Industrial Engineering Journal
ISSN: 0970-2555
Volume : 54, Issue 7, No.1, July : 2025

ANALYSING THE EFFECTS OF DATASET EXTENT ON THE EFFECTIVENESS OF DATA MINING APPROACHES.

Ashwinee Patil[1], MTech Scholar[1], Department of Computer Science & Engineering, Technocrats Institute of Technology & Science, Bhopal, India

Onkar Nath Thakur[2], Assistant Professor[2], Department of Computer Science & Engineering, Technocrats Institute of Technology & Science, Bhopal, India

Rakesh Kumar Tiwari[3], Assistant Professor[3], Department of Computer Science & Engineering, Technocrats Institute of Technology & Science, Bhopal, India

Vikas Gupta[4], Professor[4], Department of Electronics & Communication Engineering, Technocrats Institute of Technology & Science, Bhopal, India

ABSTRACT

Rising accident rates and the growing demand for blood and organ donations require efficient, data-driven systems for donor identification and management. This study investigates the impact of dataset size on the effectiveness of various data mining methods for donor classification, clustering, and forecasting. Eligible donors are filtered using criteria such as blood group, age, medical history, and geographical proximity. Clustering improves matching between donors and recipients. Rapid decision-making during emergencies demands robust outlier detection and noise reduction to enhance data quality. Experiments employ the open-source KEEL platform, which supports evolutionary algorithms and a graphical interface. Using Bayesian-D preprocessing, classification models including C4.5-C, AdaBoost-C, and C4.5_Binarization-C are evaluated across datasets ranging from 1,000 to 6,000 records. Results indicate that C4.5-C achieves higher accuracy and lower variance with larger datasets. For donor clustering, k-Means and k-Medoids improve grouping by proximity, eligibility, and urgency. Feed-forward backpropagation artificial neural networks (ANNs) predict donor availability using features such as blood type, age, and location. Mean Squared Error validates prediction performance, confirming that ANNs effectively forecast donor behavior. Larger datasets consistently strengthen model reliability, supporting scalable, data-driven decision frameworks.

Keywords: Data Mining, Clustering, KEEL Binarization Algorithm.

1. Introduction

In recent years, data analysis has transformed multiple fields, including medicine, marketing, and advertising, by extracting actionable insights from large-scale data through advanced techniques. Within medical practice, rising accident rates and complex health conditions demand timely availability of blood and organ donors. Traditional database retrieval systems increasingly fall short when processing growing donor records, highlighting the need for advanced analytical methods to support efficient and precise donor identification. Data mining plays a central role in this context, employing techniques such as classification, clustering, regression, and association to uncover patterns and relationships in extensive datasets. Classification enables supervised learning, using labeled data to train models for future predictions. In contrast, clustering, an unsupervised technique, groups data based on similarities and reveals natural structures without prior labels [1]. K-means clustering remains widely used across domains like pattern recognition and economic modeling, yet its efficiency depends heavily on initial centroid selection. Poor initialization often results in suboptimal clusters, driving research toward improved centroid optimization. Similarity measures, such as Euclidean distance, and robust centroid selection significantly enhance clustering

performance. Validating clusters ensures meaningful groupings; internal, external, and relative indices assess clustering quality. Regression, another supervised learning method, predicts continuous variables and helps forecast donor trends and demand. Artificial Neural Networks (ANNs), inspired by the biological brain, offer strong pattern recognition and handle noisy or incomplete data effectively. Attributes such as age, blood type, and location enable ANNs to predict donor behavior. Backpropagated multilayer feedforward networks prove especially valuable for classification and prediction tasks in donor management [2].

The Knowledge Extraction based on Evolutionary Learning (KEEL) tool supports the design and evaluation of computational intelligence methods, including classification, clustering, and regression. This open-source platform allows graphical configuration, algorithm performance comparison, and experiments involving preprocessing, hybrid modeling, and feature selection. Preprocessing tasks such as noise filtering and dimensionality reduction improve model performance by transforming raw data into relevant features [3]. Advanced techniques like random projection and SVD-based factorization simplify computations while preserving essential data characteristics. Recent studies propose alternative clustering approaches using dimensionality matrices and Huffman tree construction to overcome limitations of earlier algorithms [4]. These methods demonstrate higher stability and efficiency, particularly for large datasets, due to improved centroid initialization and consistent clustering outcomes. Such data mining techniques apply well in real-world donor systems. A centralized donor management platform integrates data from NGOs, hospitals, blood banks, and online systems, employing clustering for donor grouping by eligibility and proximity, and classification to identify frequent donors and predict demand trends. Web-based interfaces enhance communication among institutions, streamline requisition processes, and reduce emergency delays. Common classification methods, including decision trees, support vector machines, and ensemble techniques such as AdaBoost and Random Forests, offer advantages in interpretability, accuracy, and computational performance. Algorithm selection depends on dataset characteristics and application requirements [5]. Hybrid models combining classification and clustering (e.g., CNN-SVM or K-means-CART) hold potential for improved prediction accuracy and reduced latency, delivering powerful solutions for donor identification and stock management [5-7].

1.1 Motivation

An organized study of blood donor databases and blood bank repositories remains essential to meet the rising demand driven by increasing accidents and related conditions. Conventional database procedures often fail to extract critical donor information efficiently, creating an urgent need for more advanced repository systems. Another relevant area of study, xeno-transplantation, has also gained prominence in recent years. This work specifically focuses on identifying suitable donors who match the required blood group and reside within a defined location. To address this, a partitioning clustering technique is applied to locate and organize donor information effectively. Initially, a standard k-means clustering method partitions data based on geographic proximity. However, the selection of initial centroids strongly influences the final cluster quality. To reduce clustering complexity, enhance cluster quality, evaluate different data mining methods, and ensure robust validation, a novel system has been developed and tested.

1.2 Research contribution

Here are four short and effective research contributions:

- To evaluate how varying dataset sizes impact the performance of classification algorithms such as C4.5 and AdaBoost.
- To benchmark and validate data mining algorithms using the KEEL tool with consistent and reliable performance metrics.
- To cluster blood and organ donors effectively using improved K-Means and K-Medoids algorithms based on key eligibility parameters.
- To develop an artificial neural network model for predicting donor availability using age, blood group, and location features.

1.3 Paper Organization

This paper includes seven structured sections. The introduction highlights the growing need for efficient donor management driven by rising demand. Related work summarizes existing studies on data mining methods for donor identification and classification. The problem statement defines key challenges in managing and analyzing large-scale donor datasets. The research gap outlines the limitations of current approaches in addressing data volume and accuracy. The proposed methodology describes preprocessing, classification, clustering, and predictive modeling strategies. Data mining techniques, including C4.5-C, AdaBoost-C, k-means, k-medoids, and neural networks, are discussed for scalable donor management. The result analysis evaluates performance across varying dataset sizes. The conclusion summarizes main findings and proposes directions for integrating real-time data streams and advanced learning models to further improve donor management systems.

2. Related work

First, the capability of data mining methods has been broadly investigated in multifarious fields of applications such as healthcare, finance and social sciences. Nonetheless, one important determinant of these methods which has been given comparatively less research focus on has been the size or the extent of the dataset [8]. The size of the required dataset is a decisive factor to establish the performance, scalability and the overall generalizability of different data mining algorithms namely classification, clustering, regression and association rule mining. Some research papers have put more emphasis in the dataset size enhancement in enhancing the prediction accuracy. To illustrate, have stated that the voluminousness of the data has a direct impact on the scalability of algorithms, and efficient computational strategy was needed so that the stabilization of the performance was provided. With a high number of data, the decision tree and k-nearest neighbors (KNN) algorithms are computationally cumbersome and as such, there is slowness in learning and predicting results [9-10].

When classifying, by using decision tree algorithms such as C4.5 and CART, the results have been inconsistent based on datasets of different volumes. C4.5 has been shown to operate best on small to medium dataset but incredibly inefficient on large data where tree depth and redundancy assumes the master seat [11]. Conversely, the ensemble learning curve e.g., AdaBoost and Gradient Boosting Machine (GBM) has generally been strong across the various sizes of the dataset, owing to the fact that they use a collection of weak learners. k-means and k-medoids clustering methods are very sensitive to the size of data, particularly during initialization of the centroid, and iterative convergence, k-means scales well for small-data, but not in noisy and large-scale data without optimization by smarter initialization schemes (e.g. k-means++ or hybrid centroid estimators). A new study by [12], introduced the density-based k-means clustering algorithm that fits on large datasets by guiding the selection of the centroid on high-density areas, resulting in better accuracy and faster last iteration.

The size of dataset is also a factor that influences both linear and non-linear regression models of regression analysis. Increasing datasets tend to provide more accurate estimates of parameter and essential resources as well as optimization algorithms that are effective. Other researchers, including those who studied scalable regression methods like those by [13]. Into scalable regression techniques that are able to perform efficiently within large-scale data spaces, including stochastic gradient descent (SGD) or mini-batch processing. Pre-processing of data especially dimensions reduction is important in dealing with massive datasets. Reducing dimensionality over feature space is a common problem and is often implemented to retain data variance by using methods of Principal Component Analysis (PCA), t-SNE, and autoencoders [14]. To boost performance in classification techniques, feature selection and feature extraction techniques are also applied such as Information Gain and Chi-square tests. The methods have been applied successfully in the tools like WEKA and KEEL,

which are offered to the experimental environments of testing the method of classification, clustering, and regression activities at a different size of variable commandments [15].
KEEL, in particular, has been a valuable platform for analyzing how preprocessing impacts algorithm performance with increasing dataset sizes. By using Bayesian-D pre-processing in conjunction with classification algorithms like C4.5-C and AdaBoost, researchers have demonstrated significant improvements in accuracy and reduced variance, even with datasets exceeding 6000 records [16]. These findings validate the necessity of preprocessing when scaling data mining tasks to larger datasets. Artificial Neural Networks (ANNs), particularly multilayer perceptron's and feedforward backpropagation models, have shown improved performance with larger datasets due to their high capacity for pattern recognition. However, they also risk overfitting if the training data is not properly balanced or regularized. [17-18], deep learning models tend to generalize better with more data, but computational efficiency becomes a limiting factor, necessitating the use of GPUs and optimized training algorithms like Adam or RMSprop.
In anomaly detection, which is the setting where the class imbalance problem can seriously arise, the size of the data can have a great impact on the accuracy of the detection. Minor data sets might not give significant representative samples of abnormal behavior, which results in limited generalization. In order to overcome this, implementation of techniques such as SMOTE (Synthetic Minority Over-sampling Technique), and ensemble learning have been implemented in order to balance classes artificially and increase learning [19]. More recently studies have focused on lightweight and edge-optimized networks such as MobileNet or Tiny-YOLO and pruned neural networks, particularly in the context of relatively latency-sensitive applications such as healthcare and emergency response. Such models are specifically intolerant to the size of datasets because memory and processing constraints imply proficient learning with compressed representations. As an example, MobileNetV3 is capable of training successfully and reaching high accuracy rates using less data since it adopts effective architecture blocks such as depthwise separable convolutions [20]. Cloud and distributed environments have also been studied in terms of relationship between the size of a dataset and the performance of an algorithm. Frameworks based on Hadoop and Spark can support distributed computing of thick data volumes so that the traditional algorithms can scale. shown that distributed decision trees and support vector machines (SVMs) can be scaled with size of the data almost linearly when the computational resources are well managed [21].

3. Problem Statement

Incidences of medical emergency are more frequent and the need to donate blood and organs is on the rise creating a serious challenge to the healthcare logistic and donor management platforms. The conventional methods of donor identification based on data extraction in a manual way and simple data query are expeditious and unable to extent with the increase in the amount (or frequency) of information gathered by wide assortment of institutions like hospitals, blood banks, NGOs and internet-based registries. In an emergency situation, the inability to properly locate the right donors because of the unstructured nature of data as well as noise and absence of predictive ability will lead into life-threatening consequences. Moreover, there are several existing systems that are not as smart they identify their donors, evaluate them in terms of eligibility in real time, cluster them according to their proximity and medical parameters, predict the availability according to the previous trends, etc. Such systems are also heavily dependent on the amount of data and quality of data it receives, however the role the extent of a dataset has on algorithmic results still has minimal analysis. The urgency is necessitated by the need to create scalable, intelligent and adaptive data mining, which processes datasets of different sizes power, great classification reliability, fast clustering and solid predictions of donor availability. The study replaces this gap by conductively comparing the performance of different data mining methods on the basis of the extent of individual dataset. The aim is to create a strong data-based scheme through classification, clustering, and prediction models

that are confirmed by various tools such as KEEL and ANN architectures to assist donor decisions in health facilities in the most accurate and timely manner

4. Research Gap

Although data mining techniques have a sufficient number of applications in the healthcare sector, not much consideration has been given to determining the effects of the extent of datasets on the effectiveness of these techniques in donor identification and management systems. Current research efforts concentrate on the performance of algorithms when using fixed sizes of datasets without considering scalability and flexibility when using different amounts of data. Moreover, the traditional systems of donors do not include clustering and prediction models that could improve in-time decision-making under critical situations.

In addition to that, although the application of such tools as KEEL and ANN models demonstrates effective analysis opportunities, their joint use to assess the classification, clustering, and prediction in large and small datasets has not received a fair amount of attention yet. Strong pre-processing techniques that help in the handling of data noise, outliers and holes in the data have also not been discussed in the existing literature and their exclusion has a major impact on the reliability and good performance of the models. Thus, the gap in the research is evident regarding developing and validating an extensive data mining framework complex that integrates classification, clustering, and relationship prediction strategies as well as optimizes the process of donor identification by covering the limited issue of impacting the size of the datasets with the resounding outcomes of the performance effectiveness.

5. Proposed Methodology

To address the growing complexities in healthcare logistics—particularly blood and organ donor identification—this study presents a structured methodology to examine how dataset extent influences the effectiveness of data mining approaches. The framework consists of three core components: donor filtering, clustering, and prediction.

- Data Preprocessing and Filtering: Donor records are initially filtered based on key eligibility criteria such as blood type, age, medical history, and geographical proximity. Outlier detection and noise-handling mechanisms are implemented to ensure data quality and reliability, crucial in high-stakes, real-time medical scenarios.
- Dataset Scaling and Experimental Design:Experiments are conducted across datasets ranging from 1,000 to 6,000 records to evaluate performance variation. The open-source KEEL (Knowledge Extraction based on Evolutionary Learning) tool is employed for this analysis due to its comprehensive support for preprocessing, evolutionary algorithms, and performance evaluation. Bayesian-D preprocessing is integrated with classification algorithms including C4.5-C, AdaBoost-C, and C4.5_Binarization-C.
- Classification and Performance Assessment: Each algorithm's classification accuracy and variance are assessed across the scaled datasets. Findings reveal that C4.5-C offers the most stable and accurate predictions, particularly as dataset size increases, highlighting its scalability.
- Clustering and Donor Grouping: For donor-recipient matching, enhanced k-means and k-medoids algorithms are applied to cluster donors based on spatial and health-related attributes. This improves emergency responsiveness by enabling localized matching.
- Prediction Using ANN Models: Feed-forward backpropagation Artificial Neural Networks (ANNs) are trained to predict donor availability. Features like age, blood group, and location are used as input, and model performance is validated using Mean Squared Error (MSE), confirming ANN's robustness.

The methodology underscores the importance of scalable, data-driven systems in healthcare, demonstrating that larger datasets significantly enhance the effectiveness and reliability of data mining models, and below shown is Proposed architecture in Fig. 5.1.

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 54, Issue 7, No.1, July : 2025

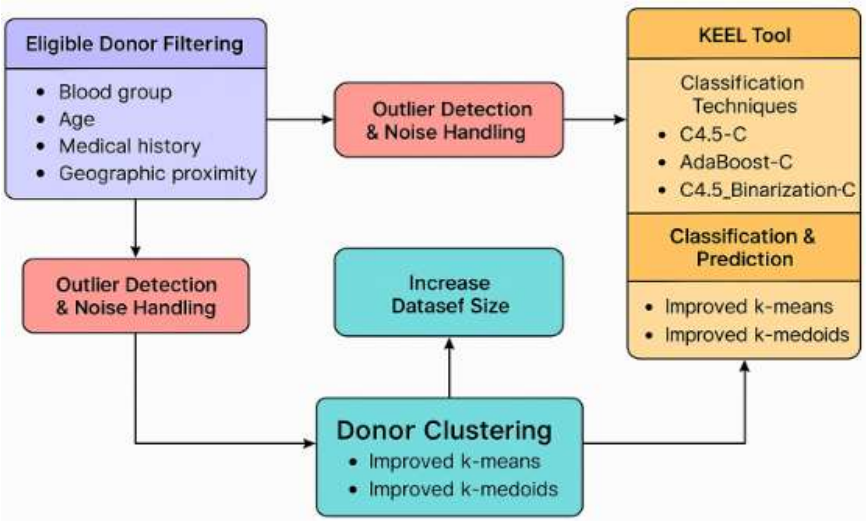

Fig. 5.1 Proposed architecture

5.1 Evaluation Metrics and Performance Analysis

In data mining, evaluating model performance is essential to determine its reliability and predictive power [22]. Four key evaluation metrics are commonly used: Accuracy, Precision, Recall, and F1-score, all derived from the confusion matrix, which consists of:

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

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

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

$$\text{F1} - \text{score} = 2 * \frac{Precision * Recall}{Precision + Recall}$$

Where, A prediction system labels positive cases correctly as True Positives while negative cases correctly get classified as True Negatives. A false positive scenario occurs when the algorithm misidentifies negative cases as positive instances and false negative events develop when positive cases get mistakenly classified as negative instances. Model performance evaluation depends on these performance metrics [23].

Model performance across several categorization thresholds is also frequently assessed using Receiver Operating Characteristic (ROC) curves and the Area Under the Curve (AUC). A higher AUC shows more discriminative capacity. Data volume and resolution have also important effects. As dataset size increases, models generally improve due to more learning instances, but may also face overfitting or increased computation [24]. Higher data resolution improves feature richness, which results in more exact classifications. All things considered, integrating these measures lets one fully grasp the strengths, shortcomings, and appropriateness of models for practical data mining uses [25].

Algorithm

Input	***Donor dataset D with features {Age, Blood Group, Medical History, Location}***
Step-1	*Load and clean dataset D (handle nulls, noise, and outliers)*

Step-2	*Normalize relevant features (e.g., age, distance)*
Step-3	*Partition D into subsets of increasing size (e.g., 1K to 6K records)*
Step-4	*For each subset Di:* *a. Apply Bayesian-D preprocessing* *b. Train classification models: C4.5-C, AdaBoost-C, C4.5_Binarization-C* *c. Evaluate models using F1-score, Accuracy, and select the best* *d. Cluster eligible donors using improved K-means and K-medoids* *e. Label clusters based on proximity, urgency, and eligibility*
Step-5	*Train ANN with input {Age, Blood Group, Location}*
Step-6	*Predict donor availability*
Step-7	*Validate prediction using Mean Squared Error (MSE)*
Step-8	*Compare performance across dataset sizes*
Step-9	*Plot accuracy, cluster stability, and prediction trends*
Step-10	*Summarize findings on how dataset size impacts model effectiveness*
Step-11	*Return classification results, cluster assignments, and predictions*
Output	***Classified donor status, clustered donor groups, predicted availability***

6. Data Mining Techniques for Scalable Donor Management Systems

Data mining techniques play a pivotal role in transforming raw donor data into actionable insights for healthcare logistics. These techniques enable the classification, clustering, and prediction of donor behavior based on features such as age, blood group, medical history, and geographic proximity. Among the most widely used techniques, classification helps identify suitable donors by assigning them to predefined categories—such as eligible or ineligible—using algorithms like C4.5, AdaBoost, and decision trees [26-27]. K-nearest neighbor also known as k-medoids, as well as improved k-means, cluster donors on the basis of proximity and urgency and improves the efficiency of matching donors and recipients. Moreover, one of the models, which is used to predict donor availability, is Artificial Neural Network (ANN); allowing prompt assistance in emergency situations [28-29]. Pre-processing algorithms such as Bayesian-D enhance data quality by sorting out noise and outliers which is essential when training a certain model. As dataset sizes increase, the performance of data mining algorithms typically improves in terms of stability and accuracy [30], below shown is Data mining technique in Fig. 5.2

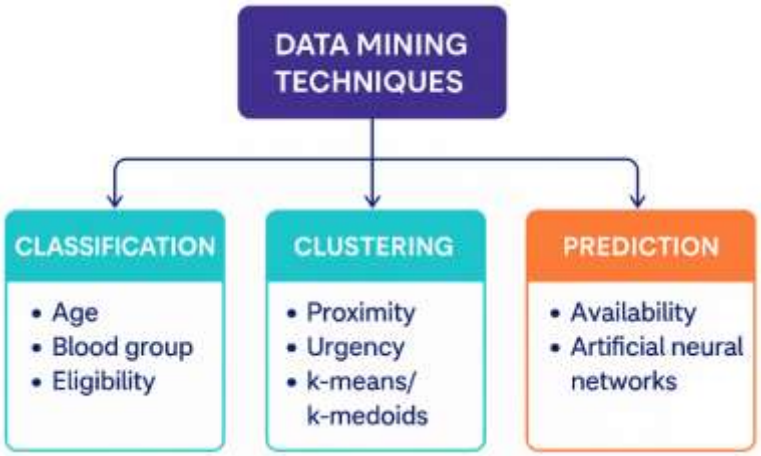

Fig. 5.2 Data mining technique

Tools such as KEEL (Knowledge Extraction based on Evolutionary Learning) support comprehensive evaluation of these techniques through visual workflows and statistical analysis. Overall, the integration of scalable data mining methods is critical for building intelligent, data-driven donor management systems that optimize healthcare delivery and save lives in real-time critical situations [25].

7. Result analysis

7.1 Dataset Description

A brief description of the data types commonly used in this topic is provided below

Ecoli Dataset:

The Ecoli dataset is an example of bioinformatics and machine learning, used to classify protein localization sites in E. coli cells. It contains 336 samples, each with eight biochemical characteristics, such as scores from sequencing signal detection methods, discriminative amino acid content analysis and elements including mcg, alm. alm1, vh, vh2, gs, imL, and imL2. Class labels indicate different regions of the cell including cytoplasm, inner membrane, outer membrane, periplasmic, extracellular, fimbrial, and lipoproteins This data set is necessary to encode subcellular proteins prediction, contribute to the understanding of protein function, and assess the performance of distribution systems in the development of new drugs and therapies Available.

Pima Dataset:

The Pima Indians Diabetes Dataset, sourced from the National Institute of Diabetes, Digestive and Kidney Diseases and accessed through the UCI Machine Learning Repository contains medical records for 768 Pima Indian women It has 8 statistical characteristics: pregnancy rate, plasma glucose concentration , Diastolic blood pressure, triceps skinfold thickness, 2-hour serum insulin, body mass index (BMI), diabetes pedigree function, and age are two target variables, which predict the presence (1) or absence of diabetes there (0). Despite the obvious missing values, there are some null values that must be assumed missing and handled during pre-processing. This dataset is widely used for binary classification tasks, exploratory data analysis, model evaluation, educational purposes, which is a benchmark for machine learning algorithms and data pre-processing, feature engineering, model evaluation It is also a valuable resource for learning methods

7.2 Experimental Setup (KEEL)

We need to calculate the Global Classification Error, Standard Deviation Global Classification Error, and appropriately categorize donor samples. We have 5310 data sample attribute values in a file. The data file has 6 characteristics.

Select data management from the KEEL tool interface. Choose the dataset sourcefile format. Valid formats include CVS, TXT, PRN, C4.5, Excel, DIF, Property List, and Weka. The import option converts TXT, Excel, XML, and other files to KEEL.

The source file path must be supplied after the file format. Click save to generate keel file. KEEL experiments employ that csv-to-keel file. Then we will split the entire data file for cross validation classification into training and testing partitions.

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 54, Issue 7, No.1, July : 2025

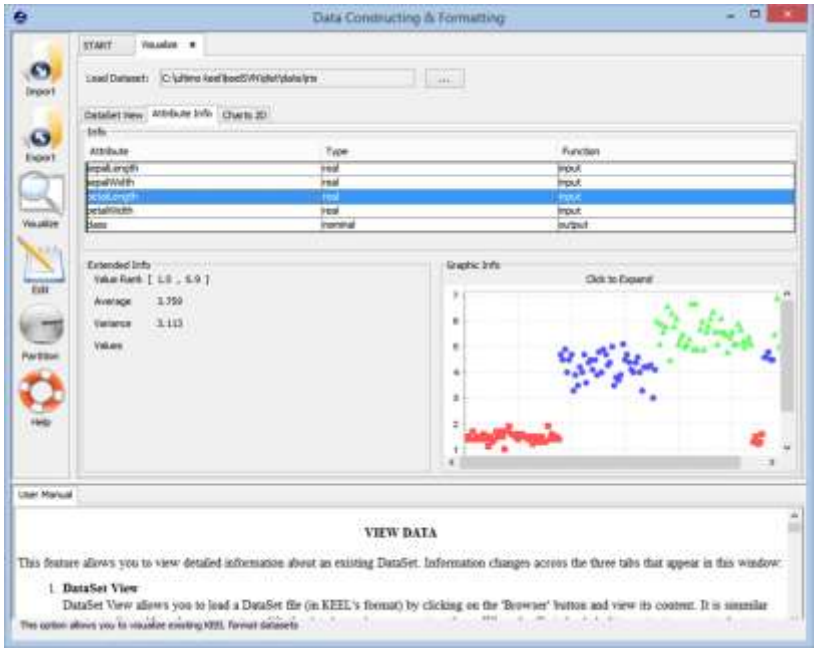

Fig. 7.1 KEEL dataset visualization

Fig.5.1 shows KEEL dataset visualization possibilities. This graphical information lets users view a dataset, attribute information, or compare two attributes using charts. In the main window of the visualization menu, pick the KEEL source dataset path to visualize. Depending on the option selected, the file loads with Dataset view, Attribute info, Charts 2D, or Edit data. The Experiments Design section lets users build experiments using a graphical interface.

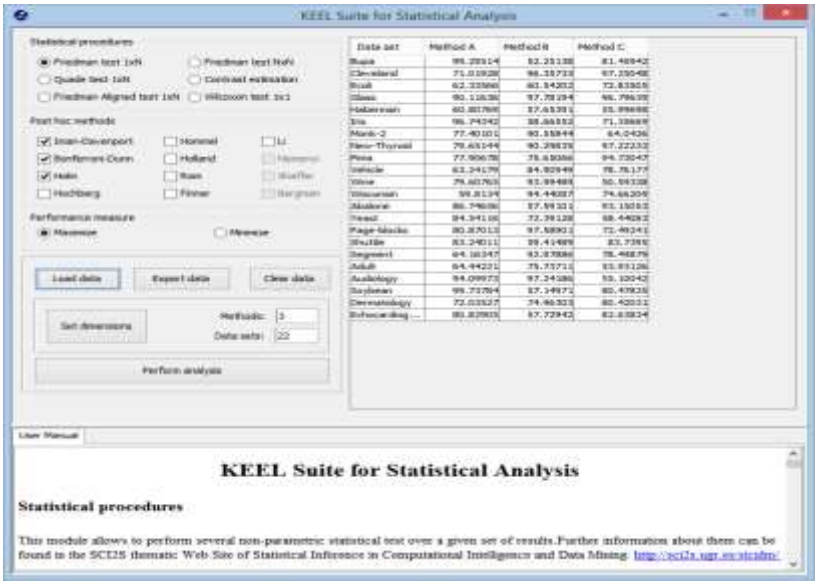

Fig. 7.2 KEEL Suite for Statistical Analysis

Experimental and Statistical Analysis can be used to design experiments, start by opening the experimental module. After importing all essential datasets, experiment design can begin. Click the white graph panel to set the dataset node of the experiment. Make designs try various, data-C45-VisClasCheck data- BayesianD-C45 VisClasCheck, data- AdaBoost.NC-C, data- BayesianD, data-C45_BinarizationC, and data- BayesianD. We must design the experiment using the following settings after configuring the datasets node. The "Tools" menu's "Run Experiment" option generates

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 54, Issue 7, No.1, July : 2025

an experiment once it is developed. Press the tools bar. Here, the software program will assess experiment completion numerous times.

Choose a path for the experiment's zip file. The generating method creates a ZIP file with all experiment components. Generation of the experiment is successful. First, we must unzip the indicated ZIP file on the experiment system. We will get a directory called “experiment Name” (How we named our experiment). We must put him in the “experiment Name” folder and the “scripts” subfolder. Simply execute “java -jar RunKeel.jar” to perform the tests. The experiment is done. After an experiment runs, its result files are in the results subfolder. The experiment must be executed using RunKeel.jar from “experiment/scripts”.

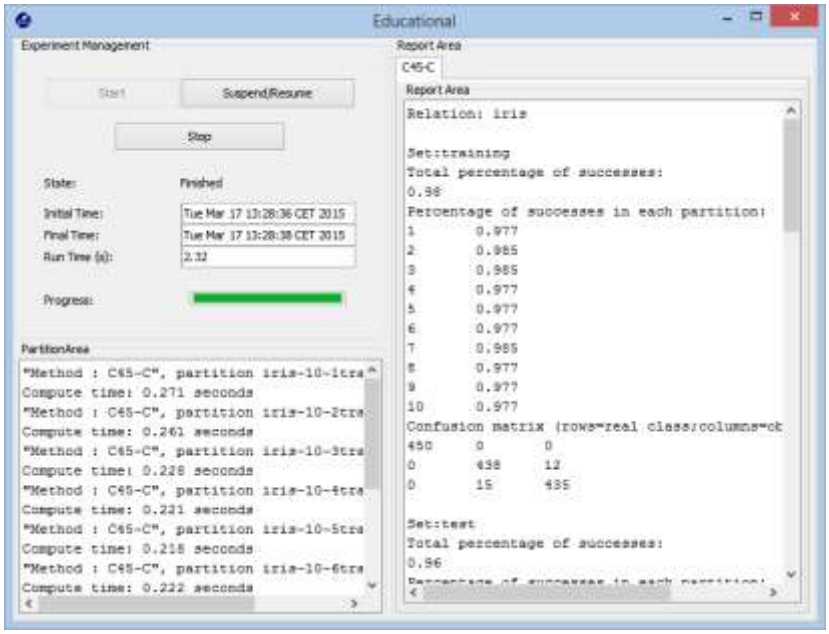

Fig. 7.3 KEEL Educational experiments.

We must monitor the experiment in the second step. The experiment design procedure can proceed after importing all essential datasets in the Education module. To do so, we choose the experiment dataset node. Design experiments using diverse combinations: C45, BayesianD-C45, AdaBoost.NC-C, C45_BinarizationC, and C45_BinarizationC. After setup, push to run the experiment. Selecting start starts the experiment. The partition area shows experiment progress and the tool window report shows results. Educational experiments indicate their duration. Each division success of training and test results is shown. Running time is provided for properly and erroneously categorized results, depending on dataset size. Time also relies on dataset size. It displays the experiment's confusion matrix.

7.3 Result Discussion

By combining the classification algorithm with the Bayesian-D pre-processing method of the KEEL tool, we evaluate the performance of this algorithm on data sets of varying sizes, especially when analysing data sets ranging from 1,000 to 6,000 records. This study enables us to examine how classification accuracy and efficiency change with a large dataset, and provides insight into the scalability and robustness of algorithms Bayesian-D pre-processing techniques help to increase data quality, and provide reliable results and it is constantly going on. This comprehensive approach helps to understand the strengths and limitations of classification algorithms in different data environments and show the test result in table 7.1.

Table 7.1 Classification Error Rates for Different Dataset Record Sizes

S.N.	Algorithms	No of datasets Records	Average

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 54, Issue 7, No.1, July : 2025

		1000	2000	3000	4000	>6000	
1	‘C45-C’	0.0032	0.0035	0.0	0.0007	0.0006	0.001592
2	‘AdaBoost.NC-C’	0.5001	0.9125	0.941	0.9559	0.9669	0.85527
3	‘C45_Binarization-C’	0.1872	0.003	0.004	0.0032	0.0028	0.040044

As a consequence of the experimental findings, we have seen that the influence of the size of the data set is enhanced, and the error rates vary depending on the kind of algorithm under consideration. Due to the fact that the size of the data set expanded, the error rate progressively climbed and then gradually dropped, as shown in table 5.1. We made the observation that the error rate is reduced to a minimum when the size of the data set is huge. As a result of such findings, classification error rates for various dataset record sizes were determined. There is an observation made on the average error rate of several algorithms. 0.001592 is the average Global Classification Error for "C4.5-C," while 0.85527 is the average for "AdaBoost.NC-C," and 0.040044is the average for "C45_Binarization-C." The best result was achieved by the C4.5-C algorithm out of all the algorithms. An obvious analysis and influence of the size of the data set is detected, as shown in Fig. 7.4.

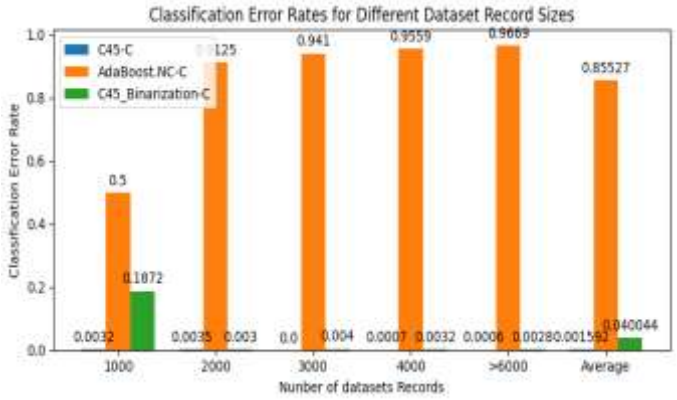

Fig. 7.4 Classification error rate of different datasets

This study analyzes the performance of classification algorithms on synthetic datasets and selected UCI Machine Learning standard datasets, namely Ecoli and Pima. The experiments employ a combination of classification methods with Bayesian-D preprocessing using the KEEL tool. Synthetic and benchmark datasets are compared to evaluate classification performance under varying conditions. Experimental results, presented in Fig. 5.5 and Table 5.2, report classification error rates for different dataset sizes. These rates are shown in the table. Based on the findings, we have determined that the performance analysis of the C4.5-C classification algorithm for the synthetic data set yields the most favourable outcomes. In comparison to other Standard datasets, the performance of the C4.5-C algorithm that we have used on our dataset is satisfactory. Within the C4.5-C method, the Global Classification Error is calculated to be 0.

Table 7.2 Classification Error Rates for Different Dataset Record Sizes.

S. N.		Number of Dataset Records					Average
	Algorithms	1000	2000	3000	4000	>6000	
1	‘C45-C’	0	0	0	0	0	0
2	‘AdaBoost.NC-C’	0.5	0.91	0.94	0.96	0.967	0.855
3	‘C45_Binarization-C’	0.18	0	0	0	0	0.0368

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 54, Issue 7, No.1, July : 2025

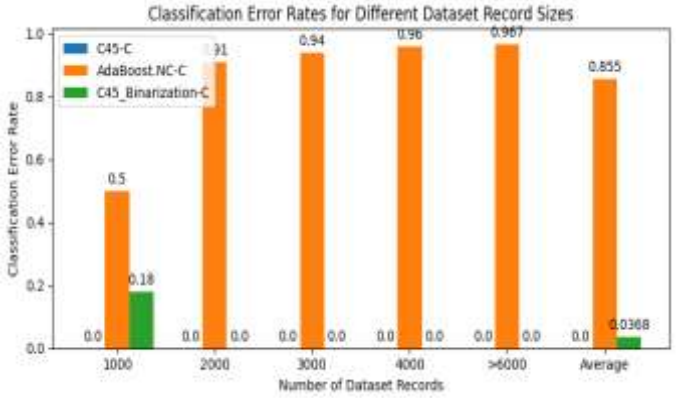

Fig. 7.5 Classification error rate of different datasets

We found that with larger dataset size, the error rate decreased, allowing a more detailed.

Table 7.3 Result comparison of test and train data

Table	Algorithms	No. of datasets Records					Average
		1000	2000	3000	4000	>6000	
Table-5.1	'C45-C'	0.003	0.004	0.0	7E-04	0.0006	0.001592
	'AdaBoost.NC-C'	0.5	0.913	0.94	0.956	0.9669	0.85527
	'C45_Binarization-C'	0.187	0.003	0.0	0.003	0.0028	0.04004
Table-5.2	'C45-C'	0.0	0.0	0.0	0.0	0.0	0.0
	'AdaBoost.NC-C'	0.5	0.91	0.94	0.96	0.967	0.855
	'C45_Binarization-C'	0.18	0.0	0.0	0.0	0.0	0.0368

analysis of the classification error rates of dataset records. The average global classification error of "C45-C" is 0.001592, 0.85527 for "AdaBoost.NC-C", and 0.040044 for "C45_Binarization-C".

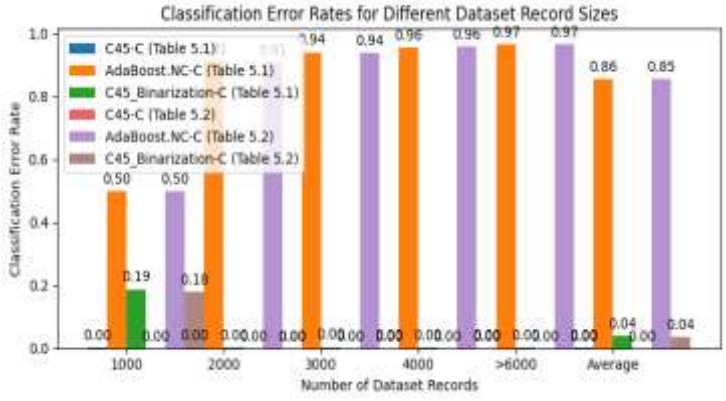

Fig. 7.6 Result comparison of test and train data

The C45-C algorithm obtained good results, indicating that its performance on synthetic data sets is the best among the tested algorithms The satisfactory performance of the C45-C algorithm on our data set, compared to standard data for other sets, it means difficulty. Importantly, the global classification error of the C45-C method is considered as 0, highlighting its effectiveness in reducing classification errors.

8. Conclusion and Future work

Data mining extracts valuable knowledge from large databases and presents it in an accessible format. A key task within this field involves classification techniques, which categorize data based on defined attributes and classes. This study reviews and compares classification algorithms such as C4.5, C4.5_Binarization, AdaBoost, and a Bayesian-D preprocessor combination. These methods are applied to donor datasets to evaluate accuracy and error rates under varying dataset sizes. Results indicate that C4.5 achieves the highest accuracy, while C4.5_Binarization performs moderately.

Future work will focus on integrating real-time data streams into donor management systems to support rapid, dynamic decision-making during emergencies. Deploying the proposed framework on cloud and edge platforms can further improve scalability and accessibility. Upcoming studies will also investigate deep learning models to enhance prediction accuracy and develop adaptive learning approaches for evolving donor data.

References

[1] comparative study of clustering techniques for electrical load."
[2] "A data mining-based framework for the identification of daily electricity."
[3] "Guest Editorial: Blockchain and AI Enabled 5G Mobile Edge Computing," IEEE Trans. Ind. Inf., vol. 16, no. 11, pp. 7067–7069, Nov. 2020, doi: 10.1109/TII.2020.2983764.
[4] A. Al-Wakeel, J. Wu, and N. Jenkins, "k -means based load estimation of domestic smart meter measurements," Applied Energy, vol. 194, pp. 333–342, May 2017, doi: 10.1016/j.apenergy.2016.06.046.
[5] A. Djellouli et al., "A datamining approach to classify, select and predict the formation enthalpy for intermetallic compound hydrides," International Journal of Hydrogen Energy, vol. 43, no. 41, pp. 19111–19120, Oct. 2018, doi: 10.1016/j.ijhydene.2018.08.122.
[6] A. E. Ghazi and A. Moulay Rachid, "Machine learning and datamining methods for hybrid IoT intrusion detection," in 2020 5th International Conference on Cloud Computing and Artificial Intelligence: Technologies and Applications (CloudTech), Marrakesh, Morocco: IEEE, Nov. 2020, pp. 1–6. doi: 10.1109/CloudTech49835.2020.9365895.
[7] A. Gamazo and F. Martínez-Abad, "An Exploration of Factors Linked to Academic Performance in PISA 2018 Through Data Mining Techniques," Front. Psychol., vol. 11, p. 575167, Nov. 2020, doi: 10.3389/fpsyg.2020.575167.
[8] A. Pratelli, M. Petri, M. Ierpi, and M. Di Matteo, "Integration of Bluetooth, Vehicle Count Data and Trasport Model Results by Means of Datamining Techniques," in 2018 IEEE International Conference on Environment and Electrical Engineering and 2018 IEEE Industrial and Commercial Power Systems Europe (EEEIC / I&CPS Europe), Palermo: IEEE, Jun. 2018, pp. 1–6. doi: 10.1109/EEEIC.2018.8493997.
[9] A. Soltani, C. J. Pettit, M. Heydari, and F. Aghaei, "Housing price variations using spatio-temporal data mining techniques," J Hous and the Built Environ, vol. 36, no. 3, pp. 1199–1227, Sep. 2021, doi: 10.1007/s10901-020-09811-y.
[10] A.-C. Paola et al., "GlyphReader App: A support game for the application of the Orton-Gillingham Method with DataMining Techniques.," Procedia Computer Science, vol. 191, pp. 373–378, 2021, doi: 10.1016/j.procs.2021.07.071.
[11] B. Robson, S. Boray, and J. Weisman, "Mining real-world high dimensional structured data in medicine and its use in decision support. Some different perspectives on unknowns, interdependency, and distinguishability," Computers in Biology and Medicine, vol. 141, p. 105118, Feb. 2022, doi: 10.1016/j.compbiomed.2021.105118.
[12] C. Savaglio and G. Fortino, "A Simulation-driven Methodology for IoT Data Mining Based on Edge Computing," ACM Trans. Internet Technol., vol. 21, no. 2, pp. 1–22, Jun. 2021, doi: 10.1145/3402444.
[13] D. Bari, M. Ameksa, and A. Ouagabi, "A comparison of datamining tools for geo-spatial estimation of visibility from AROME-Morocco model outputs in regression framework," in 2020

IEEE International conference of Moroccan Geomatics (Morgeo), Casablanca, Morocco: IEEE, May 2020, pp. 1–7. doi: 10.1109/Morgeo49228.2020.9121909.
[14] E. L. Ofetotse, E. A. Essah, and R. Yao, "Evaluating the determinants of household electricity consumption using cluster analysis," Journal of Building Engineering, vol. 43, p. 102487, Nov. 2021, doi: 10.1016/j.jobe.2021.102487.
[15] F. Biscarri, I. Monedero, A. García, J. I. Guerrero, and C. León, "Electricity clustering framework for automatic classification of customer loads," Expert Systems with Applications, vol. 86, pp. 54–63, Nov. 2017, doi: 10.1016/j.eswa.2017.05.049.
[16] F. Es-Sabery et al., "A MapReduce Opinion Mining for COVID-19-Related Tweets Classification Using Enhanced ID3 Decision Tree Classifier," IEEE Access, vol. 9, pp. 58706–58739, 2021, doi: 10.1109/ACCESS.2021.3073215.
[17] F. Saadi, B. Atmani, and F. Henni, "Integration of datamining techniques into the CBR cycle to predict the result of immunotherapy treatment," in 2019 International Conference on Computer and Information Sciences (ICCIS), Sakaka, Saudi Arabia: IEEE, Apr. 2019, pp. 1–5. doi: 10.1109/ICCISci.2019.8716415.
[18] H. N. K. AL-Behadili, K. R. Ku-Mahamud, and R. Sagban, "Hybrid Ant Colony Optimization and Genetic Algorithm for Rule Induction," Journal of Computer Science, vol. 16, no. 7, pp. 1019–1028, Jul. 2020, doi: 10.3844/jcssp.2020.1019.1028.
[19] J. K. Mandal and D. Bhattacharya, Eds., Emerging Technology in Modelling and Graphics: Proceedings of IEM Graph 2018, vol. 937. in Advances in Intelligent Systems and Computing, vol. 937. Singapore: Springer Singapore, 2020. doi: 10.1007/978-981-13-7403-6.
[20] L. Song, Z. Han, P.-W. Shum, and W.-M. Lau, "Enhancing the accuracy of blood-glucose tests by upgrading FTIR with multiple-reflections, quantum cascade laser, two-dimensional correlation spectroscopy and machine learning," Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, vol. 327, p. 125400, Feb. 2025, doi: 10.1016/j.saa.2024.125400.
[21] M. K. Gupta and P. Chandra, "A comprehensive survey of data mining," Int. j. inf. tecnol., vol. 12, no. 4, pp. 1243–1257, Dec. 2020, doi: 10.1007/s41870-020-00427-7.
[22] M. Rahman, S. A. Khushbu, and A. K. Mohammad Masum, "Associative datamining Survey on Modern Era People's engagement of Gaming addiction," in 2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT), Kharagpur, India: IEEE, Jul. 2021, pp. 01–05. doi: 10.1109/ICCCNT51525.2021.9579980.
[23] M. S. Geetha Devasena, R. Kingsy Grace, and G. Gopu, "PDD: Predictive Diabetes Diagnosis using Datamining Algorithms," in 2020 International Conference on Computer Communication and Informatics (ICCCI), Coimbatore, India: IEEE, Jan. 2020, pp. 1–4. doi: 10.1109/ICCCI48352.2020.9104108.
[24] M. Taktak and S. Triki, "A spatiotemporal datamining approach for road profile estimation using low-cost device," Procedia Computer Science, vol. 207, pp. 2767–2781, 2022, doi: 10.1016/j.procs.2022.09.335.
[25] S. Batool, J. Rashid, M. W. Nisar, J. Kim, H.-Y. Kwon, and A. Hussain, "Educational data mining to predict students' academic performance: A survey study," Educ Inf Technol, vol. 28, no. 1, pp. 905–971, Jan. 2023, doi: 10.1007/s10639-022-11152-y.
[26] S. Umamaheswari and K. Harikumar, "Analyzing Product Usage Based on Twitter Users Based on Datamining Process," in 2020 International Conference on Computation, Automation and Knowledge Management (ICCAKM), Dubai, United Arab Emirates: IEEE, Jan. 2020, pp. 426–430. doi: 10.1109/ICCAKM46823.2020.9051488.
[27] T. M. Ghazal et al., "Performances of K-Means Clustering Algorithm with Different Distance Metrics," Intelligent Automation & Soft Computing, vol. 29, no. 3, pp. 735–742, 2021, doi: 10.32604/iasc.2021.019067.
[28] W. Xiong, "Initial Clustering Based on the Swarm Intelligence Algorithm for Computing a Data Density Parameter," Computational Intelligence and Neuroscience, vol. 2022, pp. 1–8, Jun. 2022, doi: 10.1155/2022/6408949.
[29] X. Ruhang, "Efficient clustering for aggregate loads: An unsupervised pretraining based method," Energy, vol. 210, p. 118617, Nov. 2020, doi: 10.1016/j.energy.2020.118617.

[30] Y. Ma, Y. Lei, and T. Wang, "A Natural Scene Recognition Learning Based on Label Correlation," IEEE Trans. Emerg. Top. Comput. Intell., vol. 6, no. 1, pp. 150–158, Feb. 2022, doi: 10.1109/TETCI.2020.3034900.