2023 2nd International Conference on Vision Towards Emerging Trends in Communication and Networking Technologies (ViTECoN) | 979-8-3503-4798-2/23/$31.00 ©2023 IEEE | DOI: 10.1109/VITECON58111.2023.10157765

Breast Cancer Prediction Using Unsupervised Learning Technique K-Means Clustering Algorithm

Soumyalatha Naveen[1], Nachiketh.V.Kashyap[2], Varun P Kulkarni[3], Sandeep A[4], Meruva Sai Chakradhar[5]
[1]*School of Computer Science and Engineering*
[2,3,4,5] *School of Electronics and Communication Engineering*
REVA UNIVERSITY, Bangalore

Abstract- **Breast cancer is a highly deadly and common cancer that is spreading over the world and claiming many lives. Breast cancer develops in the glandular tissue of the breast in the lining of epithelial cells of ducts or (15%) lobules. These tissues heal with time. These in situ tumours may develop over time and infect the breast cells' immediate environs (stage 1), impact the lymph nodes, and finally totally spread throughout the body's organs (stage 3). There were 685,000 deaths worldwide in 2020 and a later rise in the number of malignant women. Cancer was a common disease that affected 7.8 million of the female population.**

The k-means algorithm is a popular data clustering algorithm. As the main analytical routine in data mining, the techniques of the clustering algorithm will impact the clustering outcome directly. This paper examines the shortcomings of the standard k-means algorithm and discusses them. This paper reviews existing methods for selecting the number of clusters for the algorithm. However, one of its drawbacks is the requirement that the number of clusters, K, be specified before the algorithm is applied. Therefore, we obtained an accuracy of 85% after executing the program, and we were able to depict the difference between a benign and a malignant tumour. And we were able to see the centroid between the two tumours.

***Keywords* - Cancer prediction, Data mining, machine learning, k-means clustering, Breast Cancer early diagnosis, Future prediction.**

I. INTRODUCTION

Breast cancer is the second-most heterogeneous disease after skin cancer. Breast cancer is a serious problem among women in the United States. Each year in the United States, about 264,000 cases among women are diagnosed. One of the major risk factors for breast cancer that cannot be changed includes getting older and genetic mutations. Hence, to reduce the risk and for early detection, many data mining and machine learning techniques can be used to identify breast cancer.

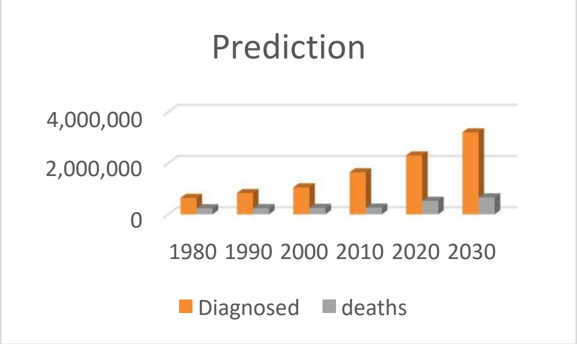

Fig 1 depicts the survey on possible victims of breast cancer over a period of 10 years.

Fig 1 shows the possible victims of breast cancer over a period of 10 years and the increase in death rates over a period.

Data mining is the process of extracting or mining data to obtain useful information. Due to the massive amount of data that is readily available and the pressing need to transform that data into information and knowledge, data mining has received a lot of attention recently in both the information industry and society. Because of the natural evaluation of information technology, data mining can be seen. The following functionalities have developed along an evolutionary route in the database system sector. gathering information and building databases. In data mining, many machine learning algorithms are used to find the solution to a given problem. With the massive amount of data that is readily available and the pressing need to transform that data into information and knowledge, data mining has received a lot of attention recently in both the information industry and society. Because of the natural evaluation of information technology, data mining can be seen. The following functionalities have developed along an evolutionary route in the database system sector: gathering information and building databases. In data mining, many machine learning algorithms are used to

979-8-3503-4798-2/23/$31.00 ©2023 IEEE

find the solution to a given problem. One of the most prominent unsupervised machine learning algorithms among them is k-means clustering.

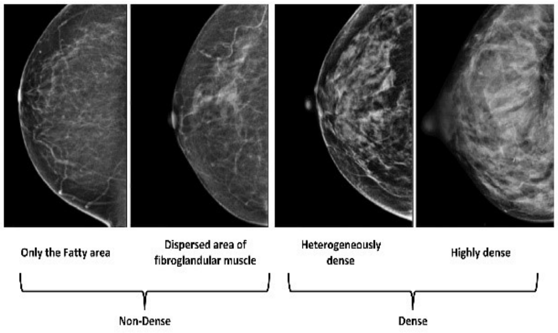

Fig. 2: X-ray mammography

Fig 2 depicts mammography. It is a technique in which the image uses low doses of x-rays to scan images of breast tissue. It is used in the detection of early-stage cancers that have no symptoms. In this process, the breast is placed on the plate, another plate is placed on the breast, and it is compressed to see the breast images. X-rays are then used to create images of breast tissue from different angles.

The K-means algorithm works by initially selecting K random centroids as the initial cluster centres, where K is the number of clusters that we want to create. The centroid of each cluster is then recalculated as the mean of the data points assigned to that cluster after iteratively assigning each data point to the nearest centroid. The letter k means clustering, which has a variety of applications, including image segmentation and anomaly detection. The most significantly used application in our context is the image segmentation application. The image segmentation application is used widely and distinctly in many real-time applications, among which the noted application includes the detection of breast cancer. Breast cancer[1-6] is one of the most substantial cancers, and it is spreading among the generations. A woman in India is diagnosed with breast cancer every four minutes, and there are 90,000 annual deaths from the disease. Tragically, a woman in the nation passes away from breast cancer every eight minutes. One woman with breast cancer dies from it for every two who are diagnosed

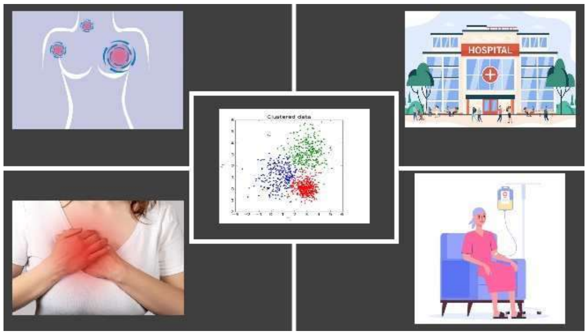

Fig 3. Scenario of early detection and diagnosis of breast cancer by implementing clustering algorithm.

Fig. 3: Processes for diagnosing cancer using the K-means clustering algorithm.

II. K-MEANS CLUSTERING

We usually classify data by grouping them based on the similarities they share among them. And K-means Cluster is such an unsupervised grouping algorithm that segments the data upon different traits and places the data like one another in various clusters.

The algorithm mainly performs these two tasks. Determining the most suitable value for centroids (K center points).

Arranging the data values closest to these centroids and further creating clusters

K-Means Clustering[7] is utilized when there is no particular outcome variable to be found. Instead, it is used when there is a set of criteria or features to be used to find a group of observations that include similar properties.

K-means Clustering[8] can be used practically in any possible domain, from disease analysis to image segmentation[9-13]. With the appropriate data set and K value, the program can be seamlessly used in any field.

A. Types of Clustering

i. Hierarchical Clustering

This kind of clustering creates clusters that have a predetermined ordering from top to bottom.

ii. Partitional clustering

It separates the object data into non-overlapping sets, meaning no unique data can be present in more than one cluster.

iii. Density-Based Clustering

It arranges the data based on the density of data points on a site. Clusters are created where the density is higher compared to other regions.

B. Pseudocode of K-Means Clustering.

1) Installing the required packages.
2) Importing data files.
3) Choosing the number of Centroids(K).
4) Obtaining the value of Data Points.
5) Placing the Centroids $K_{1,} K_{2} \ldots K_{n.}$
6) For every Centroid find the nearest data point; assign the points to those centroids and form a cluster.
7) Find the mean for every cluster and make it the new Centroid.

Perform 4 & 5 steps again until all the data gets clustered for a required number of iterations.

C. Flowchart of K-Means Clustering

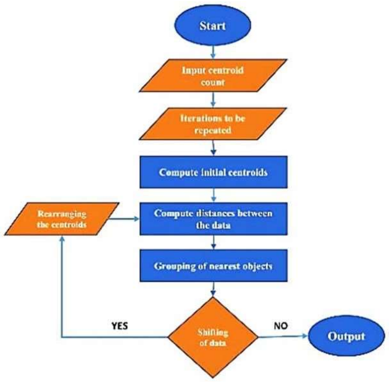
Start
Input centroid count
Iterations to be repeated
Compute initial centroids
Rearranging the centroids
Compute distances between the data
Grouping of nearest objects
YES
Shifting of data
NO
Output

Fig 4. Flowchart of K-Means Clustering

Figure 4 represents the methods of k means clustering that are followed during the execution of the program.

D. Advantages and Disadvantages of K-Means Clustering

Advantages:

1. It can cluster large size data.
2. It guarantees convergence of the Data Sets.
3. Highly flexible to new kinds of data examples.
4. Seamlessly adapts to various types of data.

Disadvantages:

1. Finding the value of K.
2. Entirely depending on the first values.
3. K-means has trouble clustering data of varying sizes and density.
4. Sometimes the outliers will also be included in the cluster.

E. Limitations of K-Means Clustering

1. Managing Empty Clusters, sometimes null clusters occur if no points are allocated to a cluster during the assigned step.
2. Since the program is an unsupervised algorithm, the data objects are relatively harder to find similarities, among the unlabeled data set.
3. With various kinds of clustering methods present, each one has different outcome results and percentage.

F. Application of K-Means Clustering

Many applications[14-17] used Machine learning algorithms for predictive analysis. Few example as follows.

Image segmentation- The program groups the pictures based on similar values of their attributes.

Market research – Analysis of the customer repetitive purchases, suggestion of new product requirement ideas.

Spam detection – The program clusters the previous spam occurrence and detects it earlier in future scenarios.

III. PROPOSED WORK

We accomplished the work with Python Idle using the packages that have been imported in the command prompt. We imported packages like pandas, which are used for data framing; NumPy, which is used to execute a wide collection of mathematical operations on arrays; matplotlib, which helps to create animated and visualizations in python; mpl_toolkits, which helps to create basic 3D plotting tools; and sklearn, which is a machine learning library for python and helps clustering algorithms that are designed to work with python NumPy.

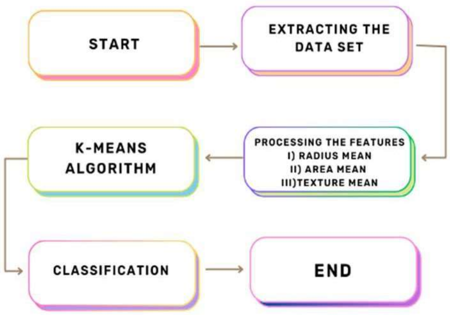

Fig 5. Algorithm Flowchart of proposed work

Figure 5 shows the algorithmic flow of events depicting the process carried out across the program.

The data is imported from Wisconsin and consists of over 569 individuals. Some parameters were used, like the radius mean, compactness mean, smoothness mean, area mean, and texture mean. The term compactness mean is used because many important properties of continuous functions only hold when the domain is a compact set. Smoothness Mean is used to give a general idea of relatively slow changes in value, with little attention paid to the close matching of data values.

To find malignant vs. benign, the Keans cluster module is used for setting to find two clusters; the number of clusters is 2 and the maximum iterations are 300. K-Means clustering is then achieved using Scikit-Learn's 'KMeans()' function, with the number of clusters set to 2. The 'fit()' method is called on the systematized to train the k-means model. Eventually, the resulting clusters are visualized using a scatter plot created via Matplotlib's first two features of the dataset are plotted on the x and y axes, and the points are shaded accordingly to their cluster operations using the 'labels attribute of the k-means model.

The outcome that we achieved by using some parameters like area mean, radius mean, and texture mean Area Mean is often used to define the amount of space taken up by a 2D shape or surface in square units. Radius means it represents the ratio of a circle's circumference to its diameter and gets the radius of the circle from the user using the input function. Texture segmentation, also known as texture, mean analysis, can be used to identify the boundaries of textures. When an object's texture best describes it in a picture, texture analysis can be applied.

In our usage case, we already knew the number of clusters to be formed, i.e., malignant and benign cancer. Though primarily the K value (the number of clusters) should be manually found as per the features of the data to find the accurate grouping.

The methods used to acquire K values are the elbow method, silhouette, gap statistic, and information criterion. Often the elbow method is implemented to find the suitable number of clusters.

The Elbow method plots the sum of square distances of every data point from the assigned centroid to obtain the optimal value of k. As the plotted graph begins to flatten, the convergence point is the K value, this curve resembles an elbow.

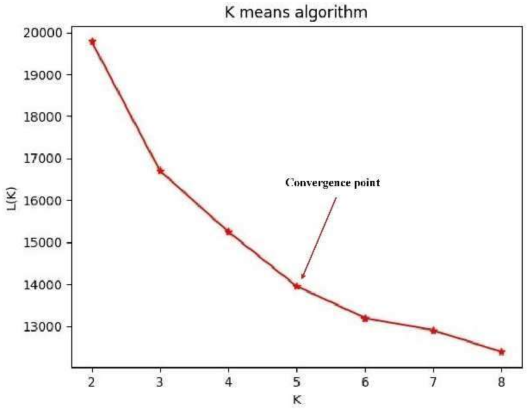

Fig 6. The curve plotted using the Elbow method.

Fig 6 intimates the point where the curve seems to bend or flatten is the optimum K value.

IV. RESULT

Through the K-means algorithm, the Euclidean formula is accustomed to calculating the distance between each data point and the centroids of the K clusters. The previously mentioned distance calculation is used to allow each data point to reach its closest centroid and to restore the centroids. The Euclidean distance formula for calculating the distance between a data point X_i and a centroid Y_j can be expressed as

$$\text{Euclidean distance} = \sqrt{(X_1 - Y_1)^2 + (X_2 - Y_2)^2} \quad (1)$$

Where,

X_1 & X_2: Observation value of Data X_i

Y_1 & Y_2: Observation value of Centroid Y_j.

The square root of the sum of squared differences between each feature of the data point and centroid is determined to achieve the Euclidean distance.

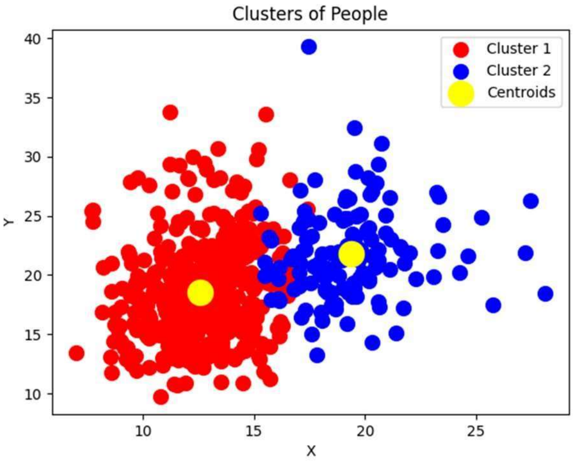

Fig. 6 Clustering of benign and malignant cancers by the algorithm.

Fig 6 indicates the differentiation of benign and malignant tumours and the identification of the centroid.

The k-means clustering algorithm was used to obtain the results, which are explained along with several instances that take variable parameters into account. The computations were based on the iteration, threshold, attribute, and centroid distance split. This method yielded an average positive prediction accuracy of about 85%. Using the same centroid and the largest variance, better results were built. Python and Idle were used to arrive at the results. We obtained almost 77% accuracy using unsupervised hierarchical clustering and 85% using K-means Clustering.

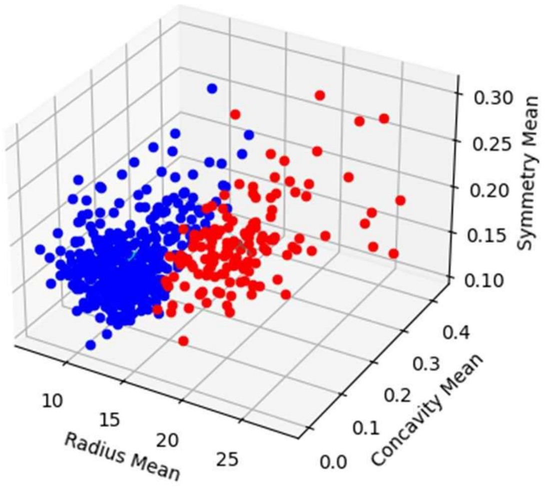

Fig. 7 A 3D pictorial view of how the data set is scattered within the clusters.

As shown in Fig 7, the data set is scattered, and there is no difference between the benign and malignant tumors.

K-Means Clustering is utilized to find almost all correct breast cancer classifications by applying different types of distance measures, scoring methods, and initialization values. The conclusion of the radiologists can be merged with clustering to improve accuracy. It is recommended to use various clustering techniques to study breast cancer data.

Discussing the result using K-Means clustering algorithms with different cases and considering every variable is how the results were obtained. According to what had been stated previously, the computations were made using the centroid distance split, threshold, epoch II, and iteration. 85% of the predicted accuracy was achieved with this approach. This was the case with the same centroid and the highest variance.

V. Conclusion

The most prevalent malignancy among women is breast cancer. So, it becomes very crucial to be able to anticipate it in advance. The Wisconsin breast cancer is analysed to make new methods to forecast cancer using the clustering algorithm.

The research on how cancerous cells can be distinguished between malignant and benign cells using an unsupervised algorithm K-means clustering separates the data objects and gives a clear understanding of the kind of cancer the patient is suffering from. The findings in this paper show the various kinds of parameters used to group the cancer types. By performing the above program, we come to the final verdict that the cancer tissue can be separated using k-means clustering, the spreading tissue can be found and diagnosed, and the patient can be treated with the appropriate medicine that is required to cure the cancer.

Overall, the results show that this approach gives more than 85% accuracy in classifying tumours as benign or malignant.

ACKNOWLEDGMENT

Authors acknowledge the support from REVA University for the facilities provided to carry out the research.

References

[1] Dubey, Ashutosh Kumar, Umesh Gupta, and Sonal Jain. "Analysis of k-means clustering approach on the breast cancer Wisconsin dataset." International journal of computer assisted radiology and surgery 11 (2016): 2033-2047.

[2] García-Becerra, Rocío, et al. "Mechanisms of resistance to endocrine therapy in breast cancer: focus on signaling pathways, miRNAs and genetically based resistance." International journal of molecular sciences 14.1 (2012): 108-145.

[3] Giaquinto, Angela N., et al. "Breast cancer statistics, 2022." CA: A Cancer Journal for Clinicians 72.6 (2022): 524-541.

[4] Rajeswari, K., et al. "Improvement in K-means clustering algorithm using data clustering." 2015

International Conference on Computing Communication Control and Automation. IEEE, 2015.

[5] Henley, S. Jane, et al. "Invasive cancer incidence—United States, 2010." Morbidity and mortality weekly report 63.12 (2014): 253.

[6] Zheng, Bichen, Sang Won Yoon, and Sarah S. Lam. "Breast cancer diagnosis based on feature extraction using a hybrid of K-means and support vector machine algorithms." Expert Systems with Applications 41.4 (2014): 1476-1482.

[7] Kapil, Shruti, Meenu Chawla, and Mohd Dilshad Ansari. "On K-means data clustering algorithm with genetic algorithm." 2016 Fourth International Conference on Parallel, Distributed and Grid Computing (PDGC). IEEE, 2016.

[8] SIMMS, CIARAN KNUT. "Vehicle pedestrian collisions: Validated models for pedestrian Impact and Projection." (2005).

[9] Wang, Hong, and Gongping Chen. "Personalized service recommendation for mobile voice communication users based on improved k-means algorithm." International Journal of Speech Technology (2021): 1-8.

[10] Shrilakshmi, K., and Soumyalatha Naveen. "Player Rating Correlation Prediction Using Machine Learning." 2021 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT). IEEE, 2021.

[11] Shalini, Lingampally, Soumyalatha Naveen, and U. M. Ashwinkumar. "Prediction of Automobile MPG using Optimization Techniques." 2021 IEEE Madras Section Conference (MASCON). IEEE, 2021.

[12] Barkana, Duygun Erol, and Engin Masazade. "Classification of the Emotional State of a Subject Using Machine Learning Algorithms for RehabRoby." Artificial Intelligence: Concepts, Methodologies, Tools, and Applications. IGI Global, 2017. 2160-2187.

[13] Jeeva, M., E. Padmapriya, and Rajesh George Rajan. "Hybridization of ML techniques for predicting Breast Cancer." 2022 Third International Conference on Intelligent Computing Instrumentation and Control Technologies (ICICICT). IEEE, 2022.

[14] Mardi, Mahnaz, and Mohammad Reza Keyvanpour. "GBKM: A New Genetic Based K-Means Clustering Algorithm." 2021 7th International Conference on Web Research (ICWR). IEEE, 2021.

[15] Abdulla, Srwa, Ali Sagheer, and Hadi Veisi. "Breast cancer segmentation using K-means clustering and optimized region-growing technique." (2022).

[16] Kashyap, Abhishek, and Soumyalatha Naveen. "A Comparative Study on Prediction of PM2. 5 Level Using Optimization Techniques." 2021 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT). IEEE, 2021.

[17] Bhattacharjee, Ananya, R. Murugan, and Tripti Goel. "A hybrid approach for lung cancer diagnosis using optimized random forest classification and K-means visualization algorithm." Health and Technology 12.4 (2022): 787-800.