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

Research Article

Efficiency of K-Prototype and K-Mean algorithm using Support Vector Machine (SVM)

Muhmmad Sharjeel Asad Areeb[1, *] and Nabeel Asghar[1]

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

[*]Corresponding Author: Muhmmad Sharjeel Asad Areeb. Email: sharjeelasadareeb@gmail.com

Received: 16 December 2024; Revised: 1 February 2025; Accepted: 17 February 2025; Published: 20 March 2025

AID: 004-01-000050

Abstract: Clustering is a key method in unsupervised machine learning, which is commonly used to find latent patterns in unlabeled datasets. This research evaluates the efficacy of K-Means and K-Prototype clustering algorithms using five benchmark datasets that include labeled, unlabeled, and mixed-type data. After routine preprocessing, datasets were divided into 2 to 5 clusters, and a Support Vector Machine (SVM) classifier was used to check the resulting cluster assignments. Experimental results show that K-Means works better on labeled datasets, while K-Prototype works better on unlabeled and mixed-type datasets. Also, accuracy goes down as the number of clusters goes up, and the best results are shown with two clusters. These results show how the type of data and the way the clusters are set up affect how well clustering and classification tasks work.

Keywords: Clustering; Support Vector Machine; K-Prototype; K-Means;

1. Introduction

Unsupervised machine learning is very effective when there isn't much labeled data, it's too expensive, or it's not available [1]. Unsupervised approaches try to find hidden patterns, structures, or relationships in data that hasn't been labeled [2], [3]. This is different from supervised learning, which uses labeled datasets. Clustering is one of the most common methods in this group. It puts related data points into groups based on their traits [4]. Finding natural groupings by clustering is useful in many areas, including healthcare [5], marketing [6], finance [7], image processing [8], and cybersecurity [9].

K-Means is one of the most popular clustering algorithms due to its simplicity, efficiency, and ability to handle large datasets [10], [11]. It works well for numerical data by grouping points so that each cluster has minimal internal variation [12]. However, real-world datasets often contain both numerical and categorical attributes. In such cases, the K-Prototype algorithm is more suitable [1], [13], as it extends K-Means to handle mixed-type data using a different measure of dissimilarity for categorical values [14].

Evaluating clustering performance is challenging because, unlike supervised learning, there are no predefined labels to compare against [15]. One way to address this is by using a post-clustering classification approach [16], [17]. Here, the clusters formed are tested using a supervised model—such as a Support Vector Machine (SVM)—to check how well the data points can be separated based on the clusters [18], [19]. This hybrid method provides an indirect measure of clustering quality and has been explored in previous works [20], [21].

In this study, we compare the performance of K-Means and K-Prototype on both labeled and unlabeled datasets [1], [3]. We test their effectiveness under different conditions, including varying the number of clusters, and use multiple publicly available datasets [22]. Data preprocessing techniques are applied to reduce noise and improve quality before clustering [23]. We then use SVM to evaluate the separability of the clusters [18].

The paper is organized as follows: Section 2 reviews related work on clustering methods and evaluation techniques. Section 3 describes the datasets and methodology, including preprocessing, clustering, and classification steps. Section 4 presents the experimental setup, results, and comparison between K-Means and K-Prototype on labeled and unlabeled data. Section 5 summarizes key findings and suggests directions for future research in unsupervised learning and clustering evaluation.

2. Literature Review

Unsupervised machine learning, especially clustering, has been widely used in healthcare, cybersecurity, finance, and behavioral analytics because it can find hidden patterns in datasets that don't have labels. K-Means is still one of the most common clustering methods since it works well with big sets of numbers and is easy to scale [1]. However, real-world datasets frequently encompass both numerical and categorical variables, constraining the direct use of K-Means. To solve this problem, Huang [2] came up with the K-Prototype algorithm, which uses the Euclidean distance metric from K-Means for numerical characteristics and the dissimilarity measure from K-Modes for categorical attributes.

Sharma et al. [3] utilized K-Means clustering on patient medical records in healthcare applications to discern high-risk groups, obtaining enhanced prediction performance when combined with Support Vector Machine (SVM) classification. In a similar way, Singla and Bhatia [4] showed that K-Means followed by SVM classification made it much easier to accurately forecast disease categories than clustering alone. These results indicate that post-clustering classification can function as an efficient indirect assessment of clustering quality.

In cybersecurity, Aljawarneh et al. [5] put forward a hybrid intrusion detection model that used K-Means to group network traffic at first and then SVM to classify it, which led to better detection accuracy. Elngar et al. [6] also used a K-Means–SVM pipeline to find anomalies in IoT environments and said that it had lower false positive rates than other methods.

Beyond numerical datasets, Joshi and Dang [7] applied K-Means with SVM classification to predict online user preferences in e-commerce, showing enhanced personalization accuracy. Kumar et al. [8] extended this approach to educational data mining, where clustering was used to identify learning behavior patterns prior to classification.

While prior works have explored K-Means extensively, fewer studies have investigated K-Prototype in conjunction with SVM for mixed-type data. Kumari and Yadav [9] conducted a comparative analysis of K-Means, K-Modes, and K-Prototype, concluding that K-Prototype produced better clustering quality for mixed-attribute datasets. However, they did not evaluate the post-clustering classification performance, leaving a research gap in understanding how such algorithms impact separability in supervised learning. A consolidated summary of related works is presented in **Table 1**.

Table 1: Literature Review Analysis

Ref.	Algorithm(s)	Purpose	Performance	Evaluation Measure
[1]	K-Means, K-Prototype	Performance analysis for outlier detection	Produced better clusters using proposed architecture	Efficiency and accuracy
[2]	Gray Wolf Optimization (GWO), Support Vector Machine (SVM)	Compare accuracy with other methods	Achieved a 27.68% accuracy improvement over baseline methods	Accuracy

[3]	K-Prototype (combination of K-Means and K-Modes)	Measure user behavior	Clusters revealed more accurate user preferences	Clustering of mixed-attribute data
[4]	K-Means	Application in data mining and pattern recognition	Demonstrated efficiency and promising results	Algorithm behavior and performance
[5]	SVM with Sequential Minimal Optimization (SMO)	Improve SVM efficiency	SMO outperformed standard SVM	Classification accuracy
[6]	K-Means + SVM, Weighted SVM (WSVM)	Measure predictive performance using boosting	Both K-Means SVM and WSVM improved classification accuracy	Accuracy comparison
[8]	K-Means, Genetic Algorithm (GA), SVM	Optimal feature selection in data mining	Achieved 98.79% accuracy on reduced datasets	Accuracy of reduced datasets
[9]	K-Means + SVM Classifier (K-SVM)	Hyperplane separation between two classes	Selected most informative samples, improving efficiency	Time efficiency
[7]	K-Means, SVM	Intrusion detection	Achieved 90% detection accuracy	Attack detection accuracy
[10]	Basic K-Means, Enhanced K-Means	Address limitations of K-Means	Enhanced K-Means outperformed Basic K-Means	Efficiency
[11]	Improved K-Means (based on largest minimum distance)	Reduce dependence on initial points and avoid local minima	Improved K-Means outperformed Basic K-Means	Efficiency and time
[12]	K-Means	Improve time efficiency	Removed limitations of standard K-Means	Time efficiency and performance
[13]	K-Means, Euclidean Distance	Data analysis	Provided efficient clustering results	Performance evaluation
[14]	K-Means, SVM	Compare clustering and classification for categorical data	K-Means produced better results	High-dimensional feature space analysis
[15]	SVM, Derivative-Free Numerical Optimizer	Process optimization	Efficiently performed classification tasks	Efficiency
[16]	SVM	Pattern recognition, regression, and operator inversion	Achieved 22× faster results compared to baseline	Accuracy and speed

3. METHODOLOGY

The proposed framework employs a hybrid evaluation strategy that integrates unsupervised clustering with supervised classification to assess the performance of K-Means and K-Prototypes algorithms. The methodology is organized into four primary stages: (1) dataset selection and preprocessing, (2) clustering, (3) supervised classification, and (4) evaluation. The overall workflow is illustrated in **Figure. 1**.

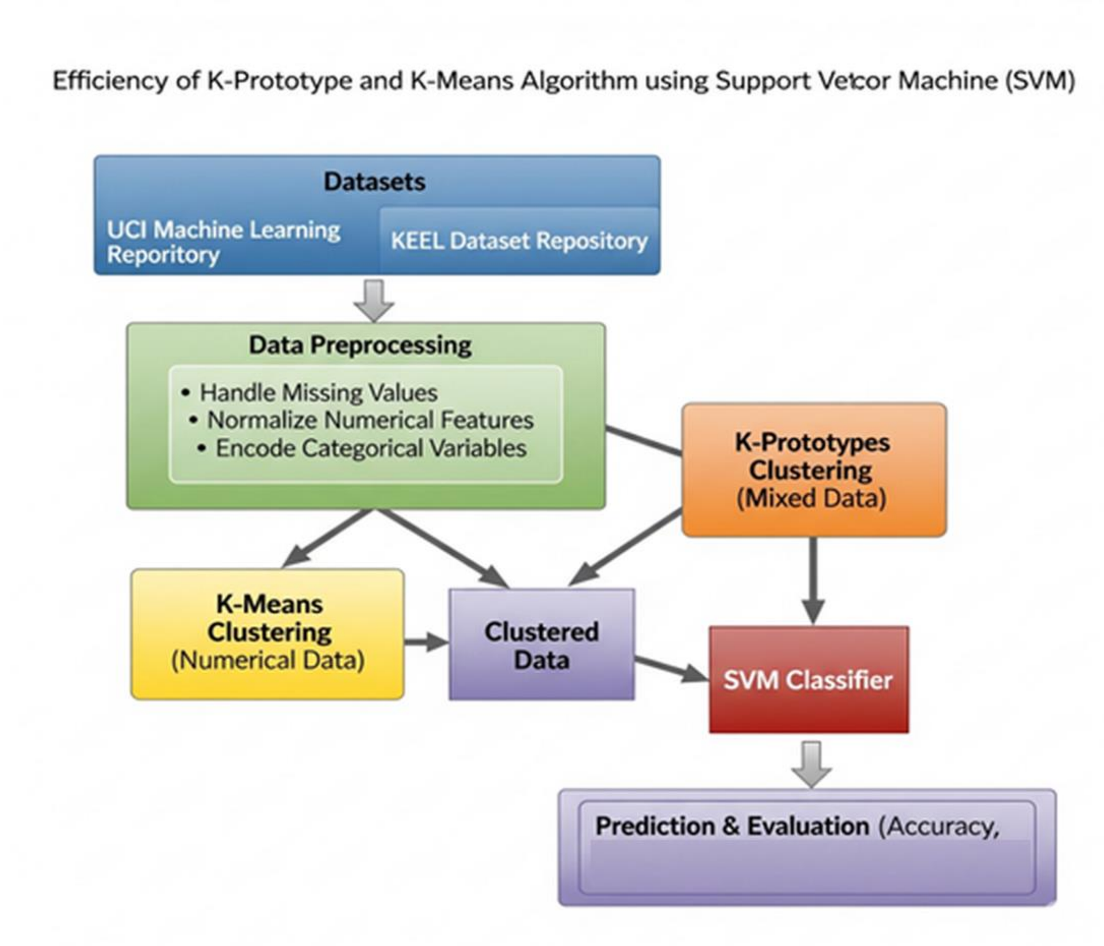
Efficiency of K-Prototype and K-Means Algorithm using Support Vetcor Machine (SVM)
Datasets
UCI Machine Learning Reporitory
KEEL Dataset Repository
Data Preprocessing
• Handle Missing Values
• Normalize Numerical Features
• Encode Categorical Variables
K-Prototypes Clustering (Mixed Data)
K-Means Clustering (Numerical Data)
Clustered Data
SVM Classifier
Prediction & Evaluation (Accuracy,

Figure 1: Proposed Methodology Diagram

3.1. Dataset Selection

Eight benchmark datasets were sourced from two widely recognized repositories: the UCI Machine Learning Repository [24] and the KEEL (Knowledge Extraction based on Evolutionary Learning) Dataset Repository [25]. These repositories were selected due to their dataset diversity, established use in prior clustering and classification studies, and suitability for evaluating hybrid algorithms. The datasets encompass both numerical and categorical features, making them appropriate for mixed-type clustering algorithms such as K-Prototypes [1], [3], which combine Euclidean and categorical dissimilarity measures.

The datasets used include: the **Adultt dataset** [24], containing 48,842 instances with 14 categorical and integer attributes, used for income classification with some missing values; the **Hepatitis dataset** [24] with 155 instances and 19 mixed-type attributes for survival prediction, also containing missing data; the **Primary Tumor dataset** [24] comprising 339 categorical attributes aimed at tumor location classification with missing values; the **Arrhythmia dataset** [24] with 452 instances and 279 mostly real-valued attributes for arrhythmia diagnosis, containing missing data; the **Monks-Problems dataset** [24] with 432 categorical instances used for binary classification, having no missing values; the **Airlines Delay dataset** [25] with 500 instances and categorical/integer features for flight status classification, without missing values; the **Credit Approval dataset** [24] containing 690 mixed-type attributes for credit approval classification, including missing values; and the **Vowel (Japanese Vowels) dataset** [24], a time series dataset with 640 real-valued attributes used for classification, without missing values. These datasets provide a comprehensive test bed for evaluating clustering algorithm performance across varied data types and domains. A summary of the selected datasets, including size, type, and missing values, is presented in Table 2.

Table 2: Summary of Selected Datasets

Dataset Name	# Instances	# Attributes	Data Types	Purpose	Missing Values
Adultt [24]	48,842	14	Categorical, Integer	Classification	Yes
Hepatitis [24]	155	19	Real, Categorical, Integer	Classification	Yes
Primary Tumor [24]	339	17	Categorical	Classification	Yes
Arrhythmia [24]	452	279	Real, Categorical, Integer	Classification	Yes
Monks-Problems [24]	432	7	Categorical	Classification	No
Airlines Delay [25]	500	7	Categorical, Integer	Classification	No
Credit Approval [24]	690	15	Real, Categorical, Integer	Classification	Yes
Vowel (Japanese) [24]	640	12	Real	Classification	No

3.2. Data Preprocessing

To ensure compatibility with the clustering algorithms and enhance performance, each dataset underwent a standardized preprocessing pipeline.

- Missing values were addressed using mean or mode imputation [26], depending on the attribute type, while records with excessive incompleteness were removed to maintain data integrity.
- Numerical features were normalized to a [0,1] range using Min–Max scaling [27] to prevent bias in distance-based clustering.
- For categorical variables, preprocessing differed based on the clustering algorithm: in the case of K-Means [2], categorical attributes were transformed into numerical form through one-hot encoding [28], whereas for K-Prototypes, categorical attributes were preserved in their native form to fully leverage the algorithm’s capability to handle mixed-type data.

This ensured comparability between the two clustering approaches and prevented distance bias.

3.3. Model Architecture

SMO classifier was used to evaluate two clustering algorithms: K-Means and K-Prototypes. Five datasets were tested both with and without the class attribute to compare classification accuracy. Initially, the K-Means technique was applied to each dataset, varying the number of clusters from 2 to 5. After clustering, the resulting grouped datasets—first including the class attribute and then with the class attribute removed—were fed into the SMO classifier to measure accuracy. Subsequently, the K-Prototypes clustering algorithm was applied under the same conditions to assess its performance with and without class labels.

3.4. Clustering Phase

In the initial stage, datasets were grouped without the use of class labels to identify inherent data structures.

- For purely numerical datasets, the K-Means algorithm [2] was employed, utilizing Euclidean distance as the similarity measure.
- In contrast, for mixed-type datasets containing both numerical and categorical attributes, the K-Prototypes algorithm [1], [3] was applied, combining Euclidean distance for numerical attributes with simple matching dissimilarity for categorical attributes.

This clustering process revealed the underlying patterns within the data, forming the foundation for subsequent supervised evaluation. In this study, the number of clusters was systematically adjusted from 2 to 5 to assess the impact of cluster granularity on classification accuracy. This range was chosen because it strikes a compromise between simplicity (fewer clusters) and granularity (more clusters), and it doesn't go too far with the number of clusters, which can lead to over-segmentation. It is important to note that no automatic cluster number discovery approach, such the elbow method or silhouette analysis, was used. Instead, the goal was to look at how clustering worked and how well the classifier worked with a set range of cluster values.

3.5. Classification Phase

To quantitatively evaluate the quality of clustering, the resulting cluster assignments were mapped to the actual class labels using a Support Vector Machine (SVM) classifier [29].

This supervised step was included for two reasons:

1. **Validation:** checking how well unlabeled clusters align with ground truth via majority voting [30].
2. **Predictive testing:** assessing if discovered clusters improve downstream classification accuracy. It facilitated predictive performance assessment, leveraging the SVM's ability to model both linear and non-linear decision boundaries [31] to achieve robust classification, even in high-dimensional and complex datasets.

3.6. Performance Evaluation

To evaluate the performance of trained SVM classifier over clustered data, we have exploited 'Accuracy' classification metric. The formulae for its evaluation have been mentioned below.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{1}$$

4. Results and Discussion

The performance of K-Means and K-Prototype clustering algorithms was evaluated across multiple datasets using the SMO classifier for accuracy measurement. The datasets used included Adultt, Hepatitis, Primary Tumor, Arrhythmia, Monks-Problems, Airlines, Credit, and Vowel datasets. For each dataset, clustering was performed with the number of clusters ranging from two to five, and experiments were conducted using both datasets containing the class attribute and those with the class attribute removed. To figure out how well the K-Means and K-Prototype clustering algorithms worked, we used the SMO classifier in a 10-fold cross-validation scenario to find the classification accuracy. This method makes sure that the results aren't biased toward either the training or the test data because each dataset is split into training and testing subsets in several folds. The accuracies that were provided are the average performance across all folds, not just the training data.

For the **K-Means** clustering technique, the Adultt dataset, comprising 48,842 instances and 14 attributes, showed the highest classification accuracy of 97.57% when clustered into two groups including the class attribute. However, at three and four clusters, the model trained on data without the class attribute outperformed that with class labels. Similarly, the Hepatitis dataset, with 155 instances and 19 attributes, achieved its peak accuracy of 92.85% at two clusters with class information included. Interestingly, for clusters of four and five, the classifier trained on data excluding the class attribute demonstrated better performance. The Primary Tumor dataset (339 instances, 17 categorical attributes) achieved perfect accuracy (100%) at two clusters for both data variants, and also at three clusters without the class attribute.

The Arrhythmia dataset, which is more complex with 452 instances and 279 attributes, attained 100% accuracy at two clusters for both dataset types. For higher cluster counts, models trained without class labels showed improved results compared to those with class information. Finally, the Monks-Problems dataset, containing 432 instances and seven categorical fields, reached a maximum accuracy of 92.3% with two clusters across both dataset types. At higher cluster counts (3 to 5), models trained without the class attribute outperformed those with it. Overall, the K-Means results consistently indicated that classification accuracy declines as the number of clusters increases, with two clusters providing optimal accuracy. Furthermore, inclusion of the class attribute generally benefitted models at low cluster counts, whereas for higher cluster counts, models trained on data without class labels sometimes achieved better performance.

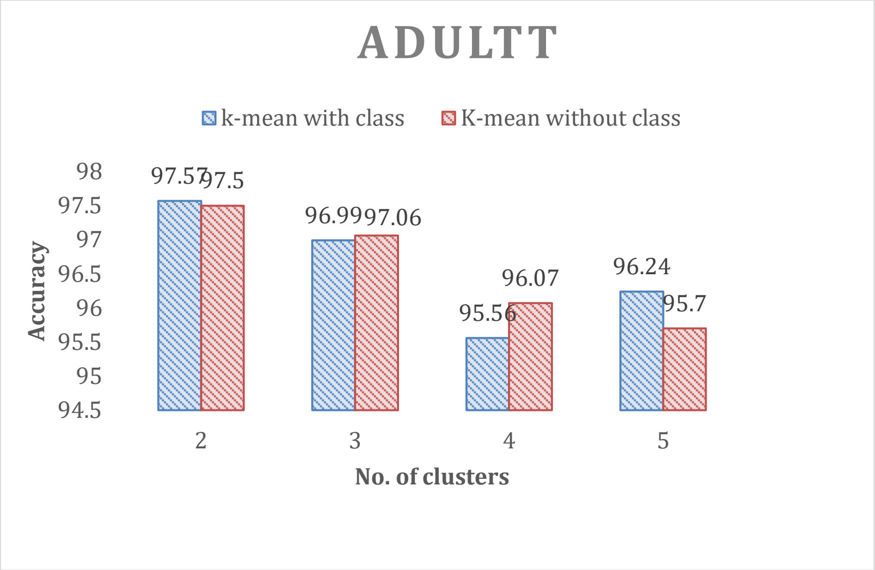

Figure 2: Performance of K-Means on ADULTT Dataset

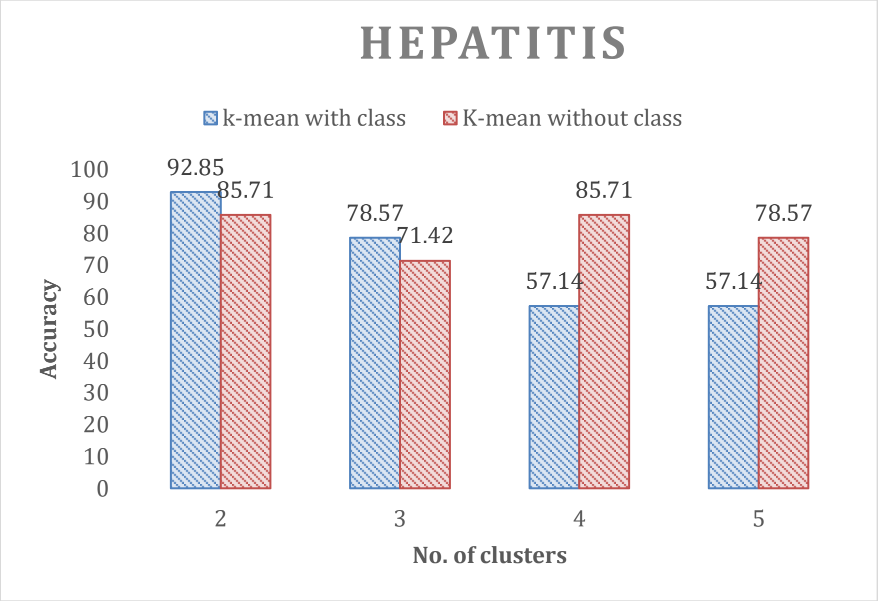

Figure 3: Performance of K-Means on HEPATITIS Dataset

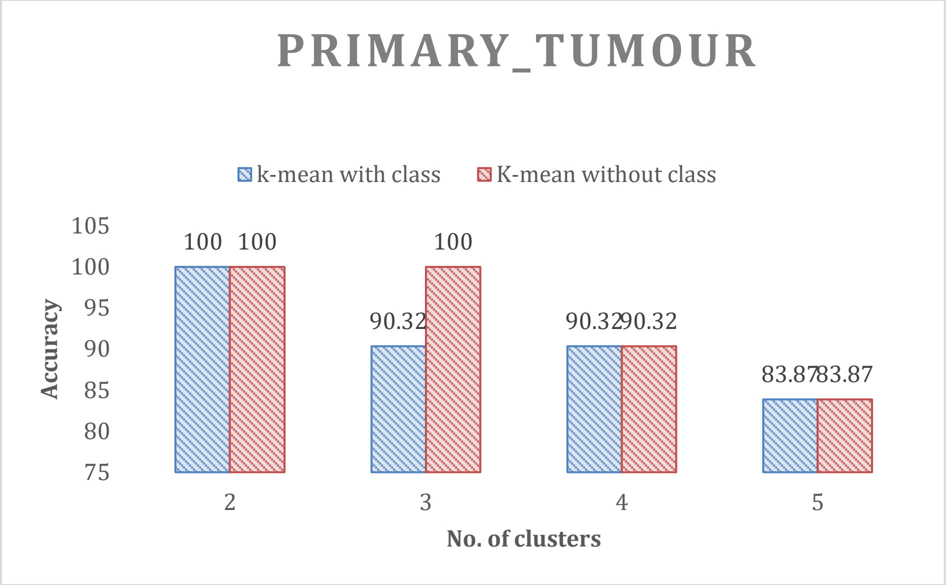

Figure 4: Performance of K-Means on PRIMARY_TUMOUR Dataset

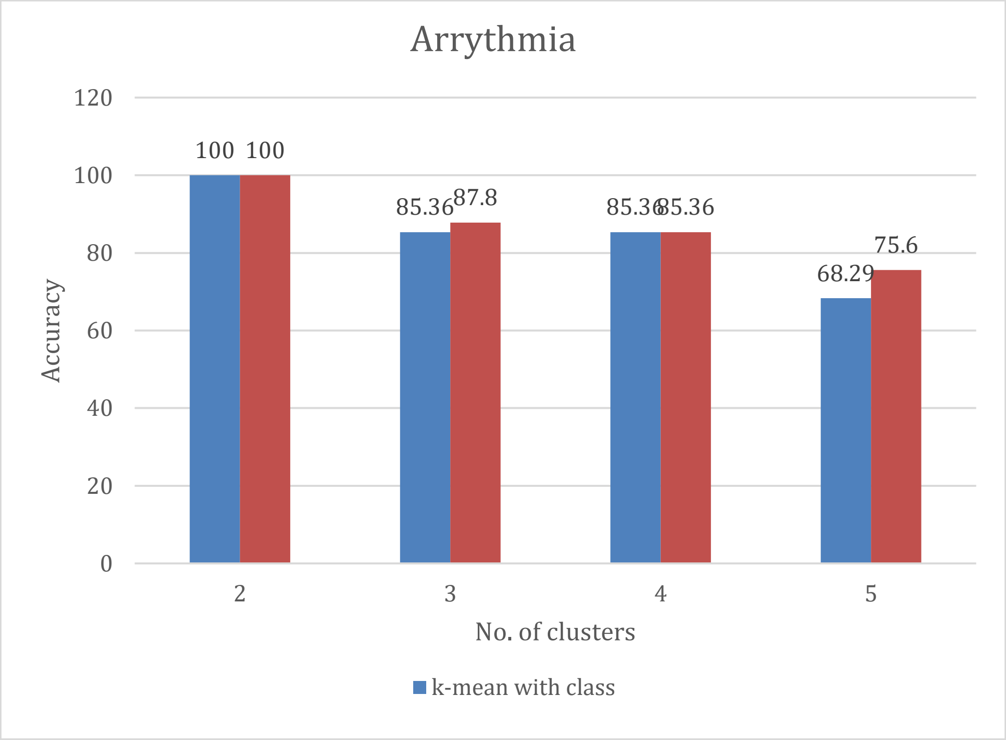

Figure 5: Performance of K-Means on ARRYTHMIA Dataset

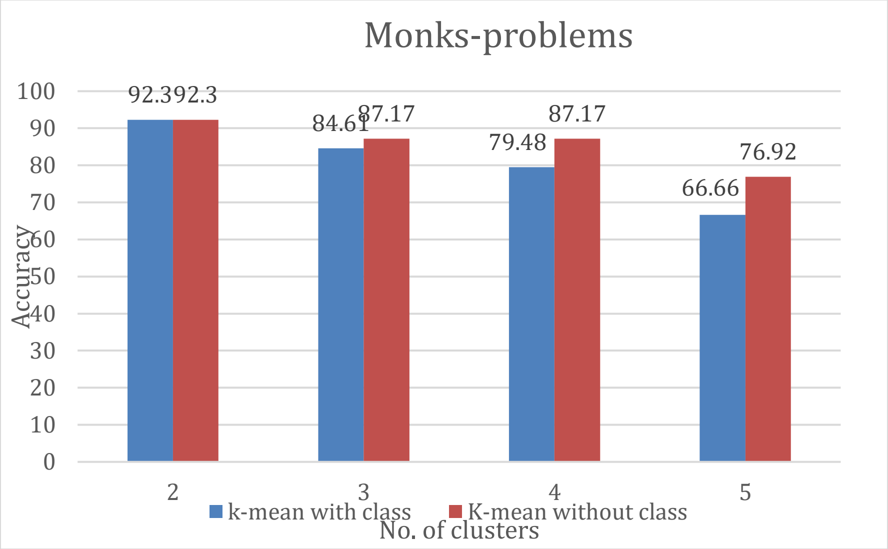

Figure 6: Performance of K-Means on MONKS-PROBLEM Dataset

The **K-Prototype** clustering algorithm showed a similar trend of decreasing accuracy with increasing cluster counts. On the Adultt dataset, the highest accuracy of 99.24% was achieved with two clusters on data without the class attribute, surpassing K-Means results. For three to five clusters, models trained on datasets including the class attribute performed better. The Airlines dataset, comprising 500 instances and seven attributes, attained its peak accuracy of 98.73% at two clusters, equally on datasets with and without class labels. The Credit dataset (690 instances, 15 attributes) achieved maximum accuracy of 96.33% with two clusters on both dataset types; however, at three clusters, models trained with class labels outperformed, while at four and five clusters, models without class labels were superior. The Hepatitis dataset yielded 97.82% accuracy with two clusters on data containing the class attribute. Lastly, the Vowel dataset, a time-series dataset with 640 instances and 12 features, showed maximum accuracy of 96.96% at two and three clusters for data without the class attribute. Models trained with class labels performed better at two and three clusters, while models without class labels excelled at higher cluster counts. These results suggest that the K-Prototype algorithm often outperforms K-Means, especially when class labels are removed, likely due to its suitability for mixed data types.

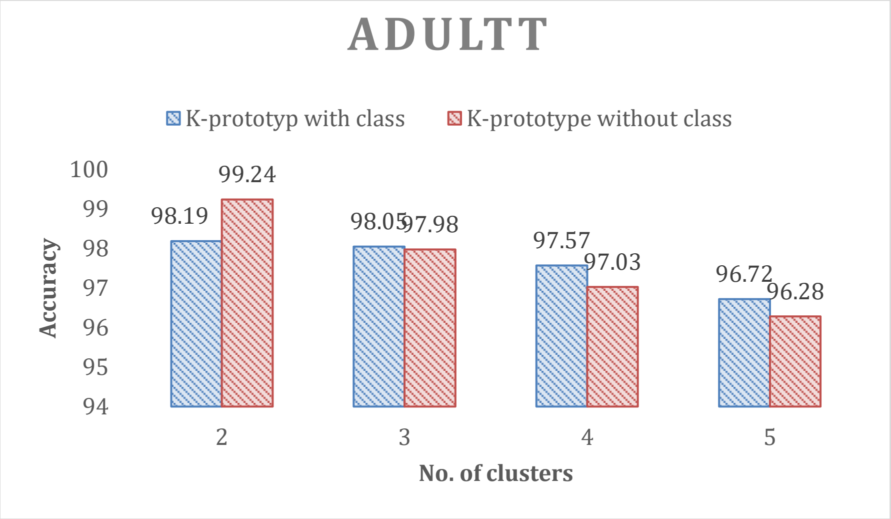

Figure 7: Performance of K-Prototype on ADULTT Dataset

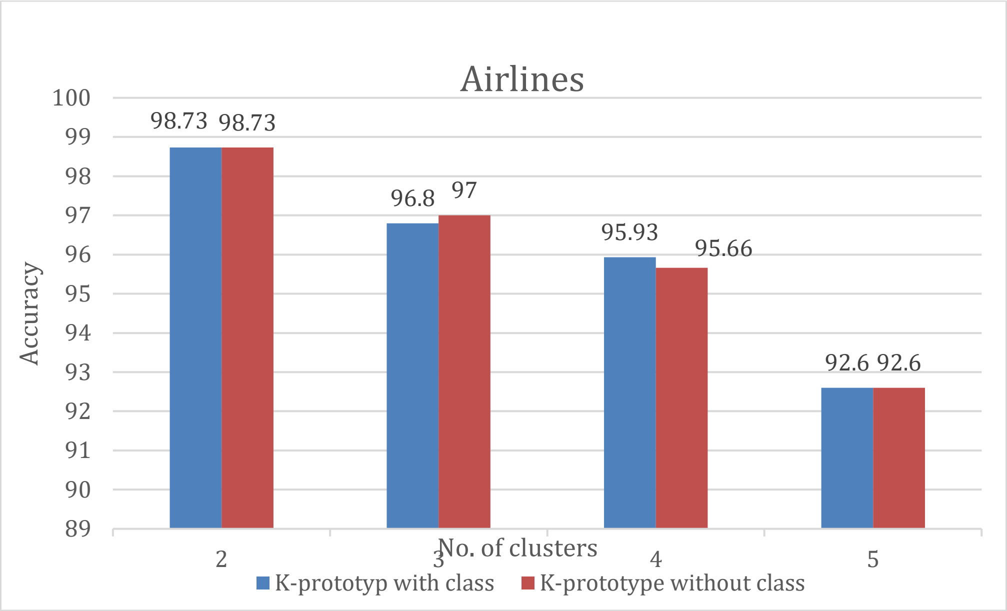

Figure 8: Performance of K-Prototype on AIRLINES Dataset

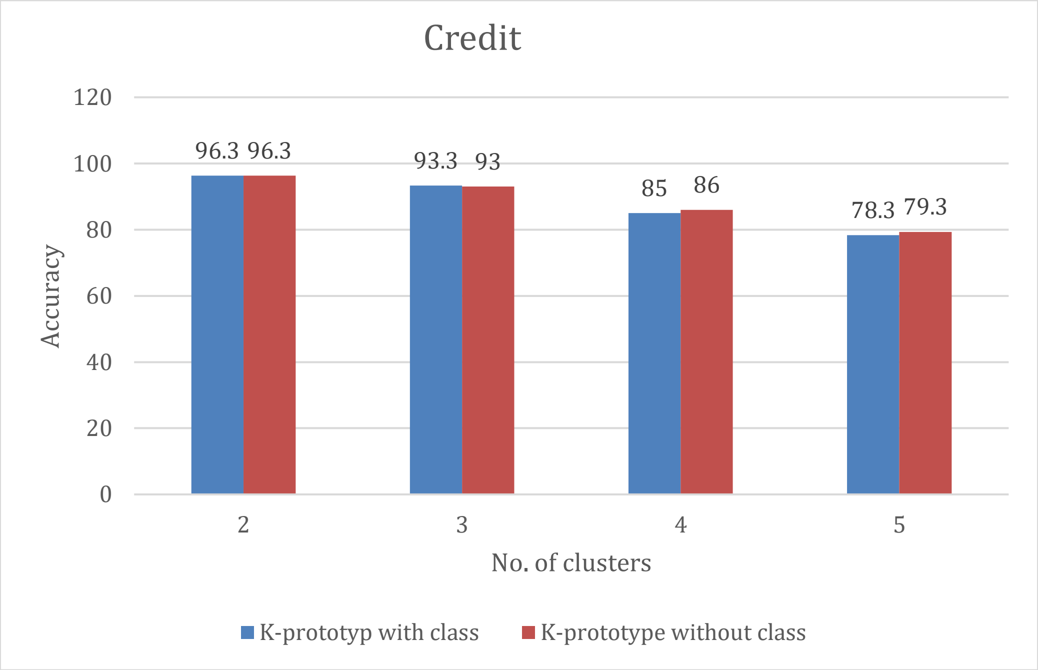

Figure 9: Performance of K-Prototype on CREDIT Dataset

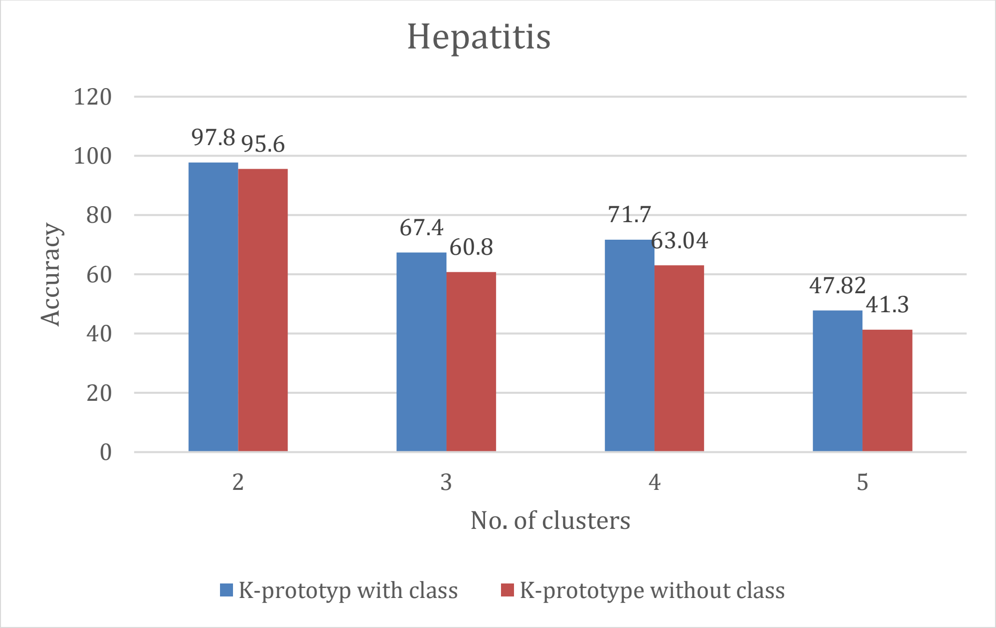

Figure 10: Performance of K-Prototype on HEPATITS Dataset

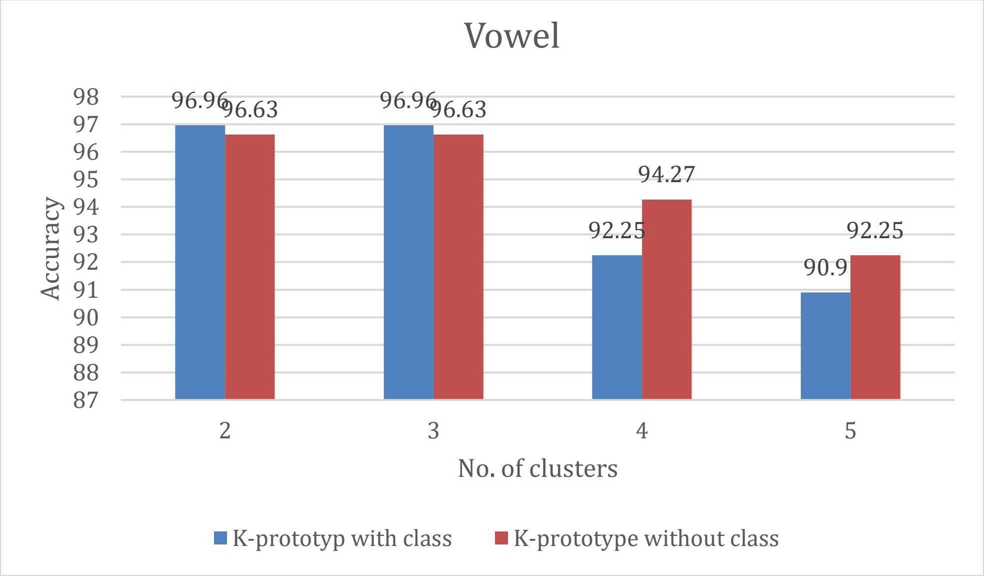

Figure 11: Performance of K-Prototype on VOWEL Dataset

Table 3 summarizes the best accuracy results achieved across all datasets for both algorithms. It is evident that both algorithms perform optimally with two clusters, and that the presence or absence of class attributes influences classifier performance differently depending on the dataset and clustering method. The general decline in accuracy with increasing cluster count may indicate over-segmentation, which negatively impacts cluster homogeneity and classifier performance.

Table 3: Results Evaluation

Dataset	Algorithm	Best Cluster Count	Accuracy with Class (%)	Accuracy without Class (%)	Majority Best
Adultt	K-Means	2	97.57	–	With Class
Hepatitis	K-Means	2	92.85	–	With Class
Primary Tumor	K-Means	2, 3	100	100	Without Class (at 3)
Arrhythmia	K-Means	2	100	100	Without Class
Monks-Problems	K-Means	2	92.3	92.3	Without Class
Adultt	K-Prototype	2	~95.28 (at 5 clusters)	99.24	Without Class
Airlines	K-Prototype	2	98.73	98.73	Equal
Credit	K-Prototype	2	96.33	96.33	Without Class
Hepatitis	K-Prototype	2	97.82	–	With Class
Vowel	K-Prototype	2, 3	~96.96	96.96	Without Class

The difference in performance between K-Means and K-Prototypes can be explained by the types of datasets and the algorithms themselves. K-Means only uses Euclidean distance, which makes it better for datasets with mostly numerical features (like Arrhythmia and Monks-Problems) but not as good for datasets with a mix of types (like Adultt or Credit Approval). K-Prototypes, on the other hand, combines numerical and categorical differences, which makes it better at finding patterns in datasets that include a lot of different types of data. This is why it works better on datasets like Adultt and Airlines, where categorical variables are very important for categorization.

The number of clusters is another thing that affects accuracy. The best results for both algorithms were at two clusters. As the number of clusters increased, the accuracy went down. This drop could be because of over-segmentation, which happens when you break data into smaller groups. This makes clusters less homogeneous and makes it harder for the SVM classifier to map clusters back to ground-truth labels. It is interesting that models trained without the class property occasionally did better than those trained with labels. This implies that in some instances, eliminating the class attribute mitigated bias and enabled clustering algorithms to identify more organic groupings, which were later confirmed using supervised classification.

In general, the comparison results show that K-Means is easier to compute and works well with numerical datasets, whereas K-Prototypes is more versatile and works better with mixed-type datasets in the real world. This supports the idea of using a hybrid framework that uses supervised learning to check the quality of clustering, making sure that unsupervised results are not only statistically sound but also useful for making predictions.

This study mainly looks at K-Means and K-Prototypes, but it is also necessary to include other sophisticated clustering approaches that have become popular in recent years. DBSCAN (Density-Based Spatial Clustering of Applications with Noise) is great at finding noise spots and dealing with clusters of any shape, but it needs careful adjustment of its parameters and has trouble with changing densities. Gaussian Mixture Models (GMM) presume that data comes from a mix of Gaussian distributions. This lets clusters take on different shapes, but GMMs are sensitive to how they are set up and need to know how many components they have. Spectral Clustering uses graph theory to group non-convex structures, but it is quite expensive to do this on big datasets.

K-Means is still fast and useful for purely numerical data, and K-Prototypes makes this efficiency work for mixed-type data as well. This means that both algorithms are good candidates for large-scale real-world datasets.

5. Conclusion and Future work

In this study, we investigated the performance of two prominent unsupervised clustering algorithms, K-Means and K-Prototype, across multiple benchmark datasets. Our experimental framework involved conducting clustering with varying numbers of clusters (from 2 to 5) and evaluating classifier accuracy using the SMO (SVM) classifier on datasets both with and without the class attribute. The results consistently showed that both clustering methods achieve their best performance when the number of clusters is set to two. Increasing the number of clusters usually made the classifier less accurate, which could mean that the clusters were too small and not strong enough to hold together. For K-Means, classifiers trained on datasets containing the class attribute usually did better with fewer clusters, while models trained without class labels usually did better with more clusters. On the other hand, the K-Prototype approach usually gave more accurate results when trained on datasets that didn't have the class property. This shows that it is good at working with heterogeneous data types. In general, the performance trends of both algorithms were similar for cluster counts greater than two.

Our results show how important it is to choose the right number of clusters and the right data set for clustering-based classification to work well. These insights enhance comprehension of the efficient application of unsupervised clustering algorithms across various data sources and classification challenges.

In future research, we intend to expand this study by examining a wider array of clustering algorithms, encompassing hierarchical, density-based, and model-based approaches. Also, using ensemble clustering methods and more advanced feature selection methods could make clustering quality and classification accuracy even better. Another good idea is to look at how different data pretreatment methods and dimensionality reduction methods affect the results of clustering. In the end, these efforts are meant to provide a stronger and more flexible clustering framework that can handle complicated real-world datasets.

Funding Statement: No grants have been received from any funding bodies to conduct this study.

Conflicts of Interest: The authors declare the absence of conflicts of interest.

Data Availability: All datasets exploited in this study are taken from open-source online data repositories.

References

[1] Ahmad, Izhar. "K-mean and K-prototype algorithms performance analysis." *International Journal of Computer and Information Technology* 3, no. 04 (2014): 823-828.

[2] Kamel, Seyed Reza, Reyhaneh YaghoubZadeh, and Maryam Kheirabadi. "Improving the performance of support-vector machine by selecting the best features by Gray Wolf algorithm to increase the accuracy of diagnosis of breast cancer." *Journal of Big Data* 6, no. 1 (2019): 90.

[3] Ranti, Kiefer Stefano, Kelvin Salim, and Abba Suganda Girsang. "Clustering Steam User Behavior Data using K-Prototypes Algorithm." In *Journal of Physics: Conference Series*, vol. 1367, no. 1, p. 012018. IOP Publishing, 2019.

[4] Ali, Huda Hamdan, and Lubna Emad Kadhum. "K-means clustering algorithm applications in data mining and pattern recognition." *International Journal of Science and Research (IJSR)* 6, no. 8 (2017): 1577-1584.

[5] Wen, Zeyi, Bin Li, Ramamohanarao Kotagiri, Jian Chen, Yawen Chen, and Rui Zhang. "Improving efficiency of SVM k-fold cross-validation by alpha seeding." In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 31, no. 1. 2017.

[6] Kim, SungHwan. "Weighted K-means support vector machine for cancer prediction." *Springerplus* 5, no. 1 (2016): 1162.

[7] Shrivastava, Alka, and Ram Ratan Ahirwal. "A SVM and K-means clustering based fast and efficient intrusion detection system." *International Journal of Computer Applications* 72, no. 6 (2013).

[8] Santhanam, T., and M. S. Padmavathi. "Application of K-means and genetic algorithms for dimension reduction by integrating SVM for diabetes diagnosis." *Procedia Computer Science* 47 (2015): 76-83.

[9] Yao, Yukai, Yang Liu, Yongqing Yu, Hong Xu, Weiming Lv, Zhao Li, and Xiaoyun Chen. "K-SVM: An Effective SVM Algorithm Based on K-means Clustering." *J. Comput.* 8, no. 10 (2013): 2632-2639.

[10] Singh, S. P., and Asmita Yadav. "Study of k-means and enhanced k-means clustering algorithm." *International Journal of Advanced Research in Computer Science* 4, no. 10 (2013): 103-107.

[11] Li, Youguo, and Haiyan Wu. "A clustering method based on K-means algorithm." *Physics Procedia* 25 (2012): 1104-1109.

[12] Chakraborty, Sanjay, and Naresh Kumar Nagwani. "Analysis and study of incremental k-means clustering algorithm." In *International Conference on High Performance Architecture and Grid Computing*, pp. 338-341. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011.

[13] Oyelade, Olanrewaju Jelili, Olufunke O. Oladipupo, and Ibidun Christiana Obagbuwa. "Application of k Means Clustering algorithm for prediction of Students Academic Performance." *arXiv preprint arXiv:1002.2425* (2010).

[14] 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.

[15] Eitrich, Tatjana, and Bruno Lang. "Efficient optimization of support vector machine learning parameters for unbalanced datasets." *Journal of computational and applied mathematics* 196, no. 2 (2006): 425-436.

[16] Burges, Christopher J., and Bernhard Schölkopf. "Improving the accuracy and speed of support vector machines." *Advances in neural information processing systems* 9 (1996).

[17] Kuswardana, Dendy Arizki, Dwi Arman Prasetya, Trimono Trimono, and I. Gede Susrama Mas Diyasa. "Comparison of Elbow and Silhouette Methods in Optimizing K-Prototype Clustering for Customer Transactions." *Jurnal Ilmiah Edutic: Pendidikan dan Informatika* 12, no. 1 (2025): 43-48.

[18] Aschenbruck, Rabea, Gero Szepannek, and Adalbert FX Wilhelm. "Initialization strategies for clustering mixed-type data with the k-prototypes algorithm: R. Aschenbruck et al." *Advances in Data Analysis and Classification* (2025): 1-30.

[19] Ping, Yuan, Huina Li, Chun Guo, and Bin Hao. "k ProtoClust: Towards Adaptive k-Prototype Clustering without Known k." *Computers, Materials & Continua* 82, no. 3 (2025).

[20] Alrasheed, Mousa, and Monjur Mourshed. "Building stock modelling using k-prototype: A framework for representative archetype development." *Energy and Buildings* 311 (2024): 114111.

[21] Mohd, Azimah, Lay Eng Teoh, and Hooi Ling Khoo. "Passengers' requests clustering with k-prototype algorithm for the first-mile and last-mile (FMLM) shared-ride taxi service." *Multimodal Transportation* 3, no. 2 (2024): 100132.

[22] Shi, Yan, Siyuan Zhang, Siwen Wang, Hui Xie, and Jianying Feng. "Multiple-perspective consumer segmentation using improved weighted Fuzzy k-prototypes clustering and swarm intelligence algorithm for fresh apricot market." *Italian Journal of Food Science* 36, no. 4 (2024): 38.

[23] Jing, Xin, and Hao Gao. "An Improved K-PROTOTYPE Clustering Algorithm and Its Application." In *Proceedings of the 2023 6th International Conference on Machine Learning and Natural Language Processing*, pp. 182-186. 2023.