

ALGORITHM FOR HEALTH CARE DATA ANALYSIS

Project report submitted in partial fulfilment of the requirements for
the degree of Bachelor of Technology

in

Computer Science and Engineering

By

ANIRUDH SHARMA (161238)

AMANDEEP SAINI(161244)

Under the supervision of

DR. YUGAL KUMAR

To



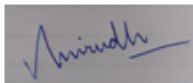
Department of Computer Science & Engineering and Information
Technology

**Jaypee University of Information Technology Wagnaghat, Solan-
173234, Himachal Pradesh**

Candidate's Declaration

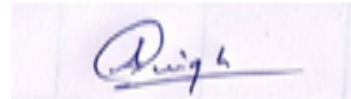
I hereby declare the work presented in this report entitled “**ALGORITHM FOR HEALTH CARE DATA ANALYSIS**” in partial fulfilment of the requirements for the award of the degree of **Bachelor of Technology in Computer Science and Engineering/Information Technology** submitted in the department of Computer Science & Engineering and Information Technology, Jaypee University of Information Technology Waknaghat is an authentic record of my own work carried out over a period from August 2019 to May 2020 under the supervision of **Dr. Yugal Kumar** (Assistant professor(Senior Grade), Department of Computer Science and Engineering).

The matter embodied in the report has not been submitted for the award of any other degree or diploma.



(Student Signature)

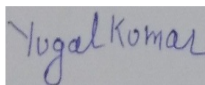
Anirudh Sharma, 161238.



(Student Signature)

Amandeep Saini, 161244.

This is to certify that the above statement made by the candidate is true to the best of my knowledge.



(Supervisor Signature)

Dr. Yugal Kumar

Assistant Professor(Senior Grade)

Department of Computer Science & Engineering and Information Technology

Dated:

Acknowledgement

We like to express our gratitude and special thanks to our project supervisor Dr. Yugal Kumar who gave us golden opportunity to do this project on the topic “ Multi Objective Algorithm For Health Care Data Analysis” which helped us to learn a lot of new things and we gained good experience from this project.

We would like to express our gratitude towards our parents and Jaypee University of Information Technology for their kind co-operation and encouragement which helped us in completion of this project.

Our thanks and appreciations also go to our colleague in developing the project and people who have willingly helped us out with their abilities.

TABLE OF CONTENT

1. Chapter-1 INTRODUCTION	1
1.1. Introduction	1
1.2. Problem Statement	8
1.3. Objectives	8
1.4. Methodology	9
1.5. Organization	9
2. Chapter-2 LITERATURE SURVEY	10
3. Chapter-3 SYSTEM DEVELOPMENT	21
3.1. Analytical	21
3.2. Computational	25
3.3. Experimental	26
3.4. Mathematical	30
4. Chapter - 4PERFORMANCE ANALYSIS	37
4.1. Performance Analysis Tools	38
5. Chapter – 5 Conclusions	39
5.1. Conclusion	39
5.2. Future Scope	39

LIST OF FIGURES

- 1) Fig:1 Difference between hard and soft clustering
- 2) Fig 2.1 KNN algorithm
- 3)Fig 2.2 Euclidian distance in KNN
- 4) Fig3.1 :Load and plot the data in matlab.
- 5) Fig 3.2: Clustering output
- 6) Fig 3.3 Diagnosis data without encoding
- 7) Fig3.4: Diagnosis data after Encoding
- 8) Figure 3.5: Visualization of dataset
- 9) Fig 4.1: confusion matrix
- 10) Fig 4.2 Accuracy comparison

LIST OF TABLES

- 1) Table 1.1 Breast cancer cases county-wise
- 2) Table 2.1 Advantages of KNN
- 3) Table 2.2 KNN and MKNN comparison
- 4) Table -3.1 Breast cancer dataset attributes.
- 5) Table 3.2: Breast cancer dataset class
- 6) Table 3.3: cluster Instance
- 7) Table 4.1: sample result analysis table

LIST OF GRAPHS

Graph 3.1: Plot of KNN accuracy output

Graph 3.2: Plot of MKNN accuracy output

Graph 3.3: Plot of both KNN and MKNN

ABSTRACT

Many computer aided diagnostic systems have been developed in order to reduce false-positives diagnosis, in breast cancer research More than ever. Through this work of ours ,we are presenting a data mining based approach which will help and support oncologists in the process of breast cancer classification and diagnosis. Mainly the objective of this research is to diagnose breast cancer in patients at early stage to help treat it better . Nowadays, not only in India but also in other countries, it has been found that the breast cancer is a major disease in many women. Clustering data mining algorithms are used for breast cancer detection, for early diagnosis of breast cancer patients. We used UCI web data repository on Breast Cancer Patients for experimental purpose. This will not only help doctors but also patients in early detection and treatment of breast cancer disease.

CHAPTER – 1 INTRODUCTION

1.1 Introduction

Bosom malignant also known commonly as Breast cancer growth is a significant sickness found in numerous Women in India. Bosom malignant growth is an infection where disease causing Cells are shaped in a lady's bosom tissue. The bosom is comprised of flaps (15 to 20 squares) and conduits. Conduits There are flimsy 3D squares to join the bosom to drain creation. The areola and areola are outside the bosom and have a darker shading. From the bosom. The most widely recognized kind of bosom malignant growth is Starts in the cells of the tubules. Projection disease Or lobules found in the two bosoms are different kinds of bosoms Cancer Hot, red and swollen bosoms are a scar to the bosom Cancer Age and wellbeing history may influence hazard Development of bosom malignant growth. Bosom malignant growth is brought about by qualities Change. Chests are X-beam, CT examine, bone sweep and PET output Used to identify the phases of bosom malignancy. Repetitive bosom Cancer is the malignancy that returns in the wake of being dealt with. Disease can return the bosom, in the chest area Wall, or any piece of the body. To take care of the issue of bosom Early research of disease and bosom malignant growth this exploration Apply bunching information mining strategies to discover Breast malignancy patients.

Early conclusion before BC can altogether improve guess and endurance, as it can elevate auspicious clinical treatment to patients. Further precise arrangement of benevolent tumors can keep patients from pointless treatment. In this manner, BC Correct determination and order of patients into dangerous or benevolent gatherings is a subject of much research. Because of its novel points of interest in distinguishing significant highlights from complex BC datasets, AI (ML) is broadly perceived as the technique for decision in BC design arrangement and forecast demonstrating.

Order and information mining techniques are a viable method of arranging information. Particularly in the clinical field, where those strategies are generally utilized in determination and examination for dynamic.

1.1.1 TYPES

Angiosarcoma

Angiosarcoma is an uncommon kind of malignant growth that structures in the covering of veins and lymph vessels. Your lymph, gather and discard microscopic organisms, infections and waste items from your body.

Angiosarcoma can happen anyplace in your body, yet it frequently happens in the skin on your head and neck. Infrequently, angiosarcoma may happen in different pieces of your body, for example, the bosom. Or then again it can shape in profound tissue, for example, liver and heart. Angiosarcoma may happen in regions recently treated with radiation treatment. Angiosarcoma treatment relies upon where the malignancy is found. Treatment alternatives may incorporate medical procedure, radiation treatment and chemotherapy.

DCS

DCS is the accumulation of unusual cells inside a milk channel in the breast. DCIS is viewed as the most punctual type of bosom disease. DCIS is noninvasive, implying that it doesn't spread past the milk pipe and has a lower danger of invasiveness. DCIS is generally proceeded as a piece of bosom disease screening or is found to analyze for bosom irregularities.

While DCIS isn't a crisis, it requires assessment and thought of treatment alternatives. Treatment may incorporate bosom rationing medical procedure joined with radiation or medical procedure to expel the entirety of the bosom tissue. A clinical preliminary contemplating dynamic reconnaissance as an option in contrast to medical procedure might be another alternative.

Inflammatory breast cancer

Provocative bosom malignant growth is an uncommon kind of bosom disease that grows quickly, making influenced bosoms become red, swollen and tender. Inflammatory bosom malignant growth happens when disease cells hinder the lymphatic vessels in the skin covering the bosom, causing a red, swollen appearance of the bosom.

Fiery bosom malignant growth is viewed as privately propelled disease - implying that it has spread from its source to close by tissue and conceivably to close by lymph hubs.

Fiery bosom disease can without much of a stretch be mistaken for a bosom contamination, which is a substantially more typical reason for bosom redness and expanding. On the off chance that you notice skin changes on your bosom, look for clinical consideration right away.

Paget disease of the nipple

This is a type of bosom Malian begins which starts to develop inside the ducts of the nipples of in a bosom, as the time passes its size increases and it starts affecting other parts such as the skin. In the future this can cause trouble and several health complications.

Invasive lobular carcinoma

Obtrusive lobular carcinoma is a kind of bosom malignant growth that begins in the bosom milk delivering organs (lobules). Intrusive malignant growth implies that the disease cells have relocated out of the lobules where they began and be able to spread to lymph hubs and different regions of the body.

Obtrusive lobular carcinoma makes up a little part of all bosom tumors. The most widely recognized sort of bosom malignancy starts in the mammary channels (obtrusive ductal carcinoma).

1.1.2 Symptoms

Signs and symptoms of breast cancer may include:

- A knot or thickening that feels secluded from nearby tissues
- Changes start to occur in size and even in the shape of bosom
- Change like dimpling can be observed.
- New types of areolas
- Red erect skin over bosom, for example, orange skin

Rank	Country	Age-standardised rate per 100,000
1	Belgium	113.2
2	Luxembourg	109.3
3	Netherlands	105.9
4	France (metropolitan)	99.1
5	New Caledonia (France)	98.0
6	Lebanon	97.6

Rank	Country	Age-standardised rate per 100,000
7	Australia	94.5
8	UK	93.6
9	Italy	92.8
10	New Zealand	92.6
11	Ireland	90.3
12	Sweden	89.8
13	Finland	89.5
14	Denmark	88.8
15	Switzerland	88.1
16	Montenegro	87.8
17	Malta	87.6
18	Norway	87.5
19	Hungary	85.5

Rank	Country	Age-standardised rate per 100,000
20	Germany	85.4
21	Iceland	85.2
22	US	84.9
23	Canada	83.8
24	Cyprus	81.7
25	Samoa	80.1

Table 1.1 Case of breast cancer.

Causes

It has been found out that the breast cancer tend to happen in situation where the cells start developing abnormally. These particular type of cells do splitting a a more rapid speed in comparison to the other solid cells and eventually develop into a knot or solid masses. These cells are quite capable of spreading to other parts of body such as lymph hubs and even other parts of body and can harm the body and can even threaten the life.

Specialists say that even the daily hygiene and the lifestyle of a person also can be responsible for the development of this type of cancer . If a person lives in unhygienic conditions and the proper hygiene is not followed then the risk further increases. In some cases the hormones or even the hereditary factors of a person are also found out to be responsible for developing this ailment and the people have to suffer a lot because of this. The lack of proper education in few countries or areas where awareness is not much, the people don't give much attention and then they have to suffer.

Inherited breast cancer

Specialists gauge that 5 to 10 percent of breast malignant growths are related with quality changes went through ages of a family.

On the off chance that you have a solid family ancestry of breast disease or different malignant growths, your primary care physician may prescribe a blood test to help recognize explicit transformations in BRCA or different qualities that your family is experiencing.

Think about approaching your primary care physician for a referral to a hereditary advisor, who can survey your family wellbeing history. A hereditary advisor can likewise examine the advantages, dangers and confinements of hereditary testing to assist you with settling on a common choice.

The risk

A breast malignant growth hazard factor is whatever makes it almost certain you will get breast disease. Yet, the presence of single or multiple breast malignant growth factors doesn't imply that one will create breast disease. A great number of women who tend to develop the tumour have no other factor than just that they are women .

Factors associated with increased risk of breast cancer :

- Being lady. Ladies are relatively more likely to create breast malignancy.
- Growing old. With the growing age the risk also increases.
- Any previous record of breast condition. On the off chance that you have a breast biopsy found in lobular carcinoma in situ (LCIS) or atypical hyperplasia of the breast, you are at expanded danger of breast malignancy.
- A Personal History of Breast Cancer. If in the past one of your breast was affected then there are quite chances of you to develop the disease in the other breast and the patients with such history need to be more careful and cautious .
- Family history of breast malignant growth. In the past in case any female family member of your family had this disease then there are quite a chance that the disease can be inherited and then you can be at risk of having this ailment in near future or anytime later and the early diagnosis is very essential in such cases.

1.2 Problem Statement

. Through this work of ours ,we are presenting a data mining based approach which will help and support oncologists for process of bosom malignant classification and detection.Mainly the objective of this research is meant for diagnosis of bosom cancer in patients at early to help in daeling and treating it better . Nowadays, not only in India but also in other countries, breast cancer is found out to be a major disease in many women. Clustering data mining algorithms are being explored for breast cancer detection, for early diagnosis of breast cancer patients. We used UCI web data repository on Breast Cancer Patients for experimental purpose. This will not only help doctors but also patients in early detection and treatment of breast cancer disease.

1.3 Objectives

The reason for this examination is to discover the highlights that are mostly helpful in foreseeing both dangerous and generous cancerous growths and search for the common patterns which may be helpful to us to make model choice as well as choice of hyper parameters. Our objective is to group is the bosom disease life threatening or amiable. To achieve that I explored the Artificial Intelligence characterization methods to fit the capacity that is capable of foreseeing discrete class of new information sources.

1.4 Methodology

Here we will use the machine learning algorithm and the most popular KNN algorithm which is K Nearest Neighbours algorithm as it works well with the cancer data set and also it generates results on each run and the need to train the model is terminated. Then we have experimented the modified version of our KNN which will enhance the accuracy and will also be helpful in comparison.

1.5 Organisation

The report comprises of different segments. The primary area gives the presentation of bosom disease and outline to the issue. It likewise gives the fundamental reasons and kinds of bosom malignancy. The subsequent segment contains the writing survey and discoveries. The third area contains the different informational collections which are utilized in the venture. The fourth segment is made out of the actualized calculations and fifth segment I generally about the normal yields and the future ramifications of task.

CHAPTER – 2 LITERATURE SURVEY

The creators of the paper concentrated on improving wellbeing Awareness. They can offer guides to accomplish this Self bosom purging and mindfulness that will help Diagnosed bosom malignant growth . Concentrate on the creator of this paper Creating a stage on world wellbeing accomplishment They have to execute counteraction and screening. Programming to diminish a disease hazard zone. They utilize a Simple k-closest neighbor calculation for ideal Performance. Paper creators realize the dangers Breast malignant growth factor. They utilize a clinical bosom care venture (CBPC) to document and examine information and build up a Prototype for information mining. They utilize the Bayesian system Analysis strategies to discover information collaborations and discover Caffeine CBCP populace. Creator of this paper center Found an approach to improve the opportunity of bosom malignant growth It must be recognized as quickly as time permits so as to endure long. They utilize Three factual methods for the determination of bosom malignant growth are the first is mammography, the other is FNA (fine needle suction) And the third is careful biopsy. Creator of this paper Breast malignant growth information examination and spotlight on testing The issue is that they utilize a grouping technique to distinguish disease in the bosom. Danger of Cancer They can apply GHSOM to 24,481 qualities DNA microarray of bosom disease tumor tests. More outcomes 17 qualities identified are probably going to be related withFour bosom malignant growth marker qualities [7]. Creator of this paper Focus is attempting to analyze ladies' initial bosom malignancy SVM, Tree Boost and Tree Forest Data Mining Classification Technique . The creator of this paper SEER centers around open - Use-information to anticipate bosom malignant growth. They use pre-grouping Method to discover and to locate a potential arrangement Breast malignant growth data. [9]. Creator of this paper Focusing on an interpretation procedure to comprehend coexpressed quality seta under regular administrative components.

They have an information pre handling strategy and two unique uses Association rule mining . Utilized paper essayist Various information digging strategies for analysis and visualization Breast disease with fundamental parameters of male and female Gene conduct, they take 311's quality articulation informational index For instance testing and approving models and key Display. They end up being characterization information mining calculation Provide progressively ideal outcomes [11]. Paper essayist Focus on different information digging arrangement calculations for Breast disease investigation.

They utilize three unique calculations with the assistance of Waikato Environmental Knowledge Analysis open source programming. They utilize a Decision tree, Bayesian System and k-closest neighbor calculations in outcome. They have accurately arranged different parameters. Models, inaccurately ordered models, mistakenly characterized Examples. Required some investment kappa information, relative total mistake, And course related squared blunder show by mining process .The creators do entrenched and similar investigations of Gene Expression Programming for Diagnosed Breast Cancer among patients. Concentrate on the creator of the paper Help a specialist in guardians conclusion and treatment plan Procedures for various classes. For the procedure they use Classification and grouping calculation and similar examination Study of the total mining calculation. Creator of The paper centers around this kind of carcinogenic sickness emerging from people. Mammary tissue cells, ordinarily from the lobules or internal fixing Milk tubes that furnish channels with milk. For They utilize diverse grouping methods for the procedure. The best malignant growth. They utilize an AI strategy They utilize two groupings to lessen the elements of the dataset. The strategy is relevant for bosom malignancy. They utilize an AI. The utilization of innovation to distinguish Barrett's malignant growth. they utilize Key part investigation strategies for segment decrease Analysis method to lessen the element of an informational collection. They Use MLP first utilizing a two arrangement procedure Prasar NN (MLP BPN) and second help vector Check machine (SVM) exactness, accuracy, review, F measure, Kappa information . Note the creator of the paper. Distinguishing qualities that are increasingly related with pregnancy of malignant growth. They utilize the K-implies bunching procedure for test information. Recognize potential bio-markers of bosom malignant growth Based on some trait and check an alternate advance Classification Technique .Focus on the creator of the paper K-implies grouping method and fluffy bunching calculation To improve the best disease [1.]. Different employments of this paper The best apparatus for diagnosing air devices and neural systems Cancer among ladies.

Our project is based on Data Clustering in Machine Learning. For this project we have considered MATLAB as our tool for implementing required algorithm.

Data Clustering

Clustering is the task of dividing the population or data points into a number of groups such that data points in the same groups are more similar to other data points in the same group than those in other groups. In simple words, the aim is to segregate groups with similar traits and assign them into clusters.

Types of Clustering

In broader prospect clustering can be divided into two subgroups :

Hard Clustering: In hard clustering, each data point is either a part of data cluster completely or it is not at all the part of that cluster.

Soft Clustering: In soft clustering, instead of putting each data point into a separate cluster, a probability or likelihood of that data point to be in those clusters is assigned.

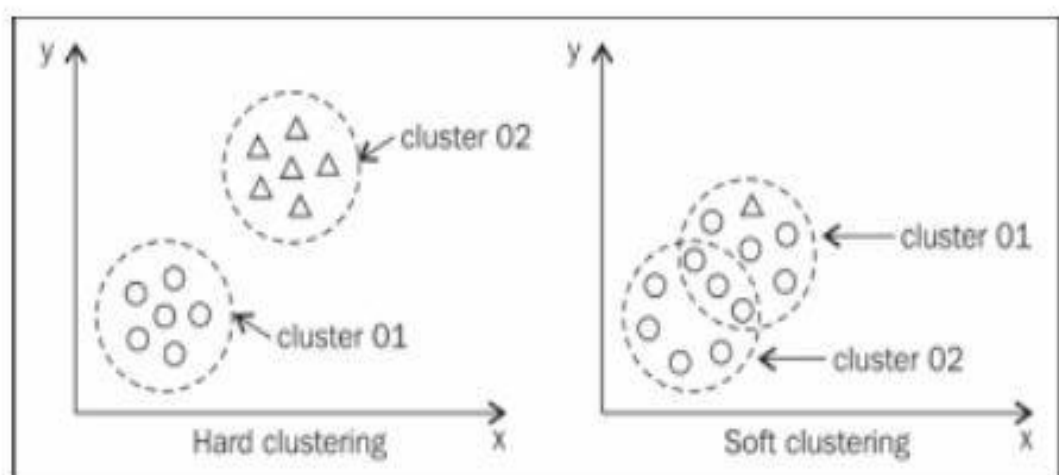


Fig:1 Difference between hard and soft clustering

Overview

- Learn about Clustering , one of the most popular unsupervised classification techniques
- Dividing the data into clusters can be on the basis of centroids, distributions , densities, etc
- Get to know K means and hierarchical clustering and the difference between the two

KNN ALGORITHM

KNN is considered as the one which is very less complex algorithm as compared to other algorithms of ML.

KNN is one of the hard clustering algo which considers the neighbours and the data is organised into classes and the class to which the new data closely resemble or we can put it this way that the class which has maximum number of neighbours to the new inserted data set, we categorise the new data entry into that very class.

Knn calculation stores all the accessible information and arranges another information point dependent on the likeness. This implies when new information shows up then it tends to be effectively characterized into a well suite class by utilizing K-NN calculation.

Knn calculation is very easy to use and it is often called as the lazy student approach as in this particular approach the model training is not required and this gives results immediately every time we run the program. The extra burden to train which requires learning from the previous results and storing the previous results is reduced by the use of this knn approach.

KNN is based on the non parametric approach which means that there is no need to make the assumptions and suppositions .

KNN calculation at the preparation stage maintains a repository of the data set every time it receives new info, at that point it groups that information into a class that is a lot of like the new information.

Model: Suppose, we have a picture of an animal that appears to be like feline and canine, however we need to know possibly it is a feline or pooch. So for this distinguishing proof, we

can utilize the KNN calculation, as it takes a shot at a closeness measures. The model has to locate comparable highlights of newly received informational index.

WHY KNN?

1. It is easy to interpret the output in knn
2. The calculation time is very less.
3. It has good predictive power.
4. We need not to train the model.

Comparison of KNN with other algorithms:

	Logistic Regression	CART	Random Forest	KNN
1. Ease to interpret output	2	3	1	3
2. Calculation time	3	2	1	3
3. Predictive Power	2	2	3	2

Table2.1 Advantages of KNN

KNN Algorithm

1. Select a value for K which is no of the neighbours
2. In second step calculate the Euclidian Distance for K numbers of neighbours.
- 3 Select the nearest neighbours K based on calculated Euclidean distance.
- 4.Now out of these K neighbours find the maximum number of neighbours belonging to a single class.
5. Now the new data has to assigned to the class in which maximum number of neighbours are belonging.
6. The model of KNN is ready.

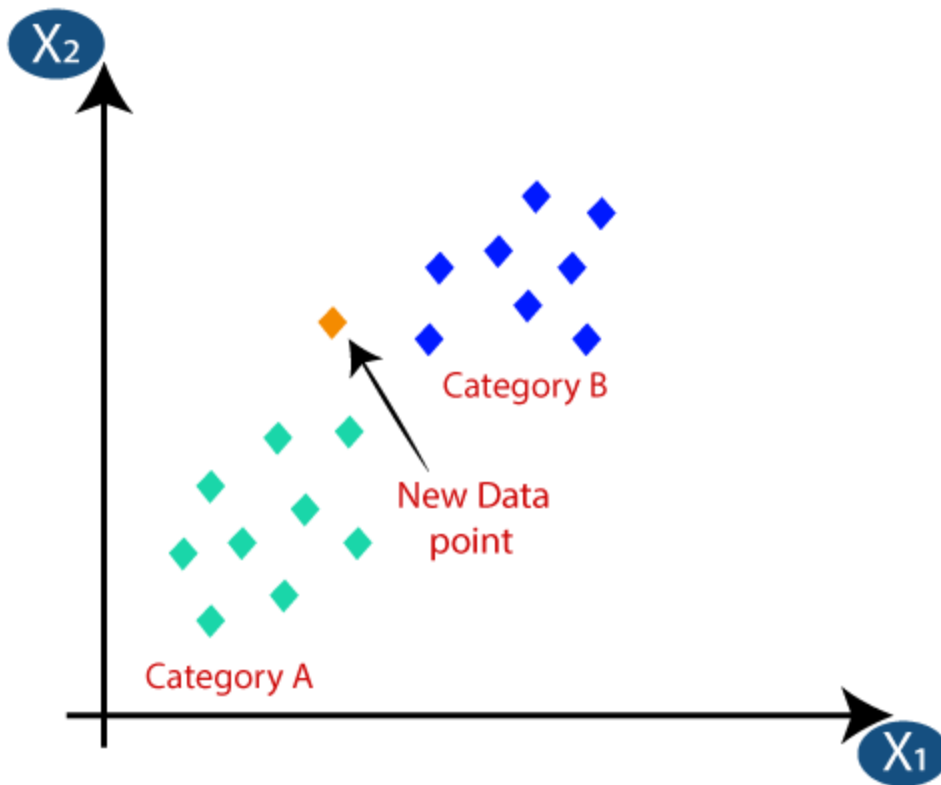
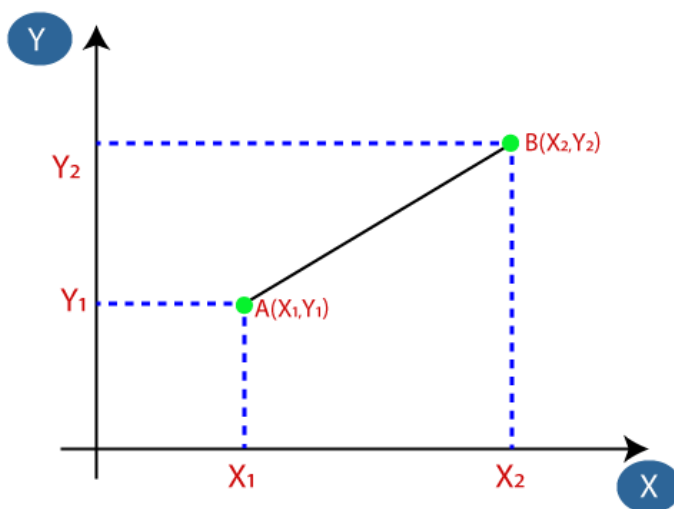


Fig 2.1 KNN algorithm

First of all we have to choose the value of k for illustration and suppose here we choose k to be 5

Then calculate the Euclidian distance, using the simple geometric formula.



$$\text{Euclidean Distance between } A_1 \text{ and } B_2 = \sqrt{(X_2 - X_1)^2 + (Y_2 - Y_1)^2}$$

Fig 2.2 Euclidian distance in KNN

BY calculating the euclidian distance we have obtained the nearest neighbours.



Since in cluster A there are 3 nearest neighbours and in B there are 2.

Hence our new data will belong to cluster A.

As it is hard clustering there is no scope for considering the probabilistic approach hence A is the required cluster for the new data input.

How to Select K?

No well defined method is available to choose k in KNN algorithm.

The most suitable value that we can set for k is 5.

The values smaller than 5 such as $k=1$, $k=2$ are not good as they can be noisy.

Larger values of k can also be considered as these are also good but in some cases these can cause some difficulties, so the ideal value is $k=5$.

To get better results the value of k can be varied.

MKNN –Modified KNN

Along with the KNN algorithm we are also implementing the modified version of the KNN algorithm which will further enhance the performance of our system and ensure early detection and also we will be able to compare the performance between KNN and MKNN

The fundamental thought of the introduced strategy is allotting the class mark of the information as per K approved information purposes for training set. In other hand, first, the legitimacy everything being equal tests in the train set is figured. At that point, a weighted KNN is performed on any test tests.

The pseudo code for modified KNN(MKNN) algorithm.

```
Output_label := MKNN ( train_set , test_sample )
Begin
  For i := 1 to train_size
    Validity(i) := Compute Validity of i-th sample;
  End for;
  Output_label:=Weighted_KNN(Validity,test_sample);
  Return Output_label ;
End.
```

Fig1. The pseudo code of modified KNN

Validation:

It is axiomatic to say that we need to validate every step in the modified KNN algorithm and for that we have devised a formula

The legitimacy of each point is processed by its neighbors. The approval procedure is performed for all train tests once. In the wake of relegating the legitimacy of every training test, it is utilized more data for the focuses.

Approve an example in thr training set, H closest neighbours for fact of the matter is thought of. Amongst H closest neighbours to the training test x, validity(x) tallies number of the focuses that have a similar name to the mark of x. Then the equation that is selected for process the legitimacy of each given incidence in the training set is numbered 1.

$$Validity(x) = \frac{1}{H} \sum_{i=1}^H S(lbl(x), lbl(N_i(x)))$$

Here we consider that H be number of the neighbours and the $lbl(x)$ gives back the label of the sample which is x . $N_i(x)$ here represents the i th nearest neighbour to the point x . The 2, defines this function.

$$S(a,b) = \begin{cases} 1 & a = b \\ 0 & a \neq b \end{cases}$$

APPLYING WEIGHTED KNN

This is one of the variations of the traditional KNN algorithm where we are using the K nearest neighbour, regardless of in which class they fall. Here we are using the weighted votes selected from each samples instead of a simple majority.

Here in this modified KNN strategy, firstly heaviness of every single neighbour is registered utilizing the function $(1/(d_e+0.5))$. At that point, legitimacy of the given preparing test is increased for its crude weight that is being decided on the basis of Euclidian separation. In this Modified KNN strategy, weight of every neighbour test to be inferred by ((.3)).

$$W(i) = Validity(i) \times \frac{1}{d_e + 0.5} \quad (3)$$

Where $W(i)$ and $Validity(i)$ stand for the weight and the validity of the i th nearest sample in the train data set.

We have evaluated the results of KNN and MKNN over an experimental problem and for a data set and the results are as follow:

		Monk 1	Monk 2	Monk 3	Isodata	Wine
K=3	KNN	84.49	69.21	89.12	82.74	80.89
	MKNN	87.81	77.66	90.58	83.52	83.95
K=5	KNN	84.26	69.91	89.35	82.90	83.79
	MKNN	87.81	78.01	90.66	83.32	85.76
K=7	KNN	79.86	65.74	88.66	80.50	80.13
	MKNN	86.65	77.16	91.28	83.14	82.54

Table 2.2 The comparison between KNN and MKNN.

As we can clearly see that the MKNN has increased and enhanced the accuracy of the output and it also increases the probability of get the correct and desired output.

LITERATURE REVIEW FINDING

Numerous specialists talk about the fundamental factor which is Responsible for bosom malignancy, for example, lethargy of cleaning, Size, shape, shading. Numerous patients languish over it Problem because of not wearing appropriate girdle. Key issues of Small rulers are not wearing girdles in the region. There is one Changes in the presence of the areola by contacting the bosom. Are you There are a few patients who have protuberance issues on the areolas or Soaked with blood from the areola.

Patients are experiencing thick tissue either in the bosom or in the areola. There Some large rashes around an areola. The fundamental element is Dimple on the bosom. It tends to be left bosom or right Bosom or both. First bosom malignancy indication The lady sees a knot or a region of their thick tissue bosom. Most knots are not hazardous for bosom malady Cancer The age factor is progressively liable for the bosom Cancer According to age gathering, bosom

disease is analyzed out of 4 1,000 ladies at the age of 30, 14 ladies out of 1,000 Out of 1,000 ladies at the age of 50, 37, 40 are 26 Out of 1,000 ladies at 60 years old . To diminish Breast disease should drink increasingly more spotless water. Most scientists state that calories ought to diminish food Reducing bosom malignant growth. Mature age, feminine cycle in a rush Age, mature age from the outset birth or never conceiving an offspring, a Personal history of bosom disease or kindhearted (non-malignant) bosom Treatment with sickness, a mother or sister, with bosom malignancy Breast/chest radiation treatment, bosom tissue that is thick On mammograms, taking hormones, for example, estrogen and Progesterone. Drinking mixed beverages, being white to be answerable for Breast Cancer.

CHAPTER – 3 SYSTEM DEVELOPMENT

3.1 Analytical

We are taking the data set from UCI Solve research objective. To do practical work WEKA is taken as an open source data mining tool And then implement separate data mining algorithms for measurement Accuracy and performance for breast cancer detection.

DATASET DESCRIPTION:

Attribute Name	Description
Age	Patient's Age in years
Menopause	the period in a woman's life when menstruation ceases
Tumor-size	Patient's tumor-size on her breast
inv-nodes	Node size in main portion of the breast.
Node-caps	Node is present or not in cap of the breast
Deg-malig	Stage of breast cancer
Brest	Left breast or Right breast or both breast
Breast-quad	Portion of the breast for example left-up, left-low, right-up, right-low, central.
Irradiate	Present or not (YES/NO)
Class	no-recurrence-events, recurrence-events (Reduce the risk of breast cancer)

Tab:(3.1) Attributes for the breast cancer dataset.

Class name	Description
Diagnosis	sick, healthy or (unpredictable) no class

Tab (3.2): Bosom cancer dataset class.

3.1.1

We are implementing the KNN algorithm on Breast Cancer Wisconsin (Original) Data Set

INFORMATION ABOUT ATTRIBUTE:

- | | |
|---------------------------------|---------------------------------|
| 1. Sample code number: | id number |
| 2. Clump Thickness: | 1 - 10 |
| 3. Uniformity of Cell Size: | 1 - 10 |
| 4. Uniformity of Cell Shape: | 1 - 10 |
| 5. Marginal Adhesion: | 1 - 10 |
| 6. Single Epithelial Cell Size: | 1 - 10 |
| 7. Bare Nuclei: | 1 - 10 |
| 8. Bland Chromatin: | 1 - 10 |
| 9. Normal Nucleoli: | 1 - 10 |
| 10. Mitoses: | 1 - 10 |
| 11. Class: | (2 for benign, 4 for malignant) |

3.1.2 Data Clustering Using Clustering Tool

Here the fuzzy clustering functions fcm and subclust, are used.

To open the tool, at the MATLAB command line, type:

```
findcluster
```

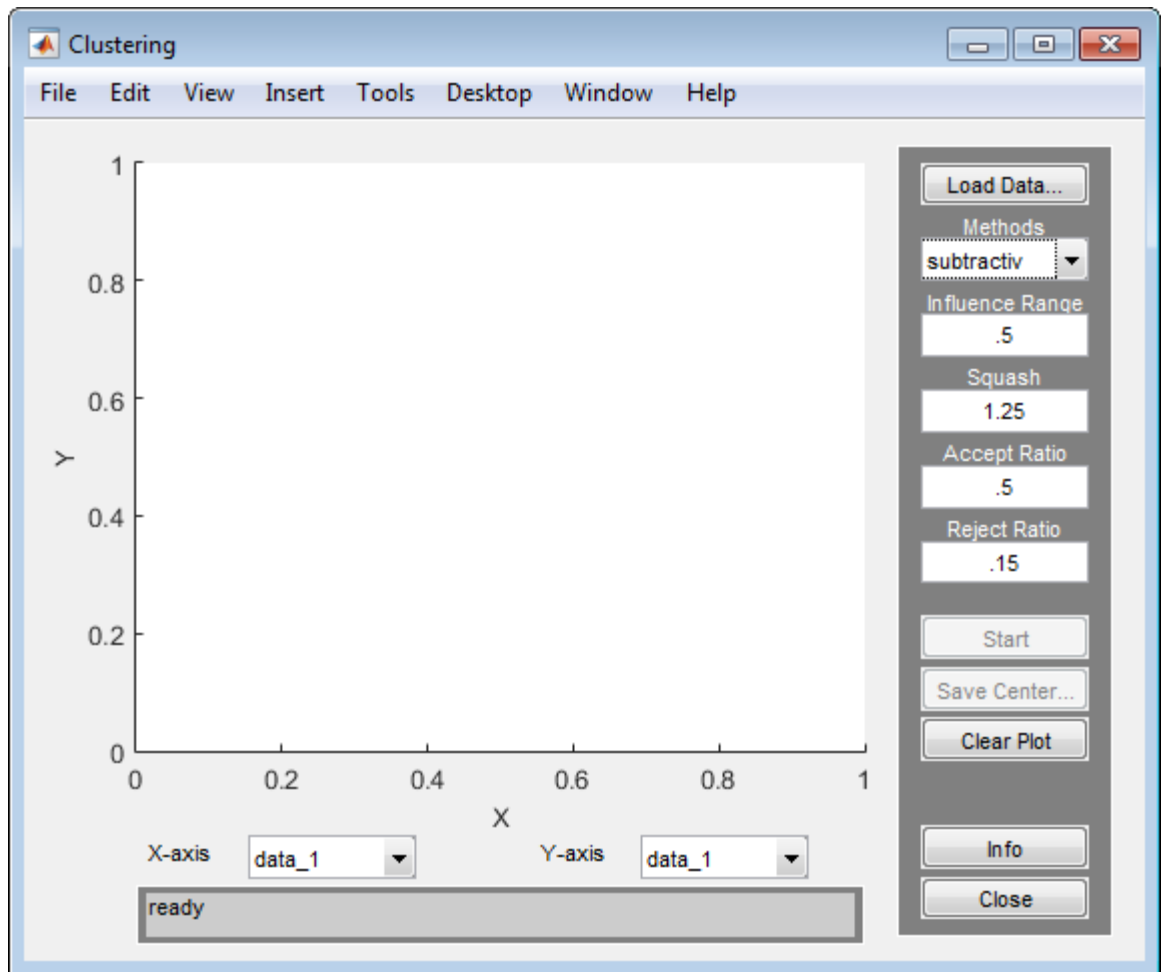


Figure3.1 :Load and plot the data in matlab.

Load and Plot Data

Open the Clustering Tool with a data set directly by calling `findcluster` with the data set as an input argument.

For example, enter:

```
findcluster('clusterdemo.dat')
```

The data set file must have the extension `.dat`. Each line of the dataset file contains one data point. For example, if you have 5dimensionaldata with 100 data points, the file contains 100 lines, and each line contains five values.

Cluster Data

To start clustering the data:

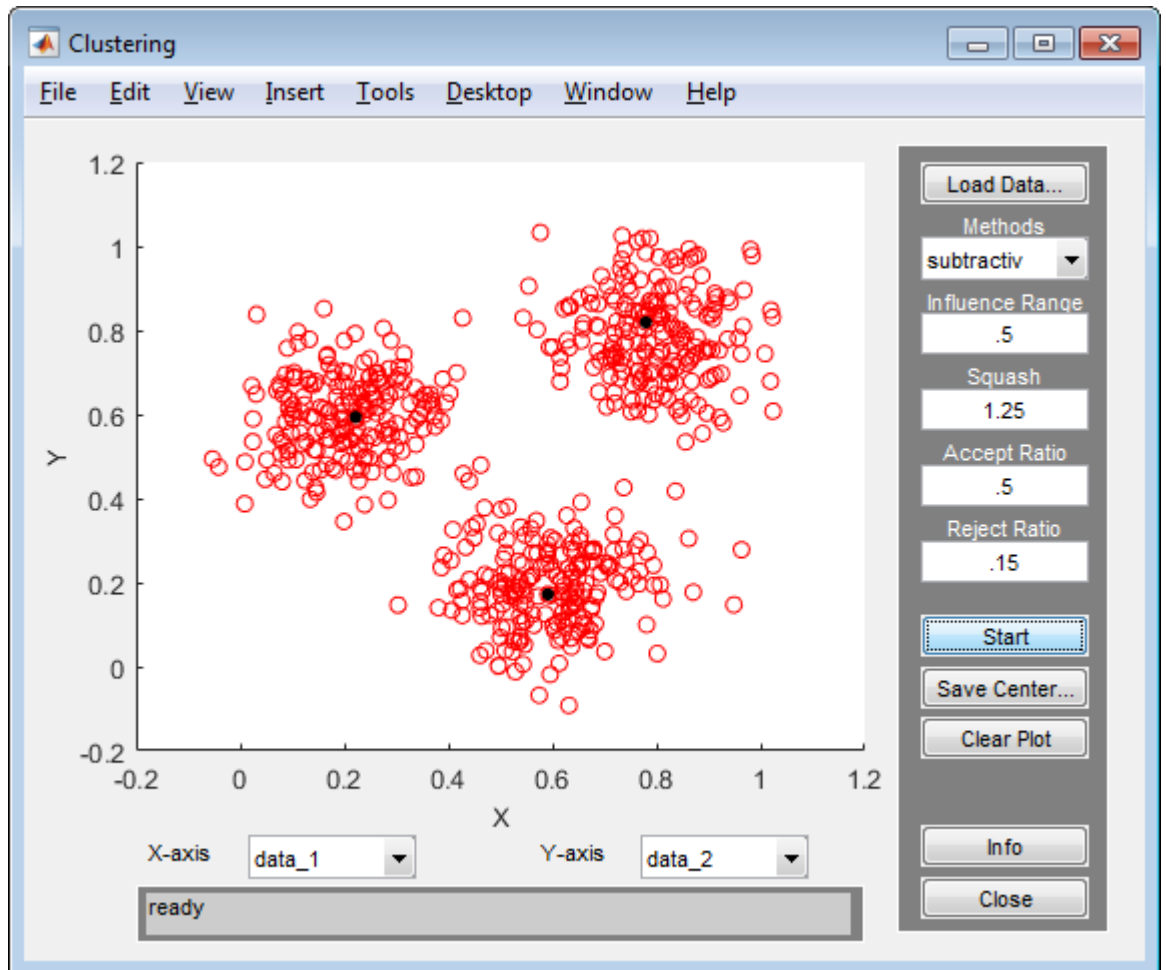


Figure 3.2: Clustering output

Using the Clustering tool, you can obtain only the computed cluster centers. To obtain additional information for:

- Fuzzy c-means clustering, such as the fuzzy partition matrix, cluster the data using `fcm`
- Subtractive clustering, such as the range of influence in each data dimension, cluster the data using `subclust`.

To use the same clustering data with either fcm or subclust, first load the data file into the MATLAB workspace. For example, at the MATLAB command line, type:

Save Cluster Centers

To save the cluster centers, click Save Center.

3.2 COMPUTATIONAL

KNN Algorithm

1. Select a value for K which is no of the neighbours
2. In second step calculate the Euclidian Distance for K numbers of neighbours.
- 3 Select the nearest neighbours K based on calculated Euclidean distance.
4. Now out of these K neighbours find the maximum number of neighbours belonging to a single class.
5. Now the new data has to assigned to the class in which maximum number of neighbours are belonging.
6. The model of KNN is ready .

Modified KNN

```
Output_label := MKNN ( train_set , test_sample )
Begin
  For i := 1 to train_size
    Validity(i) := Compute Validity of i-th sample;
  End for;
  Output_label := Weighted_KNN(Validity, test_sample);
  Return Output_label ;
End.
```

3.3 EXPERIMENTAL

3.3.1 Code for KNN

```
function [ acc ] = knn_loop( test_data, tr_data, attributes, k )
%This function applies knn algorithm for classification
%of dataset into malignant, benign classes

numoftestdata = size(test_data,1);
numoftrainingdata = size(tr_data,1);

accuracy = zeros(numoftestdata,1);
acc_ans = zeros(numoftestdata,1);
acc = 0;

for sample=1:numoftestdata

    %Computing euclidean distance
    R = repmat(test_data(sample,:), numoftrainingdata, 1);
    euclideanistance = zeros(numoftrainingdata, 1);
    for i=1:attributes-2
        euclideanistance = euclideanistance + (R(:,i+1) -
tr_data(:,i+1)).^2;
    end

    euclideanistance = sqrt(euclideanistance);

    [dist, position] = sort(euclideanistance, 'ascend');
    knneighbors = position(1:k);
    kndistances = dist(1:k);

    %Voting
    for i=1:k
        A(i) = tr_data(knneighbors(i), end);
    end

    M = mode(A);

    %Incrementing accuracy variable if truly classified
    if(test_data(sample, end) == M)
        acc = acc+1;
    end

end

acc = (acc/numoftestdata)*100; %Calculating accuracy percentage
end
```

The above code is implemented by us in the matlab for KNN algorithm.

Code for Modified KNN MKNN

```
function [ acc ] = MKNN( test_data, tr_data, attributes, k )
% This is modified KNN
    numoftrainingdata = size(tr_data,1);
    numoftestdata = size(test_data,1);
    validity = zeros(numoftrainingdata,1);
    Hnearestneighbors = 3;
    acc=0;

    %computing validity for train data set
    for i=1:(numoftrainingdata-Hnearestneighbors)
        for j=1:Hnearestneighbors
            validity(i,1) = validity(i,1) +
(tr_data(i,end)==tr_data(i+j,end));
        end
        validity(i,1) = validity(i,1)/Hnearestneighbors;
    end

    for i=(numoftrainingdata-Hnearestneighbors):numoftrainingdata
        for j=Hnearestneighbors:-1:1
            validity(i,1) = validity(i,1) + (tr_data(i,end)==tr_data(i-
j,end));
        end
        validity(i,1) = validity(i,1)/Hnearestneighbors;
    end

    for sample=1:numoftestdata

        %Computing euclidean distance
        R = repmat(test_data(sample,:),numoftrainingdata,1);
        euclideanistance = zeros(numoftrainingdata,1);
        %weight for classification
        weight = zeros(numoftrainingdata,1);

        for i=1:attributes-1
            euclideanistance = euclideanistance + (R(:,i) -
tr_data(:,i)).^2;
        end

        euclideanistance = sqrt(euclideanistance);

        for i=1:numoftrainingdata
            weight(i,1) = validity(i,1) * (1/(euclideanistance(i,1)+0.5));
        end

        %Voting
        classBenign =0;
        classMalign =0;

        for i=1:numoftrainingdata
            if tr_data(i,end) == 2
                classBenign = classBenign + weight(i,1);
            elseif tr_data(i,end) == 4
                classMalign = classMalign + weight(i,1);
            end
        end

        final_class = 0;
        if classBenign > classMalign
```

```

        final_class = 2;
    else
        final_class = 4;
    end

    %calculating accuracy
    if(test_data(sample,end) == final_class)
        acc = acc+1;
    end
end

acc = (acc/numoftestdata)*100;
end

```

Code For Main

```

dataset = csvread('D:\pdata.csv');
startValueofK = 60;
rangeOfK = 340;
accuracy = zeros((rangeOfK-startValueofK)/20,1);

%Training set is taken for first 100 datasets, rest are used for
%calculating accuracy of algorithm
for j=startValueofK:20:(startValueofK+rangeOfK)
    accuracy(floor(j/20)+1,1) =
knn_loop(dataset(1:100,:),dataset(1:599,:),11,j);
end

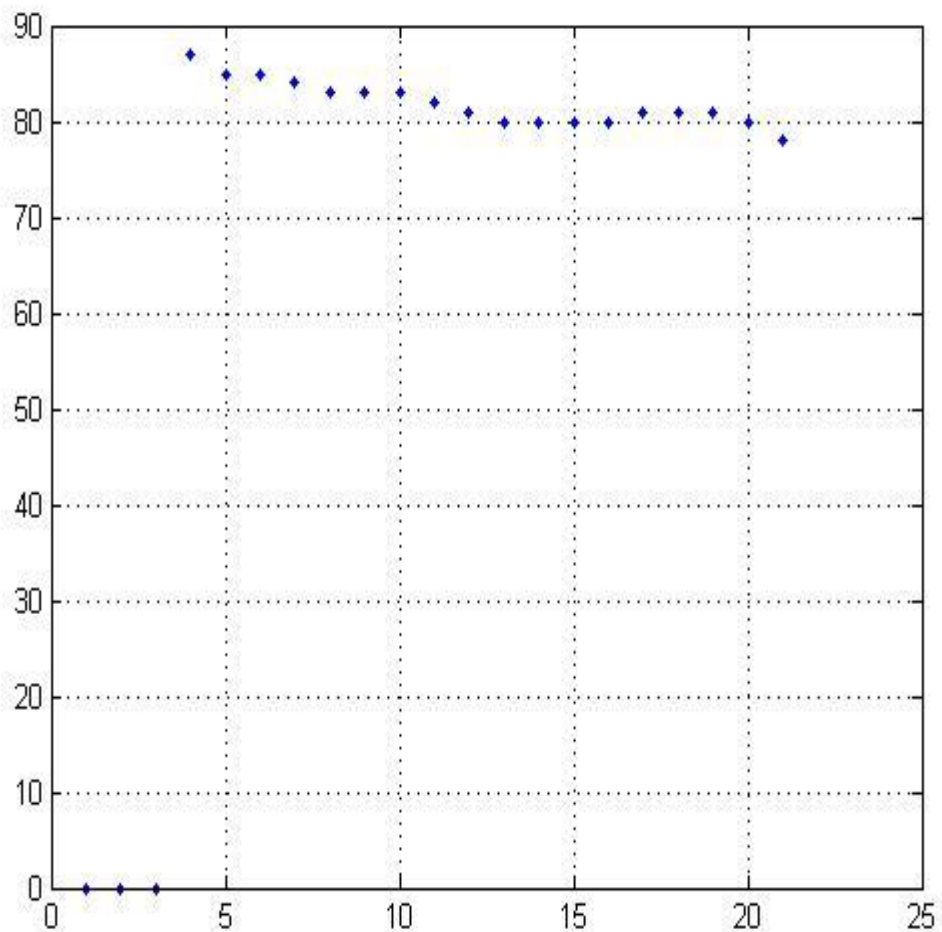
plot(accuracy','.'); %indicates knn accuracy
hold on;
grid on;

for j=startValueofK:20:(startValueofK+rangeOfK)
    accuracy(floor(j/20)+1,1) =
MKNN(dataset(1:100,:),dataset(1:599,:),11,j);
end

plot(accuracy,'k+'); %indicates modified knn accuracy

```

Experimental Output



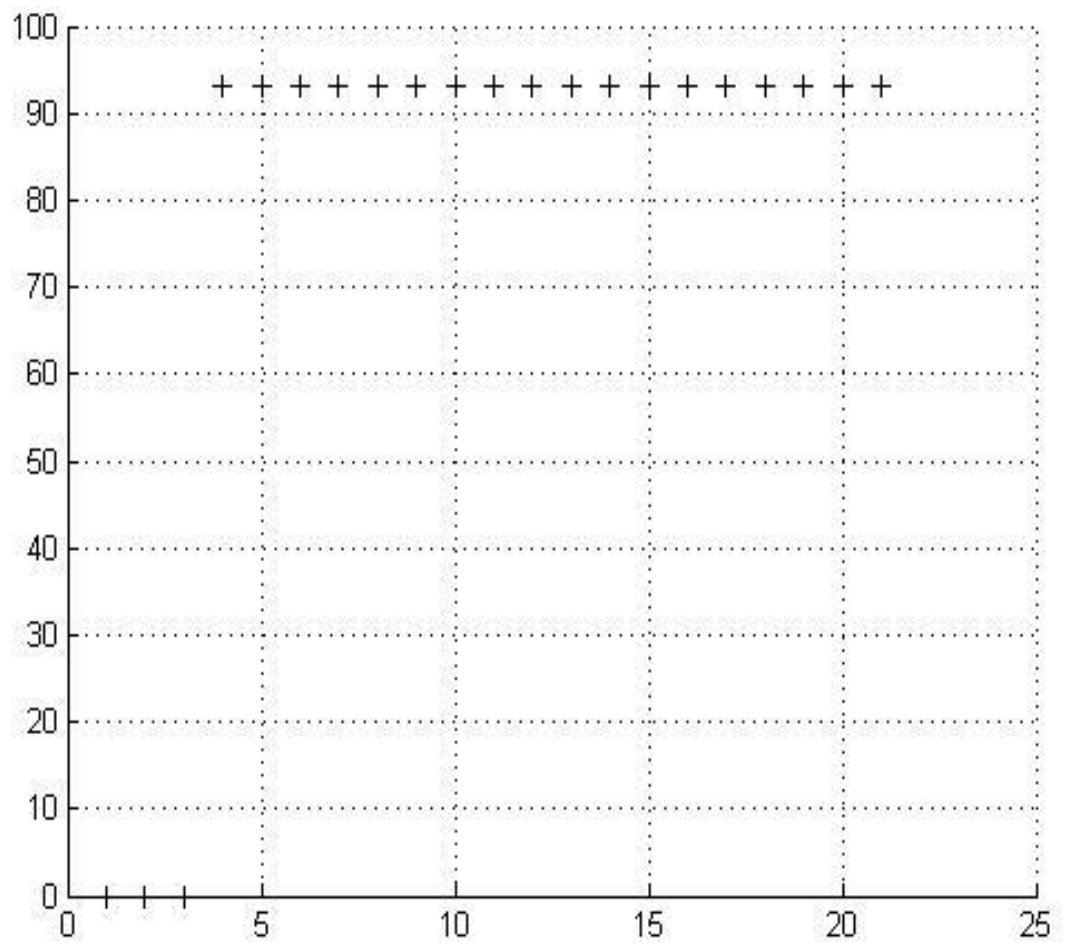
Graph 3.1: output of KNN algorithm

We obtain the above plot of graph when we run the code for KNN algorithm in matlab on our dataset which we have defined earlier in section 3.1.

From this plot we can get a fair idea of the working of KNN on our data set.

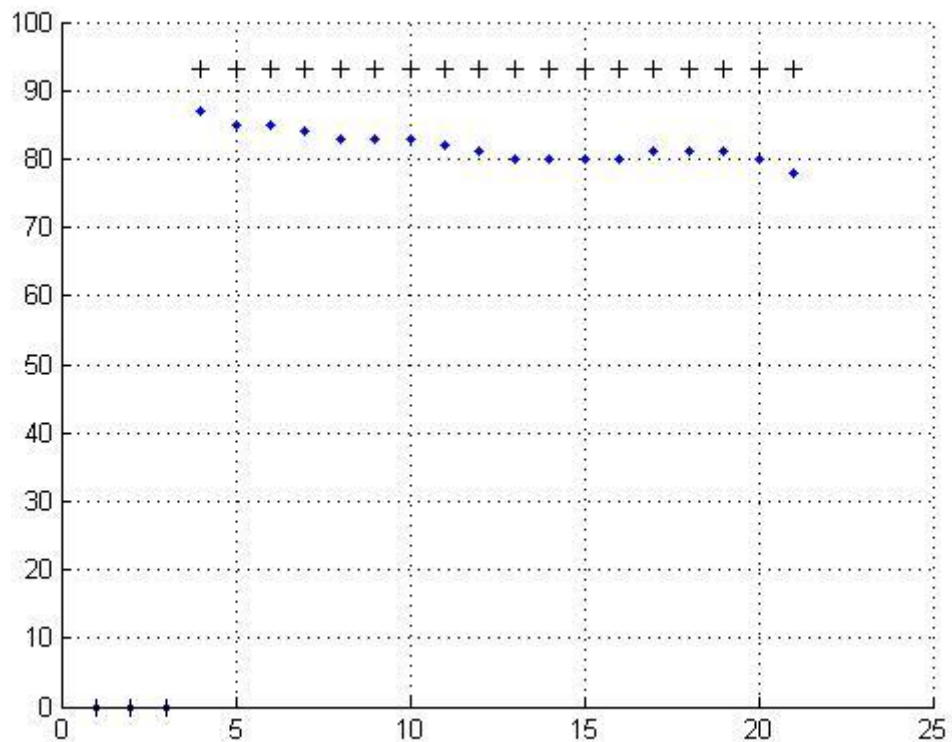
To further enhance the results and to validate and check our result we also implemented the Modified version of KNN and obtained the following output.

Output of MKNN



Graph 3.2: The result of MKNN algorithm.

When the results of both KNN and MKNN are plotted together



Graph 3.3: Output of both KNN and MKNN

Result: As clearly visible the modified version that is weighted KNN is showing enhanced accuracy. The plot of MKNN shows better accuracy and thus it is most suitable for working on our cancer data set.

3.4 Mathematical

3.4.1 Data Exploration

Our dataset contains (32) columns and (569). ‘*Diagnosis*’ column is to be predicted by us in this data set. We have taken this data set from UCI repository. M means malignant and B means benign. The value 1 indicates that the detected cancer id malignant and 0 depicts that its benign.

```
diagnosis
B    357
M    212
dtype: int64
```

We will use encoding

```
#Encoding categorical data values
from sklearn.preprocessing import LabelEncoder
labelencoder_Y = LabelEncoder()
Y = labelencoder_Y.fit_transform(Y)
```


Index	0
0	M
1	M
2	M
3	M
4	M
5	M
6	M
7	M
8	M
9	M
10	M
11	M
12	M
13	M
14	M

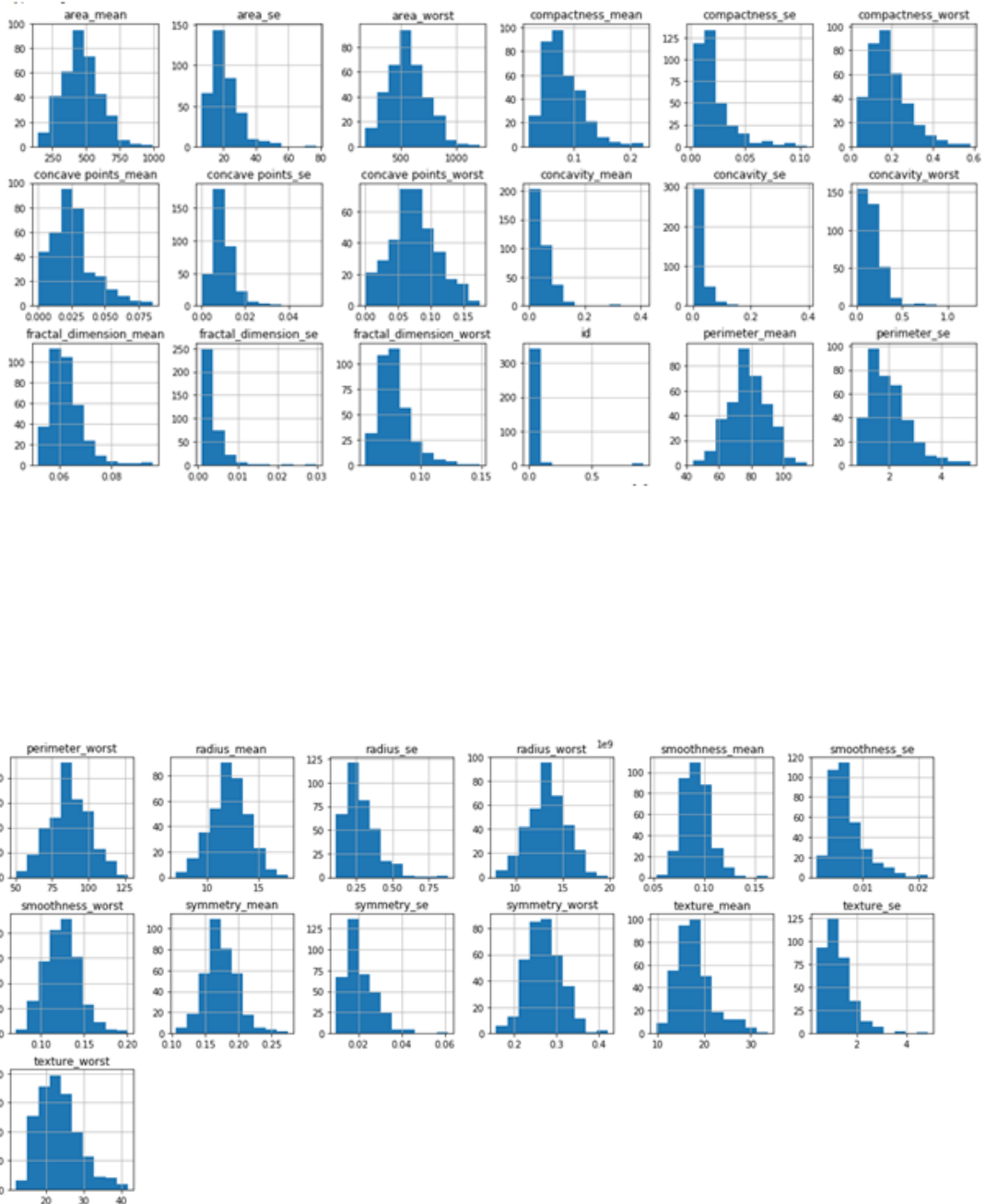
Fig 3.3 Diagnose without the encoding

	0
0	1
1	1
2	1
3	1
4	1
5	1
6	1
7	1
8	1
9	1
10	1
11	1
12	1
13	1

Fig3.4: Diagnose when the data is Encoded

3.4.2 Visualization

Visualization of data is very important and crucial aspect in data science. It helps in understanding the data and making other person understand the whole thing.



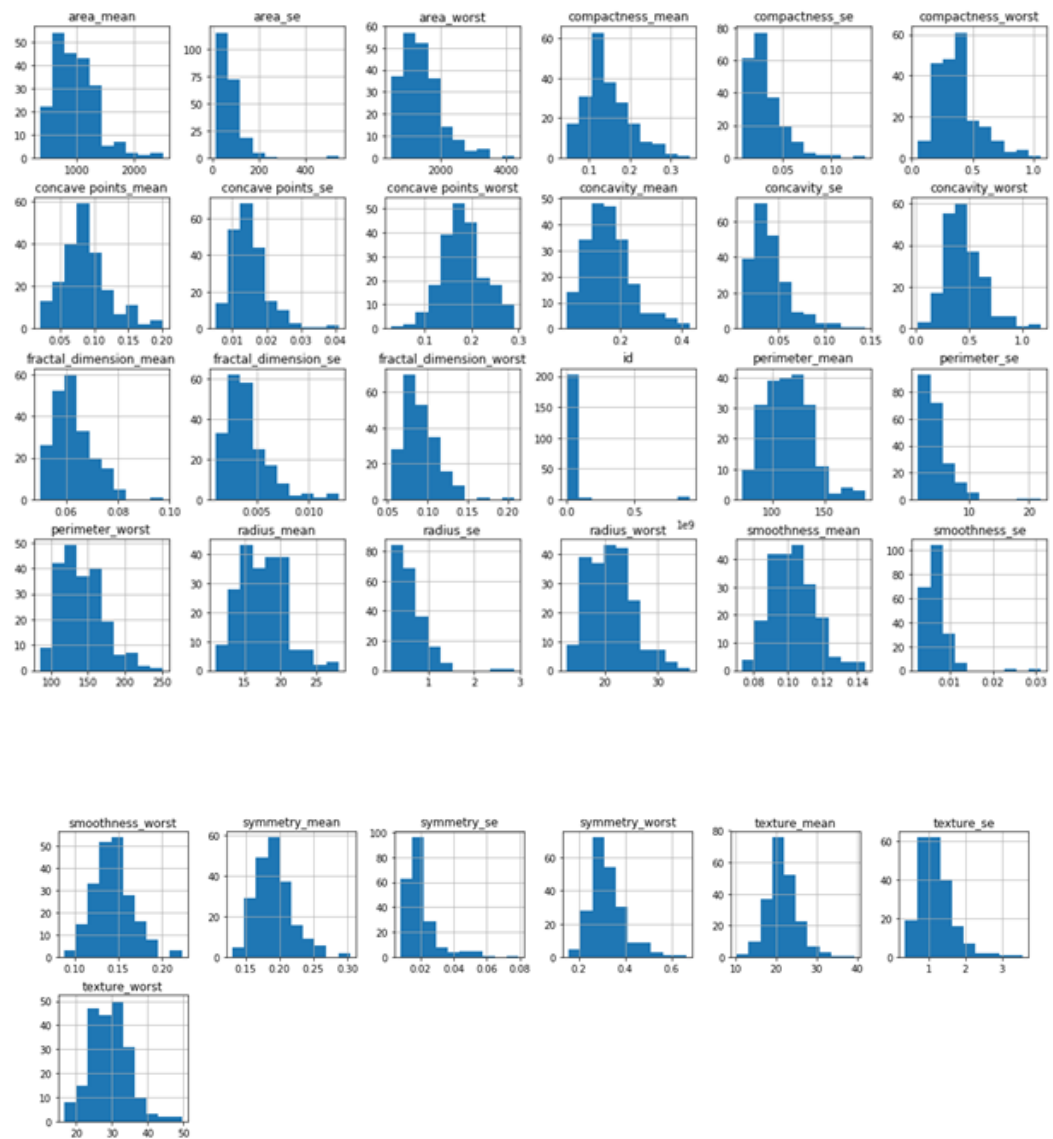


Figure 3.5: Visualization of dataset

Matlab provides various libraries to visualize the data set.

3.4.3 Model Training

```
# Let's drop the target label columns
X = df_cancer.drop(['target'],axis=1)

from sklearn.model_selection import train_test_split

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size =
0.20, random_state=5)
```

3.4.4 Evaluating the model

```
y_predict = svc_model.predict(X_test)
cm = confusion_matrix(y_test, y_predict)

sns.heatmap(cm, annot=True)
```

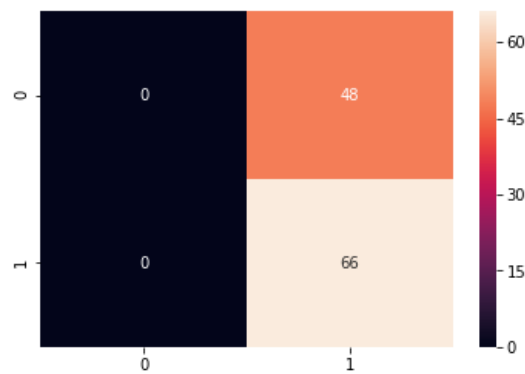


Figure3.10: Model evaluation

CHAPTER – 4 PERFORMANCE ANALYSIS

To get the fair idea of the performance of the system we are to compare the results of KNN and Modified KNN algorithm. The most appropriate results are obtained by MKNN algorithm.

We in this project are going to use the Classification Accuracy method to find out the accuracy of our developed models. Accuracy is an important factor in depicting the worth of our project .

$$\text{Accuracy} = \frac{\text{Number of Correct predictions}}{\text{Total number of predictions made}}$$

	0	1
0	87	3
1	3	50

Fig 4.1: confusion matrix

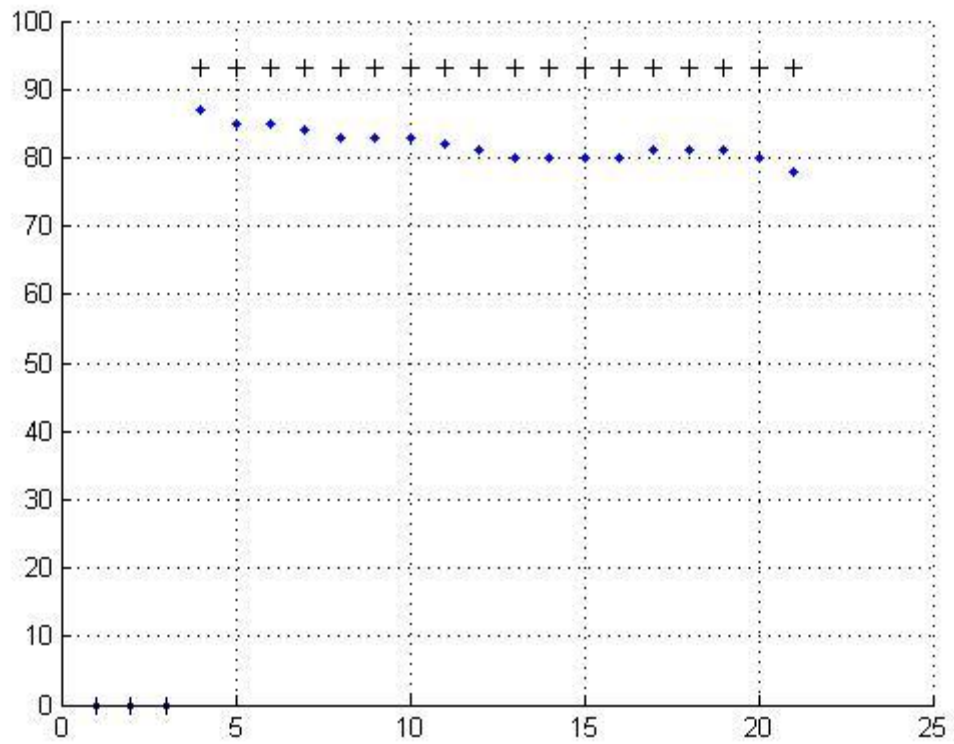


Fig 4.2 Accuracy comparison

From the above plot of MKNN against that of the KNN algorithm implemented in matlab, the accuracy of the system can be determined. The accuracy given of the MKNN is more than that of KNN and this plot verifies this claim and establishes accuracy.

CHAPTER – 5 CONCLUSIONS

5.1 Conclusions

Finally from this research we have come to the conclusion that with the help of the machine learning algorithms and by implementing those in the medical field especially if we talk about the cancer, it is possible to detect the cancer in patients. As far as the Breast cancer is concerned the Modified KNN algorithm is very helpful in early detection of cancer and thus the patient can be provided with early treatment which can potentially save many lives. KNN algorithm is also very good and provides good results but MKNN is better.

5.2 Future Scope

This investigation uses four particular clustering counts. In future this work is loosening up by applying remarkable gathering and alliance mining figuring. In this work WEKA open source data burrowing instrument is used for the explanation of the investigation. In future Orange, Tavera, Rapid Miner and other data mining instrument and relationship examination of their result give progressively perfect outcomes.

References

- [1] Rinal Doshi, “DEVELOPMENT OF PATTERN KNOWLEDGE DISCOVERYFRAMEWORK USING CLUSTERING DATA MINING ALGORITHM”, International journal of computer engineering & Technology (IJCET), ISSN 0976 – 6367(Print), ISSN 0976 – 6375(Online), Volume 4, Issue 3, May-June (2013), pp. 101-112
- [2] WEKA, “The University of Waikato”, machine learning group, weka documentation.
- [3] Wikipedia.com
- [4] Mansour, Nashat ; Department of Computer Science and Mathematics, Lebanese American University, Beirut, Lebanon ; Zantout, Rouba ; El-Sibai, Mirvat”Mining
- [5]<https://www.researchgate.net/>
- [6] Proceedings of the World Congress on Engineering and Computer Science 2008 WCECS 2008, October 22 - 24, 2008, San Francisco, USA

JAYPEE UNIVERSITY OF INFORMATION TECHNOLOGY, WAKNAGHAT
PLAGIARISM VERIFICATION REPORT

Date: 20/07/2020

Type of Document (Tick) B.Tech Project

Name: Amandeep Saini; Anurudh Sharma Department: CSE Enrolment No 161244; 161238

Contact No. 9797777033 E-mail. amandeepsinghsingh98@gmail.com

Name of the Supervisor: Dr. YUGAL KUMA Title of the Thesis/Dissertation/Project

Report/Paper (In Capital letters): ALGORITHM FOR HEALTH CARE DATA ANALYSIS

UNDERTAKING

I undertake that I am aware of the plagiarism related norms/ regulations, if I found guilty of any plagiarism and copyright violations in the above thesis/report even after award of degree, the University reserves the rights to withdraw/revoke my degree/report. Kindly allow me to avail Plagiarism verification report for the document mentioned above.

Complete Thesis/Report Pages Detail:

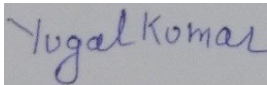
- Total No. of Pages =
- Total No. of Preliminary pages =
- Total No. of pages accommodate bibliography/references =



(Signature of Student)

FOR DEPARTMENT USE

We have checked the thesis/report as per norms and found **Similarity Index** at 0 (%). Therefore, we are forwarding the complete thesis/report for final plagiarism check. The plagiarism verification report may be handed over to the candidate.



(Signature of Guide/Supervisor)

Signature of HOD

FOR LRC USE

The above document was scanned for plagiarism check. The outcome of the same is reported below:

Copy Received on	Excluded	Similarity Index (%)	Generated Plagiarism Report Details (Title, Abstract & Chapters)	
	<ul style="list-style-type: none">• All Preliminary Pages• Bibliography/Im a ges/Quotes• 14 Words String		Word Counts	
Report Generated on			Character Counts	
		Submission ID	Total Pages Scanned	
			File Size	

Checked by
Name & Signature

Librarian