Machines and Algorithms
http://www.knovell.org/mna

Research Article

Clustering Algorithms: An Investigation of K-mean and DBSCAN on Different Datasets

Arooj Zahra[1] and Nabeel Asghar[1]

[1]Department of Computer Science, Bahuddin Zakariya University, Multan, 60000, Pakistan

[*]Corresponding Author: Arooj Zahra. Email: aroojzahra97@gmail.com

Received: 06 June 2023; Revised: 21 June 2023; Accepted: 08 August 2023; Published: 16 August 2023

AID: 002-02-000025

Abstract: The branch of artificial intelligence that studies computer techniques that allow systems to learn autonomously and deliver outcomes based on past experience without being programmed. Supervised and unsupervised machine learning are major categories. Our research focuses on unsupervised learning with unlabeled data. Clustering is an unsupervised learning method that groups unlabeled data items by similarity. Several studies have compared clustering algorithms based on complexity, performance, and the impact of cluster number on performance. To our knowledge, no study has evaluated clustering methods on small and large datasets. A detailed study was conducted to evaluate DB-SCAN and K-Means algorithms on small and large datasets. We have collected 17 open access, publicly available machine learning heterogeneous datasets from online machine learning dataset sources such as the UCI repository, Keel, and Kaggle. The datasets are divided into small and large categories based on the number of instances in each dataset. Different preprocessing techniques are used to improve the quality of datasets. The class field is removed from the preprocessed datasets and then put into the two clustering techniques outlined above. The clustered data is analyzed using three classifiers (K-Nearest Neighbor, Support Vector Machine, and Naïve Bayes) to evaluate the clustering algorithm's performance. The accuracy of the KNN, SVM, and NB classifiers was calculated as part of the final algorithm performance study. The final analysis of tests found that the K-Means algorithm performs better on large datasets, whereas the DB-SCAN clustering technique is more efficient on small datasets.

Keywords: Unsupervised machine learning; Clustering algorithms; DB-SCAN; K-Means; Classifiers;

1. Introduction

In our study, we have compared the performance of two different clustering algorithms, i.e., K-Means and DB-SCAN, over large and small datasets. Data mining refers to the process of extracting useful information from massive datasets by examining records stored in various types of repositories, databases, or data warehouses. Information management, query processing, and decision-making are all possible with this mined data. We employ both supervised and unsupervised machine learning approaches for this goal, with clustering being a common unsupervised method. Clustering organizes datasets into groups of items that share high levels of similarity. The five most commonly used clustering algorithms are as follows: Some examples of these strategies include 1) hierarchical, 2) partitioning, 3) density-based, 4) model-based,

and 5) grid-based. These groups contain algorithms that have undergone testing and evaluation based on criteria such as complexity, speed, scalability, and efficiency. Multiple analyses have contrasted various clustering methods. To illustrate the point, [1] compared the performance of EM and K-means clustering algorithms using a single dataset. In terms of accuracy, performance, and quality, another study [2] compared hierarchical clustering algorithms with soft clustering techniques like K-Means and EM. Additional research [2] compared hierarchical-based methods to partition-based ones based on dataset size, kind, and cluster count. In [3], researchers demonstrated that the CirCle method outperformed other clustering methods based on models or partitions. A comprehensive study [4] compared techniques based on density, hierarchies, grids, as well as different cluster sizes and nested cluster structures. Other research [5] has tested several clustering techniques using various criteria. There is a dearth of research on the optimal clustering methods for big and small datasets, despite the abundance of studies devoted to determining the optimal clustering algorithm in terms of performance efficiency, training time, and complexity. The goal of this research is to determine which clustering algorithms work best with large-scale datasets and which ones work best with small datasets.

In state-of-the-art comparative studies of clustering algorithms' performance, researchers have only focused on the impact of number of clusters on efficiency or compared hierarchal and partition-based clustering techniques. To our knowledge, no study has examined how dataset size affects clustering performance. Any machine learning algorithm's performance depends on dataset size. Small datasets may cause the ML model to overfit or not learn patterns. Thus, dataset size matters when comparing attributes of dataset instances or objects. However, large-scale high-quality datasets are rare, therefore researchers usually have to work with smaller datasets. Researchers benefit from finding a good ML algorithm in such instances. Our study examined two data clustering methods in light of this. This study will compare the performance of the DBSCAN and K-Means clustering methods on datasets of varied sizes. The initial stage will be to locate suitable datasets for the investigation. The datasets will then be separated into two categories: small-scale and large-scale. Finally, the core analysis will compare the performance of DBSCAN and K-Means on the categorized datasets.

In this section, we have already discussed the rationale for our research, the precise problem statement, and the primary aims and objectives. We have conducted a comprehensive literature review in the subsequent section. The methodology of our investigation is thoroughly examined in section 3. The implementation of our proposed adaptive model and the results analysis are discussed in Section 4. In Section 5, the conclusion and the intended future work of our study are discussed.

2. Literature Survey

In this article [8], a hybrid krill herd has been proposed, which includes a harmony search algorithm with a new probability value to regulate the harmony search operator during the exploration search. They have implemented the following metrics to assess their proposed methodology: precision, recall, accuracy, ASDC, and F-measure. Additionally, they have implemented the Error rate and objective function for data clustering. Their results have demonstrated that their algorithm is highly effective. In this paper [9], their objective was to suggest an optimal solution for network communications in a disaster situation that does not involve any disconnectivity, including functional and non-functional areas. They have successfully accomplished their objective; however, their proposed approach is insufficient to function independently. In order to improve the post-disaster situation and establish a stronger connection, they must incorporate additional restoration and protection techniques. This article conducted a survey [10] that employed four clustering methods: LVQ, SOM, COBWEB, and k-means. K-means clustering was the most effective algorithm in their experimentation, as it required less computational effort. The WEKA tool was employed for the experimentation, so it is uncertain whether k-means will perform as well on other tools.

A comparison of supervised and unsupervised machine learning algorithms was conducted [6] using a lung cancer dataset. The combination of Apriori and k-means algorithms results in a quicker performance, as demonstrated by their experimental results and comparison. They have introduced a novel multi-hop clustering algorithm in this article [8]. Cluster head selection mechanism for optimal cluster head selection

is a cluster model that may be presented in their scheme, which is based on priority neighbor-based technique. From their experiments, they have proposed an algorithm that enhances the reliability and stability of VANET. Jin Wang [11] has introduced the particle swarm optimization, a clustering algorithm that includes mobile sink support for WSNs. Their proposed scheme outperformed the three conventional routing algorithms for WSNs that were previously in use, as indicated by the results of their evaluation. In order to enhance the network's efficacy, they implemented the PSO algorithm and the virtual clustering technique.

The Bird Flock Gravitation Searching Algorithm (BFGSA) is a clustering algorithm that has been proposed in this article [12]. The BFGSA is implemented to monitor the progression of candidate clustering centroids in order to identify the robust data cluster in a multi-dimensional Euclidean space. The rate of error and the sum of intra-clustering distance are used to evaluate the performance using thirteen distinct datasets. The performances of k-means, PSO, and GSA are compared. The experimental results indicate that it is straightforward to implement data clusters. This article [13] proposes a routine base protocol known as LEACH-SF, which is an energy-efficient clustering. The fuzzy c-means clustering algorithm was employed to accomplish the balanced clusters. The simulation results indicate that they are capable of constructing efficient balanced clusters and maximizing the network lifetime. LEACH-SF outperformed the classic and fuzzy-clustering algorithms, which are employed to optimize the number of data packets received and minimize the intra-cluster distance. The high-dimension data was proposed to be selected by a sub-set feature clustering-base feature in this paper [7]. Clusters are features in the proposed algorithm, and each cluster is considered as a single feature. A comparison has been conducted with the renowned feature selection algorithms, including FCBF, Relief, CFS, INTERACT Consisting, and FOCUS. This algorithm has achieved the highest proportion of pre-selected features, more precise results, and a shorter runtime for RIPPER, Naïve Bayes, and C4.5, and the second-best efficacy for IB1.

This paper [12] proposes a hybrid PSO algorithm with GOs (H-FSPSOTC) for the selection of text features. The text features selected in the selection method are subsequently utilized by k-means to generate more precise clusters. The H-FSPOSTC results were the most favorable among the other comparatives. This algorithm will assist in the development of enhanced text features and text clustering techniques, such as k-means. Another algorithm has been proposed for ad-hoc networks that is based on grey wolf clustering in another study [14]. Optimize the number of clusters that have been derived from the convergence of the value of α wolf in order to achieve superior results. The simulation is conducted using MATLAB, and the results are compared to those of PSO, CLPSO, and MOPSO. The performance of the proposed method was superior to that of CLIPSO and MOPSO in terms of the number of clusters with varying transmissions, the number of nodes, and the size of the grid. Additionally, it reduces the necessary number of clusters to reduce the cost of routing for communication. This article [15] introduces a paradigm of multi-hop sensor networks known as Type 2-Fuzzy logic. The results of their simulation indicate that T2FL is more scalable, reliable, and superior to the T1FL, LEACH single hop, and LEACH multi-hop protocols.

RNN-DBSCAN, a cutting-edge clustering technique based on density, has been proposed by merging the idea of observation reachability definitions with the observation of reverse nearest neighbor. [16]. According to the evaluation results, the RNN-DBSCAN outperforms the DBSCAN; yet, it is a sophisticated method that can be improved.

Another work described an unsupervised machine learning algorithm whose main goal is to learn a finite mixture model using multivariate data. The term "unsupervised" refers to two properties of the algorithm: 1) it has the ability to choose the number of components, and 2) it must be carefully started, similar to the classic expectation-maximization (EM) technique.[17] Another disadvantage of EM mixture fitting that has been addressed by the presented method is the possibility of the algorithm converging to a unique estimate at the parameter space boundary. The suggested model is unique in that it does not require any model selection criteria. The provided approach integrates both the model selection and estimation processes. The proposed technique can be applied to any parametric mixture model for which an EM algorithm has been created. This fact was demonstrated in this work through experiments, specifically the use of Gaussian mixtures. All of the experiments in this study are designed to test the efficacy of the provided approach.

A comparison was conducted between supervised and unsupervised machine learning techniques [18] using the lung cancer dataset. The experimental results and comparison indicate that the combination of apriority and k-means algorithms results in a more efficient performance. This article conducted a survey [19] that employed four clustering methods: LVQ, SOM, COBWEB, and k-means. K-means clustering was the most effective algorithm in their experimentation, as it required less computational effort. The WEKA tool was employed for the experimentation, so it is uncertain whether k-means will perform as well on other tools. In another research study, the author has addressed the scenarios in which it is highly important to minimize the error of generalization, such as in the case of achieving excellent classification results, or in the case of the occurrence of little bit model over-fitting, which results in a critical penalty in the testing data results. In order to address these circumstances, the application of a classifier with minimal dimensions in Vapnik-Chervonenkis (VC) could result in positive distinctions. This is due to the presence of two benefits: 1) the classifier's learning power is effective even on a small number of instances, and 2) the classifier has the ability to maintain the distance between the training and testing errors. The author of this study has experimentally demonstrated that the application of a classifier known as the majority vote point (MVP) on the basis of a limited number of dimensions in VS can accomplish a lower error of generalization than any other linear classifier. A maximum bound has been established for the dimensions of the VC in the MVP classifier. In the subsequent phases, empirical analysis is employed to predict the precise dimensions of VC. The proposed method is subsequently revalidated by its application to the diagnosis of prostate cancer and the detection of machine defects, which demonstrates that the MVP classifier can achieve a significantly lower generalization error [20].

In an additional investigation [21], the computation of the distance measure for each object in the dataset dominates the computational complexity of DBSCAN. The efficiency of DBSCAN can be enhanced by reducing the complexity of nearest neighbor search for each region query and by reducing the number of region queries (DBSCAN variant). This study conducted a comparative evaluation of the efficacy and effectiveness of clustering in these region queries for DBSCAN. The study concluded that the DBSCAN variant is slightly less effective than DBSCAN, but it significantly enhances efficiency.

In an additional research study [22], the author enhanced the algorithm's global search capability and introduced a semi-supervised K clustering algorithm. Initially, the K-means clustering algorithm was implemented to manage gene data. Then, the greedy iteration was employed to identify the K mean clustering in order to obtain superior results, utilizing the enhanced semi-supervised K mean clustering. The results of the simulation experiment demonstrated that the global semi-supervised K clustering algorithm has a superior cluster effect and optimization capability in comparison to the MDO algorithm. In this context [23], the researcher conducted a systematic comparison of nine well-known clustering methods that are available in the R language, assuming that the data is normally distributed. The researcher considered artificial datasets with a variety of tunable properties, such as the number of classes and the separation between classes, in order to account for the numerous potential variations of data. The assessment of the clustering methods' sensitivity to the various parameters that have been configured. The conclusion demonstrated that the spectral approach exhibited exceptional performance when the default configurations of the adopted methods were taken into account. Additionally, they discovered that the default configuration of the implemented implementations was not consistently precise. In these instances, a straightforward method that relies on the random selection of parameter values was found to be an effective alternative for enhancing performance.

In an additional article [24], the outlier of customer data was identified in order to ascertain customer behavior. The RFM (Recency, Frequency, and Monetary) models were used to determine the customer behavior by clustering the customer data using the K-Mean and DBSCAN algorithms. The investigation has concluded that the outlier in cluster 1 had a 100% similarity in DBSCAN and K-Means. However, the aggregate similarity of the outlier is 67%. The behavior of customers was characterized by a high monetary value but a low frequency of expenditure, as evidenced by the outliers. In this study [46], the researcher proposed a novel K-Means clustering algorithm to resolve the issue of a higher probability of combining dissimilar items into the same group when the number of clusters is limited. Dynamic data clustering was

implemented by the proposed methodology. Initially, the threshold value was determined as the centroid of K-Means in the proposed method, and the number of clusters was generated from this value. A pair of data points is considered to be in the same group if the Euclidian distance between them is less than or equal to the threshold value at each iteration of K-Means. Otherwise, the proposed method will generate a new cluster that contains the dissimilar data point. It has been demonstrated that the proposed approach outperforms the original K-Means method. Clustering and other statistical tools and methods were employed to evaluate students' performance in an additional research study [26]. In this investigation, the K-mean clustering algorithm was implemented. The elbow method was employed to determine the appropriate number of clusters. The analysis was conducted on a gender basis to determine whether there was a pattern based on the gender of the students. The study's findings were that the data was clustered such that data within the same cluster were similar, while data within separate clusters were not.

3. Proposed Methodology

This section describes our research technique, which is to compare K-Means and DB-SCAN clustering algorithms on small and large datasets. Figure 1 below shows our research-adapted model for this assignment.

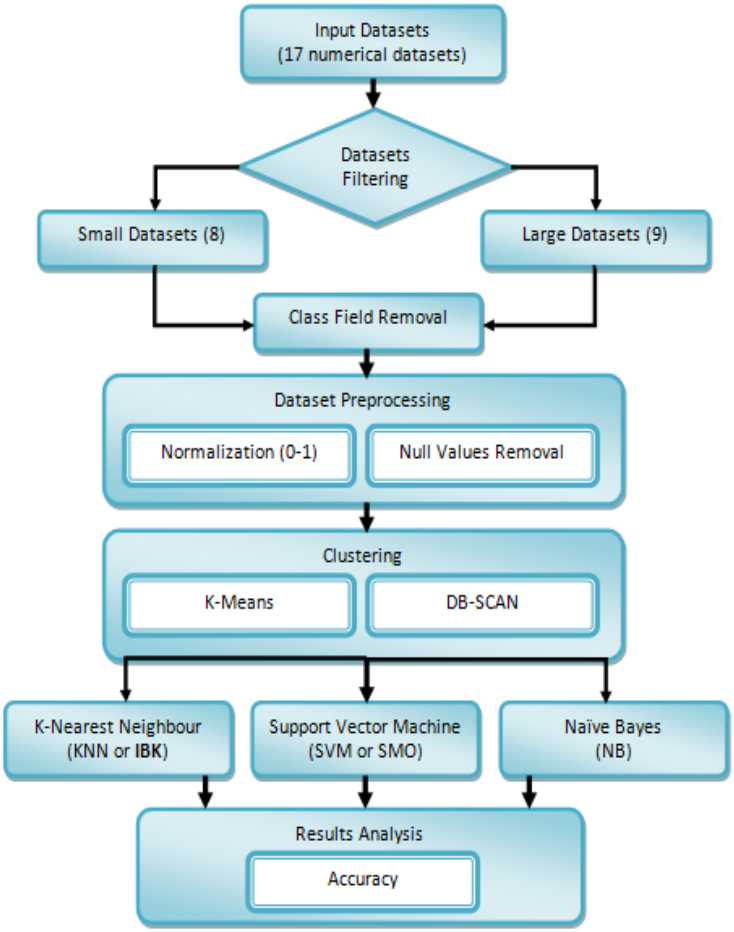

Figure 1: Proposed Methodology [24]

The detailed description of above-mentioned proposed methodology has been described below.

3.1. Input Datasets

For classification and regression, online machine learning dataset repositories offer many small and large datasets. Popular repositories include Kaggle, UCI machine learning repository, Keel, and LionBridge. Different academics use these internet databases for investigative studies and exploratory analysis. We used these repositories to retrieve 8 small and 9 large ML datasets for our investigation. Our research examined the efficiency of two popular clustering methods, K-Means and DB-Scan, utilizing these datasets.

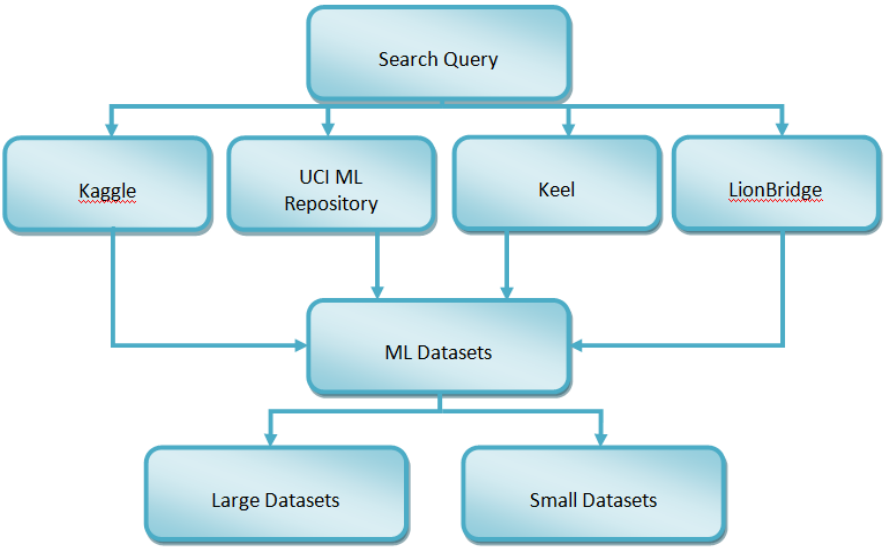

Figure 2: Datasets Searching and Filtering Process [25]

3.2. Dataset Filtering

After selecting 17 datasets, they are divided into small and large-scale datasets. This category is based on dataset instances. Small datasets have fewer than 1000 occurrences, while large datasets have more than 1000. The goal is to compare two clustering methods on large and small datasets. The figure above displays dataset collection and filtering.

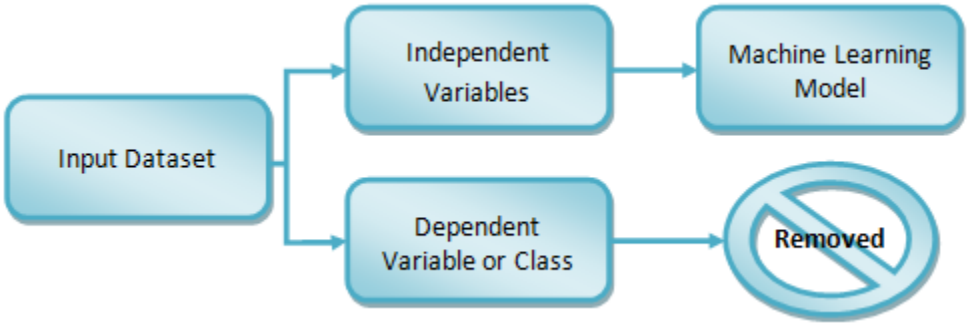

Figure 3: Class Field Removal [26]

3.3. Data Preprocessing

The use of public databases in research investigations is typically plagued by noise and missing numbers. A high-quality, noise-free dataset also boosts ML model efficiency. As we know, input datasets might be numerical, image-based, or sound-based, and each type of noise requires a distinct data mining technique to enhance. Since we are working with numerical datasets, we only investigated preprocessing approaches

that are routinely used to refine and improve numerical datasets. The table below lists two preprocessing methods we used in our investigation.

Table 1: Preprocessing Techniques

Preprocessing technique	Description
Normalization	Various data normalization techniques are used to normalize input datasets to (0-1) or (-1, 1). Min-Max, Z, and unit vector normalization are standard data or feature normalizing methods. Our study standardized input dataset numerical or quantitative attributes from 0 to 1 using min-max normalization [27].
Null Values Removal	As stated, missing values are a key concern in machine learning datasets. Common missing values removal methods include deleting rows or instances with missing data, replacing null values with column mean, median, or mode, assigning a specific value to all null cells, or using machine learning algorithms that support missing data. We used Weka's Remove-with-Filter to remove missing values from input datasets.[28]

3.4. Clustering

Data clustering is commonly used in machine learning to classify input information into two or more groups based on related properties. After data is divided into many groups or classes, each group is allocated a label. As previously stated, the goal of our research is to evaluate the efficiency of two well-known clustering techniques, DB-SCAN and K-Means. As a result, after preprocessing input datasets, these datasets are fed to the two clustering algorithms mentioned above. The number of clusters is set to two, dividing each dataset into two major categories depending on its features.

3.5. Classification

After applying clustering methods and then dataset splitting, the clustered training data is given into the classifier to determine how successfully the clustering algorithms classified the input datasets. In our study, we used three well-known machine learning classifiers: K-Nearest Neighbour (KNN), also known as Instance Based Classifier (IBK) in Weka, Support Vector Machine (SVM), and Naïve Bayes (NB). These classifiers are first trained on 70% of the training data and then tested on 30% of the testing data to determine their generalizability and the efficiency of the two clustering algorithms.

3.6. Results Analysis

There are several measures used to evaluate ML classifier performance. Accuracy, Precision, Recall, True Positives, False Positives, AUC, ROC, and others are popular evaluation measures. We used accuracy to evaluate classifier performance in our study.

The figure below illustrates a comprehensive graphical description of these steps.

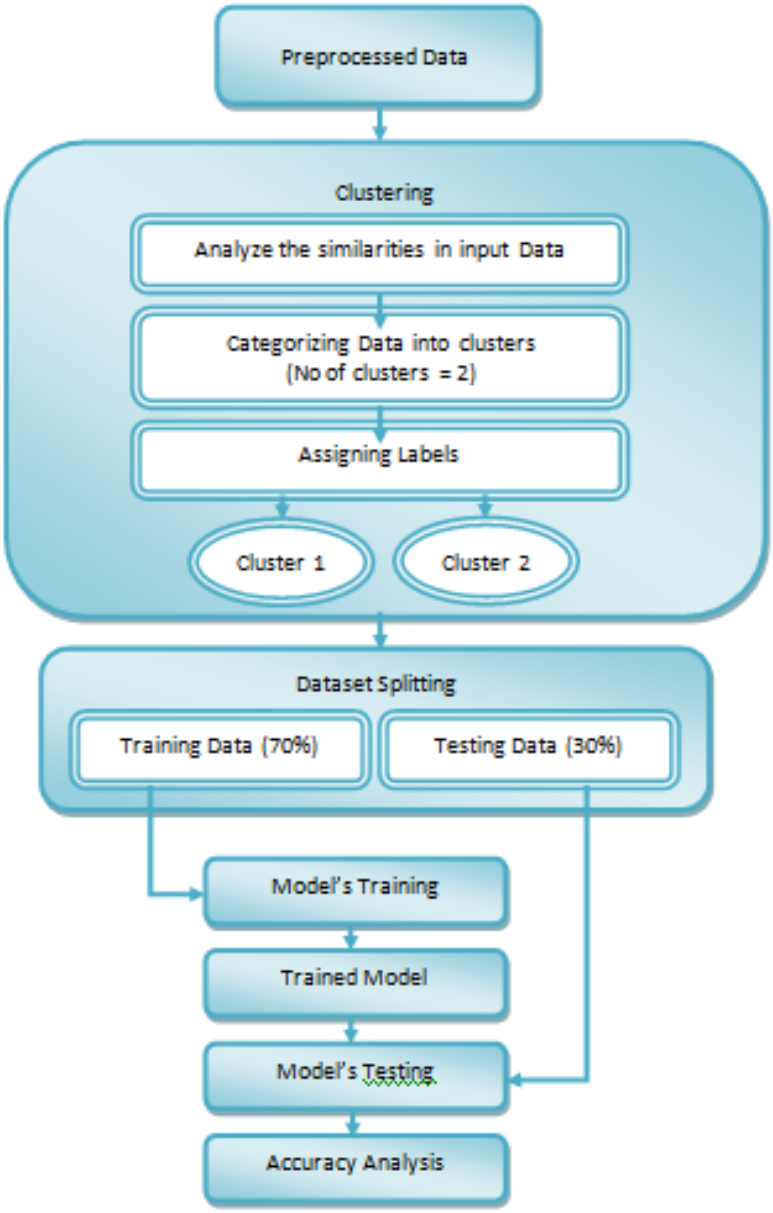

Figure 4: Detailed Model's Training Process [16]

4. Implementation and Results

In this Research, Weka 3.9.4 is used to evaluate dataset accuracy using clustering and classifiers. Data mining software Weka incorporates Machine Learning Algorithms. These algorithms can be applied on data or called from Java code. We used diverse datasets to compare K-Mean and DBSCAN clustering methods. These datasets are grouped using two algorithms, then three classifiers are deployed to evaluate clustering techniques. Results of experiments are below.

4.1. Results Analysis over Small Datasets

Below, for each dataset, are the accuracy graphs and the performance analysis (i.e., accuracy-wise) of the chosen clustering algorithms.

4.1.1. Autism Dataset

In the first scenario, we utilized the two clustering techniques previously mentioned to cluster the Autism Dataset. The table 2 below illustrates the dataset's description.

Table 2: Description of Autism Dataset

Dataset Characteristics	Value
No of Rows in Dataset	704
No of Columns in Dataset	21
Data Type of Attributes	Integer
Dataset Type	Classification
Containing Missing or Null Values	Yes

For performance analysis, the instance-based classifier (IBK), also referred to as KNN, Naïve Bayes, and SMO, is supplied the clustered dataset. Figure 16 illustrates the outcomes of the two clustering algorithms on this dataset.

Based on the results analysis, we have determined that the instance-based classifier, also known as KNN, exhibited the lowest accuracy among the two clustering algorithms. Conversely, the SMO or Support Vector Machine demonstrated the most effective performance among the two selected clustering techniques. Furthermore, the results of the experiments conducted indicate that the performance of DBSCAN is superior to that of the K-mean clustering algorithm on the Autism dataset, regardless of the classifier used.

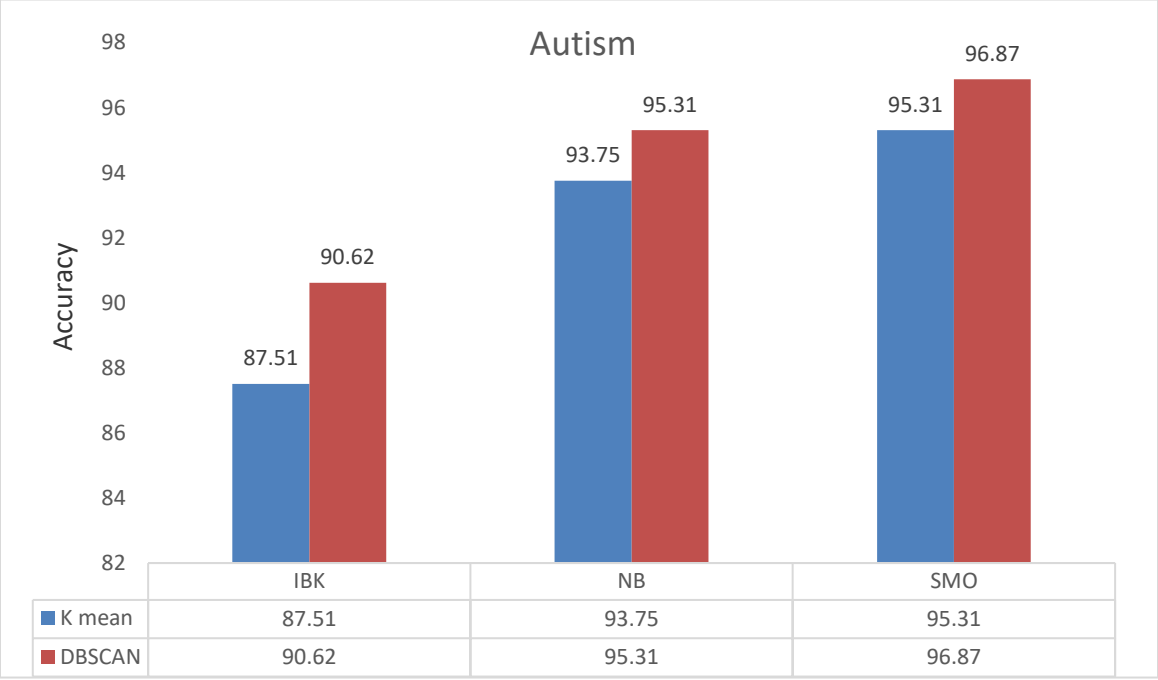

Figure 5: Performance on Autism Dataset

4.1.2. Breast Cancer Wisconsin Dataset

The second dataset that we have employed in our research is the breast cancer dataset, which is frequently employed for the automated diagnosis of breast cancer using machine learning-based techniques.

This dataset is primarily derived from digital breast images, which provide information about the characteristics of tumors or cell nuclei. As indicated in Table 3, this dataset is described below.

Table 3: Description of Breast Cancer Wisconsin Dataset

Dataset Characteristics	Value
No of Rows in Dataset	569
No of Columns in Dataset	32
Data Type of Attributes	Real
Dataset Type	Classification
Containing Missing or Null Values	No

In case of breast cancer dataset, a little variance has been noticed in accuracy of clustering techniques i.e., K-means algorithm has shown better results on two classifiers (including IBK and NB), while in case of SMO, DB-Scan has shown best accuracy of 96.15. In addition to this Naïve Bayes has depicted lowest accuracy of 73.01% on DB-Scan clustering algorithm.

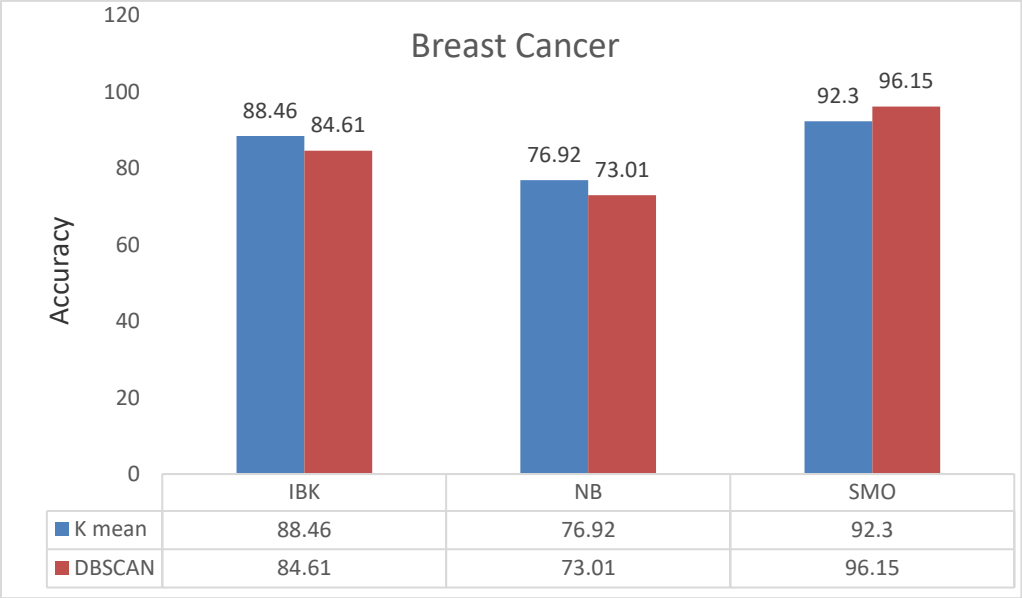

Figure 6: Performance on Breast Cancer Dataset

4.1.3. Contact Lenses Dataset

Our research has also employed an additional small-scale dataset to evaluate the effectiveness of clustering techniques on extremely small datasets. This dataset is accurate and comprehensive, as it contains no missing values. The primary objective of this dataset is to determine whether a patient requires contact lenses and whether they should be soft or firm. It comprises a total of three classes. The summary below contains additional details regarding this dataset.

Table 4: Description of Lenses Dataset

Dataset Characteristics	Value
No of Rows in Dataset	24
No of Columns in Dataset	4

Data Type of Attributes	Categorical
Dataset Type	Classification
Containing Missing or Null Values	No

The graph below illustrates the outcomes of the experiments that were conducted on this dataset.

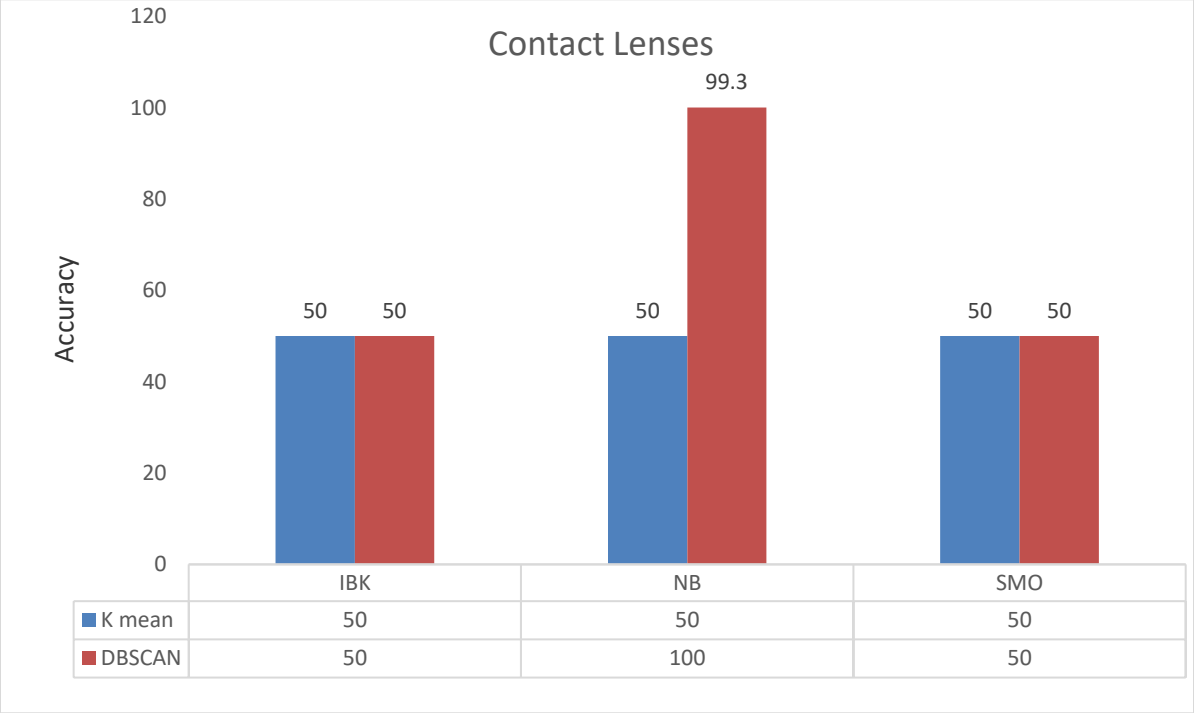

Figure 7: Performance on Contact Lenses Dataset

The classification models have not generalised well on this dataset, as evidenced by the experimental results. The primary explanation for this phenomenon may be the dataset's diminutive size. Naïve Bayes DBSCAN has demonstrated an accuracy of 99.3%, while IBK and SMO models have obtained 50% accuracy in both clustering techniques.

4.1.4. Diabetes Dataset

This dataset was compiled from two distinct sources: 1) paper documents and 2) electronic data recording devices. Various researchers have extensively employed this dataset to automate the diagnosis of diabetes disease using a variety of machine learning-based techniques. The table below provides a more detailed description of this dataset.

Table 5: Description of Diabetes Dataset

Dataset Characteristics	Value
No of Rows in Dataset	768
No of Columns in Dataset	20
Data Type of Attributes	Categorical and Integer
Dataset Type	Classification
Containing Missing or Null Values	No

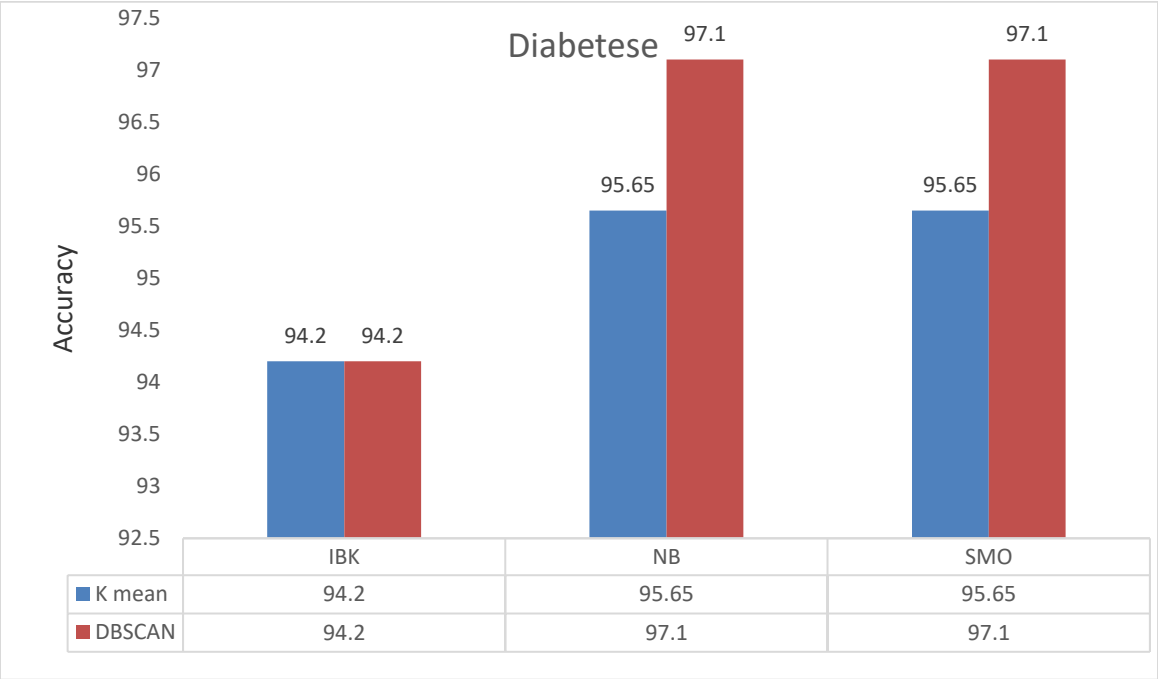

Figure 8: Performance on Diabetes Dataset

Upon analyzing the results, it is evident that both clustering techniques have demonstrated comparable results in the case of IBK (i.e., accuracy=94.2%). However, DBSCAN has outperformed K-Mean in the case of the other two classifiers, Naïve Bayes and SMO, achieving an accuracy of 97.1%

4.1.5. Titanic Dataset

Titanic dataset is an additional dataset that we have implemented in our investigation. The primary objective of this dataset is to forecast the likelihood of passenger survival aboard the Titanic. This dataset includes two classifications (YES and NO), which indicate whether the passenger has survived or not. Nine distinct risk factors have been employed to predict survival. The table below contains additional details regarding this dataset.

Table 6: Description of Titanic Dataset

Dataset Characteristics	**Value**
No of Rows in Dataset	891
No of Columns in Dataset	8
Data Type of Attributes	Categorical and Float
Dataset Type	Classification
Containing Missing or Null Values	Yes

Figure below illustrates the outcomes of the experiments conducted on the Titanic dataset. Regardless of the classifier employed, K-Mean has consistently obtained a maximum accuracy of 100%. This can be analyzed. Nevertheless, in the case of DBSCAN, IBK and NB have yielded identical results; however, SMO has attained a slightly lower level of accuracy, specifically 98.98%. In general, it is possible to infer that K-Means have demonstrated the most superior performance among all classifiers on the Titanic dataset.

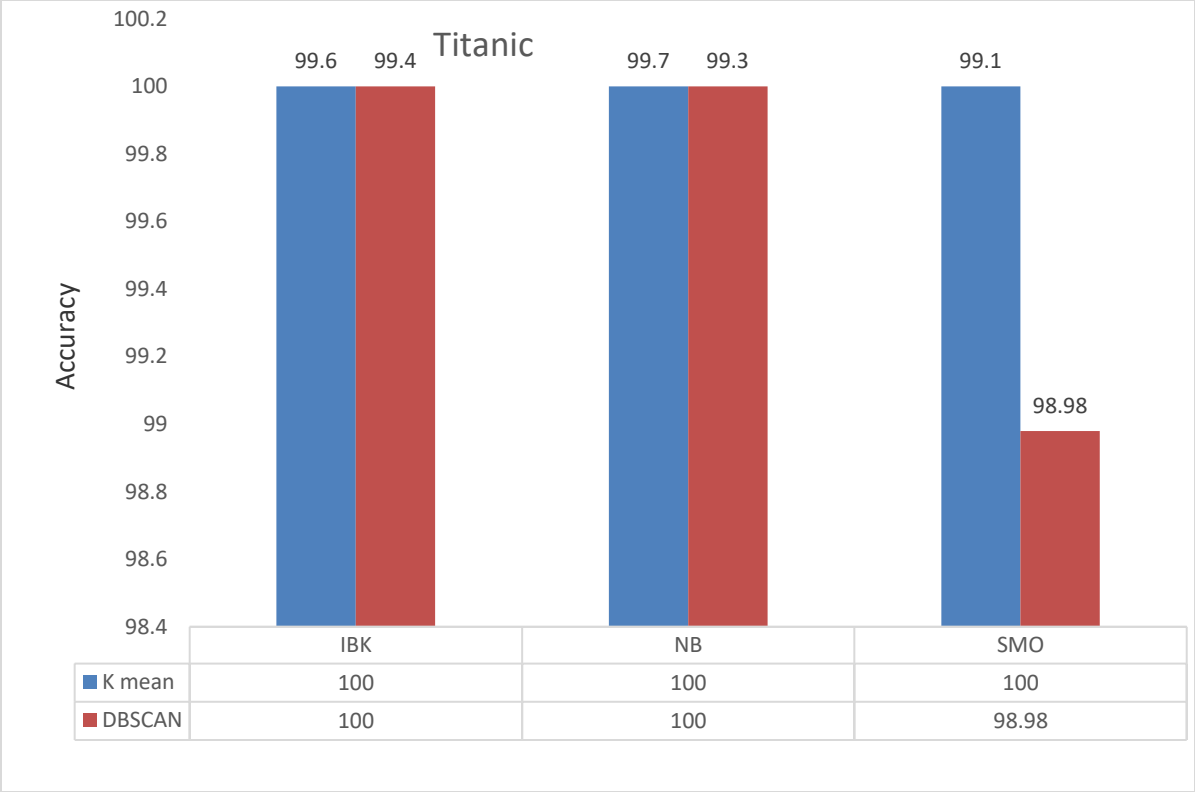

Figure 9: Performance on Titanic Dataset

4.1.6. Labor Dataset

The labor dataset is an additional dataset of limited extent that we have implemented in our investigation. This dataset has been previously employed in the literature to differentiate or categories unacceptable and acceptable contracts based on specific project attributes, such as wage, living allowance, and working hours. The table below provides a more detailed description of this dataset.

Table 7: Description of Labor Dataset

Dataset Characteristics	Value
No of Rows in Dataset	57
No of Columns in Dataset	16
Data Type of Attributes	Real, Integer and Categorical
Dataset Type	Classification
Containing Missing or Null Values	No

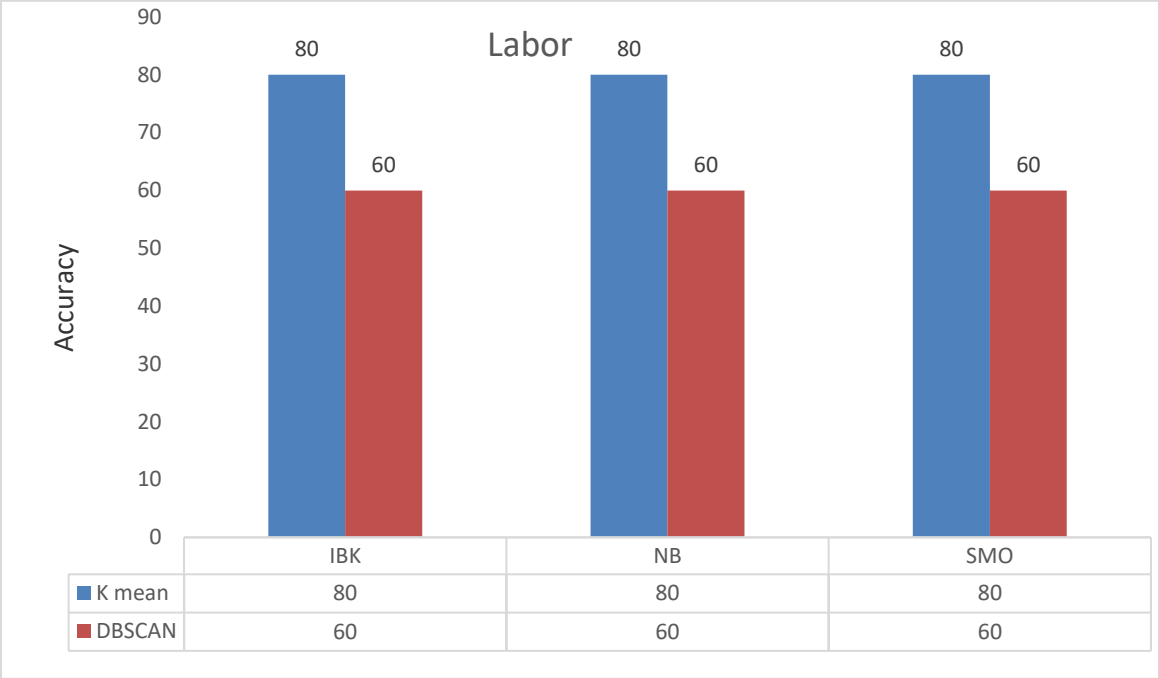

Figure 10: Performance on Labor Dataset

The labor dataset is insufficiently large to enhance the generalizability of ML classifiers. The results of the experiment conducted on this dataset indicate that the performance of K-Means is significantly superior to that of DB-Scan, despite the fact that both clustering techniques have yielded identical results across all classifiers. The maximum performance of 80% was obtained by K-Mean over all classifiers (i.e., IBK, NB, and SMO), while DB-Scan has achieved an accuracy of 60%.

4.1.7. Vote Dataset

Vote data from U.S. houses, which are emblematic of congressmen, is included in this dataset. These ballots are classified into nine distinct categories, including paired for, voted for, voted against, and paired against. Seventeen distinct categorical attributes comprise this dataset. The table below provides additional information regarding this dataset.

Table 8: Description of Vote Dataset

Dataset Characteristics	**Value**
No of Rows in Dataset	435
No of Columns in Dataset	16
Data Type of Attributes	Categorical
Dataset Type	Classification
Containing Missing or Null Values	Yes

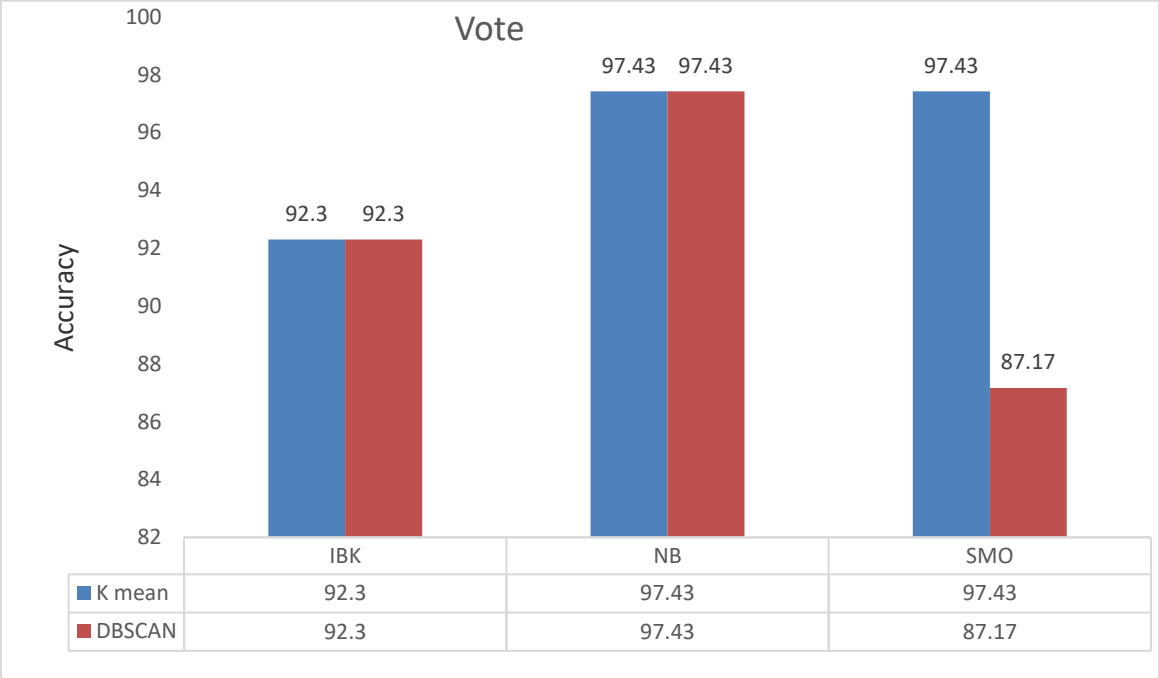

Figure 11: Performance on Vote Dataset

In the case of IBK and NB, both clustering algorithms have neared the same accuracy (92.3% and 97.43%). However, SMOK-Means have outperformed DB-Scan, with K-Means achieving an accuracy of 97.43% and DB-Scan achieving 87.17%, respectively. This is illustrated in the results graph. It is possible to infer that K-Means outperformed DB-Scan on the Vote dataset on a majority basis.

4.1.8. Soybean Dataset

The soybean dataset is the most recent dataset of a modest size that we have employed in our research. While only the first 15 classes have been utilized in prior research, this dataset contains a total of 19 classes. The class imbalance issue is caused by the relatively low number of instances pertaining to the subsequent four classes, which is why they are not being utilized. The table below provides a more detailed description of this dataset.

Table 9: Description of Soybean Dataset

Dataset Characteristics	**Value**
No of Rows in Dataset	307
No of Columns in Dataset	35
Data Type of Attributes	Categorical
Dataset Type	Classification
Containing Missing or Null Values	Yes

The graph below plainly demonstrates that DB-Scan has outperformed the two clustering techniques that were specified in all of the experiments. The highest accuracy of 98.3% was obtained by DB-Scan over the NB classifier, while the lowest accuracy was achieved over the IBK classifier at 91.8%. In summary, it could be concluded that the Soybean dataset has yielded superior results when compared to DB-Scan.

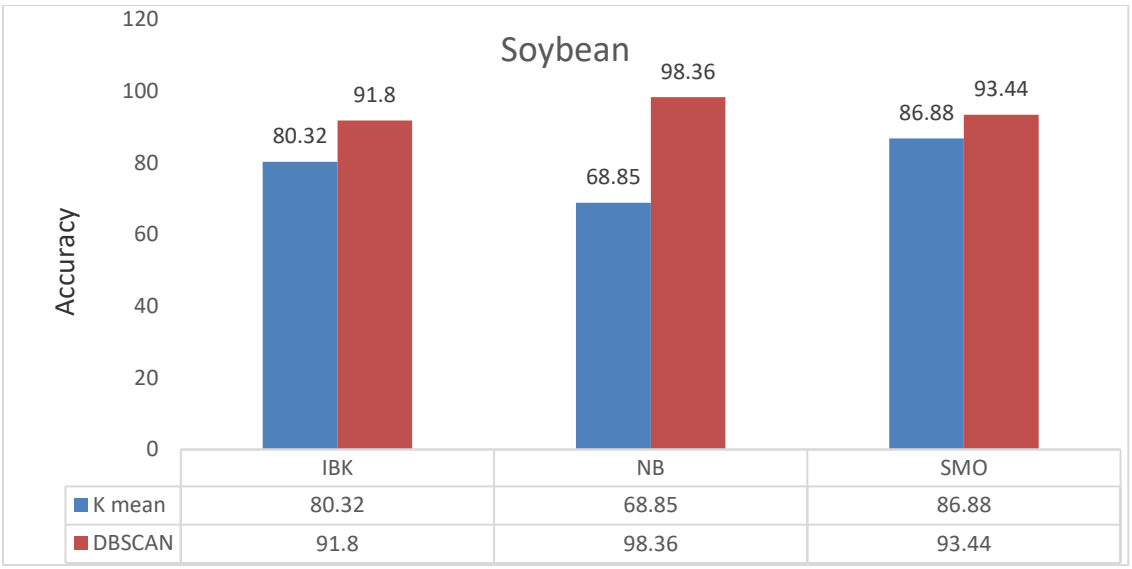

Figure 12: Performance over Soybean Dataset

4.1.9. Analysis of Small Datasets

We have conducted a collective analysis, as illustrated in the table below, to evaluate the collective impact or performance of clustering algorithms over small datasets. Specifically, we are interested in determining which clustering technique performs better when combined with which classifier.

Table 10: Analysis of Small Datasets

Small Datasets	K-Means	DB-Scan	Classifier
Autism		✓	SMO
Breast Cancer		✓	SMO
Lenses		✓	NB
Diabetes		✓	SMO, NB
Titanic	✓		SMO, NB. IBK
Labor	✓		SMO, NB. IBK
Vote	✓		SMO, NB
Soybean		✓	NB

Based on the examination of the aforementioned results, it is evident that DB-Scan has outperformed the K-Means clustering technique in the majority of instances. Specifically, K-Means has outperformed DB-Scan on three datasets, while DB-Scan has outperformed K-Means on five datasets. In addition, we have conducted an analysis to determine which classifier has yielded the most favorable outcomes when implemented with the most effective clustering technique. According to this analysis, NB and SMO have achieved the highest accuracy across six datasets, while IBK has only achieved the highest accuracy across two datasets.

4.2. Results Analysis over Large Datasets

4.2.1. Ring Dataset

The ring-norm dataset is the initial large-scale dataset that we have employed in our research. It is a 20-dimensional classification dataset that contains two classes. The summary below contains an additional description of this dataset.

Table 11: Description of Ring-norm Dataset

Dataset Characteristics	Value
No of Rows in Dataset	7400
No of Columns in Dataset	20
Data Type of Attributes	Integer, Nominal and Real
Dataset Type	Classification
Containing Missing or Null Values	No

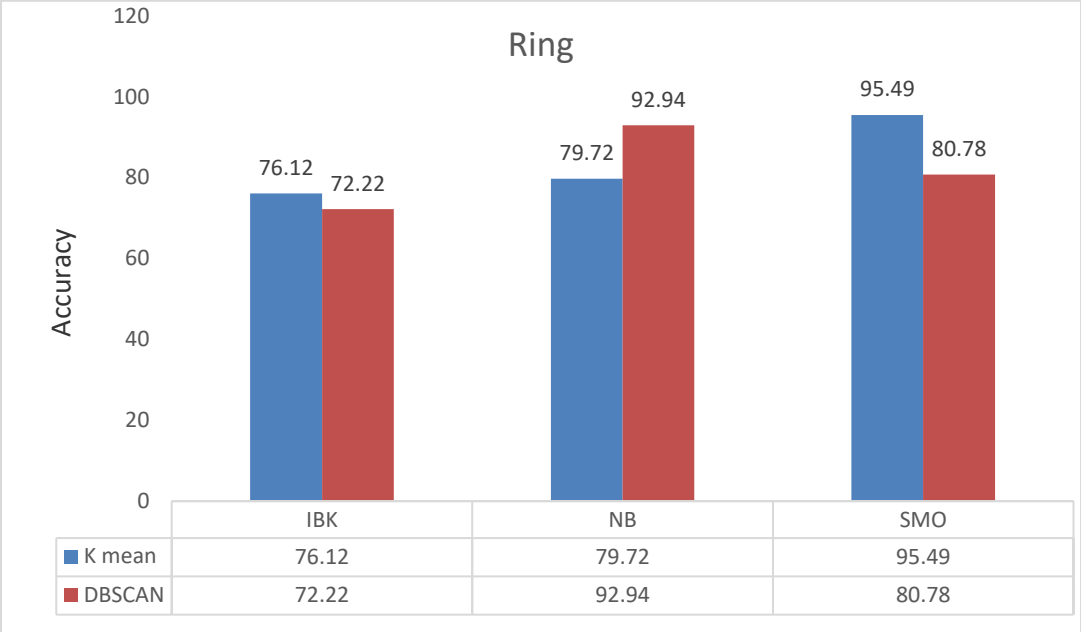

Figure 13: Performance on Ringnorm Dataset

The K-mean outperformed the ring-norm dataset in the majority of cases, as evidenced by the results analysis. Furthermore, the K-means clustering algorithm has demonstrated superior performance in the context of IBK and SMO classifiers, achieving a maximal accuracy of 95.49%. Conversely, the NB DB-Scan algorithm has outperformed the former, achieving an accuracy of 92.94%.

4.2.2. Magic Dataset

The Magic Gamma Telescope dataset is the second large-scale dataset that we have implemented in our investigation. It includes two classes: one for distinguishing gamma particles or signals from hadrons or background. Monte Carlo has made the dataset publicly available. The table below contains additional information regarding this dataset.

Table 12: Description of Magic Dataset

Dataset Characteristics	Value
No of Rows in Dataset	19020
No of Columns in Dataset	11
Data Type of Attributes	Real
Dataset Type	Classification
Containing Missing or Null Values	No

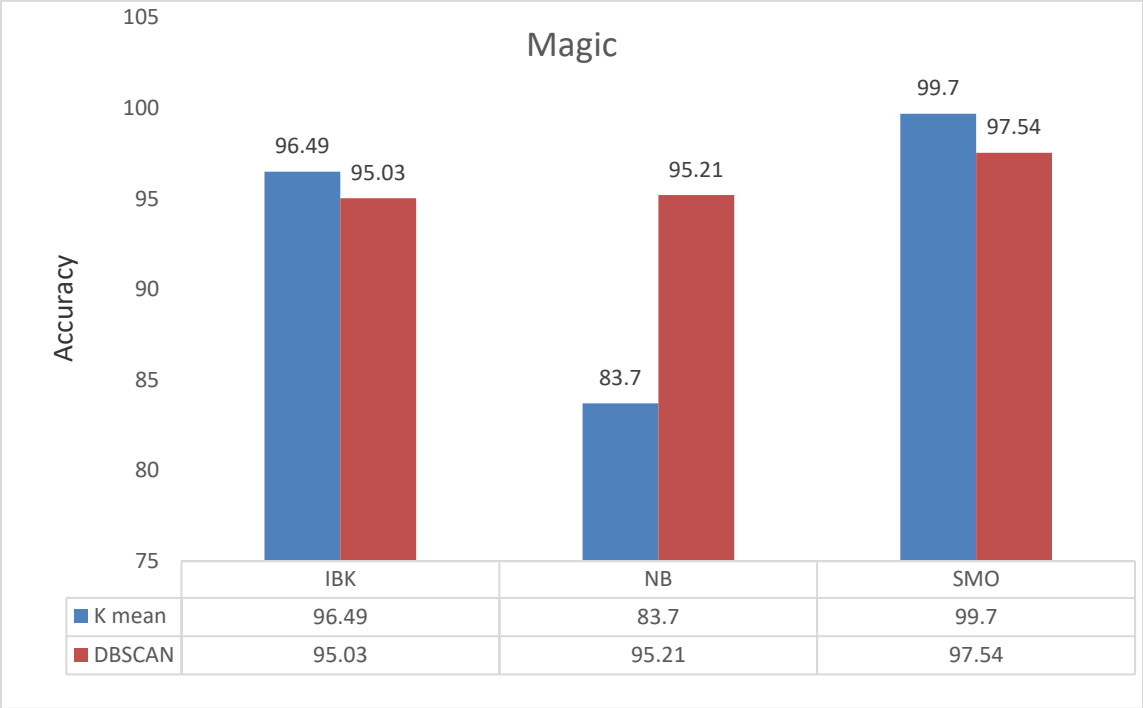

Figure 14: Performance over Magic Dataset

The experimental results reveal that K-means outperformed the two clustering strategies under consideration. K-means outperformed two of the three classifiers, IBK and SMO, with a maximum accuracy of 99.7%, although NB DB-Scan outperformed the K-mean algorithm.

4.2.3. Wine Quality Dataset

The wine quality dataset is the third large-scale dataset that we have employed in our experimental phase. The primary objective of this dataset is to evaluate the quality of wine. To achieve this, it has been compiled from two distinct varieties of Portuguese wine: red and white. This dataset has the potential to be employed to address both classification and regression issues. The quality grade is assigned within the range of 0 to 10. The summary below contains a more detailed description of this dataset.

Table 13: Description of Wine Quality Dataset

Dataset Characteristics	**Value**
No of Rows in Dataset	4898
No of Columns in Dataset	12
Data Type of Attributes	Real
Dataset Type	Regression, Classification
Containing Missing or Null Values	N/A

Again, K-Means has outperformed DB-SCAN in combination with two classifiers, NB and SMO, in the case of Wine quality in a publicly available large-scale dataset. SMO has attained the highest accuracy of 98.41% in the K-Means clustering technique. Nevertheless, DB-SCAN has outperformed K-Means in the context of the IBK classifier, achieving an accuracy of 97.73%.

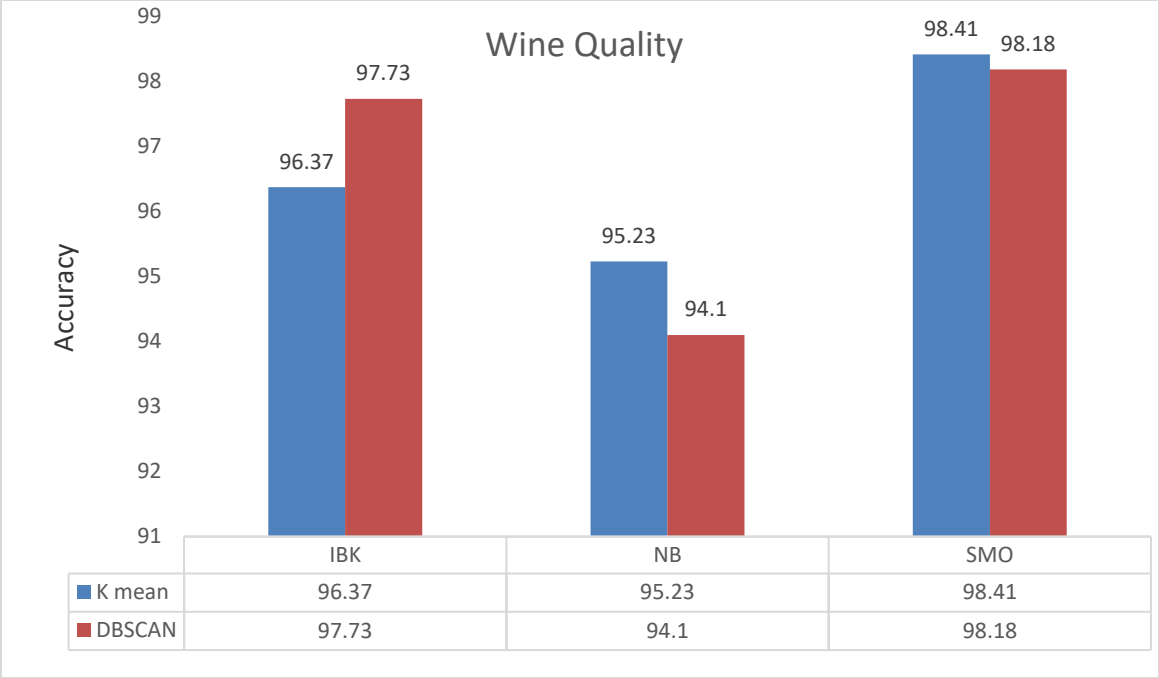

Figure 15: Performance over Win Quality Dataset

4.2.4. Shuttle Dataset

Shuttle, which is also known as the Statlog dataset, is another publicly accessible large-scale dataset. This dataset is primarily used for classification purposes and comprises a total of seven classes, which were arranged in chronological order in the original dataset. One of the seven classes is highly imbalanced, resulting in an accuracy of 80%. Consequently, the primary objective is to obtain a performance within the range of 90 to 90.9%. The summary below contains additional information regarding this dataset.

Table 14: Description of Shuttle Dataset

Dataset Characteristics	Value
No of Rows in Dataset	4898
No of Columns in Dataset	12
Data Type of Attributes	Real
Dataset Type	Regression, Classification
Containing Missing or Null Values	N/A

Similar to previous experiments, the shuttle dataset has also demonstrated superior results from two of the three classifiers, namely IBK and SMO, in comparison to the K-Mean clustering algorithm, with a maximal accuracy of 99.21%. In the case of NB, DB-SCAN has outperformed K-Means and has attained an accuracy of 93.77%.

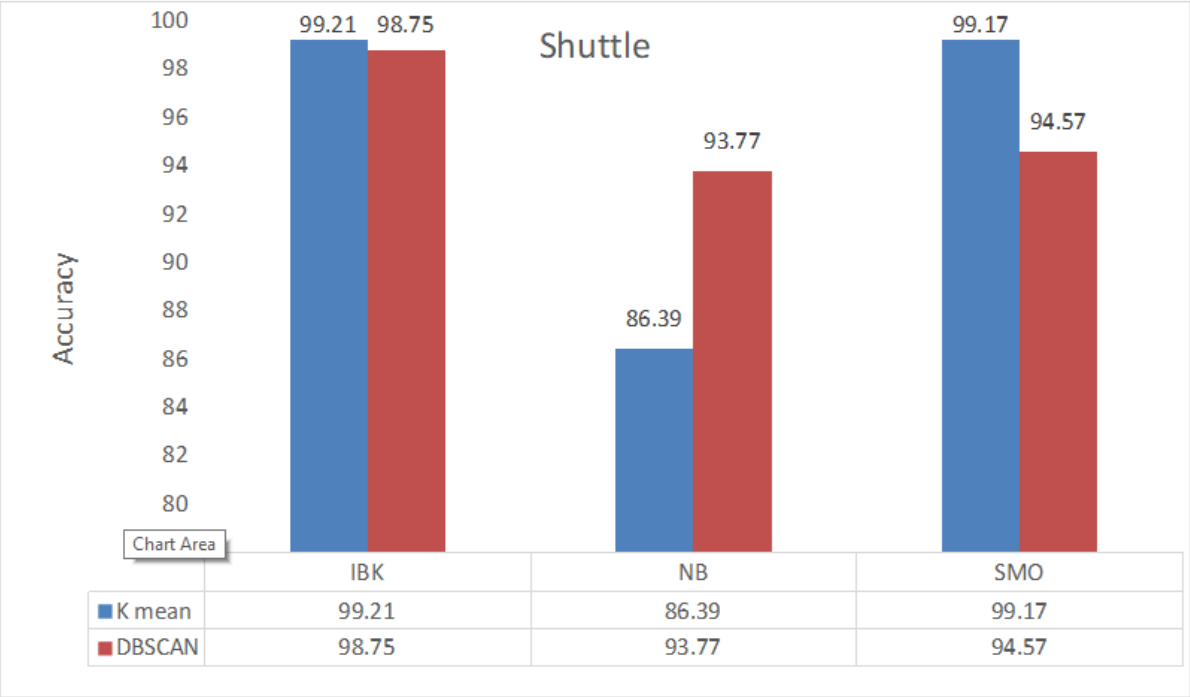

Figure 16: Performance over Shuttle Dataset

4.2.5. Thyroid Dataset

The main aim of this dataset is to differentiate between individuals with thyroid disease and those who are not. Additionally, two additional variations of this dataset have been made available to the public. The table below contains the primary attributes of this dataset.

Table 15: Description of Thyroid Dataset

Dataset Characteristics	**Value**
No of Rows in Dataset	7200
No of Columns in Dataset	21
Data Type of Attributes	Real and Categorical
Dataset Type	Classification
Containing Missing or Null Values	N/A

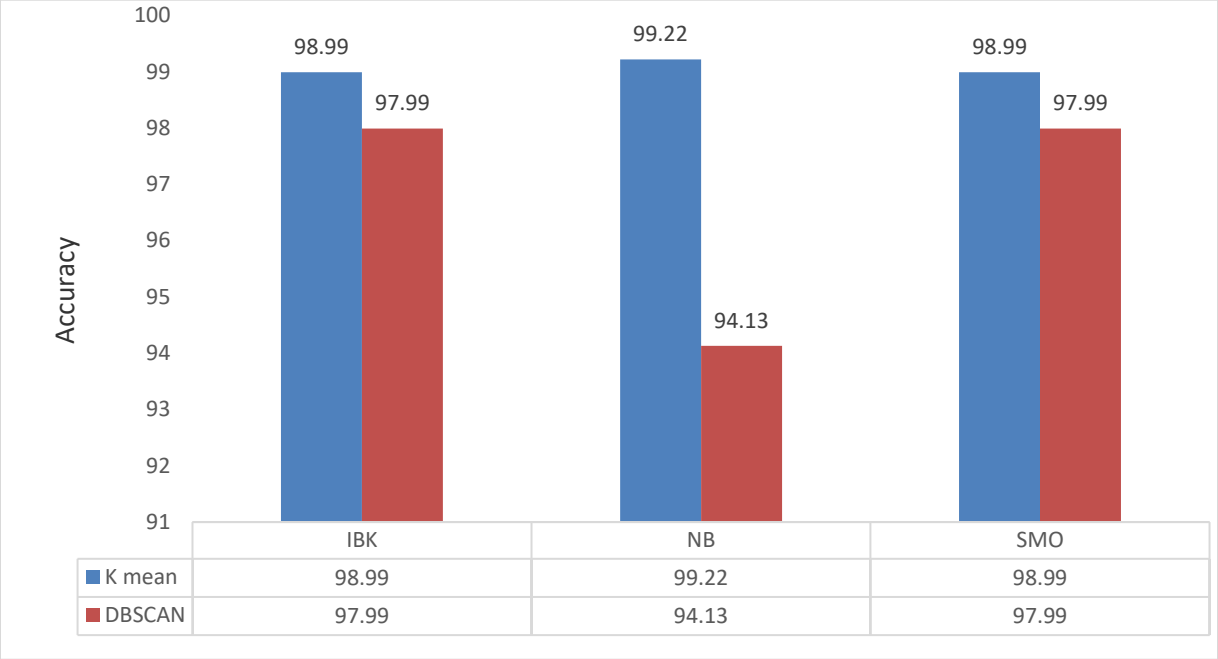

Figure 17: Performance on Thyroid Dataset

For the Thyroid dataset, all three classifiers (IBK, NB, and SMO) have demonstrated superior performance in comparison to the K-Mean clustering technique. Additionally, IBK and SMO have attained an accuracy of 98.99% in the K-Means clustering technique.

4.2.6. Texture Dataset

The texture dataset is a 40-dimensional dataset that spans a large scale and includes 11 distinct classes. Table below provides additional information regarding this dataset.

Table 16: Description of Texture Dataset

Dataset Characteristics	**Value**
No of Rows in Dataset	5500
No of Columns in Dataset	40
Data Type of Attributes	Real
Dataset Type	Classification
Containing Missing or Null Values	No

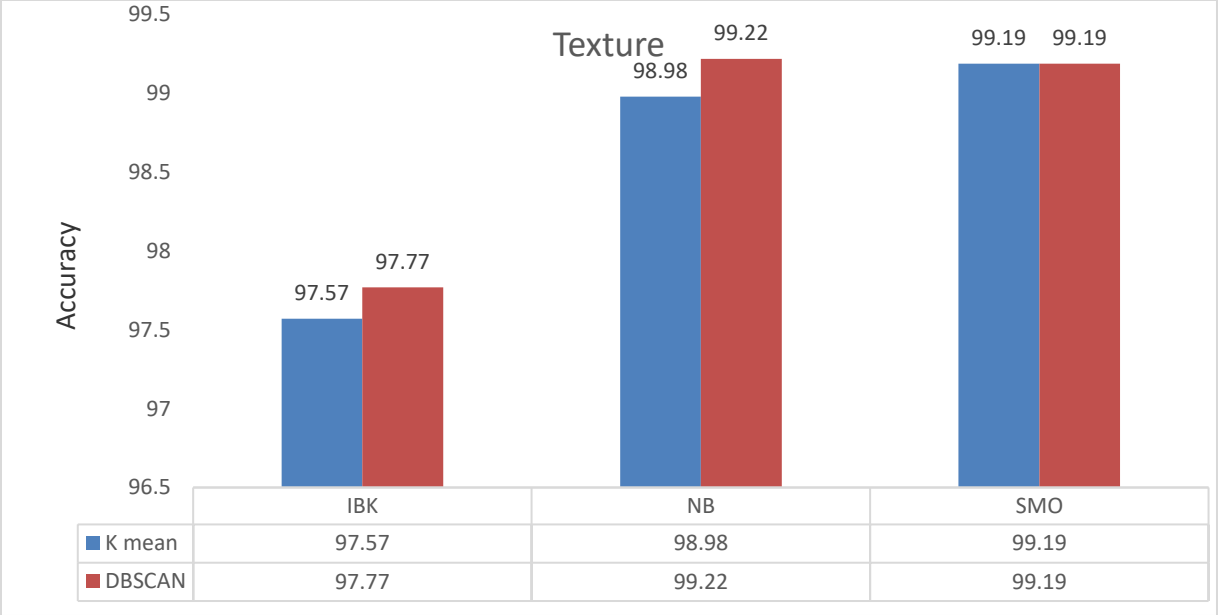

Figure 18: Performance over Texture Dataset

In contrast to previous experiments, the texture dataset has yielded the most favorable results with DB-SCAN. Additionally, both clustering techniques have attained an accuracy of 99.19% when employing SMO. DB-SCAN has obtained the highest accuracy in the case of NB and IBK, whereas NB has approached the maximum accuracy of 100%.

4.2.7. Marketing Dataset

This dataset was collected from marketing campaigns that were conducted by Portuguese financial institutes. Phone communications have been implemented during these campaigns. The primary objective of this dataset is to determine whether a client intends to enroll in a term deposit. Table below delineates additional attributes of this dataset.

Table 17: Description of Marketing Datset

Dataset Characteristics	**Value**
No of Rows in Dataset	45211
No of Columns in Dataset	17
Data Type of Attributes	Real
Dataset Type	Classification
Containing Missing or Null Values	N/A

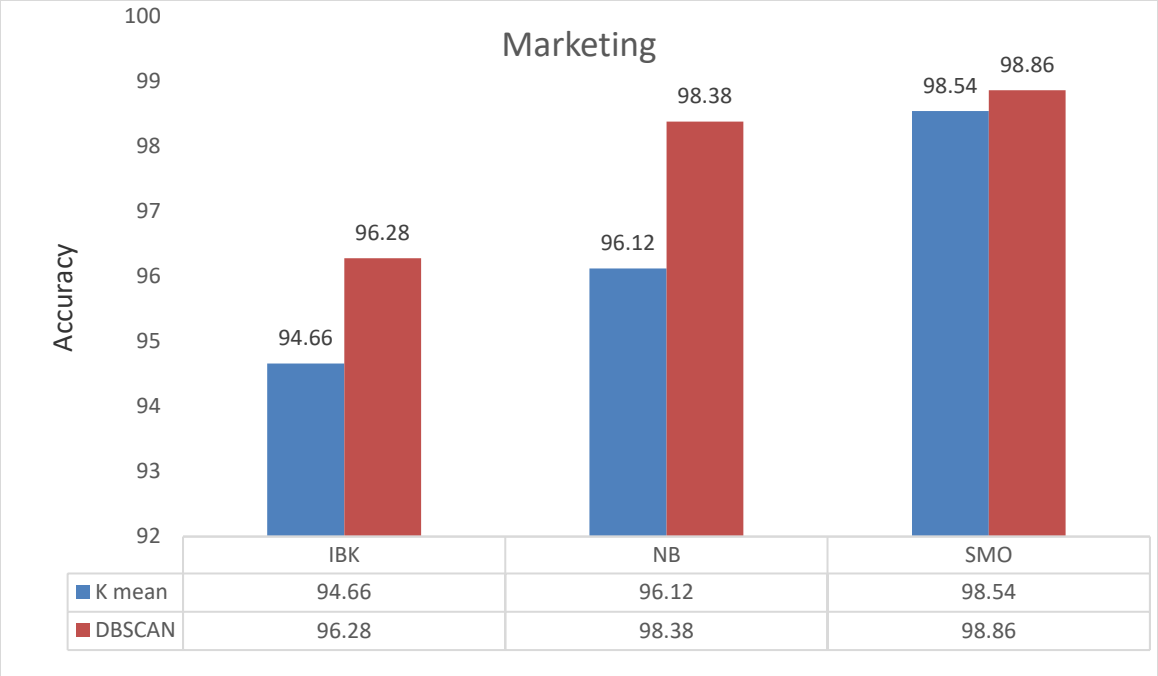

Figure 19: Performance over Marketing Dataset

The DB-SCAN clustering algorithm has demonstrated the most effective performance in the Marketing Dataset when combined with all three of the selected classifiers. SMO has obtained the highest accuracy (98.86%) when compared to data clustered by DBSCAN. IBK has attained the lowest accuracy (94.66%) when combined with K-Mean.

4.2.8. Letter Dataset

The letter recognition dataset is the second-to-last large-scale dataset that we have implemented in our investigation. The primary objective or objective of this dataset is to identify English alphabets from a rectangular grid of black and white pixels. The 20,000 distinct instances of this dataset are generated by randomly distorting the alphabetical images of 20 distinct font styles. The summary below contains a few of the dataset's most significant attributes.

Table 18: Description of Letters Dataset

Dataset Characteristics	**Value**
No of Rows in Dataset	20,000
No of Columns in Dataset	16
Data Type of Attributes	Integer
Dataset Type	Classification
Containing Missing or Null Values	No

Figure 20: Performance on Letter Dataset

DB-SCAN has obtained the highest accuracy on the letter's recognition dataset, as indicated by the experimental analysis. However, K-Means has outperformed DB-SCAN in the majority of cases. For example, K-Means outperforms DB-SCAN in the context of IBK and SMO, while DB-SCAN has obtained the highest accuracy of 98.88% in the context of NB.

4.2.9. Kr Vs K Dataset

The King-Rook Vs King-Pawn Kr-Vs-K dataset, also referred to as the chess dataset, is the most recent publicly available dataset that we have employed to evaluate the performance of clustering algorithms on large-scale datasets. The primary objective of this two-class dataset is to determine whether white individuals are capable of achieving victory. The table below provides a more detailed description of this dataset.

Table 19: Description of Kr Vs K Dataset

Dataset Characteristics	**Value**
No of Rows in Dataset	3196
No of Columns in Dataset	36
Data Type of Attributes	Categorical
Dataset Type	Classification
Containing Missing or Null Values	No

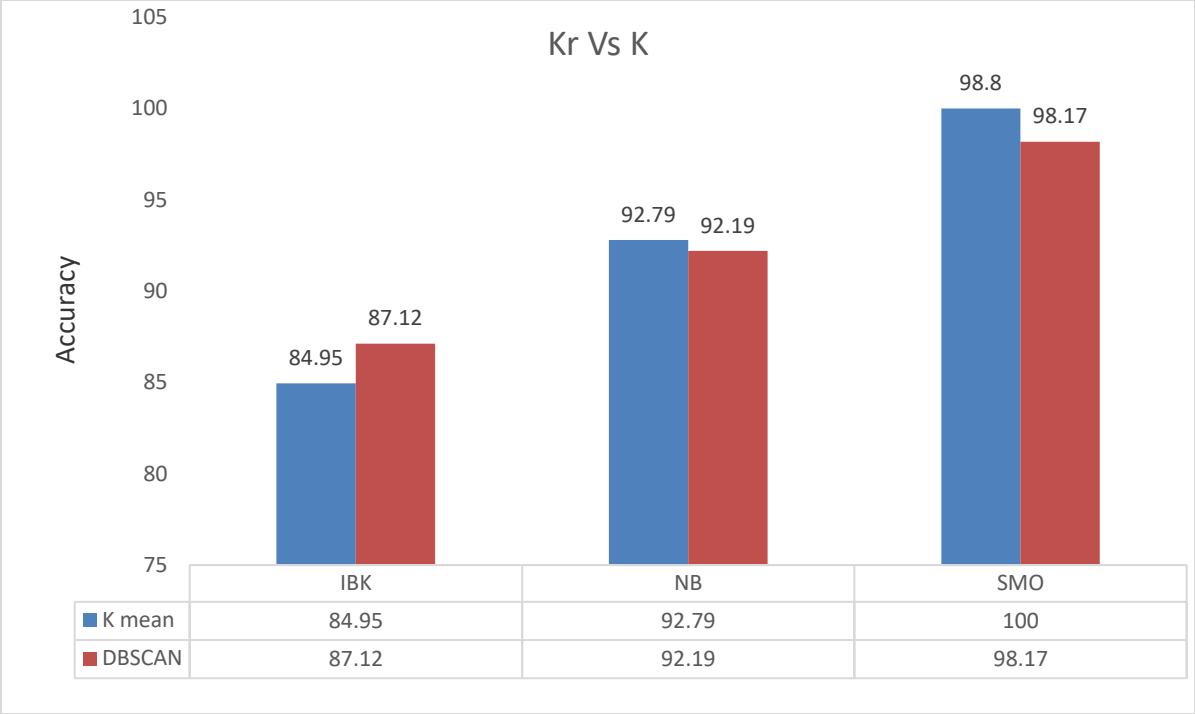

Figure 21: Performance on Kr Vs K Dataset

The efficacy of the K-Means clustering algorithm is superior to that of DB-SCAN in the case of NB and SMO classifiers, as illustrated in the graphic above. Nevertheless, the performance of DB-SCAN is superior to that of the K-Means algorithm when IBK is implemented for the classification of clustered data. As a result, it is possible to infer that the K-Means clustering technique is more accurate than the DB-SCAN clustering technique in the Kr-Vs-K dataset. Upon conducting an analysis of the classifiers, it was observed that SMO exhibited the highest level of performance among the three classifiers mentioned, with an accuracy of 100%.

4.2.10. Analysis of Large Datasets

Table 20: Results Analysis of Large Datasets

Large Datasets	K-Means	DB-Scan	Classifier
Ring-Norm Dataset	✓		SMO
Magic Dataset	✓		SMO
Wine Quality Dataset	✓		SMO
Shuttle Dataset	✓		IBK
Thyroid	✓		SMO, IBK
Texture		✓	NB
Marketing		✓	SMO
Letter		✓	NB
Kr-Vs-K	✓		SMO

Based on the comprehensive results analysis of experiments conducted over large datasets (i.e., as illustrated in Table 19), it is possible to infer that the K-Means algorithm outperforms the DB-SCAN algorithm in the context of large datasets. For example, of the nine large scale datasets utilized in our study,

K-Means outperformed DB-SCAN in the case of six datasets. Furthermore, SMO has demonstrated the highest level of accuracy in the context of datasets. Consequently, the clustering algorithm K-Means, when used in conjunction with SMO or SVM, yields superior outcomes for large-scale datasets.

Our study's limitations include the use of only 17 datasets, which may not represent the diversity of real-world data, and the focus on just two clustering algorithms, excluding many others that could yield different results. Additionally, we relied solely on accuracy for performance analysis, neglecting other important metrices.

5. Conclusion

Clustering, which groups related objects or datasets, is a popular unsupervised machine learning approach. These are clusters. various clusters of items have various features, and different similarity measures are used to compare them. Model-based, partitioning-based, hierarchal-based, grid-based, density-based, and constraint-based clustering methods are used in data analysis, image processing, pattern recognition, and market research. Given these ubiquitous clustering applications, finding the best efficient and accurate algorithm is vital. Publicly available small and big datasets are used to solve machine learning challenges and analyze algorithm performance. A large dataset is needed to fully train any machine learning algorithm, as models learned on larger datasets are more generalizable. However, most publicly accessible datasets are small or contain missing values. Small datasets often cause overfitting. Designing or developing adaptable algorithms that perform well on tiny datasets is crucial. We conducted an exploratory study to determine the best unsupervised machine learning clustering algorithm for small and large datasets. We chose two well-known clustering algorithms to test their performance in the case. The chosen clustering techniques are DB-SCAN and K-Means. To complete the study, 17 large and minor datasets were collected. Eight small and nine big datasets were preprocessed (normalized and null values eliminated). The two clustering methods receive data from the preprocessed dataset without class field. The clustered data is supplied to IBK, SVM, and NB machine learning classifiers for performance analysis. The final performance study of these algorithms used accuracy. According to the results, K-Means algorithms perform better on large datasets, whereas DB-SCAN performs better on small datasets.

In the future, we plan to broaden the scope of our analysis by incorporating additional clustering techniques into the research that has been carried out.

References

[1] Marino, Marina, and Cristina Tortora. "A comparison between K-means and Support Vector Clustering for Categorical Data." *Statistica applicata* 21, no. 1 (2009): 5-16.

[2] Namratha, M., and T. R. Prajwala. "A comprehensive overview of clustering algorithms in pattern recognition." *IOSR Journal of Computer Engineering* 4, no. 6 (2012): 23-30.

[3] Abualigah, Laith Mohammad, and Ahamad Tajudin Khader. "Unsupervised text feature selection technique based on hybrid particle swarm optimization algorithm with genetic operators for the text clustering." *The Journal of Supercomputing* 73 (2017): 4773-4795.

[4] Hinneburg, Alexander, and Daniel A. Keim. "Optimal grid-clustering: Towards breaking the curse of dimensionality in high-dimensional clustering." (1999).

[5] Dubes, Richard, and Anil K. Jain. "Clustering techniques: the user's dilemma." *Pattern Recognition* 8, no. 4 (1976): 247-260.

[6] Huang, Shujun, Nianguang Cai, Pedro Penzuti Pacheco, Shavira Narrandes, Yang Wang, and Wayne Xu. "Applications of support vector machine (SVM) learning in cancer genomics." *Cancer genomics & proteomics* 15, no. 1 (2018): 41-51.

[7] Das, Amit Kumar, Saptarsi Goswami, Amlan Chakrabarti, and Basabi Chakraborty. "A new hybrid feature selection approach using feature association map for supervised and unsupervised classification." *Expert Systems with Applications* 88 (2017): 81-94.

[8] Abualigah, Laith Mohammad, Ahamad Tajudin Khader, and Essam Said Hanandeh. "A new feature selection method to improve the document clustering using particle swarm optimization algorithm." *Journal of Computational Science* 25 (2018): 456-466.

[9] Ettouil, Monia, Habib Smei, and Abderrazak Jemai. "Particle swarm optimization on fpga." In *2018 30th International Conference on Microelectronics (ICM)*, pp. 32-35. IEEE, 2018.

[10] Arevalillo-Herráez, Miguel, Aladdin Ayesh, Olga C. Santos, and Pablo Arnau-González. "Combining supervised and unsupervised learning to discover emotional classes." In *Proceedings of the 25th Conference on User Modeling, Adaptation and Personalization*, pp. 355-356. 2017.

[11] Ibrahim, Rehab Ali, Ahmed A. Ewees, Diego Oliva, Mohamed Abd Elaziz, and Songfeng Lu. "Improved salp swarm algorithm based on particle swarm optimization for feature selection." *Journal of Ambient Intelligence and Humanized Computing* 10 (2019): 3155-3169.

[12] Idris, Adnan, Muhammad Rizwan, and Asifullah Khan. "Churn prediction in telecom using Random Forest and PSO based data balancing in combination with various feature selection strategies." *Computers & Electrical Engineering* 38, no. 6 (2012): 1808-1819.

[13] Shokouhifar, Mohammad, and Ali Jalali. "Optimized sugeno fuzzy clustering algorithm for wireless sensor networks." *Engineering applications of artificial intelligence* 60 (2017): 16-25.

[14] Xue, Bing, Mengjie Zhang, Will N. Browne, and Xin Yao. "A survey on evolutionary computation approaches to feature selection." *IEEE Transactions on evolutionary computation* 20, no. 4 (2015): 606-626.

[15] Zhang, Degan, Hui Ge, Ting Zhang, Yu-Ya Cui, Xiaohuan Liu, and Guoqiang Mao. "New multi-hop clustering algorithm for vehicular ad hoc networks." *IEEE Transactions on Intelligent Transportation Systems* 20, no. 4 (2018): 1517-1530.

[16] Xiang, Wenkun, Hao Zhang, Rui Cui, Xing Chu, Keqin Li, and Wei Zhou. "Pavo: A RNN-based learned inverted index, supervised or unsupervised?." *IEEE Access* 7 (2018): 293-303.

[17] Hofmann, Thomas. "Unsupervised learning by probabilistic latent semantic analysis." *Machine learning* 42 (2001): 177-196.

[18] Mishra, Priya, Brijesh Raj Swain, and Aleena Swetapadma. "A review of cancer detection and prediction based on supervised and unsupervised learning techniques." *Smart healthcare analytics: state of the art* (2022): 21-30.

[19] Camastra, Francesco, Marco Spinetti, and Alessandro Vinciarelli. "Offline Cursive Character Challenge: a New Benchmark for Machine Learning and Pattern Recognition Algorithms." In *18th International Conference on Pattern Recognition (ICPR'06)*, vol. 2, pp. 913-916. IEEE, 2006.

[20] Tilson, L. V., P. S. Excell, and R. J. Green. "A generalisation of the fuzzy c-means clustering algorithm." In *International Geoscience and Remote Sensing Symposium,'Remote Sensing: Moving Toward the 21st Century'.*, vol. 3, pp. 1783-1784. IEEE, 1988.

[21] Zhang, Huizhen, Fan Liu, Yuyang Zhou, and Ziying Zhang. "A hybrid method integrating an elite genetic algorithm with tabu search for the quadratic assignment problem." *Information Sciences* 539 (2020): 347-374.

[22] Mai, Xiaodong, Jiangke Cheng, and Shengnan Wang. "RETRACTED ARTICLE: Research on semi supervised K-means clustering algorithm in data mining." *Cluster Computing* 22, no. Suppl 2 (2019): 3513-3520.

[23] Rodriguez, Mayra Z., Cesar H. Comin, Dalcimar Casanova, Odemir M. Bruno, Diego R. Amancio, Luciano da F. Costa, and Francisco A. Rodrigues. "Clustering algorithms: A comparative approach." *PloS one* 14, no. 1 (2019): e0210236.

[24] Monalisa, Siti, and Fitra Kurnia. "Analysis of DBSCAN and K-means algorithm for evaluating outlier on RFM model of customer behaviour." *Telkomnika (Telecommunication Computing Electronics and Control)* 17, no. 1 (2019): 110-117.

[25] Shahriar, Nafi, SM Akib Al Faisal, Md Masfakuzzaman Pinjor, Md Al Sharif Zobayer Rafi, and Atiquer Rahman Sarkar. "Comparative Performance Analysis of K-Means and DBSCAN Clustering algorithms on various platforms." In *2019 22nd International Conference on Computer and Information Technology (ICCIT)*, pp. 1-6. IEEE, 2019.

[26] Aggarwal, Deepshikha, and Deepti Sharma. "Application of clustering for student result analysis." *Int J Recent Technol Eng* 7, no. 6 (2019): 50-53.

[27] Jansen, Aren, Manoj Plakal, Ratheet Pandya, Daniel PW Ellis, Shawn Hershey, Jiayang Liu, R. Channing Moore, and Rif A. Saurous. "Unsupervised learning of semantic audio representations." In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 126-130. IEEE, 2018.

[28] Deepajothi.S, Dr.Juliana " Survey of Clustering Algorithm of Weka Tool on Labor Dataset" International Journal of Applied Engineering Research vol. 14, no. 5, pp. 90–95, 2019