Cihan University-Erbil Scientific Journal (CUESJ)

Research Article

The Use of Clustering and Classification Methods in Machine Learning and Comparison of Some Algorithms of the Methods

Guhdar A. A. Mulla[1], Yıldırım Demir[2]

[1]Department of Economic, Faculty of Economics and Administrations, Nawroz University, Kurdistan Region, Iraq, [2]Department of Statistics, Faculty of Economics and Administrative Sciences, Van Yuzuncu Yil University, Van, Turkey

ABSTRACT

In this article, two machine learning methods such as classification and clustering are used for decision tree (DT), artificial neural network (ANN), and K-nearest neighbors algorithms. The datasets were used to evaluate the effectiveness of the clustering method and the data mining tool. Weather data were used to compare algorithms and methods in the study. This study showed that the best model was DT according to accuracy and precision measures but the best model according to F-measure and receiver operating characteristic curve area measures was ANN. Waikato Environment for Knowledge Analysis, a data mining tool, is utilized in this paper to carry out the clustering.

Keywords: Algorithms, classification, clustering, machine learning, decisions tree

INTRODUCTION

Data analysis science is one of the disciplines that was created to investigate natural phenomena and address current or upcoming issues. Data analysis science aims to shed light on future studies by employing techniques and theories that are appropriate for data structures with a small number of observations to describe the subject with a certain probability. Every science discipline in nature has a lot of data, and access to these data is becoming simpler every day. It always raises the question of how much of the collected data can be used or how significant it is. While going about their regular lives, humans not only receive information but also consume it.

Due to the extensive use of data, classification was required to make the data meaningful and produce new data. A set of data is classified when it is categorized based on algorithm-determined distinguishing characteristics. Data cannot always be categorized manually because millions of data types related to a field are created in just 1 day. The extremely convenient operation of the algorithms makes the classification of data very easy. The classification of data using many algorithms simultaneously can result in a variety of outcomes. Which algorithm is used to classify the data set effectively relies on the data set. Since not all algorithms will give the same accuracy to every data set, it is possible that the techniques employed to describe the data will also be erroneous for that data set. Machine learning and classification algorithms are prominent in this direction.[1]

Data analysis was done using the open-source Waikato Environment for Knowledge Analysis (WEKA) program, which is a tried-and-true data mining tool with a graphical user interface and regular terminal programs. This program has helped to classify the data by rebalancing and decreasing its dimension. The WEKA project seeks to offer a full range of machine learning tools and preprocessing methods to academics. This allows applications to rapidly test and contrast various machine learning process strategies on new datasets.[2]

LITERATURE REVIEW

In Soofi and Awan,[3] discussed the fundamental classification strategies in this paper. Later, he covered some of the main classification technique subcategories, including Bayesian networks, decision tree (DT), K-nearest neighbor (KNN), and support vector machine (SVM), along with their advantages, disadvantages, prospective applications, problems, and potential solutions. This study's objective is to offer a thorough evaluation of several machine learning classification methods.

Corresponding Author:
Guhdar A. Ahmed,
Department of Economic, Faculty of Economics and
Administrations, Nawroz University, Kurdistan Region, Iraq.
E-mail: guhdar.abdulaziz@gmail.com

Received: March 3, 2023
Accepted: May 19, 2023
Published: June 10, 2023

DOI: 10.24086/cuesj.v7n1y2023.pp52-59

Both experts and newbies to the field of machine learning will benefit from this research's efforts to strengthen the conceptual underpinnings of classification techniques.

In this study Sharma,[4] the many kinds of existing categorization algorithms were thoroughly explained in this work. Sharma will mostly compare and discuss in-depth the key 7 categories of categorization algorithms here. The comparison will mostly be focused on each system's assumptions, benefits, and disadvantages. Sharma will address some of the most popular classification algorithms used, different scoring schemes to gauge their effectiveness, multiclass classification strategies, and finally model selection techniques in this paper.

The study Sisodia and Sisodia,[5] have covered the prediction of diabetes using Nave Bayes, DTs, and SVMs as classification algorithms. The major goal of this study is to create a model that can accurately predict the likelihood of developing diabetes. WEKA Tool was utilized for data classification, with the Pima Indian Diabetes dataset serving as the primary dataset.

This research work by Ambigavathi and Sridharan,[6] the purpose of this work was to examine several clustering techniques from both theoretical and experimental angles. Trial findings revealed which category had the best algorithm using a set of physiological data. A variety of internal as well as stability measures were used to validate the effectiveness of each clustering technique in machine learning. This research concluded by highlighting future directions for high-dimensional health-care data sets using a suitable clustering technique.

In this study Lorena *et al.*,[7] In this research, we investigated resampling techniques and metrics that may be obtained from training datasets to characterize the complexity of the classification issues. Recent literature also looked at and addressed their approaches, opening up the possibility of new areas for study. The Extended Complexity Library R package, which includes a number of complex measurements and is finished and publicly accessible, was also defined.

This research work by Liu *et al.*,[8] centered on addressing issues with the class imbalance and identifying attribute correlations. Kaggle offers two datasets for fraud detection that may be used to create classifiers and assess the effects of various data processing methods. Through their approach, they were aware of the latest fraud detection discoveries, learned more about various data processing techniques, and created several sorts of classifiers. They supported the need of addressing class imbalance and examining attribute connections. Finally, they looked at the relationships between each attribute and marginally better performance. Their experimental findings clearly demonstrated that addressing class imbalance had a beneficial effect on unbalanced data sets and that examining the link between qualities would also be helpful.

In Zheng,[9] investigation on the application of Synthetic Minority Oversampling Technique (SMOTE) with various classifiers. To build predictive models on the heart disease data, the SMOTE method was used for comparison. Regular SMOTE, borderline-SMOTE, SVM-SMOTE, and K-means SMOTE were combined with four classification algorithms under the classical and Neyman Pearson (NP) paradigms. The performance of these models was compared. According to their findings, the SVM-SMOTE and Borderline-SMOTE perform better than other SMOTE variations, and the NP classification is better at successfully controlling type I error.

MATERIALS AND METHODS

Weather Forecasting: The Machine Learning Approach

Before the development of models, people made predictions for a given location based on observable data collected over time. For example, they used a basic approach known as climatology to predict the weather. Rough estimations of long-term trends and cumulative values of variables such as temperature or precipitation may be generated by applying basic statistics to recorded weather events over long enough periods. With the development of tools to accurately monitor the atmosphere and disseminate these readings across geographically far places, weather maps became available around the start of the 20th century. With the advent of weather forecasting techniques that anticipated the movement and effects of these pressure systems in the atmosphere, these maps first started to show the position and shape of low- and high-pressure systems. The use of statistical models in weather forecasting was first advocated in the 1950s.[10] Malone claimed that "statistics must ultimately play some part" in atmospheric simulation at the same time that the first numerical weather prediction (NWP) models were being built. The author developed a method for forecasting the sea-level pressure field using multiple linear regression.

Despite the optimism Malone expressed in the early 1950s, machine learning and statistical methods have actually been applied as a complement to NWP rather than as a replacement for it. Other approaches have not yet been able to match the power of physical equations to simulate the evolution of the atmosphere along the spatial and temporal dimensions. We discuss examples of machine learning approaches being used to address various issues in the field of weather forecasting in this section. The following categories of challenges, in which various machine learning approaches have been effectively employed to enhance our understanding of the processes in the atmosphere, can be identified depending on the task to be solved and the nature of the underlying data.

Kinds of Machine Learning

Supervised, unsupervised, and reinforcement learning methods were used to classify machine learning algorithms: Figure 1 depicts the categorization in a visual manner.[11]

Supervised Learning

For machine learning, supervised learning is crucial. Averaging input and output is the aim of supervised learning. The input data are a list of many interesting items, each of which is referred to as an attribute or an example. The output is a manager's assessment or conclusion. In classification, a form of supervised learning, averaging (or a discriminant function) is used to distinguish between several groups of occurrences. A list of the different classes is included in the output,

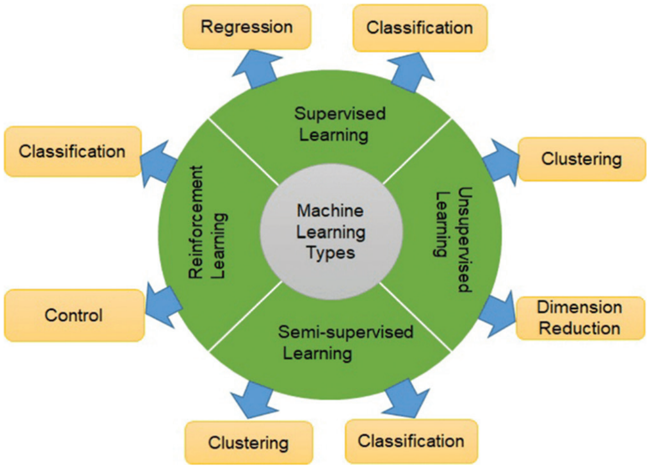

Figure 1: Overview of machine learning techniques

sometimes referred to as the class label in machine learning. The term used to describe the discriminant function is classifier or model. A training set is a set of instances for which the class label has been determined. During classification, a model is developed using a set of parameters to create a mapping between instances in the training set and labels in the training set. To identify or recognize previously undiscovered scenarios, the trained model may be used. Supervised learning is the main methodology used in real-world machine learning. In supervised learning, the function that converts input variables (X) to output variables are learned using an algorithm (Y). Y = f(X) the objective is to anticipate the output variables (Y) using new input data by as nearly approximating the mapping function as is practical (X). Since an algorithm learns from the training dataset is akin to a teacher guiding the learning process, it is known as supervised learning. When the algorithm operates satisfactorily, learning comes to an end. The algorithm iteratively makes predictions on the training data and is corrected by the teacher as we are aware of the appropriate responses. Learning comes to a conclusion when the algorithm works satisfactorily.[12]

Unsupervised Learning

It is a machine learning technique used when training data is not classified or labeled. To describe a hidden structure, unsupervised learning examines systems' capacity to infer a function from unlabeled data. The system examines the data and may utilize datasets to infer hidden structures from unlabeled data even when it is unable to select the ideal output. Unsupervised learning occurs when there are just input variables and no matching output variables (X). Unsupervised learning simulates the underlying distribution or structure of data to understand it better. These are referred to as unsupervised learning as there are no correct answers and no teachers, in contrast to the supervised learning that was previously discussed. It is up to the algorithms to find and display the fascinating structure of the data. Issues with grouping and association abound in unsupervised learning. In contrast to association rule learning problems, where you are looking for rules that adequately describe significant portions of your data, such as people who buy X also tend to buy Y, clustering problems are looking for the inherent groupings in the data, like grouping customers based on their purchasing behavior. Unsupervised learning techniques include the a priori methodology for learning association rules and K-means for classifying problems.[12]

What Kinds of Data Can Be Mined?

Any sort of data may be used in data mining, as long as it is pertinent to the application for which it is intended. The most basic sorts of data for mining applications are databases. It is also possible to mine other forms of data, including data streams, ordered/sequence, graph or networked, geographical, text, and multimedia data. Data mining will undoubtedly continue to include new data kinds as they emerge.

Machine Learning Models In Classification Method

When semantic classes for particular items are already known, classification refers to the process of determining those classes based on those objects' properties. It involves putting fresh observed samples into a specific class by comparing the sample's characteristics to an already existing class. By building a model using earlier data, it may then decide on the unobserved cases. Therefore, it finds a model (or function) that clarifies and divides groups of data or ideas. It looks at the relationship between the values of the variables for each row and the label assigned to that row using data mining techniques. Various classification methods exist, including Naïve Bayes (NB), logistic regression, SVM, KNN, DT, and artificial neural network (ANN). As a result, the extracted data may be used to predict a label for a new row when it is supplied to the classifier based on the variable values that define it. These methods use different ways to express these interactions. Because these correlations change between datasets, it is critical that the classifiers be trained using labeled training data. Classification is one of the findings and predicting process a result from a set of data. The algorithm uses a collection of variables and a training set with the property of each variable to predict the outcome. The outcome is often referred to as a target or forecast variable.[13] This article discussed DT, ANN, KNN.

DT

It is known that one of the models used in machine learning is DTs.[14] A non-parametric, intricate, and computationally costly sorting technique is the DT approach, often known as the recursive partitioning algorithm. The main idea is to keep subdividing the subsample into subgroups until it allows for decision-making, beginning with dividing the sample responses into fresh subsamples that are as similar to one another and distinct from them as practicable. While the nodes reflect the sub-samples, the root node represents the total sample.

Figure 2 shows a DT that categorizes weather conditions for those as good and bad.

The DT models are built employing the greedy method, and the best split is that which minimizes the weighted drop in impurity after evaluating a large number of variable splits at each node:

$$\Delta_i(s,t) = i(t) - p_L i(t_L) - p_R i(t_R) \tag{1}$$

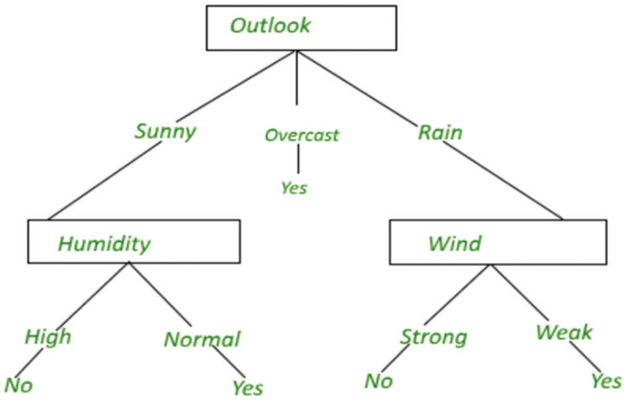

Figure 2: Diagram showing decision tree

Where P_L and P_R indicate the percentage of observations connected to the node ***t*** are sent to the left child node t_L or right t_R he relevant chide node.

KNN

A machine learning method for categorizing data is the nearest-neighbors approach, also known as pattern recognition.[15] This classifier's objective is to select a metric to compare the separation between any two applicants over the whole set of application data.[16,17] Proposed using the following metric:

$$d(x,y) = \{(x-y)^T \ (I+Dww^T \ (x-y)\}^{\frac{1}{2}} \tag{2}$$

Where X and Y are in the feature space points, I, D is a distance between parameters, w is a specific direction in the measurement space and is the identity matrix.

ANN

ANN are mathematical representations of the way the human brain works.[18] With neural networks, practically any non-linear relationship between input and objective variables may be simulated thanks to their adaptability. Despite the numerous proposed architectures, the multilayer perceptron is the one that this article concentrates on (MLP). Neurons for each input variable are located in the hidden layer of an MLP, which typically comprises three layers (in our case, one neuron). Each neuron processes the information it receives and transmits the results to the neurons in the layer beneath it. Each of these neural connections receives a weight during training. Using the weighted inputs and its bias term as an example, the logistic function, $b_i^{(1)}$ results in the computation of the hidden neuron's output:

$$h = f^{(1)}\left(b_i^{(1)} \sum_{j=i}^{n} W_{ij} X_I\right) \tag{3}$$

W_{ij} stands for the weight connecting input where W is a weight matrix. The weight *j* represents the input connection to hidden neuron *i*.

Machine Learning Models in Clustering Method

Clustering is one of machine learning's unsupervised learning strategies.[19] When no class labels are available to analyze the datasets, the clustering strategy is critical for breaking down the massive volume of data into smaller groupings of data. Each data point is classified and grouped into a distinct cluster using the clustering approach, and each cluster comprises a collection of data points. Moreover, data points within the same cluster should have comparable attributes, whereas those inside a separate cluster should have significantly different features.[20] Partitioning around medoids, also referred to as K-medoids and hierarchical K-means. The main operations of these algorithms are separated into the following groups.

K-means Clustering Algorithm

The primary function of the simple and frequently used clustering method K-means is to classify the provided unlabeled dataset. The main goal of this method is to identify clusters that are comparable to each other, as indicated by the variable k. In this approach, the cluster is described statistically using the mean or centroid. Not usually part of the collection, a centroid is a data point that symbolizes the center of a cluster. After that, the k clusters are constructed by grouping the n data points into them, with each data point joining the cluster with the closest centroid. The Euclidean distance between each data point n and the cluster centroid is then calculated precisely. A cluster's data points are always assigned to the place with the shortest Euclidean distance from the centroid point. When there are no available data points to assign, an early grouping is considered. The technique is then repeated until the "c" centroids stop migrating and the new "c" centroids are identified.[21]

Hierarchical Clustering

Hierarchical, also known as hierarchical cluster analysis, is a particular kind of unsupervised. With the use of a tree-based framework, hierarchical cluster analysis aims to merge multiple clusters made up of comparable unlabeled data points. The clusters of data points formed by the last branch of the tree are not alike. In addition, the majority of the data points inside a particular cluster coincide with those within other clusters in the data set. This approach uses a diagram, which is a diagram that resembles a tree and is based on the hierarchy. Agglomerative hierarchical clustering, also known as AGNES (Agglomerative Nesting), and divisive hierarchical clustering, also known as DIANA, are two categories of hierarchical clustering techniques (divisive analysis). This algorithm is an identical duplicate of itself. Table 1 compiles several strategies based on various criteria.[21]

Machine Learning Application Software

Using the software is a collection of one or more apps created for end users. Applications can be found in web browsers, email clients, word processors, spreadsheets, accounting software, music players, file viewers, simulators, console games, and picture editors. Machine learning software designed by professionals is generally available. The following are a few of the few notable pieces of software:

To conduct data mining tasks, the program employs a variety of machine learning algorithms. Manually apply the algorithms to a dataset or simply invoke them in your

Table 1: Summary of clustering algorithms

Algorithm	Size of dataset	Type of data	Complexity	Computation speed	Modifications corrections	Cluster shape	Result interpretation
K-means	Large	Numerical	O (nkd)	Fast	Flexible	None convex	Easy
K-medoids	Small	Categorical	O (n^2 dt)	Moderate	Difficult	None convex	Difficult
Hierarchical	Large	Numerical	O (n)	Slow	Flexible	None convex	Easy

Java code. Data pre-processing, classification, regression, clustering, association rules, and visualization are among WEKA's capabilities. At the University of Waikato in New Zealand, WEKA was upgraded. This tool is an addition to the GNU General Public License-licensed free program "Data Mining: Practical Machine Learning Tools and Methods".[22]

The Measure That Used in WEKA

Confusion matrix

A representative representation of the above-mentioned parameters in matrix form is given in Table 2, considering that advanced visualizations would be helpful

Table 2: Confusion matrix

	Predicted	
True Class	True positive	False negative
	False positive	True negative

Accuracy

To evaluate each created subset in this measurement, classification data mining methods have been offered as a fitness function.

The formulation of measuring the accuracy of each subset is described as follows:

$$\text{Accuracy}=\frac{TP+TN}{TP+TN+FP+FN}*100$$

False positive (FP) reflects the improper placement of the negative example in the positive class, whereas true negative (TN) represents the proper placement of the negative example. True positive (TP) represents the proper classifications of the positive example. False negative (FN) is the term used to describe a positive sample that was mistakenly placed in the negative category.

Precision

The precision is calculated with the following equation and the denominator consists of the total predicted positive.

$$\text{Precision}=\frac{TP}{TP+FP}$$

It is immediately apparent that precision relates to how precise/accurate your model is in terms of the proportion of successful predictions. Precision is a helpful statistic to judge whether there are high costs associated with false positives. As an example, email spam detection. When an email that is not spam (actual negative) is wrongly classified as spam, it is known as a false positive in email spam detection (predicted spam). The email user may overlook important emails if the spam detection model's precision is poor.

F-measure

In statistical analysis of binary classification, the F-score or F-measure is a measure of a test's accuracy. Even if you read a lot of other literature on precision and recall, you cannot ignore the other measure, F-measure, which is a function of those two variables. The formula is as follows:

$$\text{F-measure}=2*\frac{\text{Precision}*\text{Recall}}{\text{Precision}+\text{Recall}}$$

F-measure is necessary when seeking to balance precision and recall. What distinguishes F-measure and accuracy from one another then? If we need to find a balance between Precision and Recall and there is an uneven class distribution, F-measure would be a better statistic to utilize because False Negatives and False Positives often have business expenses (physical and intangible). As we have previously seen, a huge number of True Negatives, which are typically overlooked in corporate settings, can significantly contribute to accuracy (a large number of Actual Negatives).

Receiver operating characteristic (ROC) area

A graph that shows how well a classification model works at each classification threshold is called a receiver operating characteristic curve, or ROC area. On this curve, two parameters are plotted:

- True positive rate (TPR)
- False positive rate (FPR)

TPR is a synonym for recall and is therefore defined as follows:

$$\text{TPR}=\frac{TP}{TP+FN}$$

FPR is defined as follows:

$$\text{FPR}=\frac{FP}{FP+FN}$$

In a ROC curve, TPR vs. FPR for various classification criteria are displayed. As the classification threshold is lowered, more items are categorized as positive, which increases the quantity of both false positives and true positives.

Data

We used a dataset about the predicting whether it will rain or not using some weather conditions. And we obtained this dataset from the Kaggle side it was shown in this link. https://www.kaggle.com/datasets/ananthr1/weather-prediction.

About dataset:
Using coloms:

- Precipitation
- Tempreture min
- Tempreture max
- Wind

We will forecast the weather as follows:

- Drizzle
- Rain
- Snow
- Sun
- Fog

RESULTS AND DISCUSSION

Three distinct categorization methods were employed for the data being studied during this study. All three algorithms DT, ANN, and KNN were used in the classification process. We utilized k=1 and divided the dataset into 10 cross-validations for the KNN algorithms. The quantity of data used for reduced-error pruning when we used DT methods was defined according to the trimming confidence factor, which was 0.25. All three classifications and clustering algorithms were employed, and the training rate was 66% and the testing rate was 34%. We separated into five groups for k-mean methods, used 10 seeds, and decided to use false for debug and false for display stander division.

The effectiveness of the three classification methods in Table 3 was evaluated using the Accuracy Precision, F-measure, and ROC region metrics.

As shown in Table 3 for the classification method, the highest performance depending on accuracy and precision was 84.8, 78.7 obtained in DT, but the higher performance according to the F-Measure and ROC Area was obtained in ANN. On the other hand, the best model for DT, ANN, and KNN algorithm in this study was obtained in ROC area. Figure 3 shows one of the classification methods, so it can be decided based on Figure 3. There are 5 decisions from Figure 3.

The DT region measurement was measured for testing the efficiency of the DT algorithm as shown in Figure 3.

- The decision 1: Precipitation to sun.
- The decision 2: Precipitation to temperature min to rain.
- The decision 3: Precipitation to temperature min to wind to snow.
- The decision 4: Precipitation to temperature min to wind to temperature max to rain.
- The decision 5: Precipitation to temperature min to wind to temperature max to snow.

The k-means region measurement was measured for testing the efficiency of the clustering algorithm as shown in Table 4.

While we study the output report, we can see 5 classes (class 0, class 1, class 2, class 3, class 4). In the report appear as Number of clusters selected by cross-validation: 5

Table 3: The performance of the classification algorithms using for weather data

Summary	Accuracy	Precision	F-measure	ROC area
DT	84.8	78.7	80.7	89.5
ANN	82.5	74.4	84.9	91.7
KNN	67.3	67.8	67.6	74.5

DT: Decision tree, ANN: Artificial neural network, KNN: K-nearest neighbor

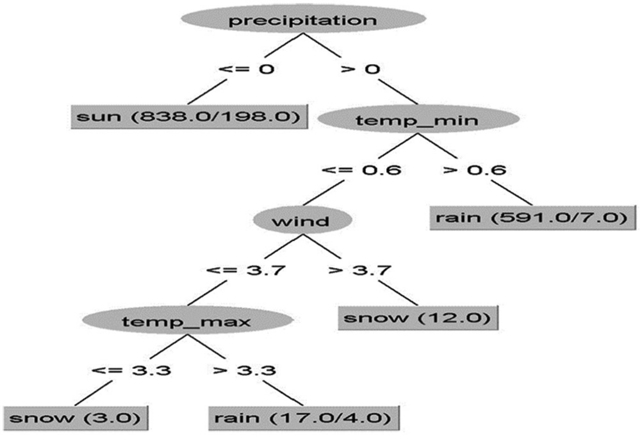

Figure 3: Decision tree visualization

You can find the centroid of the entire population in the first column. The centroids for clusters 0, 1, 2, 3, and 4 may be found, respectively, in the second, third, fourth, and fifth columns. The centroid coordinate for each row's respective dimension is provided. It is important to note that this step of the process is locating the centroids. The centroids are not unique and are the output of a particular run of the algorithm; a different run could produce a different set of centroids. There is a "clusters prior" probability for each cluster. In the estimators, each conceivable attribute value is represented by a number, which is treated sequentially. classes when we used K-mean algorithms:

- Class 0 has total 232 things, out of which majority of things (16%) data sets.
- Class 1 has total of 449 things, out of which majority of things (31%) data sets.
- Class has total 355 things, out of which majority of things (24%) data sets.
- Class 3 has total 258 things, out of which majority of things (18%) data sets.
- Class 4 has total 167 things, out of which majority of things (11%) data sets.

Classes when we used hierarchical algorithms:

- Class 0 has total 53 things, out of which majority of things (4%) data sets.
- Class 1 has total of 641 things, out of which majority of things (44%) data sets.
- Class has total 640 things, out of which majority of things (44%) data sets.
- Class 3 has total 26 things, out of which majority of things (2%) data sets.
- Class 4 has total 101 things, out of which majority of things (7%) data sets.

Table 4: Clustering method performance

Attribute	Full data (1461)	0 (232)	1 (449)	2 (355)	3 (258)	4 (167)
Precipitation	3.02	4.06	7.61	0.00	0.00	0.36
Temp-max	16.40	18.90	10.60	26.00	16.80	7.52
Temp-min	8.23	11.70	5.34	13.70	7.63	0.25
Wind	3.24	2.89	4.06	2.85	3.06	2.60
Weather	rain	rain	rain	sun	sun	sun

Table 5: Comparison result of clustering algorithms with weather dataset using WEKA tool

Hierarchical	K-mean	Name
5	5	No of clusters
1461	1461	Cluster instances
-	18	No of iterations
-	1707.637908	Root mean square errors
9.9 s	0.02 s	Time is taken to build model
0	0	Unclustered instances

WEKA: Waikato Environment for Knowledge Analysis

We examined two clustering techniques from Table 5 and separated both of them into five clusters in order to identify some distinct things. At the time, k-mean was preferred over hierarchical methods. fewer iterations in the hierarchy than the K-mean. On the other hand, when using Hierarchical, we had zero Root Mean Square Errors (RMSE), K-mean algorithms experienced a 1707.63 (RMSE).

CONCLUSION

The application of data mining to enrollment management is a relatively recent innovation. At the moment, category, and straightforward numerical data are most frequently employed in data mining. Data mining will increasingly incorporate complicated data types in the future. By taking into account more variables and how they interact with one another, any model that has been created may also be reinforced. New techniques for discovering the most exciting features, Data mining research will lead to the development of the data. As models are developed and implemented, they can be used as a tool for enrollment management. This study showed that the best model was DT according to accuracy and precision measures but the best model according to F-Measure and ROC Area measures was ANN. WEKA, a data mining tool, is utilized in this paper to carry out the clustering. The K-mean approach can cluster large data sets efficiently, and as the number of clusters rises, so does its efficiency. The K-mean method outperforms the Hierarchical Clustering Algorithm.

The performance of K-means method increases as the root mean square errors (RMSE) drops and the RMSE falls as the number of clusters. All the techniques have some noise or ambiguity in certain data when clustered. Using massive datasets significantly raises the quality of all algorithms. The K-means algorithm is quite susceptible to dataset noise. This noise interferes with the algorithm's ability to cluster data into the proper clusters and affects the algorithm's output. When working with huge datasets, the K-means method produces good clusters more quickly than alternative clustering techniques. For noisy data, the hierarchical clustering technique is more sensitive. WEKA only enables sequential single-node execution, which leads to its inability to handle very large datasets. As a result, there is little similarity between clusters when data are grouped or clustered, and significant similarity within clusters. One of the clustering methods is the K-means algorithm, which gathers data based on their traits and qualities and runs the clustering process by shortening the distances between the data centers.[5] The finally used DT, NB and SVM to make design of model in PIMA dataset but here we used DT, ANN, KNN in weather dataset.

REFERENCES

1. M. Baran. *Makine Öğrenmesi Yöntemleriyle Çoklu Etiketli Verilerin Sınıflandırılması*. (Sivas Cumhuriyet Üniversitesi, Sosya Bilimler Enstitüsü, Turkey, 2020.
2. M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann and I. Wietten. The WEKA data mining software: An update. *ACM SIGKDD Explorations Newsletter*, vol. 11, pp. 10–18, 2009.
3. A. A. Soofi and A. Awan. Classification techniques in machine learning: Applications and issues. *Journal of Basic and Applied Sciences*, vol. 13, pp. 459–465, 2017.
4. V. Sharma. Survey of classification algorithms and various model selection methods. *Journal of Machine Learning Research*, vol. 1, pp. 1–48, 2000.
5. D. Sisodia and D. S. Sisodia. Prediction of diabetes using classification algorithms. *Procedia Computer Science*, vol. 132, pp. 1578–1585, 2018.
6. M. Ambigavathi and D. Sridharan. Analysis of clustering algorithms in machine learning for healthcare data. In: A*dvances in Computing and Data Sciences*. Vol. 1244. Springer, Berlin, pp. 117–128.
7. A. C. Lorena, L. P. F. Garcia, J. Lehmann, M. C. P. Souto and T. K. Ho. How complex is your classification problem? A survey on measuring classification complexity. *ACM Computing Surveys*, vol. 52, pp. 1–34, 2018.
8. D. Liu, R. Sun and H. Ren. Efficient fraud detection classification: Class imbalanceand attribute correlations. *The Frontiers of Society, Science and Technology*, vol. 2, pp. 96–103, 2020.
9. X. Zheng. SMOTE Variants for Imbalanced Binary Classification: Heart Disease Prediction. University of California, California, 2020.
10. T. F. Malone. Application of statistical methods in weather prediction. *Proceedings of the National Academy of Sciences*, vol. 41, pp. 806–815, 1955.
11. L. Liu and J. L. Priestley. A comparison of machine learning algorithms for prediction of past due service in commercial credit. In: *Grey Literature from PhD Candidates.* DigitalCommons Kennesaw State University, Georgia, 2018.
12. G. Nakhaeizadeh and C. C. Taylor. *Machine Learning and Statistics : The Interface*. Wiley, United States, 1997.

13. G. A. A. Mulla, Y. Demir and M. M. Hassan. Combination of PCA with SMOTE oversampling for classification of high-dimensional imbalanced data. *BEU Journal of Science*, vol. 10, pp. 858–869, 2021.
14. M. A. Habara. Credit Risk Modelling in a Developing Economy: The Case of Libya. Griffith University, Australia, 2009.
15. L. Saitta and F. Neri. Learning in the “real world”. *Machine Learning*, vol. 30, pp. 133–163, 1998.
16. L. C. Thomas. A survey of credit and behavioural scoring: Forecasting financial risk of lending to consumers. *International Journal of Forecasting*, vol. 16, pp. 149–172, 2000.
17. W. E. Henley and D. J. Hand. A K-nearest-neighbour classifier for assessing consumer credit risk. *Journal of the Royal Statistical Society. Series D*, vol. 45, pp. 77–95, 1996.
18. J. A. K. Suykens and J. Vandewalle. Least squares support vector machine classifiers. *Neural Processing Letters*, vol. 9, pp. 293–300, 1999.
19. P. Nerurkar, A. Shirke, M. Chandane and S. Bhirud. Empirical analysis of data clustering algorithms. *Procedia Computer Science*, 125, pp. 770–779, 2018.
20. S. B. Tambe and S. S. Gajre. Cluster-based real-time analysis of mobile healthcare application for prediction of physiological data. *Journal of Ambient Intelligence and Humanized Computing*, vol. 9, pp. 429–445, 2017.
21. P. D. Kumar, T. Amgoth and C. S. R. Annavarapu. Machine learning algorithms for wireless sensor networks: A survey. *Information Fusion*, vol. 49, pp. 1–25, 2019.
22. *Data Mining, Machine Learning and Predictive Analytics Software Minitab. Minitab*, 2020. Available from: https://www.minitab.com/en-us/products/spm [Last accessed on 2023 Jun 07].