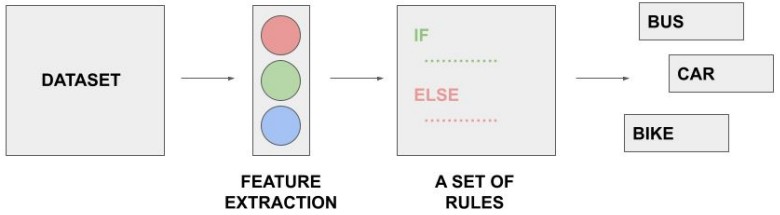
MACHINE LEARNING

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.

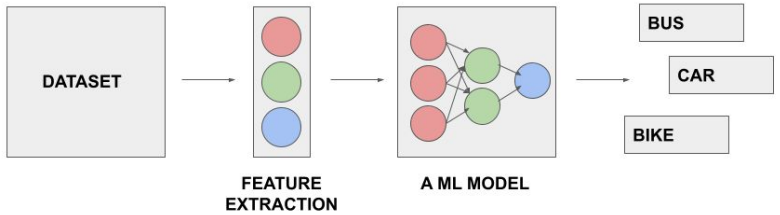
What is Machine Learning?

NORMAL PROGRAMING V/S ML

IN NORMAL PROGRAMMING

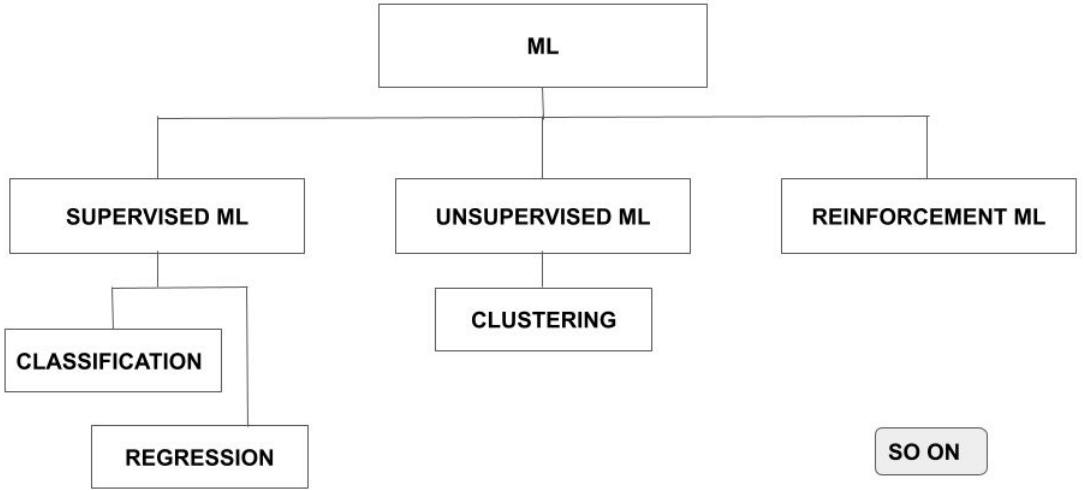


IN MACHINE LEARNING



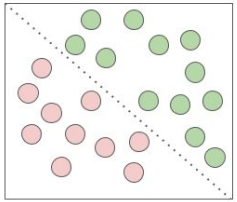
TYPES OF MACHINE LEARNING

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

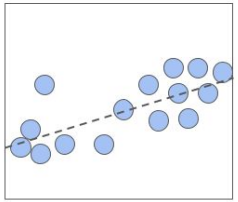


SUPERVISED ML

UNSUPERVISED ML

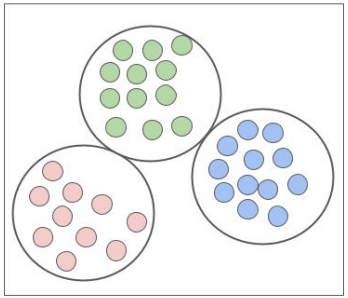


CLASSIFICATION
Classification is the process of predicting discrete class labels or categories.



REGRESSION
Regression is the process of predicting continuous values.

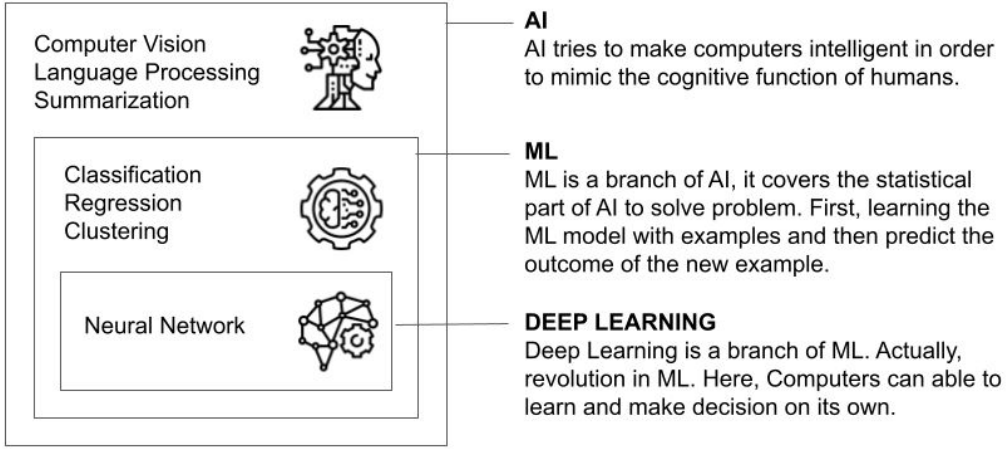
SUPERVISED ML V/S UNSUPERVISED ML



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

RELATIONSHIP OF ML WITH AI AND DEEP LEARNING

Machine Learning is a milestone to reach artificial intelligence and deep learning.



CLASSIFICATION

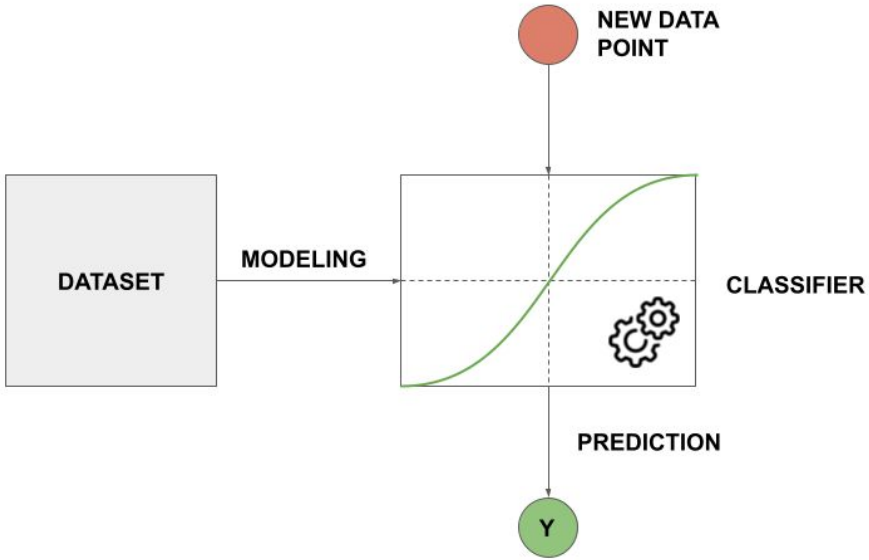
Classification is a special type of supervised Machine Learning approach, which is used to categorize some unknown items into a discrete set of categories (i.e., categorical values). Classification first attempts to learn the relationship between a set of feature variables and a target variable of interest and then determines the class label for an unlabeled test case.

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).

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	?

CLASSIFICATION MODEL

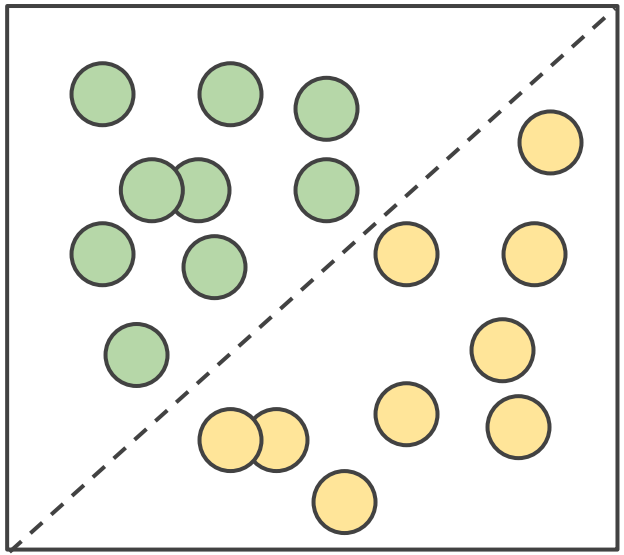
Now, we need to create a classification model which learns the relationship between the set of features and the target variable.



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

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.

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



VARIOUS CLASSIFICATION ALGORITHMS

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.

APPLICATIONS OF CLASSIFICATION

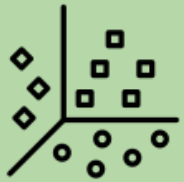
There are many uses of classifications in the real world.

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.

CLUSTERING

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. In simple words, 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.

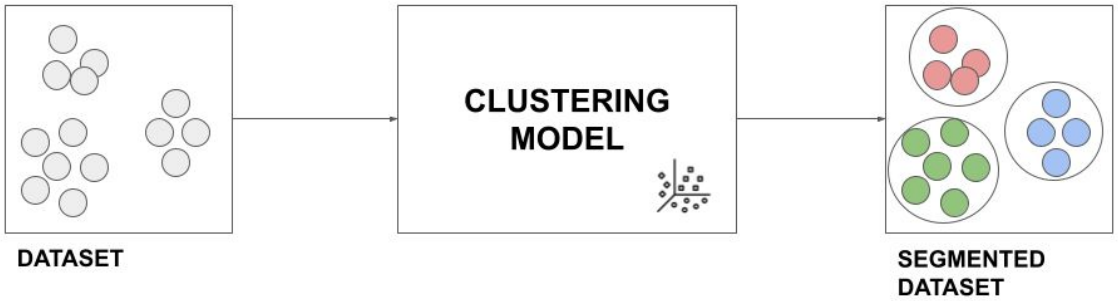


Suppose we have a patient dataset.
We have to find which patient is in which category according to their age and weight.

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

CLUSTERING MODEL

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



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

The segmented data given by the model.

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.

VARIOUS CLUSTERING ALGORITHMS

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 (down-up), Divisive Algorithms (up-down).

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 (Density Based Spatial Clustering Algorithm).

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

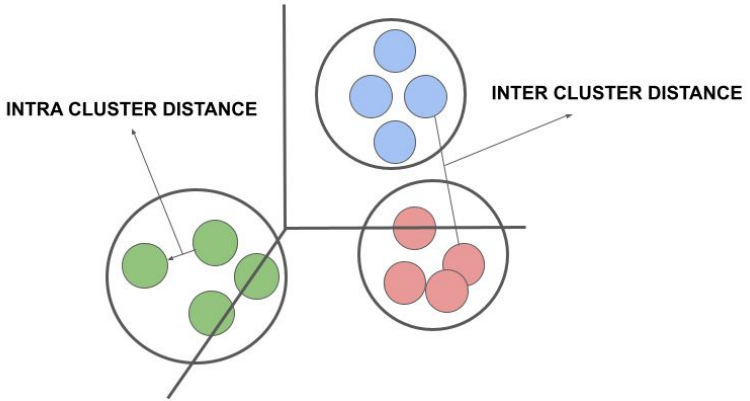
K MEANS CLUSTERING ALGORITHM

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.



DISTANCE METRIC

Here, I use Euclidean distance as distance metric to calculate the distance from one data point to another.

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)

CENTROID

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)

ACCURACY

To know how accurate our clustering model, 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.

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

Where,
xi indicates to each data points
k indicates to number of clusters
n indicates to number of data points
cj indicates to cluster centroid of j th cluster

STEPS OF K MEANS CLUSTERING ALGORITHM

I. Initialize K value and randomly/manually place K centroids, one for each cluster.

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

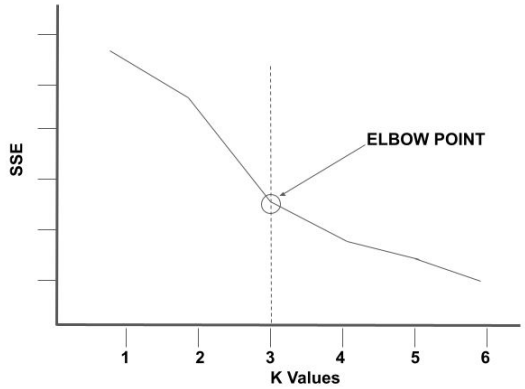
III. Assign each data point to its closest centroid and create a cluster.

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.

ELBOW METHOD



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

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 Elbow Method.

OBSERVATIONS ABOUT THE K MEANS CLUSTERING 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. **K-means is an iterative algorithm.**
c. **It is relatively efficient for medium and large size datasets.**
d. **It produces sphere-like clusters because the clusters are shaped around the centroids.**
e. **Its drawback is that we should pre specify the number of clusters, and this is not an easy task.**

— THIS CONCLUDES THE K MEANS CLUSTERING ALGORITHM.

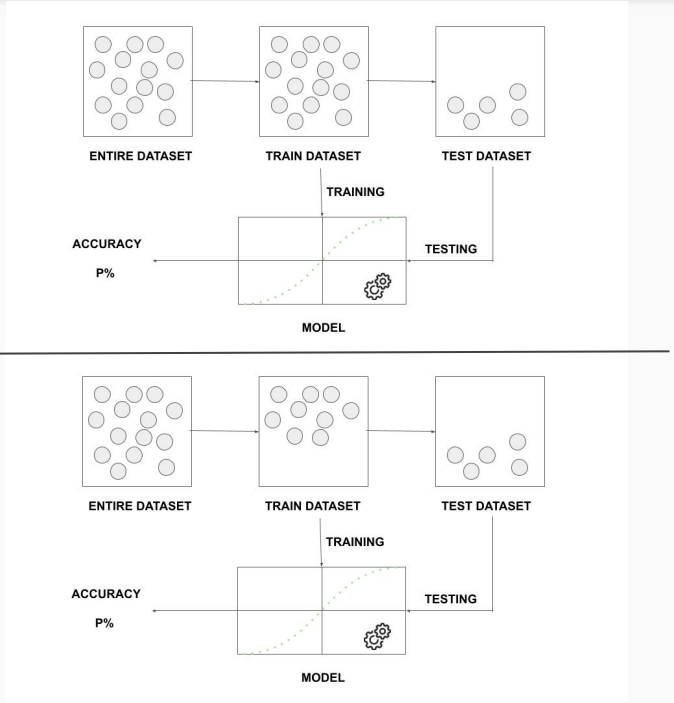
APPLICATIONS OF CLUSTERING

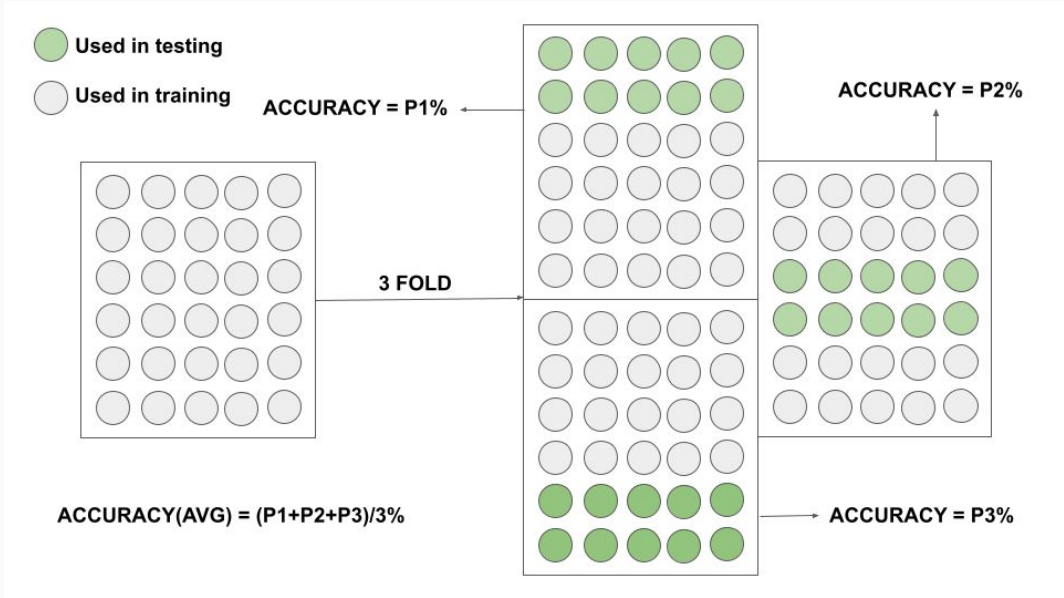
There are many uses of Clustering in the real world. One of the main uses of clustering is segmentation of data. In marketing we can identify buying patterns of customers, recommending products to the new customers, etc. In banking or insurance sector fraud detection in credit card use, identifying types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

EVALUATION MODELS FOR CLASSIFICATION

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

I. **Train and Test on the same dataset**
II. **Train Test Split**
III. **K-Fold Cross Validation, and so on.**

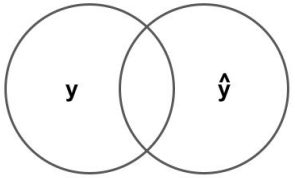




EVALUATION METRICS FOR CLASSIFICATION

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

I. **Jaccard Index**
II. **F1 Score**
III. **Log Loss, and so on.**



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

	P	N
P	TP	FN
N	FP	TN

CONFUSION MATRIX

Precision : measure of accuracy

Recall : trueness rate

F1 Score : harmonic average of Precision and Recall

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

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

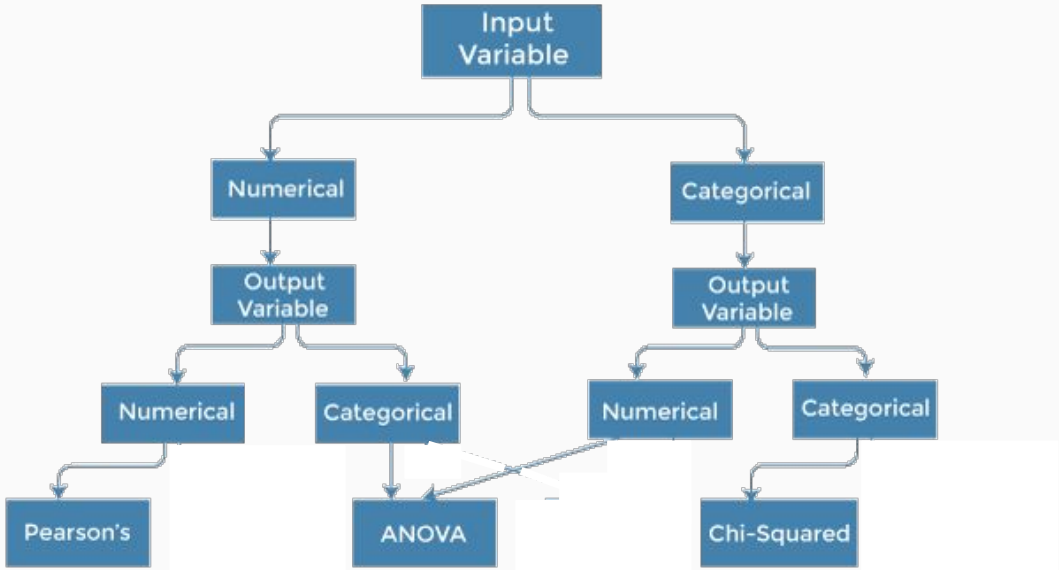
FOR P FOR N

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

FEATURE SELECTION METHODS

There are many feature selection methods according to the respective type of input variables and output variables. Some of those are,

I. **PCA (Principle Component Analysis)**
II. **Chi Squared Test**
III. **ANOVA Test, and so on.**



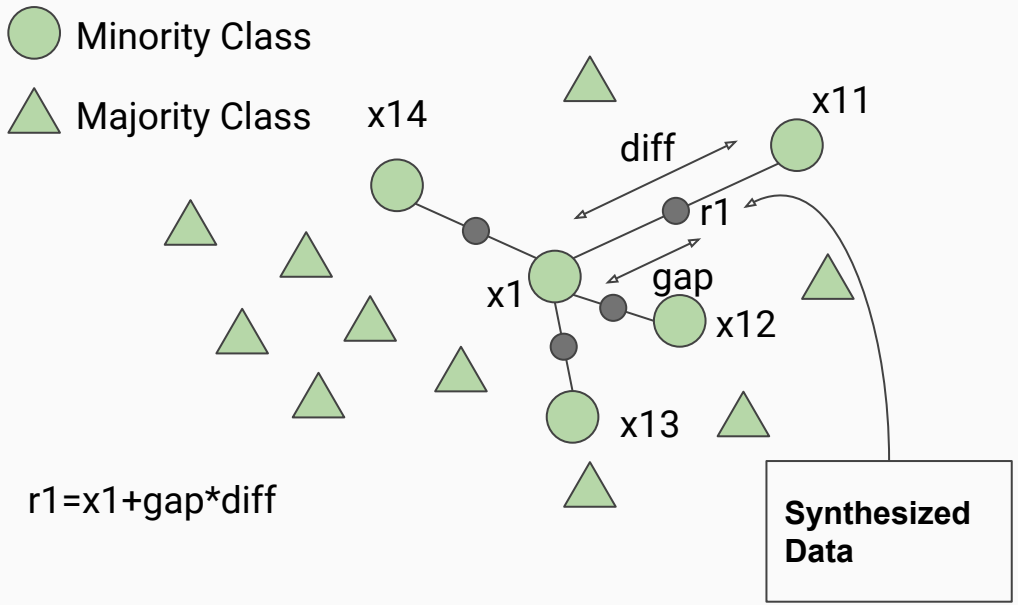
OVERSAMPLING METHODS

There are many oversampling methods for balancing the imbalanced dataset.
Some oversampling methods are:

I. **ADASYN (**Perform over-sampling using Adaptive Synthetic Sampling Approach for Imbalanced Learning.**)**
II. **Random Oversampling (**Object to over-sample the minority class(es) by picking samples at random with replacement.**)**
III. **SMOTE (**This object is an implementation of SMOTE - Synthetic Minority Over-sampling Technique**), and so on.**

SMOTE

Method that over samples the minority class by creating a new data point of minority samples.
There are some variants of SMOTE : the variants Borderline SMOTE 1, 2 and SVM-SMOTE.



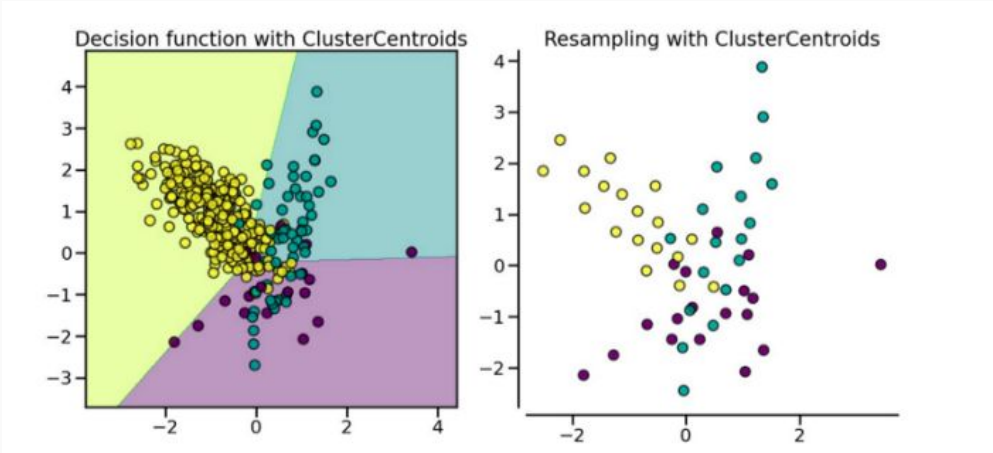
UNDERSAMPLING METHODS

There are many undersampling methods for balancing the imbalanced dataset.
Undersampling methods are actually divided into two types:

I. **Prototype Generation**
   a. Cluster Centroids
II. **Prototype Selection**

Cluster Centroids

Method that under samples the majority class by replacing a cluster of majority samples by the cluster centroid of a KMeans algorithm. This algorithm keeps N majority samples by fitting the KMeans algorithm with N cluster to the majority class and using the coordinates of the N cluster centroids as the new majority samples.

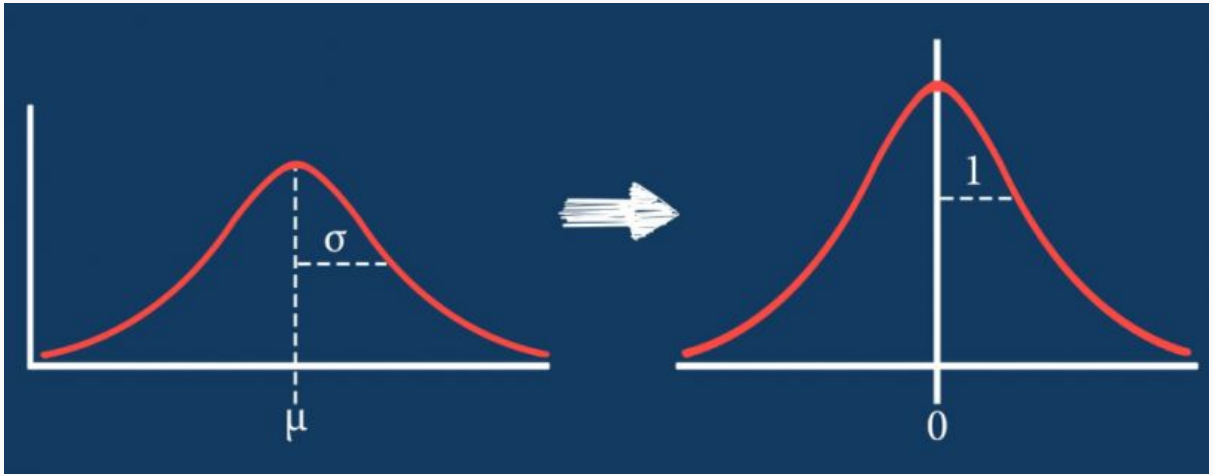


FEATURE SCALING METHODS

The two main feature scaling methods are,

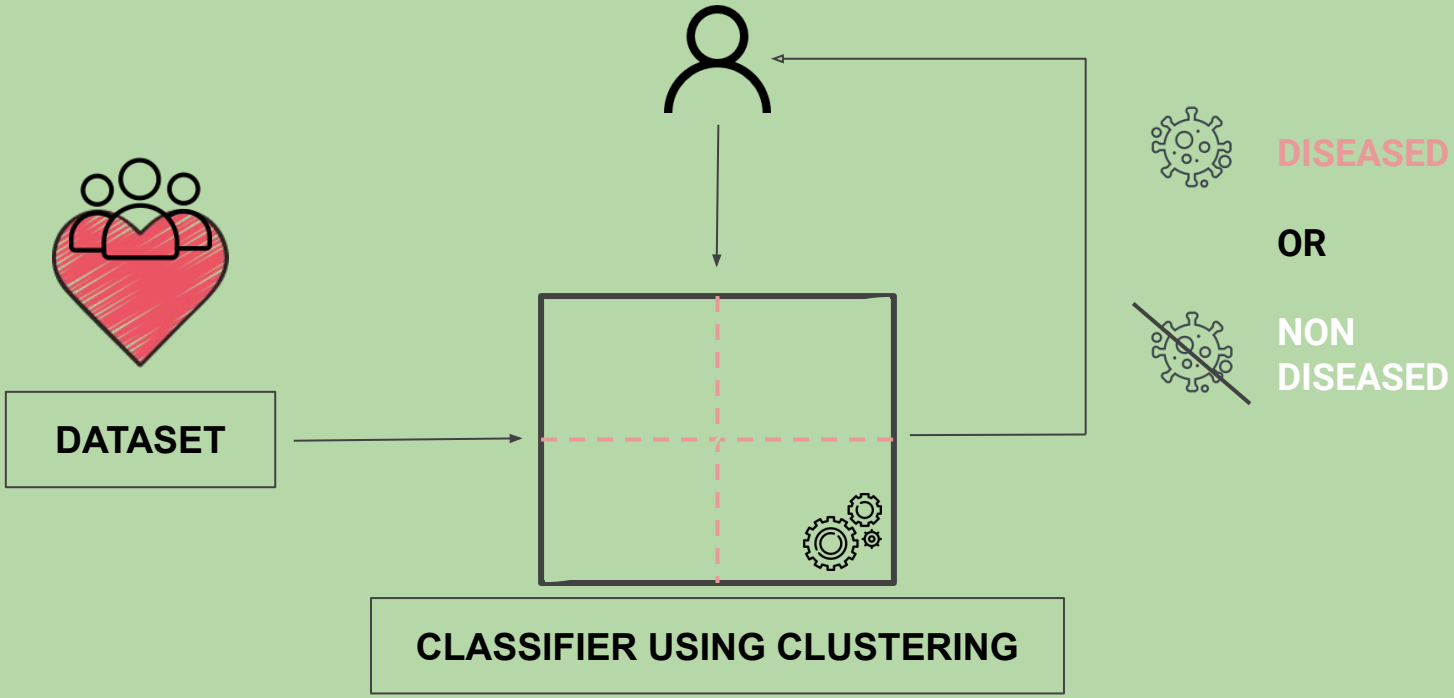
I. **Standardization**
II. **Normalization**

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



PROPOSED WORK

What is the Problem?



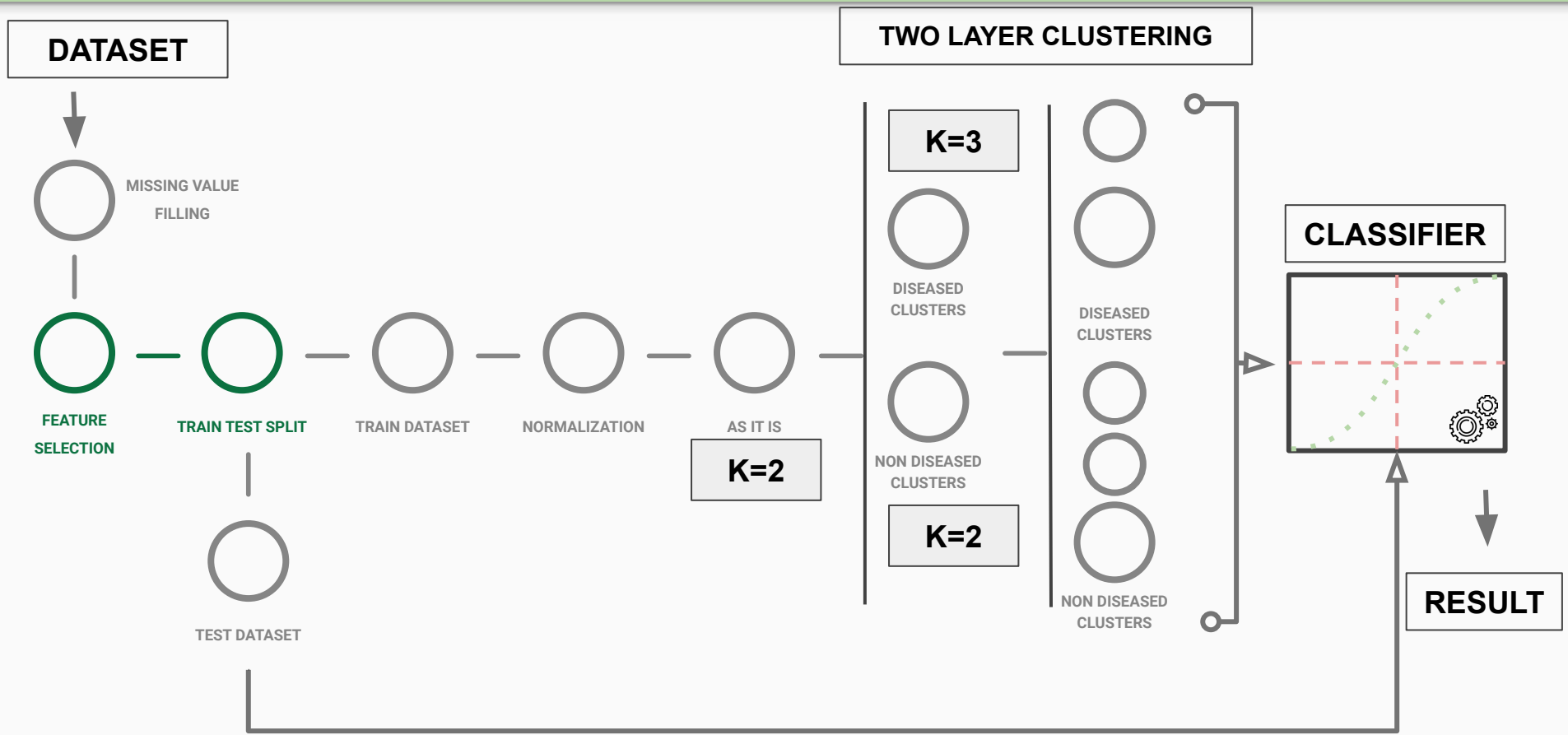
DATASET

Number of Records : 4238
Number of Features : 16

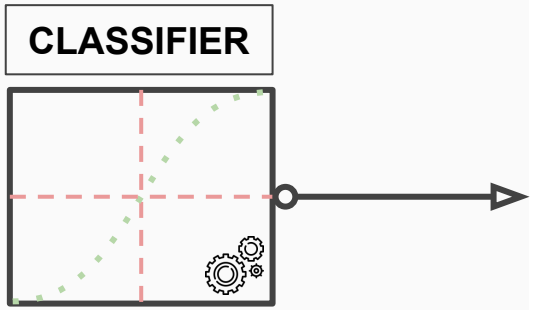
3594 of them are indicating to non diseased records and 644 of them are indicating to diseased records.

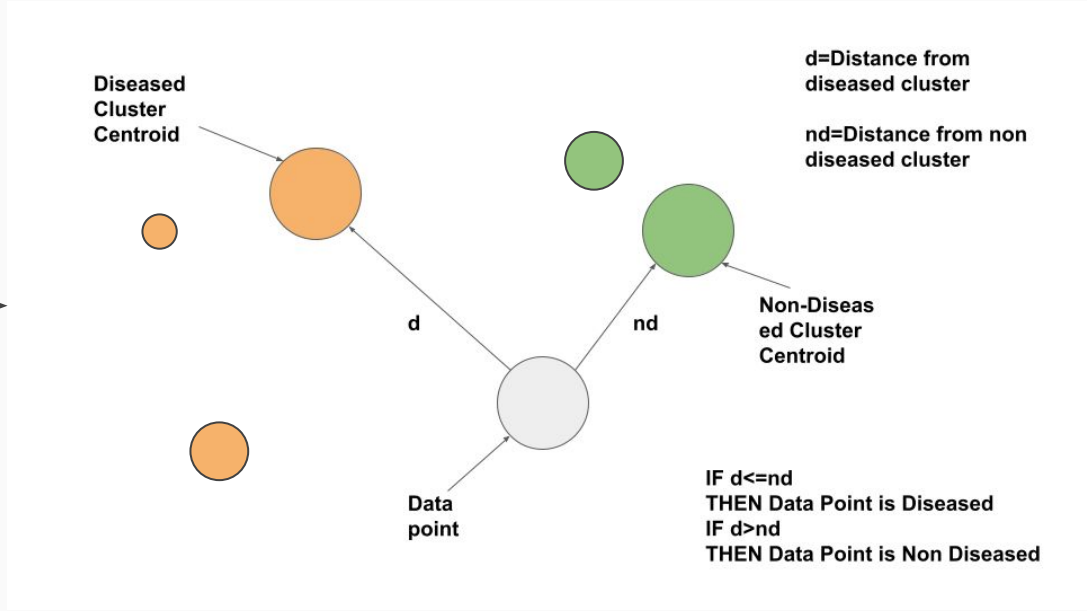
NAME OF THE FEATURE	NON-NULL COUNT	DATA TYPE	VALUE TYPE (1 => TRUE, 0 => FALSE)
sex	4238	INT64	CATEGORICAL (0/1)
age	4238	INT64	NUMERICAL
education	4133	FLOAT64	CATEGORICAL (1.0/2.0/3.0/4.0)
currentSmoker	4238	INT64	CATEGORICAL (0/1)
cigsPerDay	4209	FLOAT64	NUMERICAL
BPMeds	4185	FLOAT64	CATEGORICAL (0.0/1.0)
prevalentStoke	4238	INT64	CATEGORICAL (0/1)
prevalentHyp	4238	INT64	CATEGORICAL (0/1)
diabetes	4238	INT64	CATEGORICAL (0/1)
totChol	4188	FLOAT64	NUMERICAL
sysBP	4238	FLOAT64	NUMERICAL
diaBP	4238	FLOAT64	NUMERICAL
BMI	4219	FLOAT64	NUMERICAL
heartRate	4237	FLOAT64	NUMERICAL
glucose	3850	FLOAT64	NUMERICAL
TenYearCHD	4238	INT64	CATEGORICAL (0/1)

WORKFLOW OF THE PROJECT



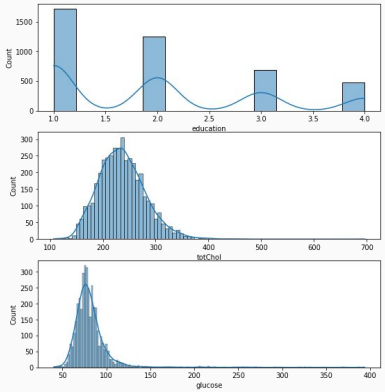
METHODOLOGY BEHIND THE CLASSIFIER

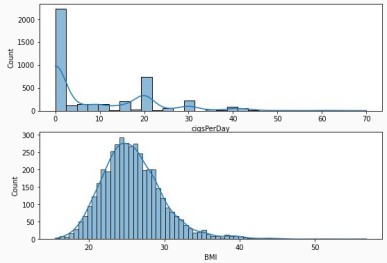


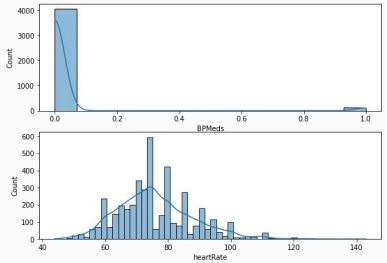


MISSING VALUE FILLING

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



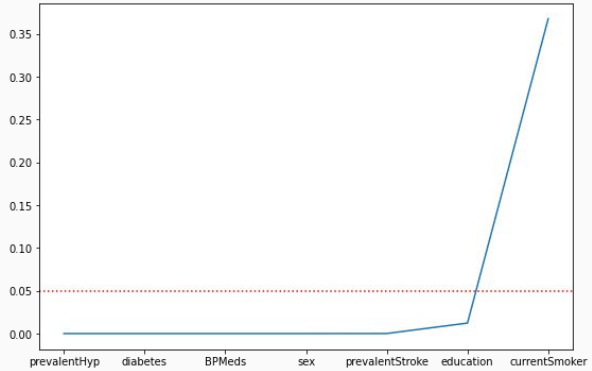




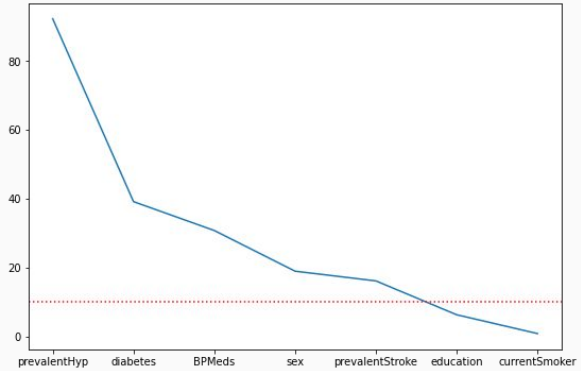
IMPUTATION METHOD

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

FEATURE SELECTION

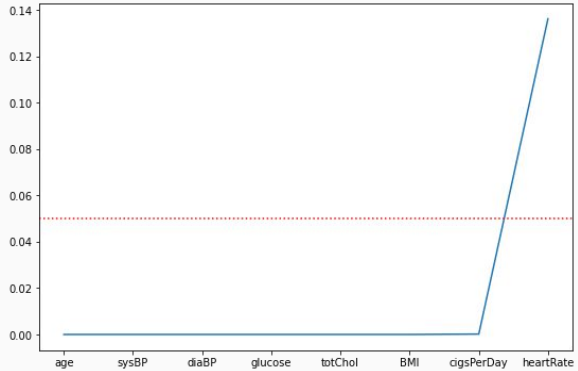


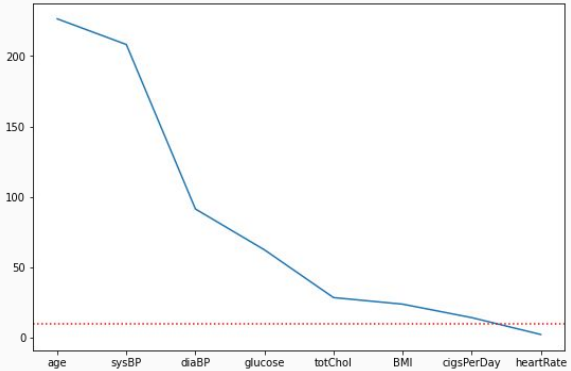
P VALUE



SCORE

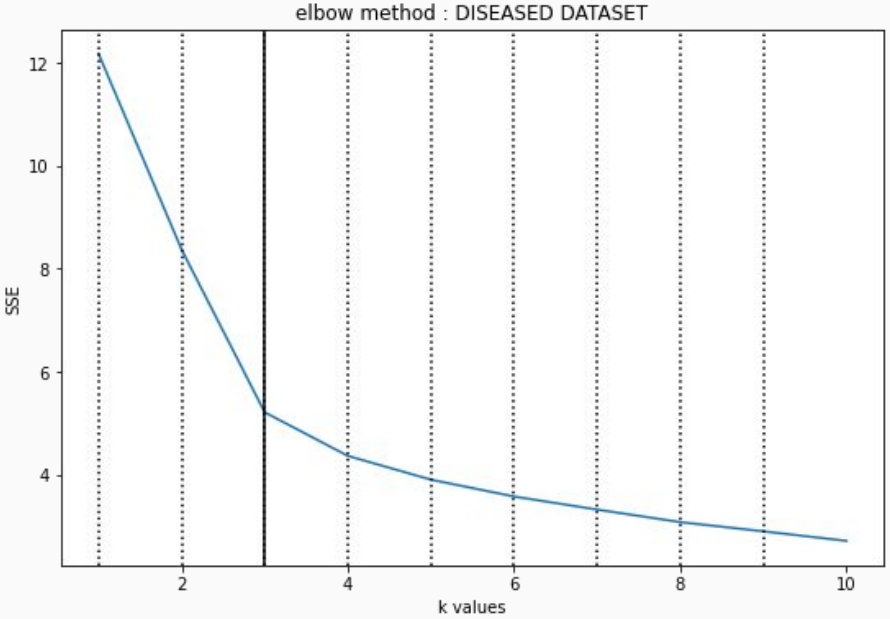
ACCORDING TO THE CHI-SQUARE FEATURE SELECTION METHOD “currentSmoker” FEATURE IS THE LEAST IMPORTANT CATEGORICAL FEATURE.

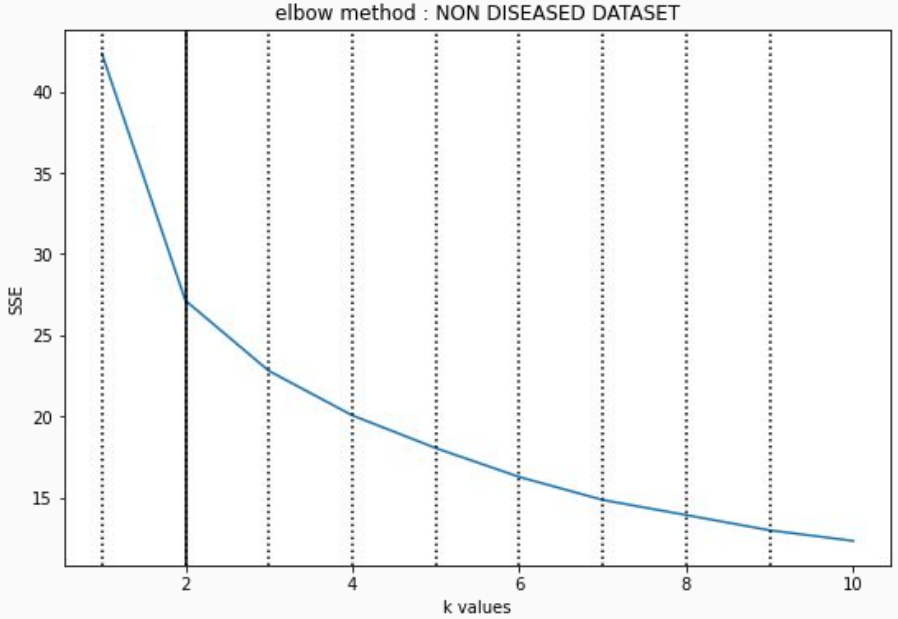




ACCORDING TO THE ANOVA FEATURE SELECTION TEST “heartRate” FEATURE IS THE LEAST IMPORTANT NON CATEGORICAL FEATURE.

BEST K USING ELBOW METHOD

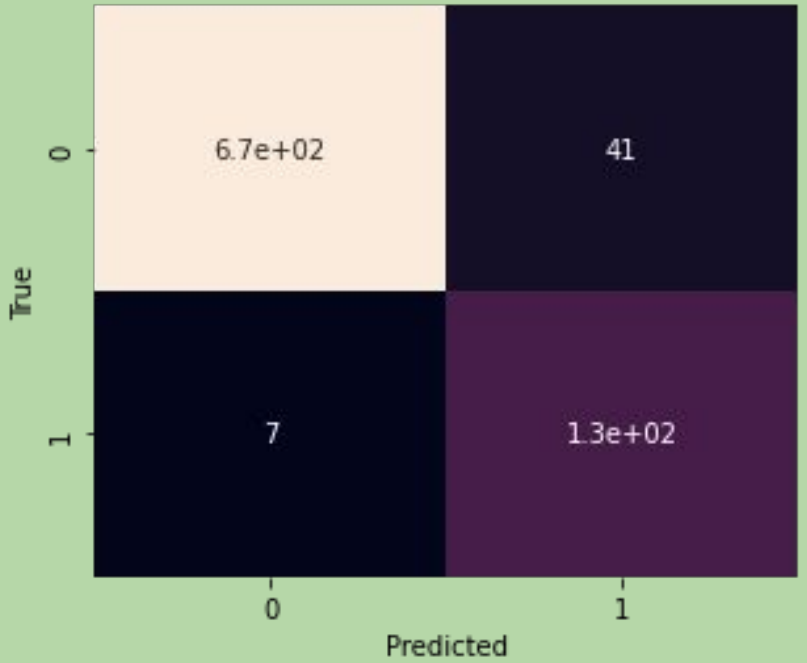




FINAL RESULT

	PRECISION	RECALL	F1-SCORE	SUPPORT
0	0.99	0.94	0.97	712
1	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

	JACCARD SCORE
0	0.94
1	0.73

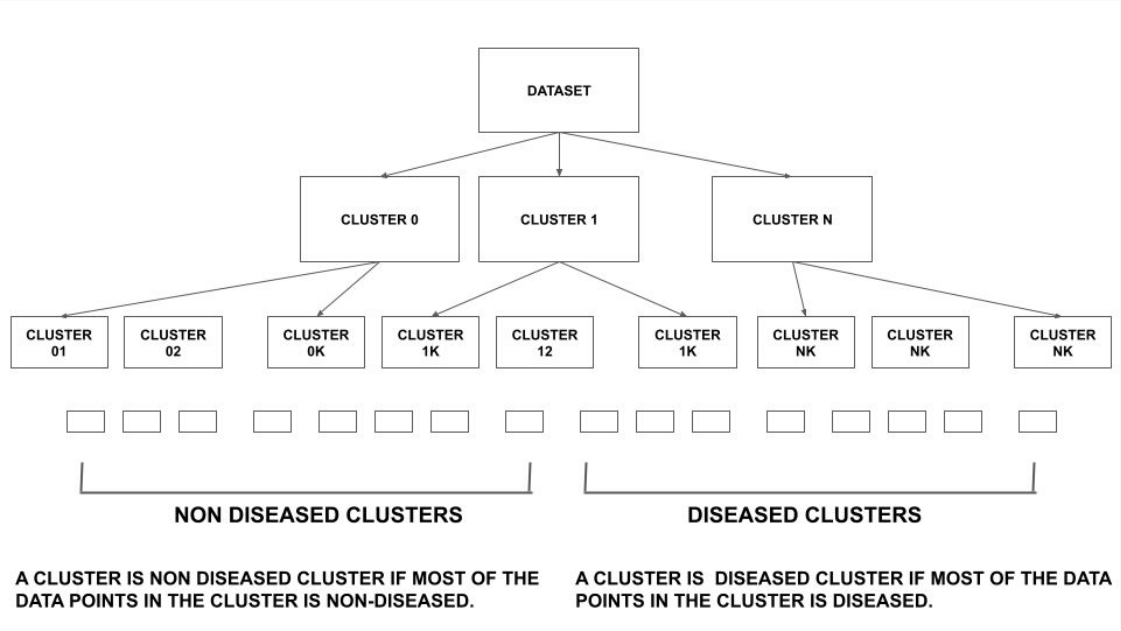


CONCLUSIONS AND FUTURE DEVELOPMENT

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.

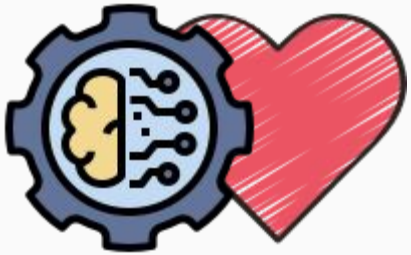
Tree-like Clustering means clustering over the clusters in a hierarchical manner.

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



END OF PROPOSED WORK

Thank You



I Conclude The Presentation Here.