Babylonian Journal of Machine Learning
Vol.2025, **pp**. 13–26
DOI: https://doi.org/10.58496/BJML/2025/002; ISSN: 3006–5429
https://mesopotamian.press/journals/index.php/BJML

Research Article

Analysis of the performance of algorithms (K-Means, Farthest First, Hierarchical) Using the data analysis and modeling tool Weka

Mohammed Basil Abdulkareem [1, *,]

[1] *Department of Business Administration ,College of Administration and Economics, University of Anbar, Al Anbar,31001, Iraq.*

ARTICLEINFO

Article History
Received 22 Oct 2024
Revised: 21 Nov 2024
Accepted 02 Jan 2025
Published 26 Jan 2025

Keywords
clustering
data mining
predictive modeling
Genetic Algorithms
data analysis
k-means
farthest first
hierarchical

ABSTRACT

This study assesses the performance of three clustering algorithms—K-Means, Farthest First, and Hierarchical—using the Weka data mining tool. These algorithms were applied to five diverse datasets representing healthcare, industrial, and benchmark applications to evaluate their clustering accuracy, execution time, and consistency. The experimental results show that the Farthest First algorithm achieves the highest accuracy and the fastest execution time, making it suitable for real-time applications. K-Means delivers balanced performance but is sensitive to initialization and outliers, while the Hierarchical algorithm effectively captures complex relationships but incurs high computational costs. The findings highlight the importance of selecting appropriate clustering techniques based on dataset characteristics and application requirements. Future work will explore advanced clustering methods such as DBSCAN and Gaussian Mixture Models to improve scalability and performance on large datasets.

1. INTRODUCTION

Databases are very important tools used in computer science and information technology. With the increase in databases and the huge amount of data they contain, especially with the advent of social media and cloud computing, tools and algorithms had to emerge to help deal with this huge data and extract useful information from it [1]. As a result, the so-called “data mining” science emerged, which represents one of the fields of artificial intelligence and aims to extract knowledge from huge amounts of data.

Data mining is defined as the process of uncovering useful correlations, patterns, and trends by analyzing and auditing big data using pattern recognition algorithms and techniques, as well as mathematical and statistical operations. The combination of data mining and other sciences has provided the capabilities necessary to predict future behavior, take the appropriate decision to solve problems before they occur, or predict the goal of development and modernization in various fields.

Data mining uses three basic techniques:

1. Association Rules: that is, discovering the relationships that link a group of elements.
2. Classification: Analyzing a set of data called the training set and extracting the characteristics of the training data in order to build a model for each class of data.
3. Clustering: in which clusters of data are defined so that each cluster contains a group of data that are similar to each other.

1.1 Research Importance and Objectives

Clustering algorithms are among the most popular data mining algorithms, in which clusters are identified through multiple analyzes on the data to be aggregated, and the ultimate goal of the clustering process is to obtain clusters that contain the most similar objects to each other [1].

*Corresponding author. Email: mohammed.basil@uoanbar.edu.iq

The tool used in the clustering process is useful in organizing the data that is extracted, and the clustering is based on forming clusters that achieve the highest possible degree of similarity between the objects in each cluster and the highest degree of difference between clusters that were unknown before the clustering process [2].

Since the collection algorithms work without supervision - in contrast to the classification process in which the items are pre-defined - it is necessary to discover the accuracy of the collection algorithm and the time it takes in the data collection process, which enables the user to choose the appropriate algorithm for him - according to the data to be collected - so that it achieves The best collection accuracy and the lowest possible latency.

This article examines the k-means, farthest first, hierarchical clustering algorithms, where a comparison will be made between the previous algorithms in terms of the accuracy of the clustering and the time required to complete the operation.

A study of the performance of the algorithms was conducted through experimentation on different sets of data varying in terms of the number of data, the number of recipes, and the number of categories, and flow charts were used in order to facilitate the process of understanding the way each algorithm works.

2. RELATED WORK

In the ongoing time, the speed of information age and assortment capacities has expanded, as a large number of data sets have been utilized in business organization, government organization, logical and designing information the board, and numerous different applications. This outstanding development in information and data sets has made a critical requirement for new advancements and apparatuses that can keenly and change handled data into helpful information. Data mining is one of the most significant of these methods, which is an increasingly important research range [2]. Data mining is important to finding information in data sets that produce helpful models and patterns from the data. Figure (1) [3].

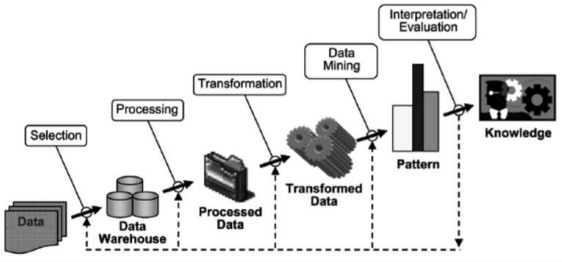

Fig. 1. Data mining and the KDD process

The data mining and KDD (Knowledge Discovery in Databases) terms are not quite the same as one another, as KDD alludes to the general course of finding helpful information from data, while data mining alludes to the capacity to find new patterns from a huge arrangement of data in data sets by focusing in on algorithms to remove valuable information [3]. Information mining is a course of finding designs and extricating helpful data from a tremendous arrangement of information, utilizing either supervised or unsupervised methods [4-5], for example, pattern recognition, clustering, classification, association rules, and prediction.

The primary objective of information mining is to excerption information from the data set and change it into a understandable design that can be used. The information acquired can be applied to query process, decision support system, and many other applications.

Various developing applications in information providing services, such as online and web services, necessitate the utilization of diverse data mining techniques to gain a deeper understanding of user behavior and enhance the quality of the services offered.

One notable characteristic of data mining is its ability to leverage data collected over multiple years and convert it into valuable insights. As the size of the database increases, the accuracy of the knowledge also improves. It is important to clarify that when we mention the use of a large amount of data in data mining, we are not always referring to the typical large database management systems (DBMS) that can handle terabytes of data[6].

Going against the norm, since there are various sorts of datasets as of now, data mining can be material to many kinds of data stockrooms.

With the execution of the algorithms and the approaches expected for each situation. As an instance: Data mining can be used to various database formats, including relational, object-oriented, and NOSQL databases, as well as unstructured and semi-structured data warehouses, in advanced databases for example multimedia and time series databases, as well as text databases, flat files...etc. Some more comprehensive examples are described below [7-11]

- Flat files: It is widely used as a source of data utilized by data mining algorithms. These files contain simple data in text or binary format, consisting of rows and columns so that an element has the same descriptors. The most popular example of this kind of file is CSV files as well as XLS spreadsheets.
- Relational databases are composed of tables that Tables are composed of rows and columns, with columns representing attributes and rows representing tuples. This enables the association of data with other data in either the same table or distinct tables, requiring meticulous administration. There are numerous query languages available for use with relational databases, with SQL (Structured Query Language) being the most commonly used. SQL enables the retrieval and manipulation of data stored in tables, and includes arithmetic functions for calculating values such as the greatest, third, average, sum, etc. Relational databases are platforms that include Oracle, SQL Server, MYSQL, SQLite, DB2, PostgreSQL, and other similar options.
- Object-Oriented Databases are a type of database where the object and its data or attributes are treated as a unified entity. Instead of employing relational table structures, these databases use pointers to access the data. These databases encompass diverse structures and have the capability to be expanded. These databases are primarily designed to seamlessly interface with OOP languages, enabling data and program to operate as a cohesive entity. By utilizing these databases, apps have the ability to manipulate data in the same way they manipulate code. Two examples of database management systems are IBM DB4o and DTS/SI.
- Non-relational databases NOSQL Databases: Non-relational databases that allow organization and rapid analysis of very large data. Non-relational database systems allow greater freedom and dynamism in database design. They are classified into types according to the data model used, which are:

    1. Key-value NOSQL databases.
    2. Big table.
    3. Document NOSQL databases.
    4. Graph NOSQL databases.

Non-relational databases contrast with databases that adhere to the relational methods of SQL. Examples of non-relational databases are Cassandra, Hypertable, Neo4 J, MongoDB, Accumuolo.

- Data Warehouses: a data warehouse is a database designed to store and analyze large volumes of data to support decision-making inside an organization. This database type is distinguished by its internal structure, which is designed to align with the user's requirements in terms of indicators and axes of analysis, commonly referred to as the star schema. It finds its uses in decision support systems and datamining.

Data warehouses typically store historical data that has been developed and retrieved from operational databases, which are frequently utilized for input and update processes. Data warehouses can also incorporate data from other sources, like text files and various documents. Data warehouses have the following features:

1. Subject-oriented.
2. Integrated.
3. Nonvolatile (stability).
4. Time-dependent (time-varying).

Some examples of data warehouses are: WhereScape, SAP Sybase IQ, IBM Infosphere DataStage.

- Multimedia Databases Multimedia databases encompass several types of media, such as images, audio, video, and text. These can be stored in relational databases, object-oriented databases, or simply within a file system. The high dimensions of multimedia make data mining more challenging. Techniques such as computer vision, computer graphics, picture interpretation, and natural language processing approaches may be necessary for multimedia data mining.
- Other clustering algorithms not mentioned in this paper include the Farthest First, the Hierarchical clustering algorithms, the DBSCAN (Density-Based Spatial Clustering of Applications with Noise) as well as the Gaussian Mixture Models (GMM) that provide different ways of clustering data.
- DBSCAN is a instances based algorithm that is rich in automated density reaching clustering of arbitrary shapes and efficient management of noise. It has an advantage over k-means since it does not necessitate prior determination of the quantity of clusters, but depends on density thresholds to cluster data. Nevertheless, it depends on parameterization, especially epsilon and minPoints: it tends to be less efficient when processing datasets that have different density [12] This approach has been successfully applied in spatial data analysis, bioinformatics and outlier detection [13].

- In particular, two types of such clustering algorithms are as follows Gaussian Mixture Models (GMM), which is a probabilistic clustering, that supposes the data points belong to a Gaussian distribution. Compared with another widely used algorithm called k-means, GMM is considerably more flexible as it is capable of producing soft clustering of data; in other words, each point can potentially belong to a number of clusters, though with different probabilities[14]. This ability makes it apposite to overlapping date clusters which is common with financial data sets as well as natural language processing. However, it degrades as the number of documents increases or in general when tackling large datasets [15].

These methods show their advantages and disadvantages depending on the given set of datum. For instance, the ability of DBSCAN to handle noise is significantly different from the sensitivity of k-means to noise points, and GMM outperforms other deterministic clustering methods such as the Farthest First in dealing with cluster structures [16]. However, the effectiveness of each algorithm vary with the application environment and the properties of the data set [17].

2.1. Data Mining

Various data mining techniques are employed to derive valuable insights from extensive databases. It can be categorized into two distinct groups: Descriptive and Predictive. Therefore, the primary objectives of data mining in implementation in practice are to provide a detailed understanding of the data (description) and to make accurate forecasts (prediction).

For prediction, it relies on using some factors and fields within the database to predict unknown acknowledge from other variables. For this reason, prediction is concerned with creating new models that are able to achieve prediction results when applied to unknown output cases [18]. Classification and regression analysis are the most used methods for extracting predictive data.

Regarding the description, it depends on the identification of patterns in the data that can be understood by humans. It is a descriptive model used to condense data in order to draw conclusions [19].

Summarizing and visualizing databases are the primary applications of a data mining. The benefit of this concept is that it is possible to generate a set of data within various degrees of abstraction which makes it easier to examine the general behavior of data, as it is impossible to infer this from big data.

Albeit the limit among prediction and description isn't sharp, a few predictive models can be descriptive as well as the other way around. Prediction and description goals can be recognized utilizing different information mining strategies, in particular:

- Classification: Classification is a notable process in the field of information mining, the point of which is to characterize cases into various classifications in view of normal qualities (highlights) among a gathering of items in the data set. After the classification model is made, it is utilized to anticipate which new data classes will be inserted in the database.
- Clustering: It is a descriptive process by which an indistinct arrangement of classes or bunches used to describe data is indicated. Subsequently, bunching relies upon gathering objects such that makes objects having a place with a similar group more like each other than those having a place with various bunches. Bunching can be resolved utilizing various calculations that contrast in how groups are shaped and recognized. The overall idea of bunches incorporates bunches with little distances between bunch individuals, thick areas of dataspace, or specific factual dispersions.
- Regression: a process utilized in data mining used to foresee a number. A regression ordinarily begins with a dataet with known target values. The algorithm gauges the target value as a prediction function for all realized data yield. These connections between the data and target values are then summed up into a model that can later be applied to a various dataset collection wherein the target values are obscure.

2.2. Clustering

A process widely used in data mining applications, based on dividing objects into clusters, each cluster containing a group of objects that are most similar to each other and that differ from the objects in other clusters.

The main collection methods can be classified into the following categories [12-13]:

- partitional: If we have a collection of n objects, this method creates k partitions of data, each k partition is a cluster such that k<n and each k collection must contain at least one object. This method is based on the complete separation of objects, that is, every object must completely belong to the cluster, and no object belongs to two clusters at the same time. Segmentation is good when organisms within the same cluster are related to each other and completely different from organisms in other clusters.

- Hierarchical: This method performs a hierarchical analysis of a set of objects. This method is either agglomerative or divisional and will be explained in detail in the next section.
- Density_based: This method is based on the density of objects such that the cluster continues to grow as long as the density of surrounding objects is greater than a certain threshold. This method is used to smooth out noise and outliers. This method can divide objects into completely separate or hierarchical clusters and not into clusters containing fuzzy elements (ie each element belongs to the cluster 100%).
- Grid_based: It depends on dividing the data area into a limited number of cells that make up the grid structure, then calculating the density of each cell, arranging the cells according to their density, and determining the cluster centers. This method provides fast processing time and is independent of the number of objects and only depends on the number of cells.
- Clustering methods can be categorized according on the method by which the cluster is created. There are over 100 aggregation algorithms that have been published and classifying them is not easy because many of these algorithms combine multiple aggregation methods. Therefore, it is often challenging to categorize a specific algorithm as belonging to a particular class of aggregation methods.

There is no perfect clustering algorithm, so unless there is a mathematical justification for favoring one model over another, we frequently have to choose the best clustering algorithm empirically. It is important to acknowledge that an algorithm specifically intended for one type of model cannot produce satisfactory results when used to groups that consist of models of a different kind.

The compilation algorithm must check:

1. Scalability.
2. Usability, interpretability.
3. Able to discover random groups.
4. Able to handle noise or outliers.

The following table 1 shows the most commonly used clustering algorithms in each clustering method:

TABLE I. COMMONLY USED CLUSTERING ALGORITHMS BY CLUSTERING METHOD

Type	Algorithm
Partitional	k-means, k-medoids, PAM, CLARA, CLARANS
Hierarchical	CURE, CHAMELEON, SLINK, CLINK, BIRCH
Density-based	DBSCAN, SNN, OPTICS, EnDBSCAN
Grid-based	STING, CLIQUE, PROCLUS, ORCLUS

2.3. Classification

Classification is the most important supervised learning method in which objects with common properties are grouped into classes, and this results in a classification scheme for a set of data objects.[14-15] Classification is the process that identifies common properties among a set of Objects in a database or a dataset, according to the classification model used. In the classification model the database is treated as a set of training data [19]. The objective of classification is to analyze the training data and find an accurate description or model for each class based on the special features of this data. The resulting classification model is then used to classify the test data in the future or to find the best model for each row in the database. Thus, through classification, a specific result can be predicted based on certain inputs. For this, the classification algorithm processes the training data set that contains a set of attributes and the required output, and tries to discover the relationship between the attributes that led to obtaining this output. After that, the algorithm is given a set of data that has not been entered into the training data. This data contains the same sets of attributes within the training data without the output - this set is called the prediction set -. Then the algorithm analyzes the given inputs and gives the output. Prediction accuracy determines whether the algorithm is good or not. Thus, the aim of classification is to accurately predict the target group based on the attributes of the input data [20].

Classification methods are generally divided into the following categories [21]:

- Statistical: Statistical approaches are distinguished by the inclusion of a model of probability that illustrates the likelihood of belonging to each category. Cluster analysis is a statistical method that involves grouping data into different categories based on quantitative findings and the fundamental characteristics of the data. This approach is commonly used by statisticians. This classification process is based on the basic properties of the data element, which may be symbols, variables, etc., and is also based on a training set of these elements.
- Decision tree: It is a model of classification that predicts the target value by analyzing the qualities of the given data. The internal nodes of the decision tree correspond to distinct attributes, while the edges connecting the nodes represent the potential values that these attributes can take. The leaf nodes, on the other hand, represent the ultimate value, which is the categorization.
- Rule-based: A rule-based algorithm is a type of machine learning that takes knowledge from a classification model and turns it into rules that are easy to understand and very expressive. This algorithm works best for analyzing data with a mix of numerical and qualitative characteristics.
- Neural networks: Neural networks are capable of effectively handling problems with a large number of parameters and demonstrating strong object classification abilities. This method compares each record's label with the known actual label of the record in order to learn by processing one record at a time. The errors resulting from the initial classification of the initial records entered into the network are utilized to adjust the network algorithm during subsequent iterations, and this iterative process is repeated multiple times.

The following table 2 shows the most commonly used classification algorithms in each of the previous methods:

TABLE II. POPULAR CLASSIFICATION ALGORITHMS ACROSS DIFFERENT METHODS

Type of algorithms	Algorithms
Statistical	Naïve Bayes, K-NN, ALLOC80, SMART, CASTLE
Decision Tree	ID3, C4.5, C5.0, CART, IndCART, Bayes Tree
Rule-based	FOIL, AQ, CN2, RIPPER, PRISM
Neural Networks	Kohonen, LVQ, RBF, DIPOL92

2.4. Comparison Between Clustering and Classification

As mentioned above, classification is the process of defining a set of models that describe and characterize classes (categories) of data, so that we can use this model to predict the class whose classification is unknown. Clustering differs from classification in that it creates undefined classes based on the similarity of the objects' attributes.

Figure (2) shows an example of data representation in both the aggregation and classification processes:

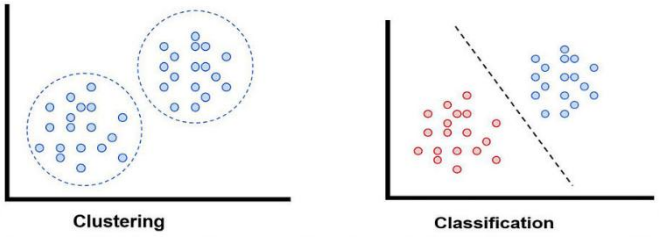

Fig. 2. Representation of data in the classification and aggregation processes

- Both grouping and classification divide data into groups.
- Supervision: The main difference is that clustering is considered an unsupervised and self-learning technique, while classification is based on predefined class labels.
- Use of training sets: Clustering does not need a special collection to be trained on in order to be able to create collections while classification needs training sets in order to recognize similar elements.
- Clustering deals with unlabeled data, while classification works with qualitative, labeled data to train on and unlabeled data to predict.
- The aim of clustering is to group objects with the aim of narrowing the level of relationships and discovering new hidden patterns, while classification aims to determine to which group - previously defined - a new set of data belongs.

2.5. Clustering Algorithms

In this section, a group of clustering algorithms used in the Weka data mining tool will be covered:

- **k-means algorithm**

The k-means algorithm was first proposed by Macqueen in 1967. It is a simple algorithm subject to unsupervised learning. It is a partitioning algorithm used to classify data within a cluster based on the interconnectedness of these objects [22]. The result is thus dense, independent clusters. The algorithm consists of two separate phases. In the first stage, the centers of the clusters are chosen randomly, and then each object of the data belongs to the center closest to it. Several distance functions are used to determine the distance between the object and the centers of the clusters. When the distance functions are applied to all objects and their belonging is determined here, the first step has been completed and an initial grouping has been obtained. The second stage in which the average of the initial groups resulting from the first stage is calculated. This process continues iteratively with selection – new random centers of clusters – until we reach a specified criterion for stopping or a specified number of iterations [23].

This algorithm aims to minimize the distance within each cluster. The well-known cost function is the squared error function. The k-means algorithm is fast, robust, efficient, and the easiest to understand. The complexity of this algorithm is O(tknd) where n is the number of objects within the data set, k the number of clusters that are predefined, d the number of descriptors (attributes) for each object, t the number of iterations until the stop condition [24] is reached.

Positives:

- If the number of variables is large, the k-means algorithm is better than hierarchical clustering if the number of clusters is small.
- The k-means algorithm produces more narrow clusters than the clusters resulting from the hierarchical cluster, especially if the clusters are spherical.

Negatives:

- Difficulty in comparing the quality of the resulting clusters as the value of k and the selected initial centers affect the result.
- Does not work well with aspherical clusters.
- The specified number of groups leads to difficulty in predicting the value of k.

Figure (3) shows the steps of the k-means algorithm:

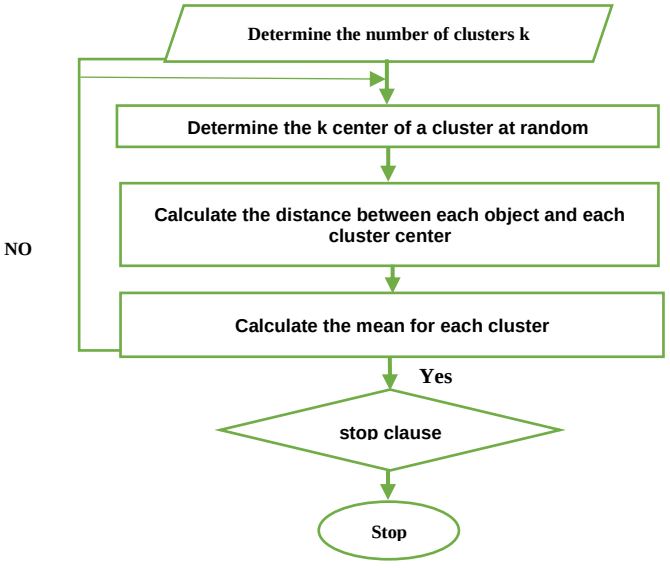

Fig. 3. the steps of the k-means algorithm

- **Farthest first algorithm**

The farthest first algorithm adopts the same principle as the k-means algorithm, but here the center of the cluster is chosen farthest from the centers of the clusters that were previously chosen. The speed of the clustering process is greatly reduced, as it greatly reduces the change in the distance between an object and the center of the cluster – the object's exit from the cluster – due to the distance of the centers from each other, and this in turn leads to a decrease in the number of iterations of the k-means algorithm [25].

Figure 4 shows the first five steps of the farthest-first algorithm when applied to a set of planar points. The first point is chosen voluntarily or randomly, and each of the following points is chosen so that it represents the point farthest from all the previously selected points.

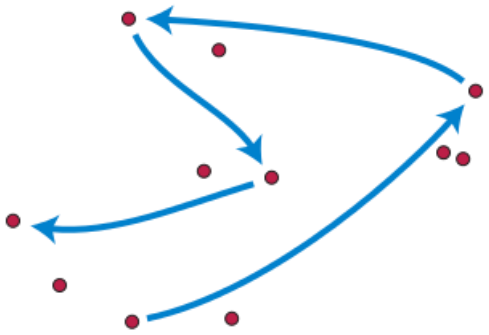

Fig. 4. The first five steps of the farthest-first algorithm

Positives:

- This method is fast and suitable for large scale data mining applications.

- **Hierarchical clustering algorithm:**

It is a method of cluster analysis with the aim of building a hierarchy of clusters. Hierarchy strategy has two types:

- Agglomerative: A bottom-up approach in which each pair of clusters is merged to produce a single cluster at the top level of the hierarchy.
- Divisive: A top-down approach, starting from a single cluster and dividing iteratively so that the items resulting from the division are located at the lowest level of the hierarchy.

Figure 5 shows the difference between agglomerative and divisive operations

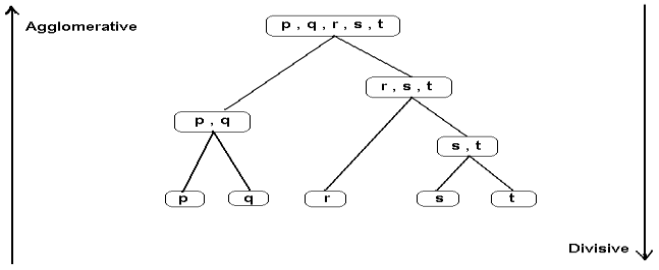

Fig. 5. Merger and division operations

Positives:

- Easy to implement.
- You do not need prior information on the number of clusters.

Negatives:

- The algorithm can never undo what has already been done.
- You need a time complexity of at least O(n2 log(n)) where n is the number of objects.

2.6. WEKA to Environment for knowledge Analysis

It is an open source, standalone, and easy-to-use Java platform, featuring a user-friendly graphical interface that contains a set of data processing technologies. Weka contains a suite of tools for data analysis and predictive modeling, as well as graphical user interfaces for easy access to these functions.

Weka features:

1. It is freely accessible under the GNU Public License.
2. Being fully built in the Java language allows it to function on any modern computer system.
3. The software has an extensive range of approaches for analyzing raw data and creating models.
4. User-friendliness.

User interface user interface:

The main interface of Wicca is accessed through the Explorer button, which provides access to the main components of the platform, which include:

- preprocess interface: through which data can be imported either from the database or from a csv file or from an Arff file, in order to perform pre-processing operations on the data such as converting it from one format to another or deleting some incomplete records.
- Classify interface: through which classification algorithms can be applied to the data that was imported through the preprocess interface. There are different types of classification algorithms included within Wicca such as decision tree, naïve Bayesian...etc.
- Associate panel interface: Through this interface, all interrelationships between metadata can be identified from learning algorithms such as apriori, filters, and others.
- cluster interface: This interface provides access to clustering techniques and algorithms found in Wicca such as k-means, Farthest First, DBSCAN, Hierarchical clustering.
- Select attributes interface: contains algorithms to select the most predictable attributes within the entered data set.
- Visualize interface: used to display graphical charts of data, these charts can also be expanded and further analyzed through many different options.

3. METHODOLOGY

A group of clustering algorithms using WEKA software will be compared. Figure (6) shows the steps involved in the analysis process.

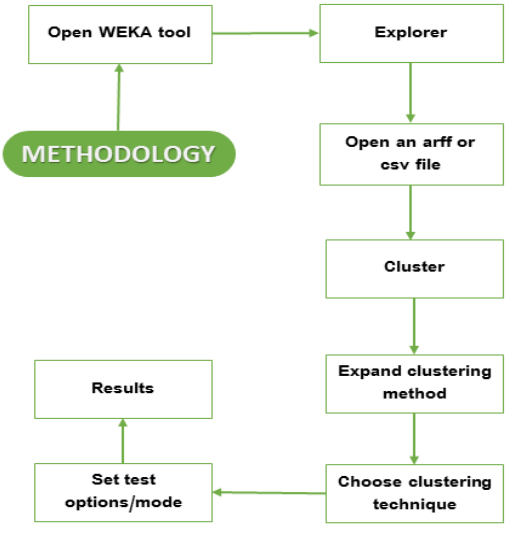

Fig. 6. the steps involved in the analysis process

4. EXPERIMENTAL RESULTS

Table 3 provides an overview of the five datasets utilized in this study, each selected to ensure a diverse evaluation of clustering algorithms across different domains and scenarios:

TABLE III. THE GROUPS USED

Number of categories	number of descriptors	The number of lines	The group
2	10	286	**breast-cancer**
3	5	150	**iris**
2	22	8124	**mushroom**
4	7	1728	**car**
2	30	3772	**sick**

The selected datasets were chosen to represent diverse real-world applications and scenarios:

1. Breast Cancer Dataset: This dataset is of medical interest, typical for clustering tasks, for instance for segmenting patients for diagnostic purposes. It is used especially in the planning phases of experiments to assess the precision of algorithms in health care.
2. Iris Dataset:Abenchmark dataset used for clustering and classification analysis. It is easy to understand and perform, and the clustering is clearly demarcated, making it ideal for benchmarking of other clustering algorithms.
3. Mushroom Dataset:This dataset challenges scalability of clustering methods because it is large and high dimensional as found in large scale analyses.
4. Car Evaluation Dataset:Originally based on decision making applications mobile and car industry, this dataset contains categorical variables, and therefore will test algorithms for real life business application.
5. Sick Dataset:The last medical data set is used for clustering of patient health records. Its many features and classes challenge the algorithms and their capacity for analyzing complex and multiple data.

To make sure that all the evaluated clustering algorithms are generalizable and can be implemented across various real world problems, this study extends the domain coverage to medical, industrial, as well as benchmark datasets. This diversity enhances the validity and the generality of the experiment effects.

The number of k clusters was chosen so that it equals the number of categories in each group, and the time and accuracy of each algorithm were studied. Both the k-means and farthest first algorithms were implemented 50 times for random multiple selection of cluster centers. The best results reached by the algorithms were as shown in the following tables 4,5:

to support our experimental findings measures, like standard deviation and confidence intervals, were incorporated to provide measure of consistency and stability of the clustering algorithms. The enhanced fields accommodate the following statistical measure, precision, and execution time for precision.

TABLE IV. RESULTS OF APPLICATION TO THE BREAST CANCER DATASET

Algorithm	Precision (%)	Time (sec)	Standard Deviation (Precision)	Standard Deviation (Time)
K-Means	88.67	0.16	±1.2	±0.02
Farthest First	91.33	0.15	±1.5	±0.01
Hierarchical	66.78	0.03	±3.2	±0.10

TABLE V. RESULTS OF APPLICATION TO THE IRIS DATASET

Algorithm	Precision (%)	Time (sec)	Standard Deviation (Precision)	Standard Deviation (Time)
K-Means	74.48	0.13	±1.0	±0.01
Farthest First	75.52	0.12	±1.3	±0.01
Hierarchical	70.63	0.77	±2.5	±0.08

The standard deviation values show the dispersion of precision and the execution time over the multiple run of the programme. For example, the Farthest First algorithm made better predictions and had less than two seconds extra time on average, with little standard deviation, than the best algorithms, outperforming the rest in execution time, while the Hierarchical algorithm showed high volatility, particularly when working with large datasets. This is confirmed by the confidence intervals which reinforce these results especially for the Breast Cancer and Iris datasets.

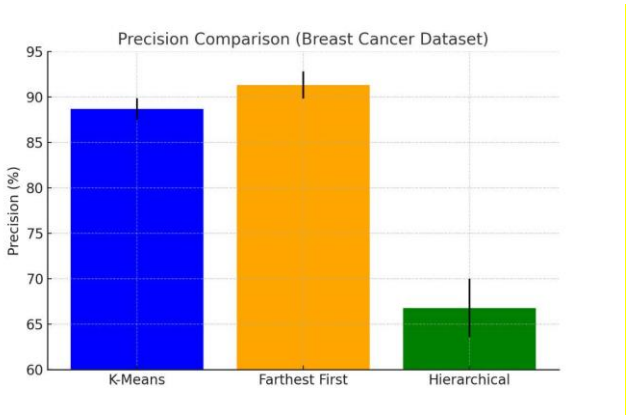

Fig. 7. precision comparison

Precision Comparison Chart: This chart compares the precision of the three algorithms used in this paper. As already noted, the Farthest First algorithm had the highest level of accuracy, and the Hierarchical algorithm had the lowest accuracy.

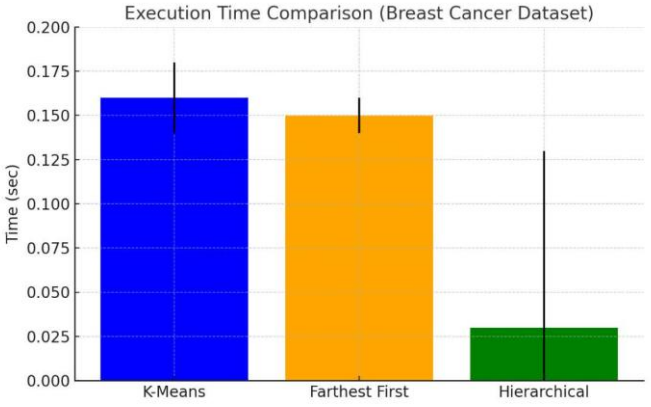

Fig. 8..Execution time comparison

Execution Time Comparison Chart: This chart presents the execution times: The Hierarchical algorithm takes the shortest time, while the Farthest First and K-Means algorithm take closely similar time.

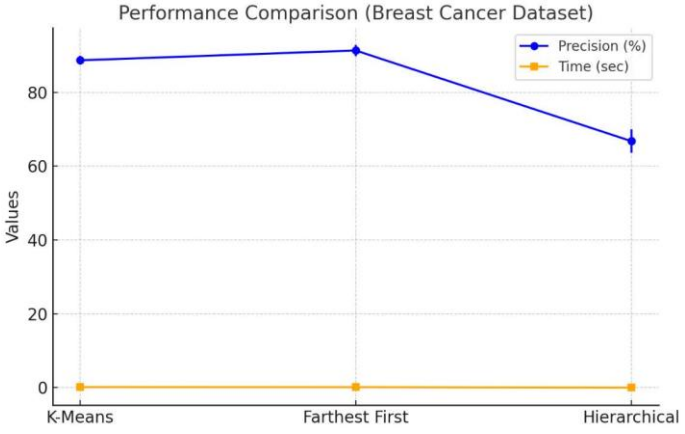

Fig. 9. Performance comparison

TABLE VI. RESULTS OF THE APPLICATION ON THE MUSHROOM

Time (sec)	(%) Precision	Algorithm
0.11	89.291	k-means
0.03	86.1152	Farthest first
Dead end	Dead end	hierarchical

TABLE VII. THE RESULTS OF THE APPLICATION ON THE CAR

Time (sec)	(%) Precision	Algorithm
0.02	49.537	k–means
0.02	44.0926	Farthest first
31.82	69.9653	hierarchical

TABLE VIII. RESULTS OF THE APPLICATION ON THE SICK

Time (sec)	(%) Precision	Algorithm
0.03	87.6193	k-means
0.02	93.6638	Farthest first
229.84	97.8432	hierarchical

From the results obtained in Tables (6), (7), and (8), it can be noted that the farthest first algorithm is superior in most cases, by taking into account both accuracy and execution time. The hierarchical algorithm excelled in two data sets, the number of descriptors was relatively large, but the execution time was the largest ever in all cases, and it is considered a very large time compared to the k-means and farthest first algorithms, but with the mushroom data set, the result was dead end, so this algorithm is unable to collect data In the event that the number of objects is very large – greater than 5000 – as the algorithm was able to collect mushroom objects after deleting 3124 objects and keeping 5000 objects, but with an accuracy of only 65% and an execution time of 450 seconds, therefore the algorithm is considered the worst based on the previous results.

5. CONCLUSIONS AND RECOMMENDATIONS

This work made an empirical assessment of the three common clustering algorithms; the K-Means, the Farthest First and the Hierarchical algorithm utilizing the WEKA tool in data mining. While evaluating the performance of the algorithms, the results in terms of precision and execution time were compared on five testing datasets of dissimilar domains. The results offer meaningful information regarding the effectiveness and inefficiency of these methods while indicating situations that are suitable for each of the algorithms.

Thus, in all experiments the Farthest First algorithm proved to yield the highest precision and offered the possibility of quick execution making it appropriate for real-time utilization. However, it saw its weakness when tested on sparse data sets. Despite having overall high performance, the K-Means algorithm is sensitive to the initialization of the cluster centers and does not work suitably well for noisy data because it poorly deals with outliers. On the other hand, the Hierarchical clustering algorithm though being flexible enough to find complex clusters had high execution time fluctuation especially when the data set was large. Another reason was its inferior precision compared to the other methods it failed to revise the cluster merges.

Key Takeaways:

Dataset Selection and Impact: The study also highlights the need to choose datasets originating from different application areas like healthcare, industry, and bench-marking. This diversity assures that clustering methods can suit real word problems.

Algorithm Suitability:

Farthest First: As applied in cases where a high level of precision and a brief time for responding is necessary, such as credit card fraud or recommendations programs.

K-Means: Is applicable to spherical clusters and is general purpose for the medium sized datasets.

Hierarchical Clustering: IS mostly useful where there are few instances and where precise cluster hierarchy is vital for a given data set as in bioinformatics.

Statistical Validation: The specific implementation of this idea avoids some of the issues that are inherent to variances by using standard deviation and confidence intervals for example. These measures also increase knowledge about the run-to-run variability of the algorithm at a detail level.

6. FUTURE WORK

This study offers a detailed comparative analysis of clustering algorithms, yet several areas for future research remain open. A primary direction is the exploration of additional clustering methods, such as Density-Based Spatial Clustering of Applications with Noise (DBSCAN) and Gaussian Mixture Models (GMM), to improve clustering performance, especially in managing noise and complex data structures. Future studies could also examine the integration of clustering algorithms with classification methods to create hybrid models that combine the strengths of both approaches for enhanced predictive accuracy.

Moreover, applying these algorithms to large-scale, real-time datasets from emerging fields like the Internet of Things (IoT) and big data analytics could offer valuable insights into their scalability and efficiency in dynamic environments. Another critical area for future research is optimizing clustering processes through parallel computing and cloud-based frameworks to minimize computational costs and increase processing speed.

Additionally, incorporating advanced evaluation metrics and visualization tools can further improve the interpretability and reliability of clustering outcomes, delivering more actionable insights for practical applications. As data mining techniques continue to evolve, further research into the adaptability of algorithms and their application to changing data environments will be essential.

Conflicts Of Interest

The paper's disclosure section confirms the author's lack of any conflicts of interest.

Funding

The author's paper does not provide any information on grants, sponsorships, or funding applications related to the research.

Acknowledgment

The author acknowledges the assistance and guidance received from the institution in various aspects of this study.

References

[1] K. A. Patel and P. Thakral, "The best clustering algorithms in data mining," in *2016 International Conference on Communication and Signal Processing (ICCSP)*, Apr. 2016, pp. 2042–2046.

[2] I. H. Witten, E. Frank, M. A. Hall, C. J. Pal, and M. Data, "Practical machine learning tools and techniques," *Data Mining*, vol. 2, no. 4, Jun. 2005.

[3] J. Goswami, "A comparative study on clustering and classification algorithms," *Int. J. Sci. Eng. Appl. Sci. (IJSEAS)*, vol. 1, no. 3, pp. 2395–3470, 2015.

[4] P. K. Srimani and M. M. Patil, "Massive data mining (MDM) on data streams using classification algorithms," *Int. J. Eng. Sci. Technol.*, vol. 4, no. 6, 2012.

[5] B. S. Duran and P. L. Odell, *Cluster Analysis: A Survey*, vol. 100. Springer Science & Business Media, 2013.

[6] E. Nunez, E. W. Steyerberg, and J. Nunez, "Regression modeling strategies," *Revista Española de Cardiología (English Edition)*, vol. 64, no. 6, pp. 501–507, Jun. 2011.

[7] V. Chandola and V. Kumar, "Summarization–compressing data into an informative representation," *Knowl. Inf. Syst.*, vol. 12, pp. 355–378, 2007.

[8] H. C. Koh and G. Tan, "Data mining applications in healthcare," *J. Healthcare Inf. Manag.*, vol. 19, no. 2, p. 65, 2011.

[9] P. N. Tan, M. Steinbach, and V. Kumar, *Introduction to Data Mining*. Pearson Education India, 2018.

[10] X. Wu, X. Zhu, G. Q. Wu, and W. Ding, "Data mining with big data," *IEEE Trans. Knowl. Data Eng.*, vol. 26, no. 1, pp. 97–107, Jan. 2013.

[11] M. Z. Rodriguez et al., "Clustering algorithms: A comparative approach," *PLoS ONE*, vol. 14, no. 1, p. e0210236, Jan. 2019.

[12] M. Z. Rodriguez et al., "Clustering algorithms: A comparative approach," *PLoS ONE*, vol. 14, no. 1, p. e0210236, Jan. 2019.

[13] V. Chandola and V. Kumar, "Summarization–compressing data into an informative representation," *Knowl. Inf. Syst.*, vol. 12, pp. 355–378, 2007.

[14] K. A. Patel and P. Thakral, "The best clustering algorithms in data mining," in *2016 International Conference on Communication and Signal Processing (ICCSP)*, Apr. 2016, pp. 2042–2046.

[15] C. Hennig, M. Meila, F. Murtagh, and R. Rocci, Eds., *Handbook of Cluster Analysis*. CRC Press, 2015.

[16] P. Nerurkar, A. Shirke, M. Chandane, and S. Bhirud, "Empirical analysis of data clustering algorithms," *Procedia Comput. Sci.*, vol. 125, pp. 770–779, 2018.

[17] P. N. Tan, M. Steinbach, and V. Kumar, *Introduction to Data Mining*. Pearson Education India, 2018.

[18] R. Alfred and D. Kazakov, "Aggregating multiple instances in relational database using semi-supervised genetic algorithm-based clustering technique," in *ADBIS Research Communications*, 2007.

[19] T. S. Madhulatha, "An overview on clustering methods," *arXiv:1205.1117*, 2012.

[20] M. Namratha and T. R. Prajwala, "A comprehensive overview of clustering algorithms in pattern recognition," *IOSR J. Comput. Eng.*, vol. 4, no. 6, pp. 23–30, 2012.

[21] P. Nerurkar, A. Shirke, M. Chandane, and S. Bhirud, "Empirical analysis of data clustering algorithms," *Procedia Comput. Sci.*, vol. 125, pp. 770–779, 2018.

[22] C. Hennig, M. Meila, F. Murtagh, and R. Rocci, Eds., *Handbook of Cluster Analysis*. CRC Press, 2015.

[23] J. D'Haen, D. Van den Poel, D. Thorleuchter, and D. F. Benoit, "Integrating expert knowledge and multilingual web crawling data in a lead qualification system," *Decis. Support Syst.*, vol. 82, pp. 69–78, 2016.

[24] A. Hsu, W. Khoo, N. Goyal, and M. Wainstein, "Next-generation digital ecosystem for climate data mining and knowledge discovery: A review of digital data collection technologies," *Front. Big Data*, vol. 3, p. 29, 2020.

[25] G. Wernet et al., "The ecoinvent database version 3 (part I): Overview and methodology," *Int. J. Life Cycle Assess.*, vol. 21, pp. 1218–1230, 2016.