MATICS : Jurnal Ilmu Komputer dan Teknologi Informasi   ISSN : 1978-161X(p); 2477-2550(e)
(*Journal of Computer Science and Information Technology)*   *Volume* 15, No. 2(2023), pp 72-78
Received : September 25th 2023; **Accepted** : October 9th 2023 ; **Avalaible Online** : October 23th 2023

Utilizing the K-Means Algorithm for Breast Cancer Diagnosis: A Promising Approach for Improved Early Detection

Nur Fitriyah Ayu Tunjung Sari, Maharini Nabela, Muhammad Falah A

Abstract—**Breast cancer is a pressing non-communicable disease, especially affecting women, with its incidence on the rise. In 2020, it ranked among the most common cancers in Indonesia. Timely detection and precise diagnosis are pivotal for effective breast cancer management. To enhance diagnostic accuracy, the K-means clustering method is applied to group patients based on shared attributes. This research aims to contribute significantly to breast cancer diagnosis by leveraging the K-means method, potentially improving patient survival rates.**

The research process involves data collection, preprocessing, K-means application, evaluation, and visualization. A dataset of 569 breast cancer patient records with 32 attributes from Kaggle is utilized. The K-Means algorithm is assessed using accuracy, yielding a value of 0.8457, signifying good performance. Malignant cases (211) and benign cases (301) are visualized in a scatter plot, distinguishing between them.

In conclusion, this study presents an initial step in utilizing the K-means algorithm for breast cancer diagnosis, offering promising results. Further research and the development of more advanced models are imperative to address the global health challenge posed by breast cancer among women.

Index Terms—**breast cancer; clustering; K-Means Algorithm**

I. Introduction

Breast cancer is one of the non-communicable diseases that has a significant impact on public health. Breast cancer has emerged as one of the most prevalent diseases, particularly among women, with an increasing incidence rate in recent years, making it a significant public health concern.

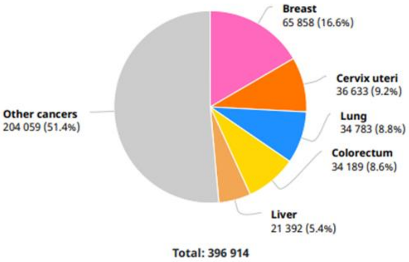
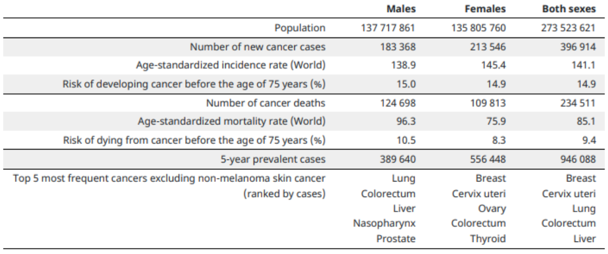


Fig. 1. Cases of cancer in Indonesia in the year 2020

According to data officially released by the Global Cancer Statistics (GLOBOCAN) in 2020 (Figure 1), breast cancer is one of the most common types of cancer in Indonesia. The chart below illustrates the cases of cancer in Indonesia in the year 2020. Based on the statistical table released by GLOBOCAN (Figure 2), breast cancer is a type of disease that is frequently found in women, and the statistics indicate that its prevalence is higher among the female population compared to males.



	Males	Females	Both sexes
Population	137 717 861	135 805 760	273 523 621
Number of new cancer cases	183 368	213 546	396 914
Age-standardized incidence rate (World)	138.9	145.4	141.1
Risk of developing cancer before the age of 75 years (%)	15.0	14.9	14.9
Number of cancer deaths	124 698	109 813	234 511
Age-standardized mortality rate (World)	96.3	75.9	85.1
Risk of dying from cancer before the age of 75 years (%)	10.5	8.3	9.4
5-year prevalent cases	389 640	556 448	946 088
Top 5 most frequent cancers excluding non-melanoma skin cancer (ranked by cases)	Lung Colorectum Liver Nasopharynx Prostate	Breast Cervix uteri Ovary Colorectum Thyroid	Breast Cervix uteri Lung Colorectum Liver

Fig. 2. Statistical data on cancer cases in the year 2020

Early detection and accurate diagnosis play a crucial role in the management of breast cancer. The identification of breast cancer types continues to be improved through research and the development of new methods, with the aim of enhancing efficiency and accuracy in the diagnostic process. One of the methods that can be applied is the K-Means method. The K-means method is one of the clustering algorithms used to group data based on their similar characteristics [1]. The K-means method is employed to identify patterns and categorize patient data based on relevant attributes.

This research aims to apply the K-means method in the diagnosis of breast cancer types. Through the analysis of collected patient data, the K-Means method is used to group patients into clusters with similar

Manuscript received September 25, 2023. This work was supported in part by Informatics Engineering Department of Maulana Malik Ibrahim Islamic State University.

Nur Fitriyah Ayu Tunjung Sari is with the Informatic Engineering Departement of Maulana Malik Ibrahim Islamic State University, Malang, Indonesia (email nur.fitriyah@ti.uin-malang.ac.id)

Maharini Nabela is with the Informatic Engineering Departement of Maulana Malik Ibrahim Islamic State University, Malang, Indonesia (email maharininabela4@gmail.com).

Muhammad Falah Abdurrohman is with the Informatic Engineering Departement of Maulana Malik Ibrahim Islamic State University, Malang, Indonesia (email mfalah16@gmail.com).

characteristics of cancer types. This research is expected to make a significant contribution to the field of breast cancer diagnosis by leveraging the advantages of the K-means method. Higher diagnostic accuracy will enable better early detection and timely treatment, which, in turn, can improve patient survival rates and reduce the adverse impact of breast cancer.

II. MATERIALS

A. *Breast Cancer*

Breast cancer is the most common type of cancer occurring in women worldwide [2]. This disease occurs when cells within breast tissue undergo uncontrolled growth. Breast cancer can be categorized into several types based on the characteristics of cancer cells, such as invasive ductal carcinoma, invasive lobular carcinoma, ductal carcinoma in situ, and so on. It is crucial to achieve an accurate diagnosis and differentiate between these types of breast cancer to plan appropriate treatment. The process of diagnosing breast cancer involves a series of steps aimed at identifying the presence of cancer, determining its type, and assessing its stage. These stages include gathering information from the patient, physical examinations, the use of imaging technology, and taking breast tissue samples for biopsy. The significance of timely and precise diagnosis is paramount in determining the course of treatment to be administered.

B. *Data Mining*

Data mining became known in the 1990s when the utilization of data became essential in various fields, ranging from academia to business and medicine [3]. Data mining is the process of exploring and uncovering knowledge to extract valuable patterns from large and complex datasets. In the context of breast cancer diagnosis, data mining is employed to identify useful patterns for distinguishing breast cancer types, detecting risk factors, or predicting treatment responses. The application of data mining techniques can provide profound insights from breast cancer patient data and enhance the accuracy of the diagnostic process.

C. *K-Means Clustering Algorithm Mining*

Clustering is a technique in data mining [4] where this algorithm operates in an unsupervised manner, meaning it doesn't require training or guidance. In data mining, there are two types of clustering methods for data grouping, namely hierarchical clustering and nonhierarchical clustering [5].

The K-means algorithm is one of the commonly used clustering algorithms in data mining [6]. The K-means algorithm falls under nonhierarchical clustering, which works by grouping data into clusters with similar characteristics based on the distance between the data points. Data grouped within a cluster exhibit high similarity and significant differences compared to other clusters [7].

The K-Means algorithm has been widely applied across various fields for data analysis and clustering purposes. Examples of its applications encompass domains such as agriculture [8], environmental science [9][10], education [11][12], and healthcare [13]-[15]. Several studies have even noted that, in some cases, the K-Means algorithm has demonstrated greater effectiveness in data clustering compared to alternatives like the K-Medoids algorithm [14].

In breast cancer diagnosis, the K-means algorithm can be applied to group patients into clusters with similar cancer type characteristics. The steps of the K-Means algorithm in data clustering are illustrated in Figure 3.

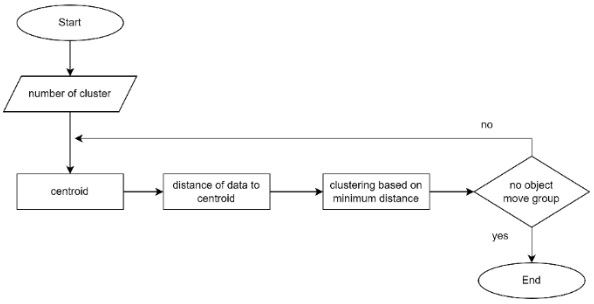

Start
number of cluster
no
centroid
distance of data to centroid
clustering based on minimum distance
no object move group
yes
End

Fig. 3. K-means Algorithm Flowchart

Based on Figure 3 above, the stages of the clustering algorithm process using the K-Means method [16] are as follows:

a. Determine the number of clusters, k.

b. Initialize k cluster centers or centroids. There are several common methods used in this process, but the primary method chosen at this stage is typically random initialization.

c. Allocate all data/objects to the nearest cluster. To perform data grouping around each centroid, the Euclidean distance (d) theory is used, formulated as follows:

$$d_{Euclidean}(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \tag{1}$$

where x_i = *centroid* and y_i corresponds to the number of n attributes (columns).

d. Calculate the new centroid using the mean formula, which involves computing the average values of the points within each cluster.

e. Repeat steps c through d until there are no further changes in cluster membership or data convergence.

D. *K-Means Algorithm Implementation*

The implementation of the K-means algorithm in diagnosing breast cancer types involves the following steps: 1) Gathering breast cancer patient data that includes relevant attributes such as age, family history, tumor size, and other examination results; 2) Data Preprocessing, this step includes data normalization or standardization and handling missing or incomplete data; 3) Application of the K-means Algorithm, applying the K-means algorithm to cluster patient data into groups with similar characteristics. 4) Clustering Results Evaluation, evaluating the clustering results to determine the accuracy of the K-means algorithm in diagnosing breast cancer types. 5) Visualization of Clustering Results, creating scatter plots or other visualizations to represent the clustering results, aiding in the interpretation and understanding of the data.

III. METHOD

In this research phase, a systematic explanation is provided regarding the sequence of processes employed in the study. The stages outlined in this sequence can be understood starting from needs analysis to research outcomes. This research involves several phases, including data needs analysis, data collection, literature review, data preprocessing, data analysis using the K-Means algorithm using the Python programming language on the Google Colab platform, evaluation, and visualization of the results. Based on this sequence, the research stages will be illustrated in Figure 4 below.

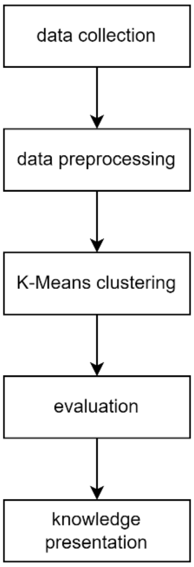


Fig. 4. Research Phases

A. Data Collection

Data collection aims to gather the necessary data to test the research hypothesis. By collecting relevant and valid data, the research can provide stronger, more accurate, and more meaningful results, which, in turn, can support better decision-making and contribute to scientific knowledge.

B. Data Preprocessing

Before proceeding with the analysis of the dataset using the K-Means algorithm, the dataset must undergo a preprocessing stage to clean the data and remove irrelevant data. Data preprocessing is a crucial step in data preparation before analysis or modeling is performed. By conducting data preprocessing, the data becomes easier to interpret.

C. K-Means Clustering

In this study, clustering analysis on breast cancer data is conducted using the K-Means algorithm. The K-Means algorithm operates by grouping data into clusters that exhibit similar characteristics based on the distance between the data points.

D. Evaluation

The evaluation stage is conducted to assess the accuracy of the clustering results generated by the K-Means algorithm. The evaluation of clustering results is performed by comparing the actual labels from the dataset with the labels generated by the K-Means algorithm.

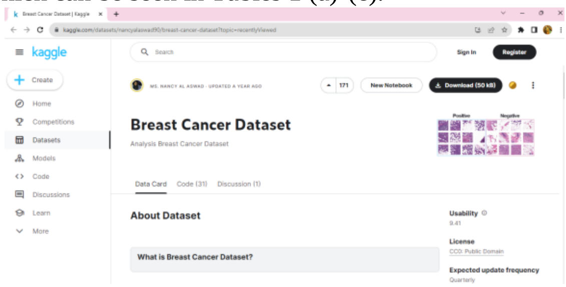
E. Knowledge Presentation

The resulting clusters are subsequently represented in the form of graphs for ease of understanding by users.

IV. RESULT AND DISCUSSION

A. Data Collection

The data utilized in this research is a breast cancer patient dataset obtained from the Kaggle website (https://www.kaggle.com/datasets/nancyalaswad90/breast-cancer-dataset?topic=recentlyViewed) as shown in Figure 5. The dataset used comprises 32 attributes and 569 data samples related to breast cancer patients, which can be seen in Tables 1 (a)-(e).



Breast Cancer Dataset

Analysis Breast Cancer Dataset

About Dataset

What is Breast Cancer Dataset?

Usability 9.41

License CC0: Public Domain

Expected update frequency Quarterly

Fig. 5. Breast cancer dataset in Kaggle

Table 1 (a). Breast cancer dataset

id	diag nosis	radius_ mean	texture_ mean	perimeter_ mean	area_ mean	smoothness_ mean	compactness_ mean
842302	M	17.99	10.38	122.8	1001	0.1184	0.2776

842517	M	20.57	17.77	132.9	1326	0.08474	0.07864
84300903	M	19.69	21.25	130	1203	0.1096	0.1599
84348301	M	11.42	20.38	77.58	386.1	0.1425	0.2839
…	…	…	…	…	…	…	…
…	…	…	…	…	…	…	…
927241	M	20.6	29.33	140.1	1265	0.1178	0.277
92751	B	7.76	24.54	47.92	181	0.05263	0.04362

Table 1 (b). Breast cancer dataset

id	concavity_mean	concave_points_mean	symmetry_mean	fractal_dimension_mean	radius_se	texture_se	perimeter_se
842302	0.3001	0.1471	0.2419	0.07871	1.095	0.9053	8.589
842517	0.0869	0.07017	0.1812	0.05667	0.5435	0.7339	3.398
84300903	0.1974	0.1279	0.2069	0.05999	0.7456	0.7869	4.585
84348301	0.2414	0.1052	0.2597	0.09744	0.4956	1.156	3.445
…	…	…	…	…	…	…	…
…	…	…	…	…	…	…	…
927241	0.3514	0.152	0.2397	0.07016	0.726	1.595	5.772
92751	0	0	0.1587	0.05884	0.3857	1.428	2.548

Table 1 (c). Breast cancer dataset

id	area_se	smoothness_se	compactness_se	concavity_se	concave_points_se	symmetry_se	fractal_dimension_se
842302	153.4	0.006399	0.04904	0.05373	0.01587	0.03003	0.006193
842517	74.08	0.005225	0.01308	0.0186	0.0134	0.01389	0.003532
84300903	94.03	0.00615	0.04006	0.03832	0.02058	0.0225	0.004571
84348301	27.23	0.00911	0.07458	0.05661	0.01867	0.05963	0.009208
…	…	…	…	…	…	…	…
…	…	…	…	…	…	…	…
927241	86.22	0.006522	0.06158	0.07117	0.01664	0.02324	0.006185
92751	19.15	0.007189	0.00466	0	0	0.02676	0.002783

Table 1 (d). Breast cancer dataset

id	radius_worst	texture_worst	perimeter_worst	area_worst	smoothness_worst	compactness_worst	concavity_worst
842302	25.38	17.33	184.6	2019	0.1622	0.6656	0.7119
842517	24.99	23.41	158.8	1956	0.1238	0.1866	0.2416
84300903	23.57	25.53	152.5	1709	0.1444	0.4245	0.4504
84348301	14.91	26.5	98.87	567.7	0.2098	0.8663	0.6869
…	…	…	…	…	…	…	…
…	…	…	…	…	…	…	…
927241	25.74	39.42	184.6	1821	0.165	0.8681	0.9387
92751	9.456	30.37	59.16	268.6	0.08996	0.06444	0

Table 1 (e). Breast cancer dataset

id	concave_points_worst	symmetry_worst	fractal_dimension_worst
842302	0.2654	0.4601	0.1189

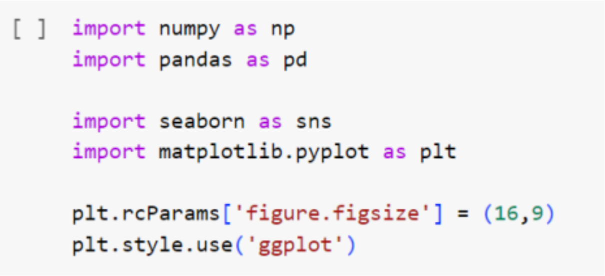
842517	0.186	0.275	0.08902
84300903	0.243	0.3613	0.08758
84348301	0.2575	0.6638	0.173
…	…	…	…
…	…	…	…
927241	0.265	0.4087	0.124
92751	0	0.2871	0.07039

B. Data Preprocessing

The data preprocessing stage encompasses a series of data processing steps aimed at enhancing data quality and preparing it for use in clustering using the K-Means algorithm. Before commencing the data preprocessing stage, the initial step is to import several libraries and read the dataset in Google Colab. Subsequently, data preprocessing steps such as dropping data, data cleaning, data transformation, feature selection, and data reduction are carried out.

Import Libraries

Import several libraries such as numpy, pandas, seaborn, and matplotlib in Google Colab.

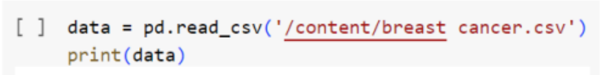


```
[ ]  import numpy as np
     import pandas as pd

     import seaborn as sns
     import matplotlib.pyplot as plt

     plt.rcParams['figure.figsize'] = (16,9)
     plt.style.use('ggplot')
```

Fig. 6. Import Libraries

Read Dataset

The next step is to read the dataset.

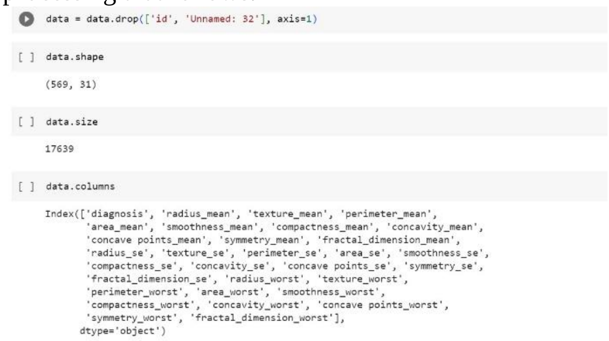


```
[ ]  data = pd.read_csv('/content/breast cancer.csv')
     print(data)
```

Fig. 7. Read Dataset

Data Dropping

Performing data dropping of unused data. In this dataset, the features "id" and "Unnamed 32" are dropped because both features have no significance in the data processing that follows.

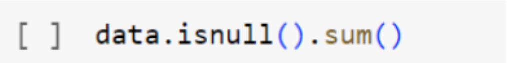


```
     data = data.drop(['id', 'Unnamed: 32'], axis=1)

[ ]  data.shape

     (569, 31)

[ ]  data.size

     17639

[ ]  data.columns

     Index(['diagnosis', 'radius_mean', 'texture_mean', 'perimeter_mean',
            'area_mean', 'smoothness_mean', 'compactness_mean', 'concavity_mean',
            'concave points_mean', 'symmetry_mean', 'fractal_dimension_mean',
            'radius_se', 'texture_se', 'perimeter_se', 'area_se', 'smoothness_se',
            'compactness_se', 'concavity_se', 'concave points_se', 'symmetry_se',
            'fractal_dimension_se', 'radius_worst', 'texture_worst',
            'perimeter_worst', 'area_worst', 'smoothness_worst',
            'compactness_worst', 'concavity_worst', 'concave points_worst',
            'symmetry_worst', 'fractal_dimension_worst'],
           dtype='object')
```

Fig. 8. Data Dropping

Data Cleaning

This stage is carried out to check for missing values in the dataset. Upon completion, it is determined that there are no missing values in the data.

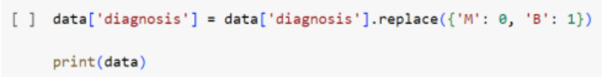


```
[ ]  data.isnull().sum()
```

Fig. 9. Data Cleaning Process

Data Transformation

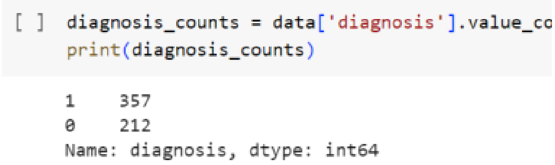
The diagnostic feature in the dataset used in this research is not yet in numeric form. Therefore, the researcher needs to transform the values in the diagnostic feature into numeric values through data transformation.



```
[ ]  data['diagnosis'] = data['diagnosis'].replace({'M': 0, 'B': 1})

     print(data)
```

Fig. 10. Data Transformation Process

From the data transformation process into numeric values mentioned above, "Malignant," represented by "M" indicating malignancy, is converted to the number "0," while "Benign," represented by "B" signifying benign characteristics, is represented by the number "1." The initial results obtained are shown in Figure 11 below.



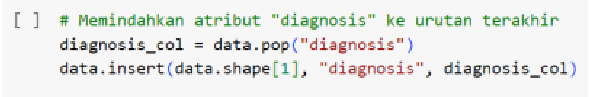
```
[ ]  diagnosis_counts = data['diagnosis'].value_co
     print(diagnosis_counts)

     1    357
     0    212
     Name: diagnosis, dtype: int64
```

Fig. 11. Diagnosis Feature in The Dataset

Figure 11 above indicates that there are 212 data points related to malignant breast cancer, while benign breast cancer comprises 357 data points.

Feature Relocation

The relocation of the diagnosis feature is carried out because this feature will later serve as the label or target for the implementation of the K-Means algorithm.

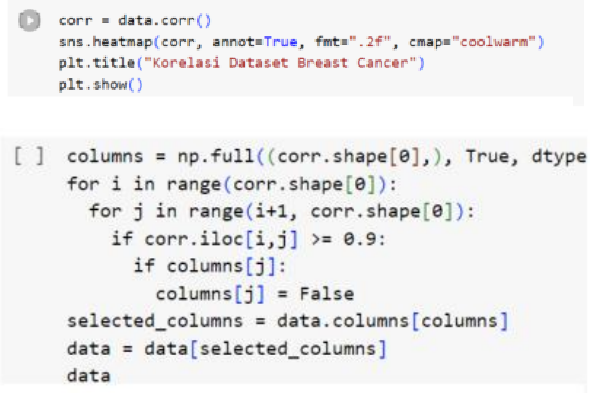


```
[ ]  # Memindahkan atribut "diagnosis" ke urutan terakhir
     diagnosis_col = data.pop("diagnosis")
     data.insert(data.shape[1], "diagnosis", diagnosis_col)
```

Fig. 12. Diagnosis Feature Relocation

Data Reduction

After conducting an analysis of data correlations, it was found that there are several features showing high correlations. To address this situation, the researcher decided to reduce the number of features with high correlations. As a result, after the reduction, there are 21 features remaining out of the initial 32 features in the entire dataset.



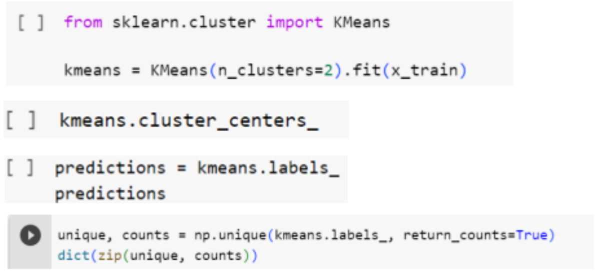
```
corr = data.corr()
sns.heatmap(corr, annot=True, fmt=".2f", cmap="coolwarm")
plt.title("Korelasi Dataset Breast Cancer")
plt.show()
```

```
[ ]  columns = np.full((corr.shape[0],), True, dtype
     for i in range(corr.shape[0]):
       for j in range(i+1, corr.shape[0]):
         if corr.iloc[i,j] >= 0.9:
           if columns[j]:
             columns[j] = False
     selected_columns = data.columns[columns]
     data = data[selected_columns]
     data
```

Fig. 13. Data Reduction

C. K-Means Clustering

At this step, the K-Means algorithm will be implemented on the dataset resulting from the data preprocessing step. The process of implementing the K-Means algorithm is illustrated in Figure 14 below.



```
[ ]  from sklearn.cluster import KMeans

     kmeans = KMeans(n_clusters=2).fit(x_train)
```

```
[ ]  kmeans.cluster_centers_
```

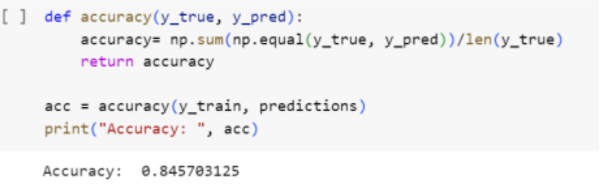
```
[ ]  predictions = kmeans.labels_
     predictions
```

```
unique, counts = np.unique(kmeans.labels_, return_counts=True)
dict(zip(unique, counts))
```

Fig. 14. K-Means Algorithm Implementation

D. Evaluation

The performance evaluation of the K-Means algorithm is carried out using the accuracy metric. This evaluation is conducted to measure the precision of the K-Means algorithm in classifying breast cancer types. The accuracy value is calculated by comparing the actual labels from the dataset with the labels resulting from the K-Means clustering. The accuracy value obtained is 0.8457. A higher accuracy indicates better performance of the algorithm in predicting breast cancer types.



```
[ ]  def accuracy(y_true, y_pred):
         accuracy= np.sum(np.equal(y_true, y_pred))/len(y_true)
         return accuracy

     acc = accuracy(y_train, predictions)
     print("Accuracy: ", acc)

     Accuracy:  0.845703125
```

Fig. 15. Evaluation of The Results of The K-Means Algorithm Analysis

E. Knowledge Presentation

Based on the earlier implementation of the K-Means algorithm, it was identified that 211 data points correspond to malignant breast cancer, while 301 data points correspond to benign breast cancer. The researcher used a scatter plot to provide a visual representation of the clustering results, as shown in the image below. When examining the graph, the yellow-colored points indicate breast cancer with benign characteristics, while the purple-colored points represent malignant breast cancer.

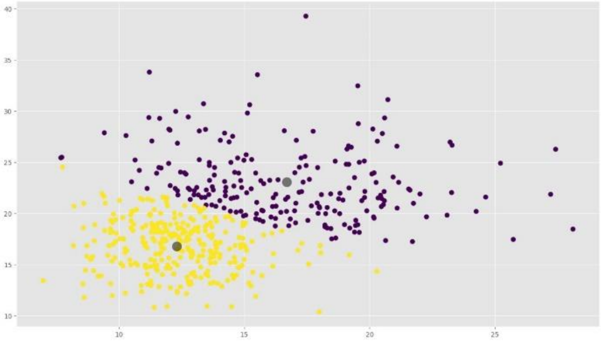


Fig. 16. Knowledge presentation

V. CONCLUSION

In this research, the implementation of the K-Means clustering algorithm was conducted for diagnosing breast cancer types. The dataset used contains relevant features related to breast cancer. Prior to the data analysis using the K-Means algorithm, data preprocessing was performed. The analysis revealed that 211 data points correspond to malignant breast cancer, while 301 data points were identified as benign breast cancer. Accuracy measurement was conducted by comparing the actual labels in the dataset with the labels generated by the K-Means algorithm during the clustering process. The result indicated an accuracy of 0.845703125, signifying a good level of accuracy.

This research represents an initial step, and in the medical context, further research and the development of more complex models are needed to address the health challenges faced by millions of women worldwide.

REFERENCES

[1] Agusta Yudi, “K-Means – Penerapan, Permasalahan dan Metode Terkait,” J. Sist. dan Inform., vol. 3, no. Februari, pp. 47–60, 2007.

[2] H. Dewi, “Analisis risiko kanker payudara berdasar riwayat pemakaian kontrasepsi hormonal dan usia,” J. Berk. Epidemiol., vol. 3, no. 1, pp. 12–23, 2015.

[3] Mehmed Kantardzic, “Data Mining: Concepts, Models, Methods, and Algorithms,” p. 360, 2003.

[4] F. Gullo, “From patterns in data to knowledge discovery: What data mining can do,” Phys. Procedia, vol. 62, pp. 18–22, 2015, doi: 10.1016/j.phpro.2015.02.005.

[5] B. Santosa, T. Conway, and T. Trafalis, “A hybrid knowledge based-clustering multi-class svm approach for genes expression analysis,” Springer Optim. Its Appl., vol. 7, pp. 231–274, 2007, doi: 10.1007/978-0-387-69319-4_15.

[6] A. K. J. A. Harding, M. Shahbaz, Srinivas, “Data Mining in Manufacturing: A Review,” vol. 128, no. 4, 2006, doi: https://doi.org/10.1115/1.2194554.

[7] P. Eko, “Data Mining - Konsep dan Aplikasi Menggunakan Matlab,” 2012.

[8] A. K. Singh, S. Mittal, P. Malhotra and Y. V. Srivastava, "Clustering Evaluation by Davies-Bouldin Index(DBI) in Cereal data using K-Means," 2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC), Erode, India, 2020, pp. 306-310, doi: 10.1109/ICCMC48092.2020.ICCMC-00057.

[9] E. Tasia and M. Afdal, “Comparison of K-Means And K-Medoid Algorithms For Clustering of Flood-Prone Areas In Rokan Hilir District,” IJIRSE: Indonesian Journal of Informatic Research and Software Engineering, vol.3. No.1, 2023.

[10] N. F. Mustamin, A. F. Zulkarnain, and M. R. B. Ramadhan, “Sistem Informasi Geografis untuk Sebaran Titik Panas (Hotspot) di Kalimantan Selatan Menggunakan Metode Clustering,” In Prosiding Seminar Nasional Lingkungan Lahan Basah, Vol. 6, No. 3, 2021.

[11] M. R. Putri, G. S. Nugraha, and R. Dwiyansaputra, "Pengelompokan Provinsi di Indonesia Berdasarkan Indikator Pendidikan Menggunakan Metode K-Means Clustering," Journal of Computer Science and Informatics Engineering (J-Cosine) 7.1, 2023, pp: 76-83.

[12] G. Ghufron, D. Kurniadi, and A. Sugiyono, "Clustering Menggunakan K-Means Untuk Menentukan Mahasiswa Berprestasi (MAWAPRES)," JSI: Jurnal Sistem Informasi (E-Journal) 15.1, 2023.

[13] J. Jadhav, P. H. Charan, and S. Mehrotra, "Brain Tumor Diagnosis Using K-Means and Morphological Operations," Proceedings of Data Analytics and Management: ICDAM 2022. Singapore: Springer Nature Singapore, 2023, 507-515.

[14] S. Sivakumar and T. Kamalakannan, "Performance-Based Analysis of K-Medoids and K-Means Algorithms for the Diagnosis and Prediction of Oral Cancer," Computational Intelligence for Clinical Diagnosis. Cham: Springer International Publishing, 2023. 215-226.

[15] A. Chusyairi, O. Nurdiawan, K. Sambath, R. N. Hayat and Y. Arie Wijaya, "Hepatitis Cluster Model With K-Means Algorithm," 2023 International Conference on Computer Science, Information Technology and Engineering (ICCoSITE), Jakarta, Indonesia, 2023, pp. 811-815, doi: 10.1109/ICCoSITE57641.2023.10127719.

[16] B. Harahap, “Penerapan Algoritma K-Means untuk Menentukan Bahan Bangunan Laris (Studi Kasus Pada UD. Toko Bangunan YD Indarung),” Reg. Dev. Ind. Heal. Sci. Technol. Art Life, pp. 394–403, 2019, [Online]. Available: https://ptki.ac.id/jurnal/index.php/readystar/article/view/82.

[17] S. Agarwal, “Data mining: Data mining concepts and techniques,” Proc. - 2013 Int. Conf. Mach. Intell. Res. Adv. ICMIRA 2013, pp. 203–207, 2014, doi: 10.1109/ICMIRA.2013.45.

[18] Y. P. Aritama, “Penerapan Metode K-Means Clustering untuk Mengelompokkan Data Kasus COVID-19 di Indonesia,” USD Repos., pp. 11–12, 2022.