SN Computer Science (2024) 5:1158
https://doi.org/10.1007/s42979-024-03483-z

ORIGINAL RESEARCH

Data Analytics Approach for Enhanced Sales Forecasting (DAAESF): Feature Selection and Classifier Integration Analysis

Gagandeep Kaur[1] · Harpreet Kaur[1] · Sonia Goyal[2]

Received: 10 May 2024 / Accepted: 28 October 2024

Abstract

Accurately forecasting sales has significant ramifications for producers, distributors, and investors. Sales forecasting accuracy enables businesses to enhance their manufacturing, distribution, and promotional activities. The current research intends to examine the implications of feature selection methods on enhancing the accuracy of seasonal sales forecasts. The author evaluates different feature selection methods in combination with predictive models, aiming to determine their impact on the effectiveness of predictions. Within this research, a diverse set of eight classifiers has been utilized: namely Naïve Bayes, Logistic Regression, Neural Network, Random Forest, J48, IBK, SVM, and K Star. Alongside these classifiers, four distinct feature selection techniques-namely Gainratio, Infogain, Relief, and CFS have also been employed. The effectiveness of these strategies was evaluated individually as well as collaboratively. The outcome of the proposed novel methodology DAAESF engendered a notable advancement in accuracy rates. Combining feature selection techniques with Neural Network led to a 32% accuracy enhancement compared to other classifiers for cement sales prediction, while Naïve Bayes experienced a decline in performance from 19.55% to 32.55% due to its distinct functions. Additionally, feature selection notably improved prediction accuracy across classifiers, with Neural Network achieving up to 22.68% improvement using CFS, and SVM showing gains of 26.3% with Infogain, highlighting the critical role of feature selection in model optimization. Naïve Bayes and J48 exhibited mixed results across datasets and feature selection methods. Additionally, to substantiate the robustness and validity of the observed outcomes, the Friedman test was judiciously applied.

Keywords Data analytics · Logistics · Machine learning · Classification · Feature selection techniques · Sales forecasting · Predictive models · Friedman test · Seasonal sales

Introduction

The management of supply chains (SCM) has a significant influence on sales and overall company efficiency. An effectively managed supply chain ensures a consistent supply of raw materials, manufacturing procedures, and distribution routes, ultimately impacting the accessibility, effectiveness, and rapid delivery of cement products to clients [1]. Modern sales are significantly influenced by technology, which enables businesses to increase customer engagement, optimize processes, and receive insightful knowledge about market trends. In this dynamic sector, embracing technology innovations [2–4] in the sales process can result in increased competition, better customer satisfaction, and long-term corporate growth [5].

Technological advancements and machine learning have had a significant impact on supply chain management [6–10]. Machine learning, an artificial intelligence subset, has transformed how corporations manage many areas of their supply chains as pictorially depicted in Fig. 1. Machine learning and technology have enhanced supply chain management in specific ways such as Demand forecasting, Real-time Tracking and Visibility, Warehouse Automation, Supply Chain Analytics, etc. [11].

✉ Gagandeep Kaur
phd_gagandeep8@csepup.ac.in

Harpreet Kaur
harpreet.ce@pbi.ac.in

Sonia Goyal
soniagoyal@pbi.ac.in

[1] Department of Computer Science and Engineering, Punjabi University, Patiala 147002, Punjab, India

[2] Department of Electronics and Communication Engineering, Punjabi University, Patiala 147002, Punjab, India

Published online: 10 December 2024

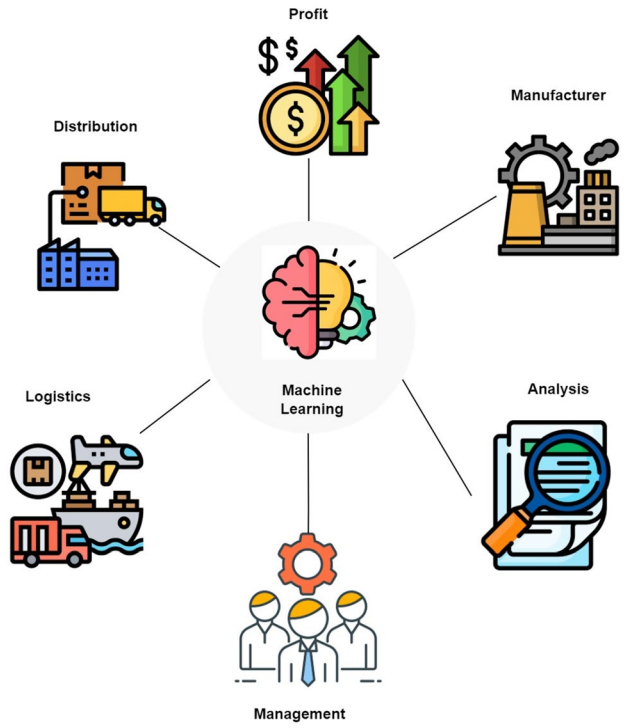

Fig. 1 Depicts the relationship of supply chain management and machine learning

The forecasting of sales provides businesses with useful information that helps them to improve customer satisfaction, minimize expenditures, and acquire a competitive edge in a changing market environment. Machine learning [2–4] plays a significant role in forecasting various aspects of supply chain management.

Companies can more efficiently anticipate demand whilst they can make accurate sales predictions. As a result, there is less chance of overproduction or stockouts because they can better manage production schedules, inventory levels, and distribution in accordance with customer requirements. Businesses may optimize their inventory levels by anticipating sales, avoiding excessive stockpiling during slow periods, and ensuring they have enough inventory during high demand, resulting in cost savings and better efficiency. Prediction of sales facilitates better resource allocation, including staff planning, equipment usage, and supply chain management [12, 13]. This research constitutes a substantial advancement by introducing a methodological framework for cement sales prediction that bestows enterprises with invaluable insights, enabling them to optimize diverse operational facets. It culminates by succinctly encapsulating the study's findings and charting out future avenues of exploration. The objectives of the research are stated below:

- Investigate the significance of accurate sales forecasting for producers, distributors, and investors.
- Enhance manufacturing, distribution, and promotional strategies through precise sales predictions.
- Demonstrating the integration of feature selection techniques with classifiers leads to a significant enhancement in accuracy rates.
- Highlight the subsequent elevation in predictive accuracy as a result of the proposed DAAESF methodology.
- Apply the rigorous Friedman test to validate and establish the reliability and validity of the enhanced predictive accuracy.

Overall, the work presents a thorough analysis showcasing the successful integration of advanced classifiers and feature selection methods to bolster prediction accuracy, reinforced by the meticulous validation using the Friedman test. This research work has been organized as follows. The related work is shown in Sect. Related Work. The proposed approach is outlined in Sect. System Architecture. The experimental results and discussions are presented in Sect. Experimental Results. The conclusion and future work have been addressed in Sect. Conclusion and Future Scope.

Related Work

There has been a surge in interest in supply chain management and logistics in recent years. Numerous studies have developed addressing various facets as scholars and practitioners attempt to enhance their understanding and address various issues in this sector. This section seeks to provide a comprehensive review of prior related work, highlighting significant findings and approaches used by earlier researchers.

In the present study, Borucka et al. (2023) examine the significance of occasional interest-determining approaches concerning manageable development for inventory network organizations. The review plunges into the challenges that organizations go up against in accomplishing manageability objectives while satisfying client needs successfully. It explores occasional anticipating approaches and their possible advantages for upgrading inventory network tasks and advancing reasonable practices. The primary objective of this work is to provide critical information to businesses aiming to improve asset utilization, minimize waste, and provide environmental responsibility without affecting revenues [5]. Feizabadi et al. (2022) illustrate the increasing significance of machine learning (ML) in handling supply chains and logistics. In this research, the author explored the potential of optimizing the administration of inventory and enhancing the accuracy of demand forecasts [6, 14]. Zhu et al. (2021) investigate how the

pharmaceutical industry utilizes forecasting techniques that encompass the supply chain and machine learning (ML) for demand forecasting. The survey evaluates the potential benefits of planning, and the creation of network data in improving demand forecasting accuracy [11, 15]. The author intends to illustrate the efficacy of ML-based demand prediction approaches in enhancing inventory management and overall supply-chain performance through empirical data and contextual evaluation in the medicine area.

Pereira et al. (2022) present a data-driven approach to dynamic supply and demand synchronization in omnichannel retail supply chains. The authors present an approach for merchants to optimize inventory levels, enhance customer service, and improve overall supply chain effectiveness by using real-time data analytics and predictive modeling [12]. Nguyen et al. (2021) give a review that looks into applying Long Short-Term Memory (LSTM) and LSTM Autoencoder algorithms for prediction and anomaly detection in supply chain management. The research study addresses the issues of predicting demand and recognizing supply chain deviations. The contributors leverage empirical and contextual information that demonstrates the usefulness of algorithms based on deep learning to enhance supply chain management by forecasting demands, identifying anomalies, and enhancing precision [13].

Ji et al. (2019) offers a novel, three-stage XGBoost-based methodology that efficiently handles the challenging matter of sales forecasting for global online retail businesses. The study provides an in-depth review of their proposed approach and provides insightful perspectives regarding its prospective results and future benefits [16]. Cheriyan et al. (2018) execute a systematic review of several different machine learning algorithms in order to address the challenge of precise forecasting of sales. It offers useful details on ensemble techniques, model determination, and information preprocessing. The research's contribution to the field of sales forecasting is enhanced by the relevant experimental results and ramifications [17]. Mohamed-Iliasse et al. (2020) evaluate the consequences of AI (ML) on the management of supply chains. The work focuses on how AI approaches are changing different components of the production network, enhancing methodology, further developing navigation, and eventually expanding proficiency [18].

Bousqaoui et al. highlight the increasing relevance of supply chain optimization and the possibilities for machine learning to enhance its efficiency. It emphasizes the significance of neural networks as effective tools for dealing with the complexities of supply chain operations such as demand forecasting, inventory management, and logistics optimization [19]. Gupta et al. (2022) focus on predicting sales in the context of a retail environment, notably targeting MegaMart, and employ a variety of machine learning approaches. The primary components, methods, results, and implications of the research are comprehensively examined in this review [20].

System Architecture

An elaborated model has been precisely developed to prognosticate cement sales and anticipate cement logistics [21]. This model encompasses a refined classification process that incorporates four advanced feature selection techniques, namely Infogain, Gainratio, CFS, and Relief [22]. The graphical representation of the proposed model is elucidated in Fig. 2.

Dataset Collection

Cement supply logistics have been explored by collecting 20,000 invoice records from various supply units of cement manufacturing companies in various districts of Punjab for the year 2018–2023. There are 20 attributes in the dataset. The dataset collected from different supply units undergoes label encoding in Python to transform the data into a suitable format for classification purposes.

Dataset Input and Preprocessing

Python label encoding is used to preprocess a dataset that has been gathered from a variety of supply units so that it can be transformed into a dataset that is used for mining. This preprocessing step allows for the appropriate representation of the data in a format that can be utilized effectively for classification tasks. In this study, data was gathered from multiple distribution locations of the industrial structure. Data gathered includes significant factors of cement logistics, such as customer supply and demand, individual housebuilder supply, and commercial-industrial infrastructure. It contains information regarding truck load factors, volumes of sales, primary freight (from manufacturing units), secondary freight (from godowns), and transportation [23].

Data Classification

In this analysis, the data has been split into testing and training phases deploying the k-fold validation technique. Data from the dataset was divided into sections with 70% used

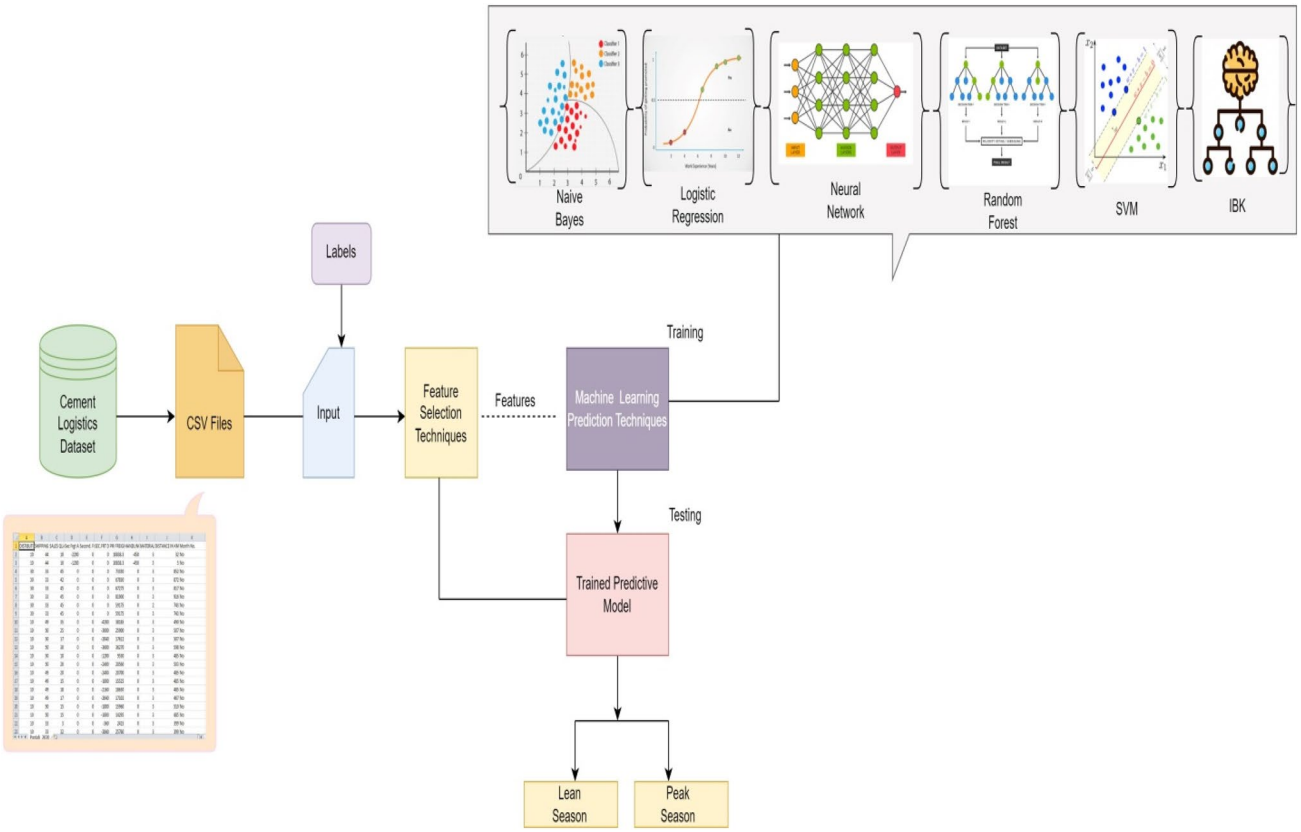

Fig. 2 Workflow of the proposed model

to train a classifier and 30% used to test it [24]. To classify based on seasonality, Python 3.7 is used. The evaluations were conducted on dual-core i5 machines having 1.60 GHz processors, Windows 10 platforms, 8 GB of RAM, 1 TB of storage on the hard drive, and 64-bit libraries. In addition, data transformations, cleaning, integration, and selection are conducted before the classification step. The data classification process is done on the data during this step. The Cement Logistics dataset is initially divided into testing and training sets [25–27].

Applying Feature Selection and Machine Learning Techniques

In this study, eight ML (Machine Learning) models such as Naïve Bayes (NB), Logistic regression (LR), Neural Network (NN), Random Forest (RF), J48, IBK, SVM, and K Star [28] are used. Further, the four feature selection techniques i.e.CFS, Infogain, Relief, and Gain ratio [29] are applied with these classifiers on four datasets that are used for developing a prediction model which is described below.

Naive Bayes

It is one of the fundamental approaches in the domain of machine learning [30–32] valued for its simplicity and effectiveness in various classification tasks [33]. It applies Bayes' theorem to generate accurate forecasts based on the probability of specific categories corresponding to observed data values. Naive Bayes is a probabilistic approach used for categorization and other tasks, based on the theorem of Bayes with the "naive" hypothesis that features are autonomous.

Logistic Regression

The logistic regression approach is a type of statistical method for accomplishing binary classification tasks. It is an effective technique for anticipating the possibility of an instance belonging to a given class [34].

Neural Network

A neural network classifier comprises a machine learning model that categorizes data incorporating artificial neural networks [35]. Neural networks constitute computations that utilize the structure and operation of the human brain, composed of connected nodes, frequently referred to as neurons, stacked in layers.

Random Forest

Random Forest (RF) is an approach that develops an ecosystem with numerous decision trees. The fundamental concept underlying the emergence of RF is the blend of forests and elections [36]. In RF, each decision tree in the forest serves as a candidate, and the final forecasts are produced by incorporating the predictions and votes submitted by each individual tree.

Support Vector Machine (SVM)

Support Vector Machine (SVM) is a prevalent agile, and robust supervised machine learning technological advance that is often used for prediction challenges.SVM works by separating data points in an n-dimensional feature space, often using a non-linear kernel function to handle complex relationships between data points. Hyperplanes are constructed using a tagged training Cement logistics dataset to separate the feature space into severity categories [37]. This allows SVM to assign a new category to labeled classes in a prediction dataset.

J48

The J48 classifier, additionally referred to as C4.5, is a decision tree approach originally developed by Ross Quinlan. It has been constantly employed for classification problems due to it is intuitiveness, interpretability, and success in building decision trees from data [38].

Kstar

K* is an instance-based classifier, which suggests that the class of a test instance is determined by the class of training examples that are similar to it, as determined by some similarity function. It is distinguished from other instance-based learners by its use of an entropy-based distance function [39].

IBK

The IBK (Instance-Based learning with k-nearest neighbors) classifier is a simple and intuitive machine learning algorithm for classification tasks. It makes predictions for new instances by comparing them to the nearest neighbors in the training dataset [40]. It is a simple instance-based classification algorithm that makes predictions based on the majority class of the k nearest training instances to a given test instance.

Experimental Results

This section presents the experimental results of the work. The section is divided into the following subsections as demonstrated below:

- Performance measures.
- Prediction result of proposed methodology.
- Friedman statistical test.
- Threats to validity.

Performance Measures

An evaluation of the proposed cement logistics prediction model is conducted based on a variety of performance parameters, including Accuracy, Precision, Sensitivity, and *F*-value [41].

Accuracy

Accuracy is defined as the number of correctly predicted samples divided by the total number of samples in the proposed system. It is calculated using the Eq. (1):

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

where TP: True Positive, TN : True Negative, FP: False Positive, FN: False Negative.

Precision

Precision is the ratio of accurately predicted positive samples to the total number of positive samples, including FP samples, and is determined by the following Eq. (2):

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

Table 1 Result analysis comparison of Naive Bayes with feature selection methods

Dataset	Methods	Precision	Sensitivity	Accuracy	F1 Score
D1	CFS+NB	0.9913	0.7454	0.7406	0.851
	Infogain+NB	0.9670	0.7454	0.5471	0.8542
	Relief+NB	p.9320	0.6454	0.5454	0.8642
	Gainratio+NB	0.9905	0.7454	0.6401	0.8507
D2	CFS+NB	0.4609	0.0121	0.6282	0.0236
	Infogain+NB	0.9699	0.3831	0.4101	0.5493
	Relief+NB	0.3064	0.6432	0.4293	0.4199
	Gainratio+NB	0.8699	0.3431	0.4101	0.5493
D3	CFS+NB	0.9913	0.7454	0.7106	0.851
	Infogain+NB	0.976	0.7657	0.7562	0.8581
	Relief+NB	0.9511	0.7593	0.7369	0.8444
	Gainratio+NB	0.9723	0.7454	0.7454	0.8542
D4	CFS+NB	0.9957	0.9696	0.6667	0.9825
	Infogain+NB	0.9957	0.9696	0.5666	0.9825
	Relief+NB	0.9723	0.7654	0.7454	0.8542
	Gainratio+NB	0.9541	0.9689	0.6928	0.9615

Table 2 Result analysis comparison of logistic regression with feature selection methods

Dataset	Methods	Precision	Sensitivity	Accuracy	F1 Score
D1	CFS+LR	0.9582	0.7454	0.7324	0.8385
	Infogain+LR	0.9782	0.7454	0.7471	0.8542
	Relief+LR	0.9541	0.9689	0.4928	0.9615
	Gainratio+LR	0.9913	0.7454	0.7406	0.851
D2	CFS+LR	0.959	0.7497	0.7366	0.8415
	Infogain+LR	0.9588	0.7484	0.7378	0.8406
	Relief+LR	0.9957	0.9696	0.6667	0.9825
	Gainratio+LR	0.9599	0.7539	0.7432	0.8446
D3	CFS+LR	0.8389	0.9685	0.8259	0.8991
	Infogain+LR	0.8464	0.9701	0.8333	0.904
	Relief+LR	0.8572	0.9725	0.8443	0.9113
	Gainratio+LR	0.8656	0.9743	0.5852	0.9167
D4	CFS+LR	0.8905	0.9796	0.8779	0.9329
	Infogain+LR	0.8891	0.9793	0.8765	0.932
	Relief+LR	0.7838	0.9553	0.7699	0.8611
	Gainratio+LR	0.9733	0.9589	0.5388	0.966

Sensitivity (True Positive Rate)

Sensitivity, also known as True Positive Rate (TPR) or Recall, quantifies the fraction of actual positive events that the model adequately predicts as positive. It is calculated using the following Eq. (3):

$$Sensitivity = \frac{TP}{TP + FN} \tag{3}$$

F-Value

The harmonic mean of recall and precision is defined as the F-value and is calculated in Eq. (4):

$$F - value = 2 * \frac{Recall * Precision}{Recall + Precision} \tag{4}$$

Prediction Result of Proposed Methodology

In this section, the experimental results of the proposed DAAESF methodology are illustrated. Eight classifiers, including Naïve Bayes, Logistic Regression, Neural Network, Random Forest, J48, IBK, SVM, and K* are used to predict cement sales according to the season. Additionally, four feature selection techniques, namely CFS, Infogain, Relief, and Gain Ratio, are applied with these classifiers on four datasets, as shown in Tables 1, 2, 3, 4, 5, 6, 7, and 8.

QR1: Which Classification technique gives the best performance?

In this study, we have used 8 Classifiers such as Naïve Bayes, Logistic regression, Neural Network, Random Forest, J48, IBK, SVM, and K Star for predicting the seasonality of cement sales. Further, these classification techniques are assimilated with different feature selection methods. These methods are Infogain, Gainratio, Relief, and CFS. Performance of the system is measured using various performance parameters such as Precision, Sensitivity, Accuracy, and F1 Score but we considered accuracy as the critical parameter. The experimental results are shown in Tables 1, 2, 3, 4, 5, 6, 7, and 8. The graphical representation of the performance of these classifiers is shown in Figs. 3, 4, 5, 6, 7, 8, 9, and 10. The experimentation results stated that Neural Networks obtained better results than the other classification techniques. From Fig. 3. it has been concluded that the accuracy results of CFS and Neural Network, Infogain and Neural network, Relief and Neural Network, and Gainratio and Neural Network are 97%, 97.75%, 94.93%, 93.24% respectively for Dataset 1. For Dataset 2 the accuracy results are 96.5%, 94.98%, 96.67%, and 94.32% respectively. For Dataset 3 the accuracy results are 92.59%, 96.03%, 96.65%, and 94.12% respectively and for Dataset 4 the accuracy results are 96.67%, 98.06%, 93.66%, and 73.78% respectively. Cooperating the Neural Network with different feature selection techniques can improve the results of sales prediction by approximately 2% from SVM,

Table 3 Result analysis comparison of neural network with feature selection methods

Dataset	Methods	Precision	Sensitivity	Accuracy	F1 Score
D1	CFS+NN	0.9879	0.9806	0.9731	0.9842
	Infogain+NN	0.8903	0.979	0.8775	0.9325
	Relief+NN	0.9746	0.9643	0.9439	0.9694
	Gainratio+NN	0.9582	0.7454	0.9324	0.8385
D2	CFS+NN	0.9725	0.9747	0.9494	0.9736
	Infogain+NN	0.9768	0.9706	0.9498	0.9737
	Relief+NN	0.9957	0.9696	0.9667	0.9825
	Gainratio+NN	0.9599	0.7539	0.9432	0.8446
D3	CFS+NN	0.8389	0.9685	0.9259	0.8991
	Infogain+NN	0.9817	0.9768	0.9603	0.9792
	Relief+NN	0.9776	0.986	0.9665	0.9818
	Gainratio+NN	0.9687	0.9687	0.9412	0.9687
D4	CFS+NN	0.9957	0.9696	0.9667	0.9825
	Infogain+NN	0.9913	0.7454	0.7406	0.851
	Relief+NN	0.959	0.7497	0.9366	0.8415
	Gainratio+NN	0.9588	0.7484	0.9378	0.8406

Table 4 Result analysis comparison of random forest with feature selection methods

Dataset	Methods	Precision	Sensitivity	Accuracy	F1 Score
D1	CFS+RF	0.9913	0.7454	0.7406	0.851
	Infogain+RF	.8956	0.7454	0.7471	0.8136
	Relief+RF	0.9625	0.9801	0.6457	0.9712
	Gainratio+RF	0.9683	0.9651	0.6376	0.9667
D2	CFS+RF	0.9776	0.986	0.7532	0.9818
	Infogain+RF	0.8389	0.9685	0.8259	0.8991
	Relief+RF	0.9817	0.9768	0.5603	0.9792
	Gainratio+RF	0.9776	0.9860	0.6965	0.9818
D3	CFS+RF	0.9777	0.9754	0.9552	0.9765
	Infogain+RF	0.9804	0.9747	0.9569	0.9775
	Relief+RF	0.9783	0.9721	0.8531	0.9752
	Gainratio+RF	0.9913	0.7454	0.7406	0.851
D4	CFS+RF	0.959	0.7497	0.7366	0.8415
	Infogain+RF	0.9721	0.9783	0.9531	0.9752
	Relief+RF	0.9776	0.986	0.5965	0.9818
	Gainratio+RF	0.8389	0.9685	0.8259	0.8991

Table 5 Result analysis comparison of J48 with feature selection methods

Dataset	Methods	Precision	Sensitivity	Accuracy	F1 Score
D1	CFS+J48	0.978	0.9716	0.7369	0.9748
	Infogain+J48	0.9231	0.9175	0.8565	0.9203
	Relief+J48	0.8248	0.8134	0.7111	0.819
	Gainratio+J48	.8956	0.7454	0.7471	0.8136
D2	CFS+J48	0.4609	0.0121	0.6282	0.0236
	Infogain+J48	0.9699	0.3831	0.4101	0.5493
	Relief+J48	0.8799	0.4331	0.3911	0.6793
	Gainratio+J48	0.9699	0.3831	0.4101	0.5493
D3	CFS+J48	0.894	0.8865	0.8148	0.8903
	Infogain+J48	0.9688	0.897	0.8777	0.9315
	Relief+J48	0.9808	0.9639	0.7487	0.9723
	Gainratio+J48	0.9799	0.9623	0.5948	0.971
D4	CFS+J48	0.9812	0.9646	0.9506	0.9728
	Infogain+J48	0.9799	0.9622	0.9514	0.971
	Relief+J48	0.894	0.8865	0.8148	0.8903
	Gainratio+J48	0.9688	0.897	0.8777	0.9315

32% from Naïve Bayes, 21% from Logistic Regression, 19% from Random Forest, 24.02% from J48, 20.89% from IBK and 25% from K Star. This is because a Neural Network represents the complex relationship between the input and output. Further, it has also been observed that Naïve Bayes provides the worst results for sales prediction. The accuracy of results of CFS and Naïve Bayes, Infogain and Naïve Bayes, Relief and Naïve Bayes, and Gainratio and Naïve Bayes is 74.06%, 54.71%, 54.54% and 64.01% respectively for Dataset 1. For Dataset 2 the accuracy results are 62.82%, 41.01%, 42.93%, and 41.01% respectively. For Dataset 3 the accuracy results are 71.06%, 75.62%, 73.69%, and 74.54% respectively and for Dataset 4 the accuracy results are 66.67%, 56.56%, 74.71%, and 69.28% respectively. This is due to the different objective functions used by Naïve Bayes for predicting cement sales.

QR2: Is there any effect of Feature selection technique on performance measure?

Yes, the choice of feature selection technique can have a significant effect on the performance measure of a prediction model. Feature selection plays a crucial role in optimizing

Table 6 Result analysis comparison of IBK with feature selection methods

Dataset	Methods	Precision	Sensitivity	Accuracy	F1 Score
D1	CFS+IBK	0.8218	0.8084	0.722	0.8151
	Infogain+IBK	0.9793	0.9726	0.9758	0.9759
	Relief+IBK	0.9824	0.9104	0.7405	0.9451
	Gainratio+IBK	0.9817	0.9768	0.5603	0.9792
D2	CFS+IBK	0.9776	0.986	0.7934	0.9818
	Infogain+IBK	0.9777	0.9754	0.9552	0.9765
	Relief+IBK	0.9804	0.9747	0.7569	0.9775
	Gainratio+IBK	0.9913	0.7454	0.7406	0.851
D3	CFS+IBK	0.976	0.7657	0.7562	0.8581
	Infogain+IBK	0.9511	0.7593	0.7369	0.8444
	Relief+IBK	0.9913	0.7454	0.7406	0.851
	Gainratio+IBK	0.959	0.7497	0.7366	0.8415
D4	CFS+IBK	0.9588	0.7484	0.7378	0.8406
	Infogain+IBK	0.9957	0.9696	0.9667	0.9825
	Relief+IBK	0.9957	0.9696	0.8666	0.9825
	Gainratio+IBK	0.8956	0.7454	0.7471	0.8136

Table 7 Result analysis comparison of SVM with feature selection methods

Dataset	Methods	Precision	Sensitivity	Accuracy	F1 Score
D1	CFS+SVM	0.9582	0.7454	0.9524	0.8385
	Infogain+SVM	0.962	0.7799	0.7644	0.8614
	Relief+SVM	0.9529	0.7393	0.9505	0.8326
	Gainratio+SVM	0.8825	0.7393	0.9788	0.8046
D2	CFS+SVM	0.9768	0.9706	0.9498	0.9737
	Infogain+SVM	0.9957	0.9696	0.9667	0.9825
	Relief+SVM	0.9599	0.7539	0.9432	0.8446
	Gainratio+SVM	0.8389	0.9685	0.9259	0.8991
D3	CFS+SVM	0.959	0.7497	0.9366	0.8415
	Infogain+SVM	0.9588	0.7484	0.9378	0.8406
	Relief+SVM	0.959	0.7497	0.9366	0.8415
	Gainratio+SVM	0.9588	0.7484	0.7378	0.8406
D4	CFS+SVM	0.8926	0.7583	0.9799	0.82
	Infogain+SVM	0.958	0.8893	0.8593	0.9224
	Relief+SVM	0.9611	0.8938	0.9668	0.9262
	Gainratio+SVM	0.5124	0.8821	0.9506	0.9153

the model's overall performance. Therefore in this section, we are checking the impact of four feature selection techniques on eight prediction techniques. The performance comparison of prediction techniques with and without feature selection technique is demonstrated in Figs. 11, 12, 13, 14, 15, 16, 17 and 18.

The analysis presented in Fig. 11 reveals noteworthy insights regarding the performance of Neural Network models augmented with different feature selection techniques, namely CFS (Correlation-based Feature Selection), Infogain, Relief, and Gainratio, across four distinct datasets (Dataset 1, Dataset 2, Dataset 3, and Dataset 4). The results indicate substantial improvements in accuracy when employing feature selection techniques in conjunction with the Neural Network. Specifically, for Dataset 1, the accuracy is enhanced by approximately 21.93%, 22.68%, 19.32%, and 18.17% with CFS, Infogain, Relief, and Gainratio, respectively. Similarly, for Dataset 2, the improvements amount to 22.75%, 21.23%, 22.92%, and 20.57%, respectively. For Dataset 3, the accuracies are boosted by 18.84%, 22.28%, 22.9%, and 20.37%, respectively, while for Dataset 4, the improvements stand at 21.6%, 22.9%, and 18.59% when utilizing CFS, Infogain, and Relief. However, the accuracy results degraded when we combined Gainratio with a neural network in Fig. 12. However, without feature selection, the accuracy

Table 8 Result analysis comparison of KStar with feature selection methods

Dataset	Methods	Precision	Sensitivity	Accuracy	F1 Score
D1	CFS+KStar	0.956	0.7866	0.7792	0.8631
	Infogain+Kstar	0.4609	0.0121	0.6282	0.0236
	Relief+Kstar	0.9699	0.3831	0.4101	0.5493
	Gainratio+Kstar	0.8799	0.4331	0.3911	0.6793
D2	CFS+KStar	0.9812	0.9646	0.8506	0.9728
	Infogain+Kstar	0.9799	0.9622	0.9514	0.971
	Relief+Kstar	0.9712	0.9546	0.8506	0.9728
	Gainratio+Kstar	0.9799	0.9622	0.8514	0.971
D3	CFS+KStar	0.9086	0.9627	0.884	0.9348
	Infogain+Kstar	0.8799	0.4331	0.3911	0.6793
	Relief+Kstar	0.9812	0.9646	0.6506	0.9728
	Gainratio+Kstar	0.9576	0.8969	0.8706	0.9262
D4	CFS+KStar	0.9694	0.9244	0.899	0.9464
	Infogain+Kstar	0.8799	0.4331	0.3911	0.6793
	Relief+Kstar	0.9699	0.3831	0.4101	0.5493
	Gainratio+Kstar	0.9898	0.8506	0.8454	0.9149

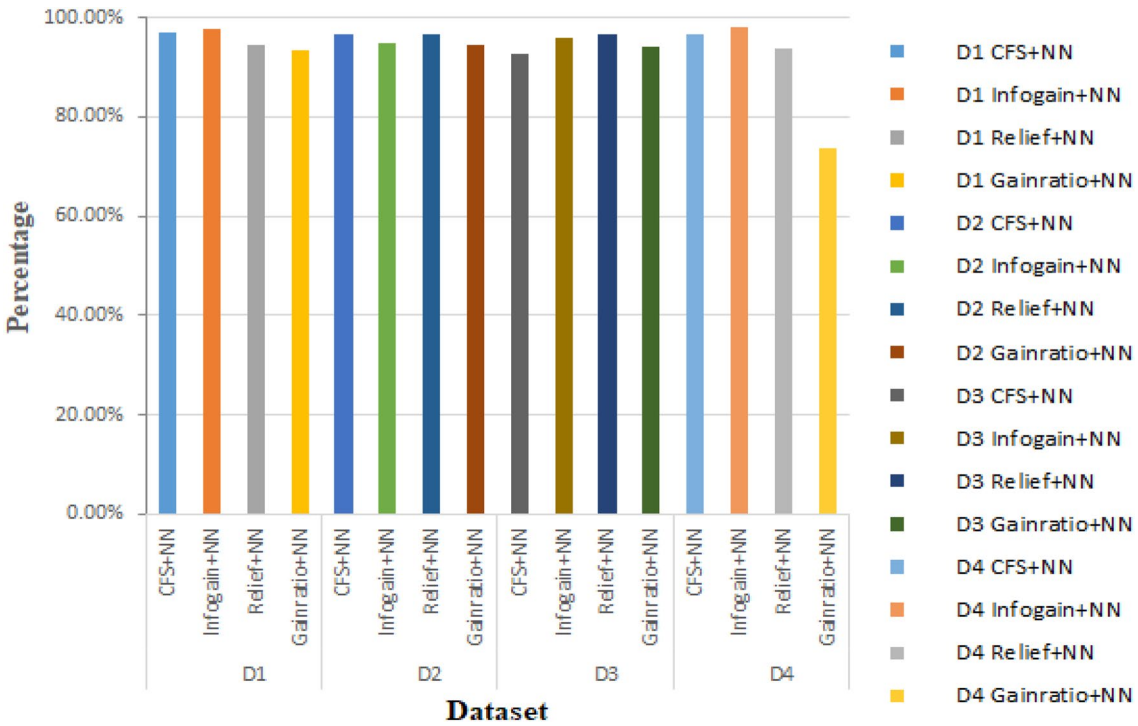

Fig. 3 Performance comparison of neural network with different feature selection techniques using accuracy parameter

is 75.07%, 73.75%, 73.75%, and 75.07%. The outcomes derived from the Naïve Bayes classifier exhibit notable variations in their performance when paired with distinct feature extraction techniques across the datasets under consideration. For Dataset 1 and Dataset 2, it is evident that the accuracy of the Naïve Bayes classifier experiences a significant boost solely with the integration of the CFS feature extraction method. Conversely, the scenario alters for Dataset 3 and Dataset 4, where the accuracy of the Naïve Bayes classifier displays improvement through the utilization of not only the CFS technique but also the Infogain, Relief, and Gainratio feature extraction methodologies. The results obtained from employing the Logistic Regression classifier in Fig. 13 reveal intriguing variations in its performance when paired with distinct feature extraction techniques CFS, Infogain, and Relief,

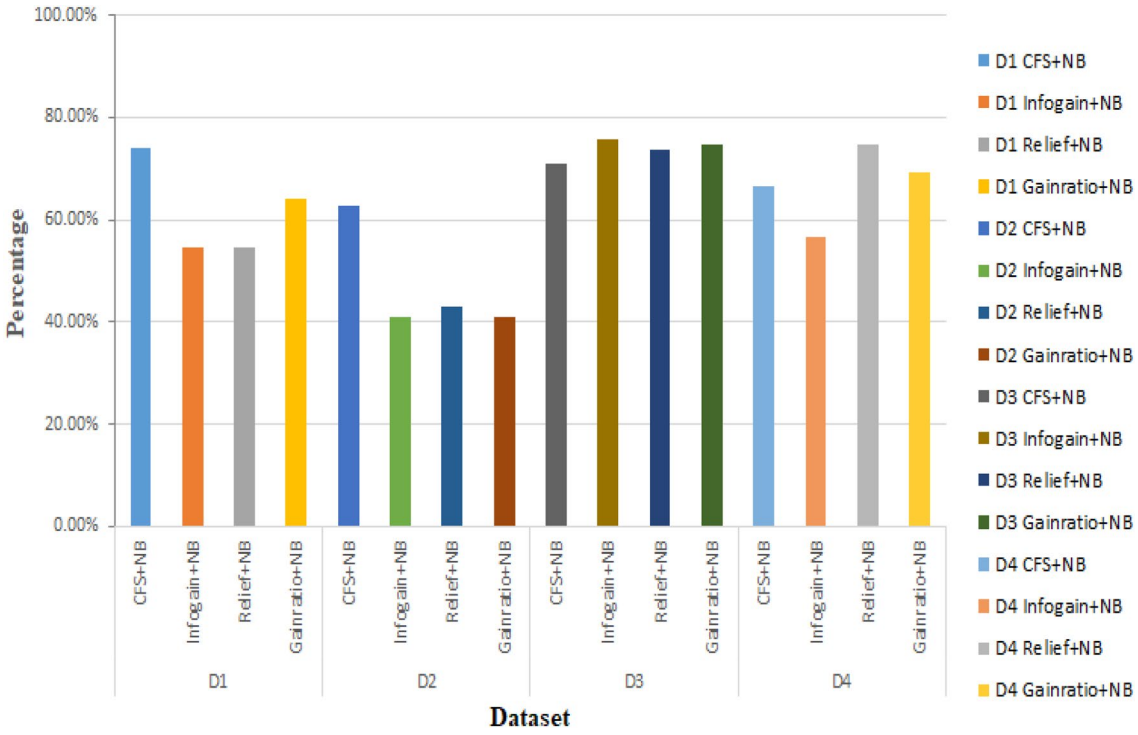

Fig. 4 Performance comparison of Naïve Bayes with different feature selection techniques using accuracy parameter

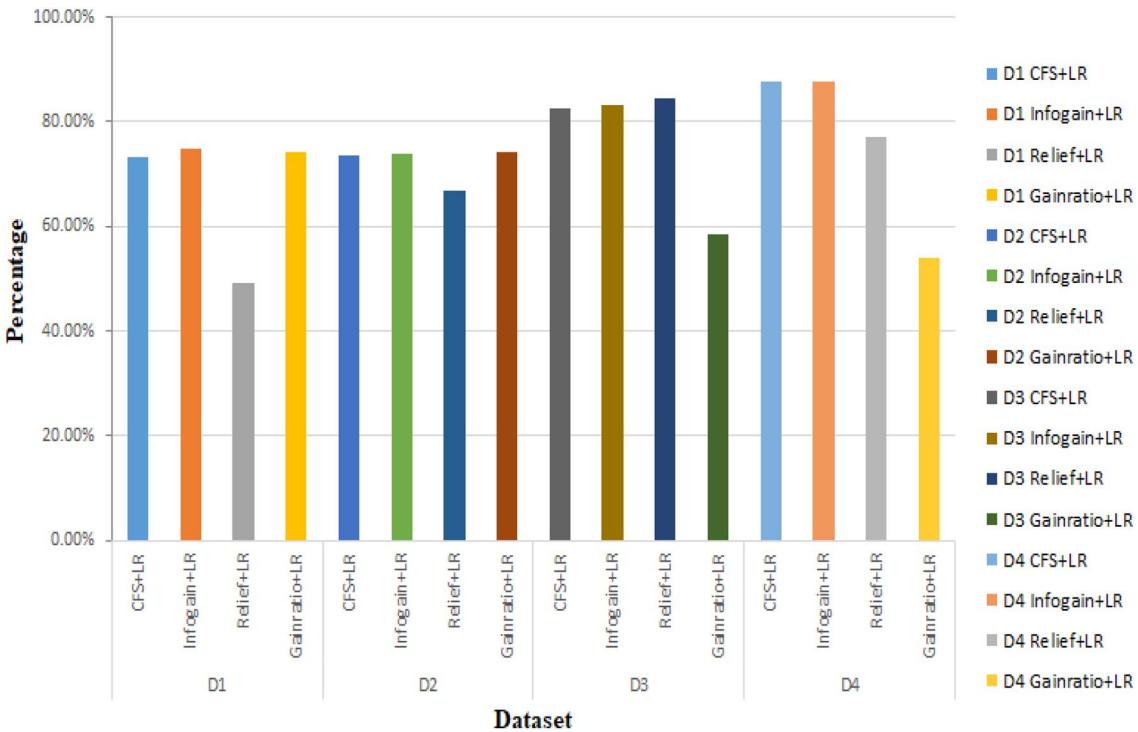

Fig. 5 Performance comparison of logistic regression with different feature selection techniques using accuracy parameter

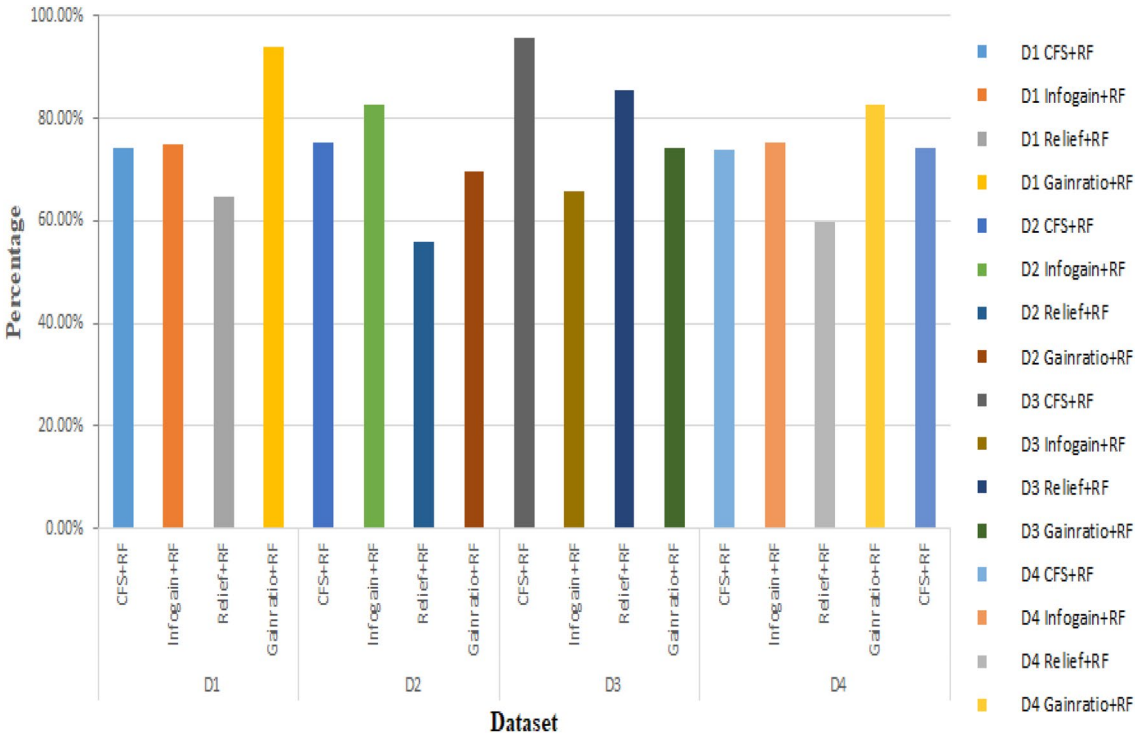

Fig. 6 Performance comparison of random forest with different feature selection techniques using accuracy parameter

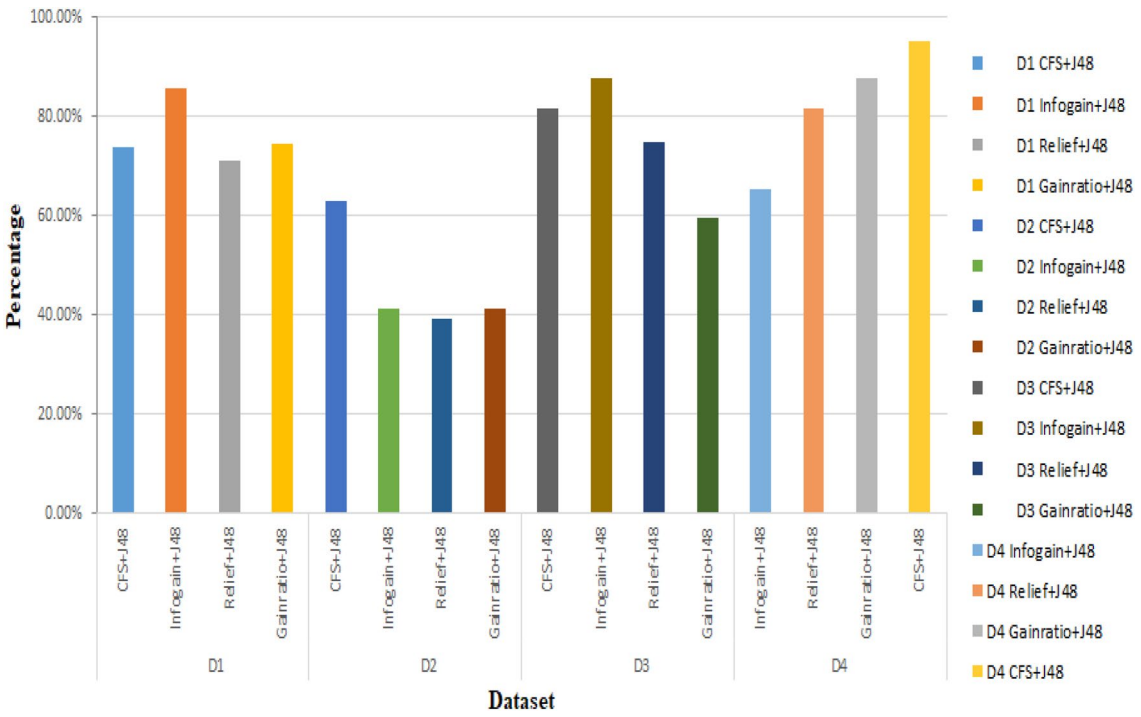

Fig. 7 Performance comparison of J48 with different feature selection techniques using accuracy parameter

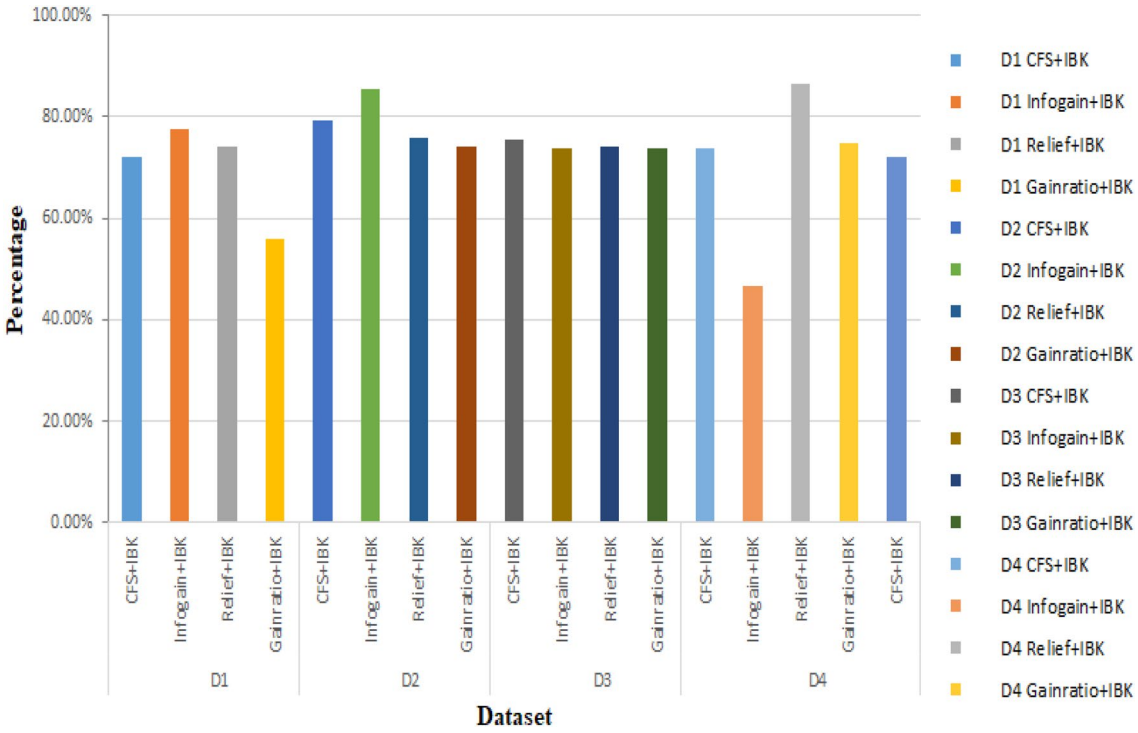

Fig. 8 Performance comparison of IBK with different feature selection techniques using accuracy parameter

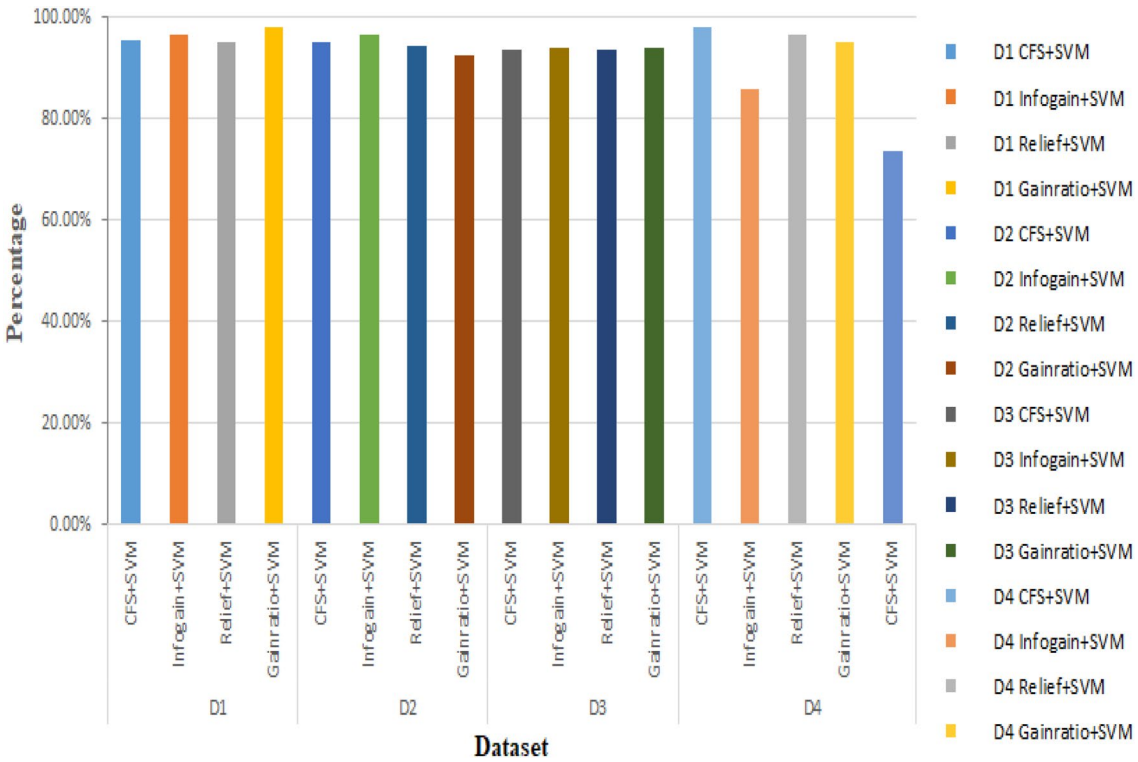

Fig. 9 Performance comparison of SVM with different feature selection techniques using accuracy parameter

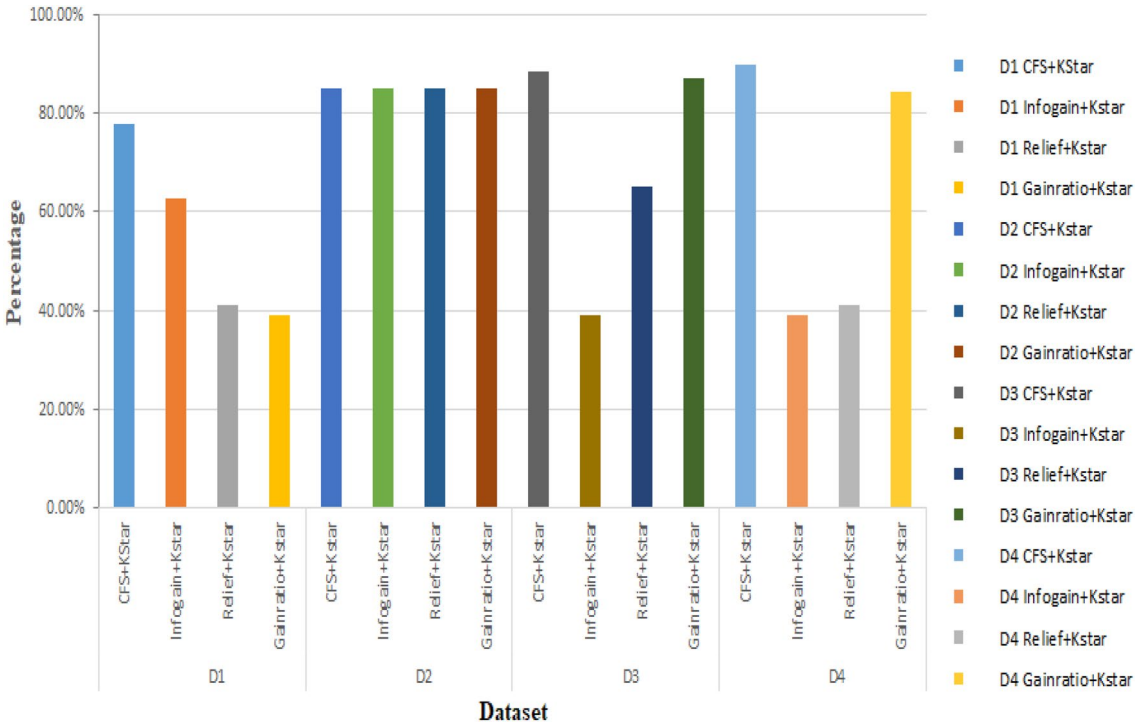

Fig. 10 Performance comparison of KStar with different feature selection techniques using accuracy parameter

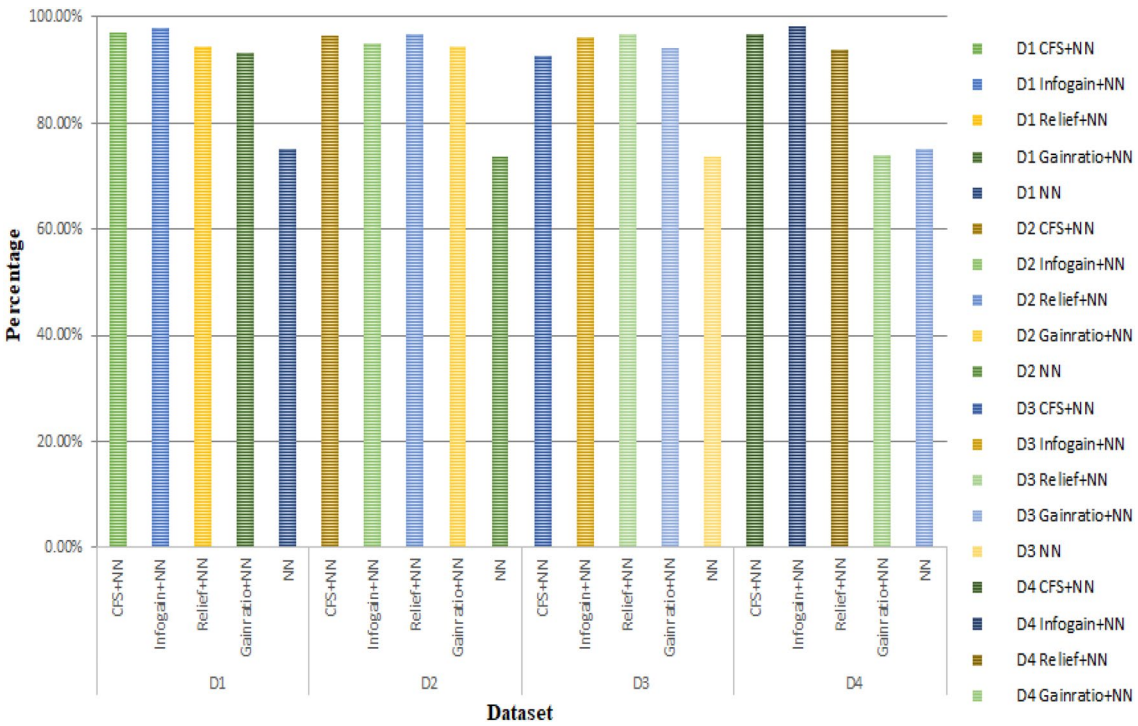

Fig. 11 Performance comparison of neural network with and without different feature selection techniques using accuracy parameter

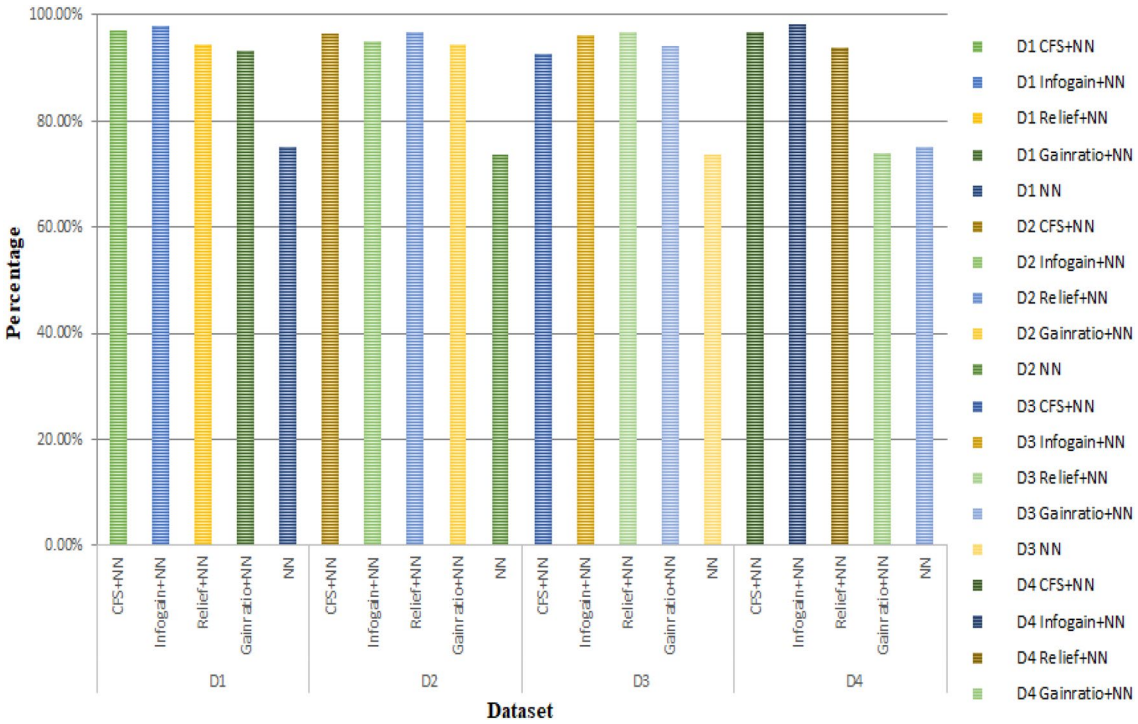

Fig. 12 Performance comparison of Naïve bayes with and without different feature selection techniques using accuracy parameter

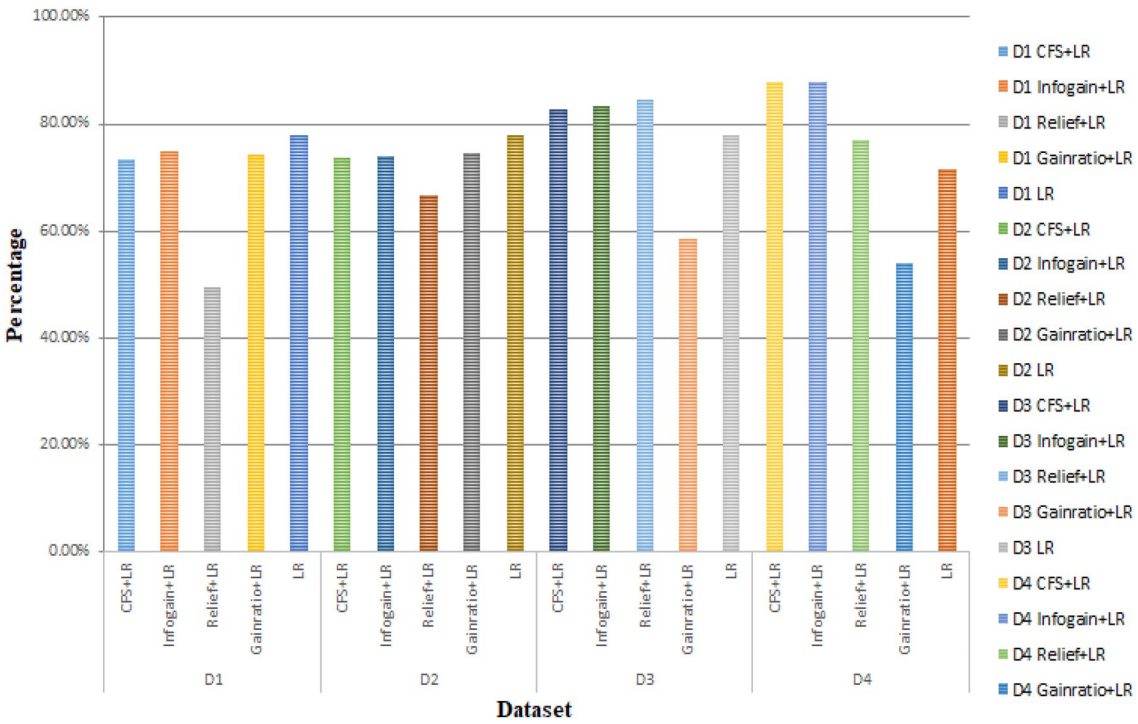

Fig. 13 Performance comparison of logistic regression with and without different feature selection techniques using accuracy parameter

particularly in the context of Dataset 3 and Dataset 4 but in the case of Dataset 1 and 2 the Logistic regression works better with accuracy 77.8%. In the context of the Random Forest algorithm, different feature selection techniques demonstrate varying levels of effectiveness depending on the specific dataset they are applied as shown in Fig. 14.

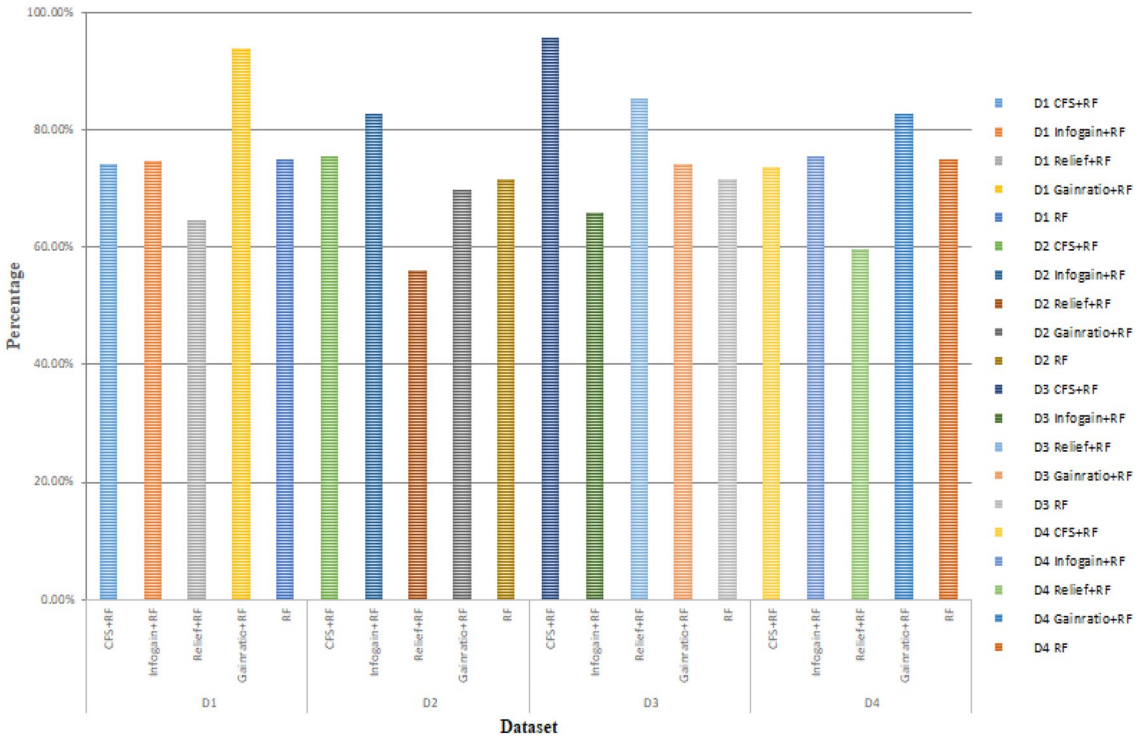

Fig. 14 Performance comparison of random forest with and without different feature selection techniques using accuracy parameter

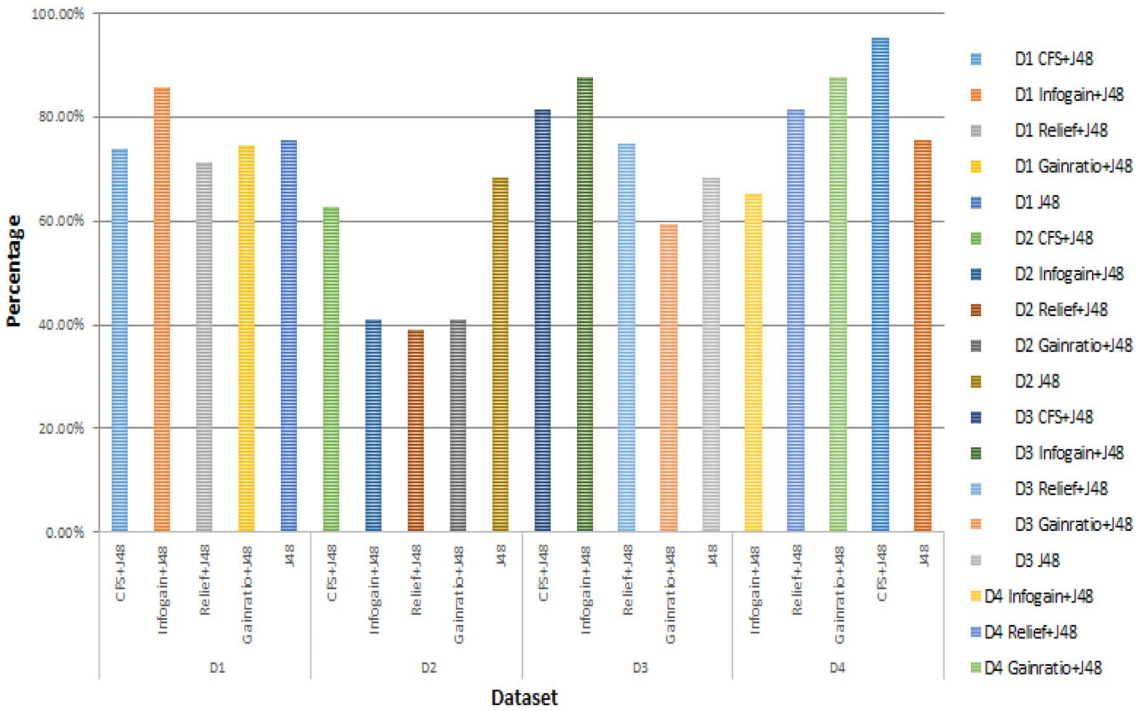

Fig. 15 Performance comparison of J48 with and without different feature selection techniques using accuracy parameter

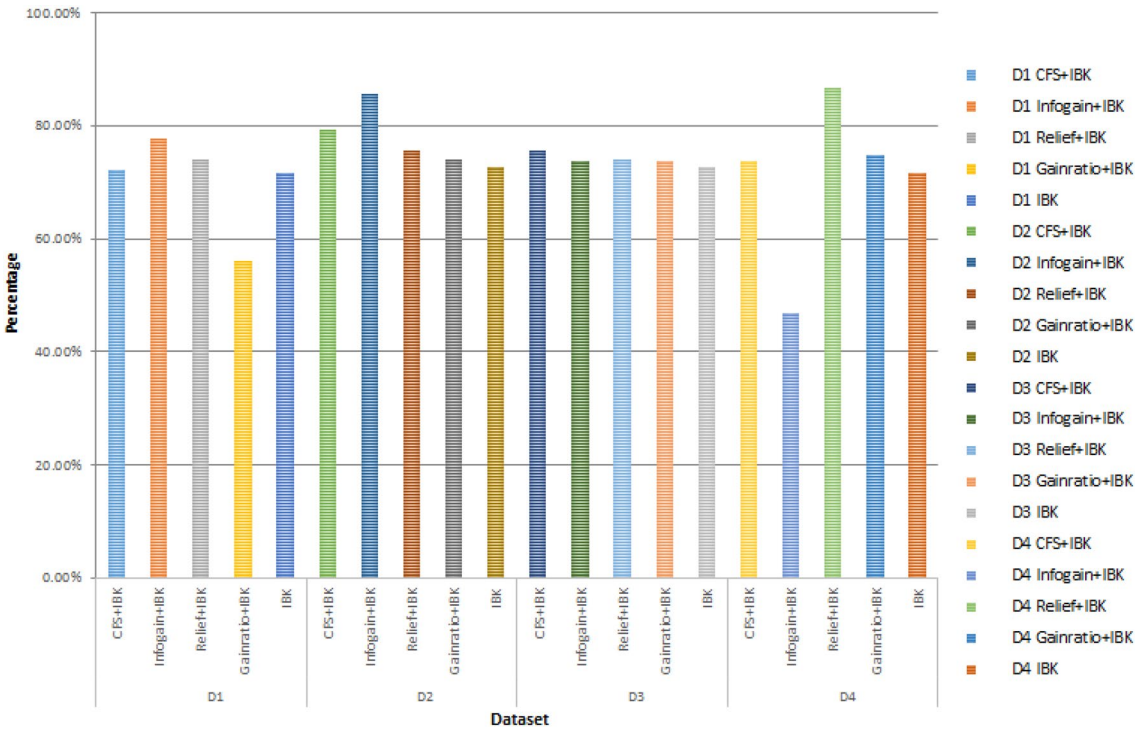

Fig. 16 Performance comparison of IBK with and without different feature selection techniques using accuracy parameter

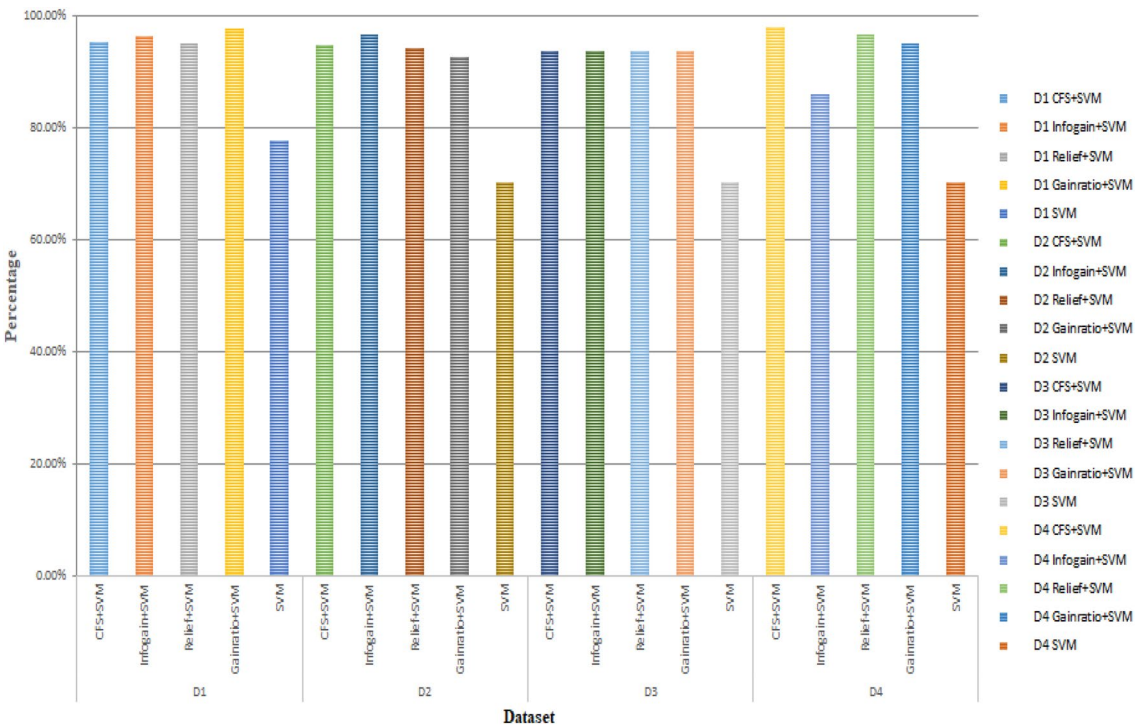

Fig. 17 Performance comparison of SVM with and without different feature selection techniques using accuracy parameter

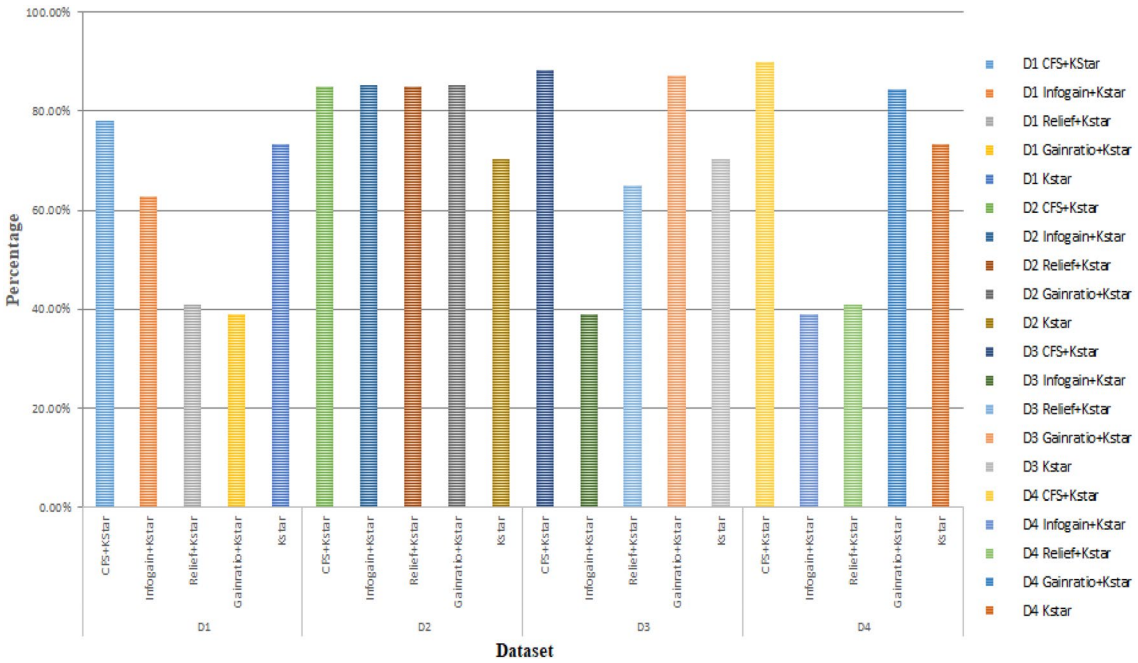

Fig. 18 Performance comparison of KStar with and without different feature selection techniques using accuracy parameter

For instance, when dealing with dataset 1, the gain ratio feature selection technique performs well, yielding satisfactory results. However, in the case of dataset 2, both CFS (Correlation-based Feature Selection) and InfoGain (Information Gain) prove to be more advantageous, outperforming other methods. Moving on to dataset 3 CFS, InfoGain, and GainRatio emerge as the superior approach for feature selection. This combination of techniques exhibits better performance compared to other alternatives when dealing with the characteristics of the third dataset. Lastly, when faced with dataset 4, it is observed that InfoGain and GainRatio once again show strong performance, making them the most suitable feature selection techniques for this particular dataset. In the J48 classifier, the accuracy rate is increased only when used with infogain as shown in Fig. 15 i.e. 85.65% for Dataset 1. For dataset 2 the feature selection technique decreased the performance of J48. For Dataset 3 the accuracy of J48 is increased when applied with CFS, Infogain and relief. In case of Dataset 4 J48 gives the worst result in the case of Infogain feature

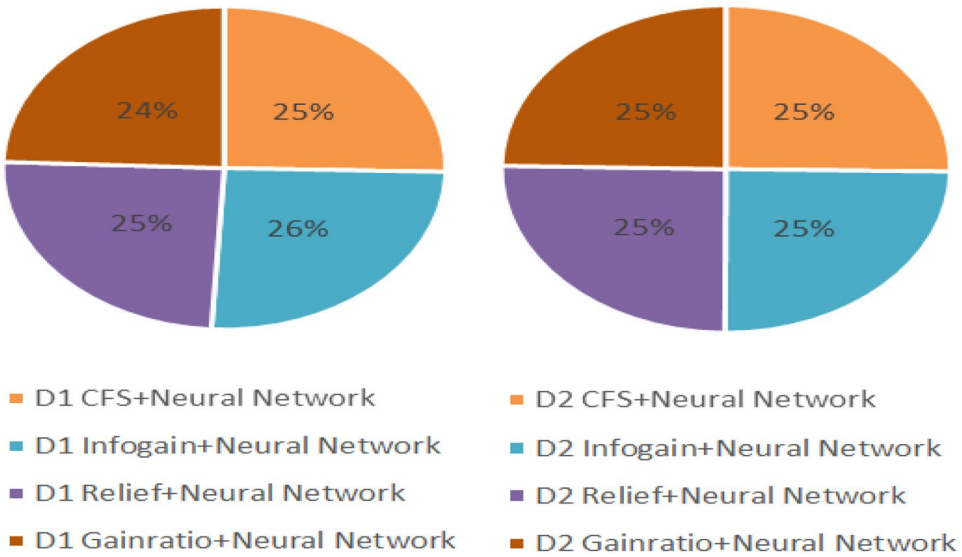

Fig. 19 Experimental result of neural network with different feature selection techniques using Dataset 1 and Dataset 2

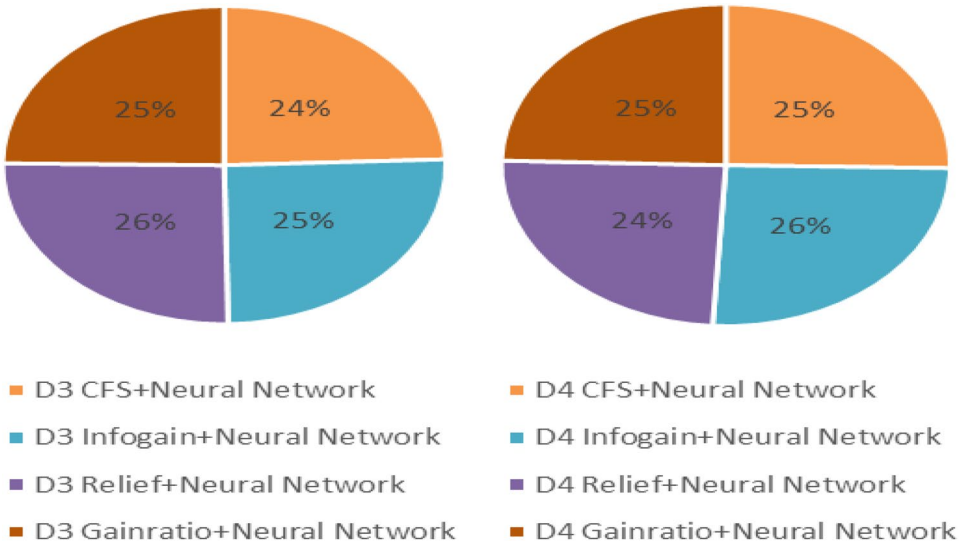

Fig. 20 Experimental result of neural network with different feature selection techniques using Dataset3 and Dataset 4

Table 9 Average ranking of techniques using Friedman test based on accuracy parameter of feature selection technique 1 (CFS)

Classifier	NN	NB	LR	RF	J48	IBK	SVM	KSTAR
Rank	7.25	2	3.25	4.625	3.375	2.75	7.25	5.5

Table 10 Statistics of Friedman test based on accuracy parameter of feature selection technique 1 (CFS)

Statistical value	Critical value	Hypothesis
18.9671	14.0671	Rejected

selection. For dataset 1, IBK's accuracy improves notably when combined with CFS, Infogain, and Relief feature selection techniques. Dataset 2 and 3 demonstrate superior accuracy when feature selection techniques are employed. However, for dataset 4, IBK performs poorly with the

Table 11 Average ranking of techniques using Friedman test based on accuracy parameter of feature selection technique 2 (Infogain)

Classifier	NN	NB	LR	RF	J48	IBK	SVM	KSTAR
Rank	7.75	2.375	4.625	3.625	4.375	4	7	2.25

Table 12 Statistics of Friedman test based on accuracy parameter of feature selection technique 2 (Infogain)

Statistical value	Critical value	Hypothesis
18.4012	14.0671	Rejected

Infogain feature selection technique as shown in Fig. 16. The findings reveal significant enhancements in accuracy by incorporating feature selection techniques in conjunction with the Support Vector Machine (SVM) model as shown in Fig. 17. Specifically, when analyzing Dataset 1, it is observed that remarkable accuracy increases by approximately 17.5%, 18.66%, 17.27%, and 20.1% with the adoption of CFS, Infogain, Relief, and Gainratio, respectively. Similarly, for Dataset 2, the improvements amounted to 24.61%, 26.3%, 23.95%, and 22.22%, respectively. Moving on to Dataset 3, the accuracies experienced substantial boosts of 23.29%, 23.41%, 23.29%, and 23.41%, respectively, while for Dataset 4, the improvements were noteworthy, standing at 27.6%, 15.56%, 26.31%, and 24.69% when utilizing CFS, Infogain, Relief, and Gainratio. In case of Kstar the accuracy rate is only enhanced with CFS as shown in Fig. 18.

QR3: Which feature selection technique gives the best results?

In RQ1 we have concluded that Neural Network gives the best prediction accuracy among all classifiers. Therefore in this work, we analysed which feature selection technique works best with neural networks (Figs. 19, 20).

Selection techniques using four datasets Pie charts serve as a valuable instrument for visually representing proportional data and affording a rapid overview of datasets. In

Table 13 Average ranking of techniques using Friedman test based on accuracy parameter of feature selection technique 3 (Relief)

Classifier	NN	NB	LR	RF	J48	IBK	SVM	KSTAR
Rank	7.5	2.5	3.75	3.75	3.75	5	7.5	2.25

Table 14 Statistics of Friedman test based on accuracy parameter of feature selection technique 3 (Relief)

Statistical value	Critical value	Hypothesis
19.333	14.0671	Rejected

the context of this research, the simulation results of neural networks have been explored, utilizing distinct feature selection techniques across four datasets.

The findings reveal noteworthy variations in the performance of neural networks concerning the employed feature selection methods and the datasets under consideration. For instance, in dataset 1, the Gainratio technique yielded the most favorable outcomes for the neural network. Conversely, in dataset 2, a state where all combinations exhibit equal weightage was observed. Furthermore, in dataset 3, the Relief method showcased superior efficacy for the neural network, while in dataset 4, the Gainratio technique again emerged as the optimal choice.

Further, it can be cogently inferred that the outcomes of the neural network simulations exhibit significant dependency on the specific characteristics and attributes of the datasets and the efficacy of different feature selection techniques therein.

Friedman Statistical Test

This section delves into the statistical results of the Friedman test. To validate the technique's trustworthiness, thorough statistical analyses were carried out. These tests were carried out to identify notable differences in the performance of the various classifiers that integrate feature selection approaches. The findings of such significant distinctions attest to the suggested algorithm's uniqueness in comparison to established algorithms, hence substantiating the experimental and statistical efficacy of the proposed technique. As a result, in this work, the Friedman test is used to justify and confirm the effectiveness of new classifiers. Further, two hypotheses are designed, i.e., H_1 and H_0. The H_0 hypothesis suggests that Feature selection techniques have no impact on the performance of the classifiers. Conversely, the H_1 hypothesis posits that Feature selection techniques do influence the classifiers' performance. Table 9 shows the average ranks of several approaches based on their accuracy parameter performance in the case of feature selection technique 1 (CFS). According to the findings, the Neural Network and SVM techniques outperform the other methods, indicating greater efficacy. The Naive Bayes technique, on the other hand, falls behind the other approaches, with a significantly lower ranking. The statistical details of the Friedman test are shown in Table 10. These statistics are evaluated using a confidence level of 0.05 and 7 degrees of freedom. F (0.05; 7) represents the computed critical value of 14.067. As a result, the null hypothesis (H_0) is rejected which implies Feature selection techniques do influence the classifiers' performance. Similar results can be seen when the author applies other feature selection techniques which can be shown in Tables 11, 12, 13, 14, 15, and 16.

In addition, a post hoc test has been carried out to assess the efficacy of the SVM and NN techniques. The findings of the post-hoc test are referenced in Table 17, 18, 19 and 20 denote the significant difference that happened between the results of different techniques. It is seen that SVM and NN techniques are fundamentally unique from Naïve Bayes, Logistic regression, Random Forest, J48, IBK, and K Star techniques. Consequently, it tends to be expressed that the Post-hoc test validates the performance of the SVM and NN. The post hoc test splits all approaches into five distinct categories when the test has been performed according to the findings of the Friedman test in the case of Table 9. The groups of techniques are given as Group 1 comprises the NN and SVM techniques; Group 2 includes NB and IBK techniques; Group 3 comprises LR and J48 techniques; Group 4 includes the RF technique, Group 5 consists of the KStar technique. It becomes apparent that RF and Kstar techniques vary significantly from others. Whereas, the combination approaches do not exhibit significant modifications. While these methods perform similarly statistically, they provide distinct outcomes in experiments. Similarly, when the post-hoc test is used to analyze the Friedman test findings in the instance of Table 11, it separates all approaches into

Table 15 Average ranking of techniques using Friedman test based on accuracy parameter of feature selection technique 4 (Gainratio)

Classifier	NN	NB	LR	RF	J48	IBK	SVM	KSTAR
Rank	7.25	2.875	2.75	4.5	3.625	3	7.5	4.5

Table 16 Statistics of Friedman test based on accuracy parameter of feature selection technique 4 (Gainratio)

Statistical value	Critical value	Hypothesis
16.9045	14.0671	Rejected

four groups. The approaches have been classified into four groups: LR, IBK, and J48 techniques are in Group 3; NN and SVM techniques are in Group 2; NB and Kstar techniques are in Group 3; and RF techniques are in Group 4. It has been found that RF varies greatly from others. On the other hand, there are no significant distinctions in the combination approaches. It is shown from a statistical standpoint that different strategies function comparably. On the other hand, experimental analysis shows that various methods provide different results. When the post-hoc test is carried out to the Friedman test findings in the example of Table 13, it classifies all approaches into four groups. Group 1 comprises NN and SVM methods, Group 2 includes NB and Kstar techniques, Group 3 encompasses LR, RF, and J48 techniques, and Group 4 consists of IBK techniques. It becomes evident that IBK differs significantly from others. The combination strategies show no significant differences from a statistical perspective. However, in real-world experiments, these methods provide a range of results. Based on the Friedman test findings in Table 15, all strategies are split into four groups whenever the post-hoc test is used. NN and SVM methods make Group 1, NB and LR strategies constitute Group 2, RF and KStar techniques form up Group 3, and J48 and IBK techniques build up Group 4. These techniques perform similarly statistically, however they provide different outcomes in experimentation.

Hence, it can be stated that the NN and SVM approach is one of the most effective approaches for forecasting cement sales seasonality categorization since it has been shown that

Table 17 Results of post-hoc test after performing Friedman test by considering feature selection technique CFS

Techniques	NN	NB	LR	RF	J48	IBK	SVM	KSTAR
NN		S	S	S	S	S		S
NB	S		S	S	S		S	S
LR	S	S		S		S	S	S
RF	S	S	S		S	S	S	S
J48	S	S		S		S	S	S
IBK	S		S	S	S		S	S
SVM		S	S	S	S	S		S
KSTAR	S	S	S	S	S	S	S	S

Table 18 Results of post-hoc test after performing Friedman test by considering feature selection technique Infogain

Techniques	NN	NB	LR	RF	J48	IBK	SVM	KSTAR
NN		S	S	S	S	S		S
NB	S		S	S	S	S	S	
LR	S	S		S			S	S
RF	S	S	S		S	S	S	S
J48	S	S		S			S	S
IBK	S	S		S			S	S
SVM		S	S	S	S	S		S
KSTAR	S		S	S	S	S	S	

Table 19 Results of post-hoc test after performing Friedman test by considering feature selection technique relief

Techniques	NN	NB	LR	RF	J48	IBK	SVM	KSTAR
NN		S	S	S	S	S		S
NB	S		S	S	S	S	S	
LR	S	S				S	S	S
RF	S	S				S	S	S
J48	S	S				S	S	S
IBK	S	S	S	S	S		S	S
SVM		S	S	S	S	S		S
KSTAR	S		S	S	S	S	S	

Table 20 Results of post-hoc test after performing Friedman test by considering feature selection technique gainratio

Techniques	NN	NB	LR	RF	J48	IBK	SVM	KSTAR
NN		S	S	S	S	S		S
NB	S			S	S	S	S	S
LR	S			S	S	S	S	S
RF	S	S	S		S	S	S	
J48	S	S	S	S			S	S
IBK	S	S	S	S			S	S
SVM	S		S	S	S	S		S
KSTAR	S	S	S		S	S	S	

it considerably differentiates from other techniques while achieving the top ranking among all techniques.

Threats to Validity

The present research includes four datasets to assess the performance of various prediction techniques. However, a few possible challenges to validity should be addressed.

Threats to External Validity

The external validity of this research may be threatened by the generalizability of the findings. The study focused specifically on cement sales [42–44] forecasting, and the results might not be applicable to other industries or domains.

Threats to Internal Validity

The internal validity of the present research might be jeopardized by numerous factors. One potential challenge is the selection of the feature selection method for selecting the features and classifiers. In case if the selections were made inadequately or if there were biases in the selection process, it could lead to an overestimation or underestimation of the actual impact of these techniques on prediction accuracy. Additionally, the way in which the classifiers were implemented and the data preprocessing steps taken could introduce sources of bias or error. Any inconsistencies or errors in these processes could compromise the internal validity of the study.

Conclusion and Future Scope

The primary objective of this research was to evaluate how various feature selection techniques influence the accuracy of cement sales predictions. There are 20 attributes in total. The results show that the features are reduced from 20 to 11, 20 to 14, 20 to 12, and 20 to 10 in the case of CFS, Infogain, Relief, and GainRatio respectively. The results of our study demonstrate that these strategies represent an essential role in minimizing the impact of noise resulting from irrelevant features, so effectively distinguishing the essential features. Additionally, the concern of overfitting was dealt with, demonstrating evidence of the effectiveness of feature selection. The inclusion of these relevant features into the classifier resulted in a significant rise in the accuracy of cement forecasts for sales thereby exerting a notable influence on the supply chain sector. The implications of this study are significant and have an extensive impact, on cement manufacturers, distributors, and investors. By using suitable feature selection methods, sales forecasting can be improved. Precise sales forecasting improves demand planning, streamlines supply chain operations, and minimizes costs by reducing overstocking and stockouts. It also enables better investment decision-making, optimizes resource allocation, and promotes strategic product development, improving customer satisfaction, securing a competitive advantage, and promoting revenue growth.

As a result, this enhancement may initiate several kinds of benefits, including improved allocation of resources, enhanced inventory management, and more informed strategic marketing decision-making. As we anticipate the future, a promising opportunity exists for further study. Advanced feature selection approaches are emerging, offering the possibility of improved prediction models. Moreover, the use of numerous predictors shows the potential for improving prediction accuracy. Extending on this concept, our objective is to develop a composite model that utilizes the combined abilities of the predictors with the highest possible degree of accuracy. In the future, our aim is to develop an integrated feature selection framework and leverage advanced deep learning models to further refine and enhance the accuracy of our predictions.With this acquisition, our objective is to bring about a time frame of enhanced and accurate forecasts, boosting the quality of decision-making in the cement business.

Author Contributions All the authors have equally contributed to this work.

Funding No funding, grants, or other aid was received during the preparation of this manuscript.

Data Availability The data used to support the findings of this study are available from the corresponding author upon request.

Declarations

Conflict of Interest The authors declare that they have no conflict of interest.

References

1. Montoya-Torres JR, Muñoz-Villamizar A, Mejia-Argueta C. Mapping research in logistics and supply chain management during covid-19 pandemic. Int J Log Res Appl. 2023;26(4):421–41.
2. Kukkar A, Sharma A, Fan J, Zhang M. Data mining applications in university information management system development; 2022.
3. Singh S, Ramkumar K, Kukkar A. Machine learning techniques and implementation of different ml algorithms. In: 2021 2nd global conference for advancement in technology (GCAT). IEEE; 2021. pp. 1–6.
4. Kumar A, Kumar Y, Kukkar A. A feature selection model for prediction of software defects. Int J Embedded Syst. 2020;13(1):28–39.
5. Borucka A. Seasonal methods of demand forecasting in the supply chain as support for the company's sustainable growth. Sustainability. 2023;15(9):7399.
6. Feizabadi J. Machine learning demand forecasting and supply chain performance. Int J Log Res Appl. 2022;25(2):119–42.
7. Kaur G, Goyal S, Kaur H. Brief review of various machine learning algorithms. In: Proceedings of the international conference on innovative computing & communication (ICICC); 2021.
8. Goswami K, Kandali AB. Machine learning algorithms for predicting electrical load demand: an evaluation and comparison. Sādhanā. 2024;49(1):1–14.
9. Chatziloizos G-M, Gunopulos D, Konstantinou K. Deep learning for stock market prediction using sentiment and technical analysis. SN Comput Sci. 2024;5(5):446.
10. Chatziloizos G-M, Gunopulos D, Konstantinou K. Deep learning for stock market prediction using sentiment and technical analysis. SN Comput Sci. 2024;5(5):446.
11. Zhu X, Ninh A, Zhao H, Liu Z. Demand forecasting with supply-chain information and machine learning: evidence in the pharmaceutical industry. Prod Oper Manage. 2021;30(9):3231–52.
12. Pereira MM, Frazzon EM. A data-driven approach to adaptive synchronization of demand and supply in omni-channel retail supply chains. Int J Inf Manage. 2021;57: 102165.
13. Nguyen HD, Tran KP, Thomassey S, Hamad M. Forecasting and anomaly detection approaches using lstm and lstm autoencoder techniques with the applications in supply chain management. Int J Inf Manage. 2021;57: 102282.
14. Knoll D, Prüglmeier M, Reinhart G. Predicting future inbound logistics processes using machine learning. Proc CIRP. 2016;52:145–50.
15. Budak A, Ustundag A, Guloglu B. A forecasting approach for truckload spot market pricing. Transport Res A Policy Pract. 2017;97:55–68.
16. Ji S, Wang X, Zhao W, Guo D. An application of a three-stage xgboost-based model to sales forecasting of a cross-border e-commerce enterprise. Math Probl Eng. 2019. https://doi.org/10.1155/2019/8503252.
17. Cheriyan S, Ibrahim S, Mohanan S, Treesa S. Intelligent sales prediction using machine learning techniques. In: 2018 international conference on computing, electronics & communications engineering (iCCECE), IEEE; 2018. pp. 53–58.
18. Mohamed-Iliasse M, Loubna B, Abdelaziz B. Is machine learning revolutionizing supply chain? In: 2020 5th International conference on logistics operations management (GOL). IEEE; 2020. pp. 1–10.
19. Bousqaoui H, Achchab S, Tikito K. Machine learning applications in supply chains: an emphasis on neural network applications. In: 2017 3rd International conference of cloud computing technologies and applications (CloudTech). IEEE; 2017. pp. 1–7.
20. Gupta G, Gupta KL, Kansal G. Megamart sales prediction using machine learning techniques. In: Proceedings of third international conference on computing, communications, and cybersecurity: IC4S 2021, Springer; 2022. pp. 437–446.
21. Albadrani A, Zohdy MA, Olawoyin R. An approach to optimize future inbound logistics processes using machine learning algorithms. In: 2020 IEEE international conference on electro information technology (EIT). IEEE; 2020. pp. 402–406.
22. Htun HH, Biehl M, Petkov N. Survey of feature selection and extraction techniques for stock market prediction. Fin Innov. 2023;9(1):26.
23. Kaur G, Kaur H, Goyal S. Correlation analysis between different parameters to predict cement logistics. Innov Syst Softw Eng. 2023;19(1):117–27.
24. Lei Y, Qiaoming H, Tong Z, et al. Research on supply chain financial risk prevention based on machine learning. Comput Intell Neurosci. 2023. https://doi.org/10.1155/2023/6531154.
25. Cheriyan S, Ibrahim S, Mohanan S, Treesa S. Intelligent sales prediction using machine learning techniques. In: 2018 International conference on computing, electronics & communications engineering (iCCECE). IEEE; 2018. pp. 53–58.
26. Mohamed-Iliasse M, Loubna B, Abdelaziz B. Is machine learning revolutionizing supply chain? In: 2020 5th International conference on logistics operations management (GOL). IEEE; 2020. pp. 1–10.
27. Bousqaoui H, Achchab S, Tikito K. Machine learning applications in supply chains: an emphasis on neural network applications. In: 2017 3rd International conference of cloud computing technologies and applications (CloudTech). IEEE; 2017. pp. 1–7.
28. Pallathadka H, Mustafa M, Sanchez DT, Sajja GS, Gour S, Naved M. Impact of machine learning on management, healthcare and agriculture. Mater Today Proc. 2023;80:2803–6.
29. Prahara PJ, Hariadi TK. Improved feature selection algorithm of electricity price forecasting using svm. In: 2022 2nd international conference on electronic and electrical engineering and intelligent system (ICE3IS). IEEE; 2022. pp. 34–39.
30. Kaur G, Kaur H. Prediction of the cause of accident and accident prone location on roads using data mining techniques. In: 2017 8th International conference on computing, communication and networking technologies (ICCCNT). IEEE; 2017. pp. 1–7.
31. Bindal R, Sarangi P, Kaur G, Dhiman G. An approach for automatic recognition system for Indian vehicles numbers using k-nearest neighbours and decision tree classifier; 2019.
32. Sharma A, Mishra PK. Performance analysis of machine learning based optimized feature selection approaches for breast cancer diagnosis. Int J Inf Technol. 2022;14(1):1–12.
33. Reddy EMK, Gurrala A, Hasitha VB, Kumar KVR. Introduction to naive bayes and a review on its subtypes with applications. In: Bayesian Reason. Gaussian Process. Mach. Learn. Appl.; 2022. pp. 1–14.

34. Karki S, Hadikusumo B. Machine learning for the identification of competent project managers for construction projects in Nepal. Constr Innov. 2023;23(1):1–18.
35. Wu D, Wang Q, Olson DL. Industry classification based on supply chain network information using graph neural networks. Appl Soft Comput. 2023;132: 109849.
36. Banik S, Islam MR, Rahman KN, Rahman MA. A comparative analysis of machine learning algorithms to predict backorder in supply chain management. SSRN. 2023. https://doi.org/10.2139/ssrn.4444976.
37. Luo J. Application of machine learning in supply chain management. In: 2022 3rd international conference on big data economy and information management (BDEIM 2022). Atlantis Press; 2023. pp. 489–498.
38. Esmaeili M, Olfat L, Amiri M, Raeesi Vanani I. Classification and allocation of suppliers to customers in resilience supply chains using machine learning. J Ind Manage Perspect. 2023;13(3):39–70.
39. Ghasemkhani B, Aktas O, Birant D. Balanced k-star: an explainable machine learning method for internet-of-things-enabled predictive maintenance in manufacturing. Machines. 2023;11(3):322.
40. Khosravi K, Golkarian A, Omidvar E, Hatamiafkoueieh J, Shirali M. Snow water equivalent prediction in a mountainous area using hybrid bagging machine learning approaches. Acta Geophys. 2023;71(2):1015–31.
41. Nguyen HD, Tran KP, Thomassey S, Hamad M. Forecasting and anomaly detection approaches using lstm and lstm autoencoder techniques with the applications in supply chain management. Int J Inf Manage. 2021;57: 102282.
42. Hasan MR. Addressing seasonality and trend detection in predictive sales forecasting: a machine learning perspective. J Bus Manage Stud. 2024;6(2):100–9.
43. Soltaninejad M, Aghazadeh R, Shaghaghi S, Zarei M. Using machine learning techniques to forecast Mehram company's sales: a case study. J Bus Manage Stud. 2024;6(2):42–53.
44. Kaur G, Kaur H, Goyal S. Strategic feature selection for precision augmentation in cement sales forecasting. In: 2023 Seventh international conference on image information processing (ICIIP). IEEE; 2023. pp. 765–770.