DOI: 10.15849/IJASCA.251130.07

Int. J. Advance Soft Compu. Appl., Vol. 17, No. 3, November 2025
Print ISSN: 2710-1274, Online ISSN: 2074-8523

Leveraging Machine Learning with Feature Selection to Enhance Business Intelligence in Predicting E-Commerce Acquiring Behaviors

Nesreen Alsharman[1*], Ismail Hababeh[2*], Deefallah Alshorman[3], Ibrahim Jawarneh[4], Mohammad Alqudah[5]

[1,2] Computer Science Department, German Jordanian University, Amman, Jordan, nesreen.alsharman@gju.edu.jo, ismail.hababeh@gju.edu.jo

[3] Department of Elementary Teacher Education, Al-Zaytoonah University of Jordan, Amman, Jordan, d.alshorman@zuj.edu.jo

[4] Department of Mathematics, Al-Hussein Bin Talal University Maan, Jordan, ibrahim.a.jawarneh@ahu.edu.jo

[5] Department of Mathematics, German Jordanian University, Amman, Jordan, mohammad.qudah@gju.edu.jo

*Corresponding Authors: Ismail Hababeh: ismail.hababeh@gju.edu.jo

Abstract

One key component of business intelligence that may greatly improve a company's strategic and operational choices is the ability to predicate eCommerce acquiring behaviors. Machine learning, a subfield in artificial intelligence, is concerned with developing methods to analyze, interpret, and forecast large data for predictions and decisions. Feature selection is an essential stride used to increase machine learning model performance that is evaluated using cross-validation techniques. This study integrates machine learning algorithms, data mining tools, and filter-based feature selection techniques to help unveil patterns in purchasing behavior and exploring consumer trends. Particularly, the algorithms for machine learning Decision tree, k-Nearest Neighbor, Naive Bayes, artificial neural network, Logistic Regression, K-means, Random Forest, and Support Vector Machine are utilized with Orange data mining tool, Chi-squared, and the Minimum Redundancy Maximum Relevance MRMR feature selection techniques to predict the customer purchase behaviors. A quantitative analysis is conducted to evaluate the accuracy of each machine learning model. The experiment results show the competence of the proposed approach and confirm its proficiency to improve business predictions on E-commerce acquiring behaviors.

Keywords: *machine learning, data mining, feature selection, behavior prediction, business intelligence, E-commerce*

Received 20 July 2025; Accepted 9 September 2025

1 Introduction

Business intelligence (BI) describes the methods, instruments, and technologies used by companies to collect, process, and display data to aid decision making [1]. Understanding consumer behavior includes examining how customers engage with a company, what influences their purchase decisions, and how businesses can optimize their tactics, which is essential to increase customer satisfaction and profitability [2]. The customer is a valuable resource and the cornerstone of any successful business. New businesses and aggressively competitive firms make investments to maintain positive customer relationships [3]. Maintain, develop, and strengthen relationships with customers is known as custom-er relationship management CRM [4] where consumers are connected through digital media. Online purchasing is rapidly expanding, and several e-commerce enterprises focus on gaining customers by different digital media networks. Businesses even have a variety of tactics to promote their market, but contacting new customers comes at a very high cost [5]. Therefore, organizations are concentrating on customers who demonstrate a higher tendency to buy their products to increase the effectiveness of their promotions. Observing their purchase patterns, one might establish a connection with devoted clients [6].

For enterprises anticipating staying competitive in the market, understanding and forecasting consumer purchase behavior has become a crucial objective. Business organizations are progressively using sophisticated machine learning and data mining analytics techniques [7] to predict consumer purchasing traditions and alter their marketing strategies in response to the expansion of e-commerce and digital communications [8]. By allowing enterprises to expect demand and customize services, predictive models not only increase customer relationship management but also enhance revenue [9]. The accuracy of predicting consumer behavior has intensely enhanced with the latest developments in machine learning, especially deep learning models [10]. Enormous levels of transactional and behavioral data are analyzed by these models to find embedded relationships that are often missed by traditional approaches. Additionally, using concurrent data from electronic devices and social media stands feeds more contextually conscious and vigorous forecasts, which suggests the rapidly moving styles of consumers [11]. In the current marketing and service businesses, the ability to accurately predict what customers are expected to purchase and when has become a necessary component of strategic decision making as enterprises' aspect to increase customer involvement and faithfulness.

Nonetheless, several enduring issues reduce the precision and consistency of these predictions. Quality and integration of data where mobile, web, and local store customer data are frequently dispersed across platforms, which causes problems with integration and uneven data quality, hence machine learning model performance is limited by low-quality data [12]. In addition, changing customer social trends, economic upheavals, and the deployment of new technologies cause changes in customer preferences that could not be captured by traditional models [13]. Organizations should thump a balance between personalization and customer trust because of raising data protection regulations [14]. New customers without previous data are difficult to manage by predictive methods, which renders initial predictions incorrect [15]. Furthermore, stakeholder trust and decision-making are squeezed by machine learning models, which offer significant accuracy but are challenging to understand [16].

In predictive analytics, machine learning has become a potential instrument that can recognize intricate patterns and connections in data where conventional statistical techniques would miss [17]. The supervised machine learning [18-19] involves training a

model using a labeled dataset, each in-put-output pair is given in the training data, and the algorithm gain knowledge of a mapping from the input features to the output labels. On the other hand, unsupervised learning [20] is an efficient algorithm like grouping and analysis of principal components that can be used to convert unstructured data into a more structured format for use in supervised tasks. For example, clustering of K-means can be used to group comparable data points into clusters. The labels assigned to these clusters can then be used as additional features in a trained model.

This study aims to create an impressive prediction approach that can precisely predict customer purchasing behavior by employing the machine learning model on the dataset gathered from the loyalty free Kaggle repository [21] and based on the proposed feature selection technique. Specifically, we implement eight current powered machine learning models namely, K-Means KM [22], Artificial Neural Network ANN [23], Support Vector Machine SVM [24], Random Forest RF [25], k-Nearest Neighbor kNN [26], Logistic Regression LR [27], Decision Trees DT [28], and Naive Bayes NB [29]. The unsupervised machine learning model K-Means is used as a preprocessing step for the other supervised machine learning models. The foremost contributions of this work are follows:

- Developing a feature choice technique that determines the principal characteristics of the customer's purchase behavior.
- Developing a prediction technique that integrates machine learning algorithms, data mining tools, and methods for feature selection based on filtering patterns in purchasing behaviors and exploring consumer trends.

This paper is outlined as follows; section 2 discusses the materials and methods. Section 3 presents experimental results and performance evaluation. The conclusions and future directions are clarified in section 4.

2. Materials and Methods

This research aims to support data scientists and analysts in comprehending the variables driving purchasing decisions throw evaluating and testing different efficient machines' learning algorithms with features selection. The proposed method consists of four main steps: Loading input file, selecting the features that have most impact on the customer purchase behavior, preprocessing missing values, and predicting new customers' purchase behaviors. The architecture of predicting customer purchase behavior system is illustrated in figure 1.

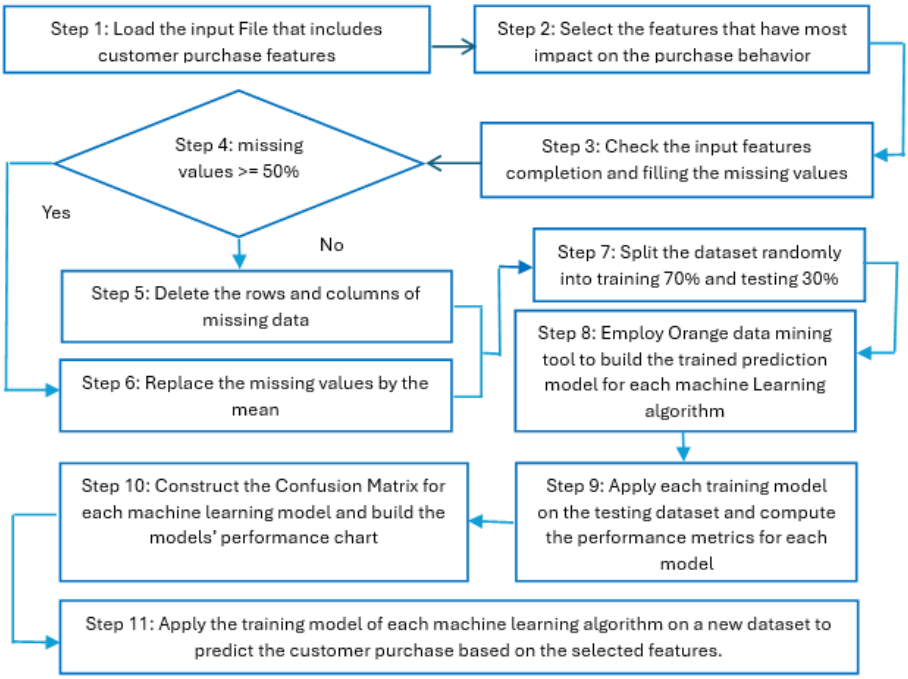

Fig. 1. The architecture of predicting customer purchase behavior system

2.1 Preprocessing Data Files

The data is imported from files that include information on customer purchase behavior across a variety of criteria and contains customer pertinent attributes, demographic data, and E-commerce behaviors. However, data entry methods could result in missing values or lack of data in datasets. The missing values have a negative impact on machine learning models' accuracy and dependability. To ensure that raw data is converted into a format that is appropriate for modeling, preprocessing datasets is a necessary phase in data analysis and machine learning. Therefore, handling the missing or lack of values is an essential phase in preparing data files for classification. We utilized the most commonly used method of dealing with missing data [30] that assumed if the proportion of misplaced values is low, the rows and columns of missing data are deleted but if the proportion of misplaced values is high, the missing numerical values are replaced by arithmetic approximations, such as the mode, median, or mean, and the missing categorical information are replaced by the most recurrent category.

2.2 Feature Selection Technique

Several feature selection techniques can be applied to improve the machine learning techniques' performance depending on the features characteristics, complexity, and occurrence during model training. Three types of feature selection techniques exist, Filter Techniques that employ statistical methods to score features; Wrapper Techniques that implement a predictive model to evaluate combinations of features; and Embedded Techniques aim to decrease the feature space dimension by removing the inappropriate and redundant features without affecting the quality of the trained model decision [31].
This study proposed a supervised classification technique that reduces the feature dimensionality by selecting the most significant features while maintaining the most

crucial characteristics. This technique utilizes two filter-based selection methods, namely, fscchi2 and fscmrmr, to compute the features' weight and determine the important characteristics that affect the customer's purchase behavior. Both methods, fscchi2 and fscmrmr, are employed since they offer rapid, scalable, and model-independent behavior of selecting pertinent features, primarily for high-dimensional data. The fscchi2 and fscmrmr methods are described as follows:

- fscchi2 is a Chi-squared [32] test statistic that assesses whether a discrepancy between actual and expected data is statistically significant. This method is best used for categorized target variables and non-negative features, and the features are ranked based on their arithmetic confidence in the target.
- fscmrmr is a minimum redundancy maximum relevance method [33] that uses reciprocal information to select features that are both minimally redundant with each other and highly relevant to the target. This approach works well for both classification and regression, succeeds with continuous features, ranks the features that maximize variety and consequence.

Notably, the suggested feature selection method conceptually resembles deep learning attention mechanisms, which teach models to dynamically concentrate on the most pertinent portions of the input. Attention mechanisms accomplish a similar goal implicitly during training by allocating weights to input components, whereas fscchi2 and fscmrmr explicitly rank features using statistical and information-theoretic criteria. As a result, our approach can be thought of as a portable, interpretable attention analog, especially for structured tabular data. The proposed feature selection technique is described in Algorithm 1.

Algorithm 1: Feature Selection Technique

Input: Customer dataset that includes feature matrix and class labels
Output: The selected set of features that determine the target eCommerce (customer purchase behavior)
Initialization: Normalize the feature matrix (non-negative values for fscchi2) and class labels are set properly
Processing:
Step 1: Delete the feature column in the customer dataset that has all values (0) or has missing values > 50% of the sum of all feature values
Step 2: If the columns' missing values <= 50% of the sum of all feature values, fill the missing values by using the mean (average) values
Step 3: Compute all remaining features' weight in the customer dataset by using the fscchi2 method
Step 4: Compute all remaining features' weight in the customer dataset by using the fscmrmr method
Step 5: Find the selected set of features that determine the target eCommerce as follows:
If the feature weight at both methods fscchi2 and fscmrmr > 0.1
(least effective feature value), then add the feature to the selected features set
else
the feature is waved from the selected features set
end if
End.

The proposed feature selection technique guarantees e-commerce systems' efficiency by determining only the features that have high impact on the customer purchase behavior, then reducing the feature selection time and thus minimizing the prediction time. The features of the customer in Kaggle dataset [21] are used as a ground truth to validate our proposed approach in predicting the target customer purchase behavior (purchase:1, not purchase: 0). The selected features that determine the target E-commerce (customer purchase behavior) include Age, Gender (0: Male, 1: Female), Annual Income, Number of Purchases, Product Category, such as Electronics (0), Clothes (1), Home appliances (2), Beauty (3), and Sports (4), Website time consuming (in minutes), Loyalty Program (not member:0, member:1), and Discounts Availed (range: 0-5).

2.3 Machine Learning Algorithms

Several standard and cutting-edge machine learning algorithms were used for facilitating and predicting customer purchase behavior, specifically K-Means, Decision Trees, Support Vector Machine, k-Nearest Neighbor, Artificial Neural Networks, Random Forest, Logistic Regression, and Naive Bayes. The description of each method is detailed as follows:

- K-Means: an unsupervised clustering method that generates a new dataset with the cluster label added as a meta-attribute to facilitate classification process of other supervised machine learning.
- Decision Tree: a supervised clustering technique that divides the data into nodes based on mean squared error to assess the model performance for continuous variables, and information gain for categories.
- k-Nearest Neighbor: a non-parametric approach to supervised learning used for classification. This method computes the average of k-nearest training samples in feature distance to accomplish predictions. In this research, we use the following kNN configuration to generate best results: the number of closest neighbors is 3, Manhattan metric, and Distance weight.
- Support Vector Machine: a categorization technique utilizing supervised machine learning. This method aims to create a higher-dimensional attribute space from the supplied data to simplify dataset categorization.
- Artificial Neural Network: a method of supervised machine learning that performs complex mathematical transformation to predict certain outcomes based on adapting data weights.
- Random Forest: a method of supervised machine learning that joins the results of various decision trees to accomplish one result. Random Forest is used for resolving regression and grouping problems.
- Logistic Regression: a one-layer supervised neural network method that is used to expect the probability of a particular category based on a given data set of independent variables.
- Naive Bayes: a method of supervised machine learning that realizes the probability of objects, features, and is used for classification problems. This method compares the conditional probabilities of the dependent variable and employs it to the categorization of new features and objects. Table 1 shows Sorting Machine Learning Techniques into Classification and Clustering Groups.

Table 1: Sorting Machine Learning Techniques into Classification and Clustering Groups

Category	Machine Learning Methods
Classification	Artificial Neural Network, Decision Tree, Support Vector Machine, k-Nearest Neighbor, Random Forest, Logistic Regression, Naive Bayes
Clustering	K-Means

2.4 Predicting Customer Purchase Behavior

Based on the results of the proposed filter-based feature selection algorithm 1, Orange data mining tool [34] is employed to predict the customer purchase behavior. Orange is a visual programming data mining tool that uses standard machine learning algorithms and lets users construct machine learning systems to predict consumer expenses behavior without prior programming knowledge. The predicting customer purchase behavior is described in algorithm 2 and illustrated in Figure 2.

Algorithm 2: Predicting customer purchase behavior

Input: The customer dataset with labels purchase (1)/ not purchase (0).
The selected set of features that determine the target E-commerce (customer purchase behavior) generated in algorithm 1
Output: Prediction of the target E-commerce (customer purchase behavior)
Processing:
Step 1: Split the customer dataset randomly into training 70% and testing 30%.
Step 2: Apply the Orange data mining tool to build the training prediction model for each machine learning
algorithm, Random Forest, Decision Trees, Naive Bayes, Artificial Neural Network, Support Vector Machine, k-Nearest Neighbor, and Logistic Regression.
Step 3: Apply each training model in step 2 on the testing dataset and determine the performance indicators for
Every model, Accuracy, AUC, F1-Score, Precision, Recall, and MCC.
Step 4: Construct the confusion matrices and performance charts for the machine learning models using Orange data
Mining tool (as shown in Figure 2).
Step 5: Apply the training model of each machine learning algorithm on a new dataset to predict the customer
Purchase based on the features selected in algorithm 1.
End.

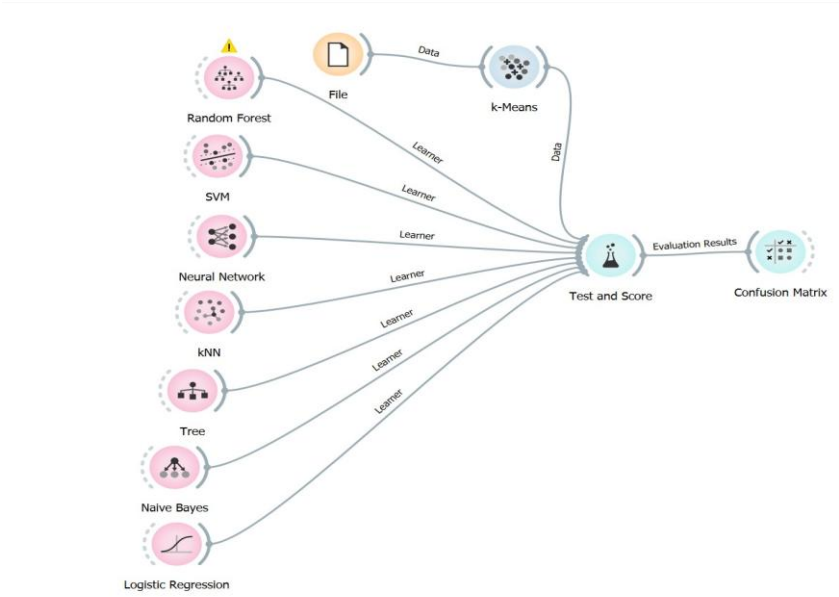

Fig. 2. Predicting customer purchase behavior using Orange data mining tool

3. Experimental Results and Performance Evaluation

The experimental results are carried out using laptop machine supported with Intel(R) G13, Core -i5, 1.3 GHz, and 8.00 GB. The machine learning algorithms: Random Forest, Decision Trees, Naive Bayes, Artificial Neural Network, Support Vector Machine, k-Nearest Neighbor, Logistic Regression, and K-means were evaluated by using the Confusion Matrix [35] analysis and the Quantitative Analysis [36]. This step measures how accurate and reliable the models are. Each experiment was performed ten folds on each machine learning model. The training to testing ratio was 70:30. The machine learning models were explicitly evaluated in terms of performance measure parameters [37] described in Table 2, where True positive (TP), True negative (TN), False negative (FN), and False positive (FP).

TABLE 2. Performance parameters

Performance Metric	Computation	Metric Results Description
Accuracy	Accuracy = (TN + TP)/(TN+TP+FN+FP)	Higher accuracy value increases feature detection
F1-Score	F1 = (recall*precision)*2 / (recall + precision)	Higher F1-Score value increases feature detection
Precision	Precision = TP / (FP+TP)	Higher precision value increases feature detection
Area under the curve (AUC)	AUC = ∫a,b f(x) dx	Higher AUC value increases feature detection
Sensitivity (Recall)	Sensitivity = TP/(TP + FN)	Higher sensitivity value increases feature detection
Matthew's Correlation Coefficient (MCC)	MCC = (TN*TP - FN*FP) / sqrt((FP+TP) * (FN+TP) * (FP+TN) * (FN+TN))	Higher MCC value increases feature detection

Table 3 presents the customer purchase behavior of each machine learning model in terms of performance measure metrics computations based on the fscchi2 feature selection method after applying the proposed predicting customer purchase behavior.

TABLE 3. The mean values generated by the machine learning models based on fscchi2 method

Machine Learning Model	Accuracy	F1-Score	Precision	AUC	Sensitivity	MCC
Decision Tree	0.900	0.900	0.900	0.925	0.900	0.796
k-Nearest Neighbor	0.643	0.640	0.639	0.684	0.643	0.264
Support Vector Machine	0.662	0.663	0.665	0.722	0.662	0.316
Artificial Neural Networks	0.864	0.863	0.864	0.916	0.864	0.722
Random Forest	0.928	0.928	0.928	0.950	0.928	0.853
Logistic Regression	0.729	0.728	0.728	0.801	0.729	0.446
Naive Bayes	0.889	0.888	0.889	0.931	0.889	0.773

Figure 3 depicts the machine learning models performance based on FSCCHI2 filter-based method after applying the proposed predicting customer purchase behavior.

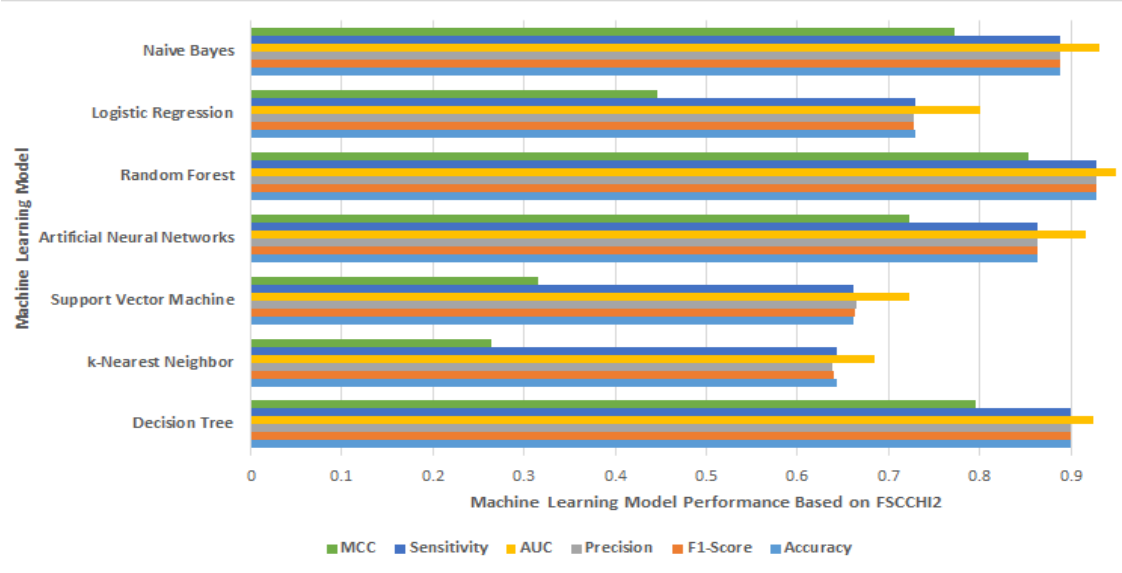

Fig. 3. The machine learning algorithms performance based on FSCCHI2 method

Table 4 shows the customer purchase behavior of each machine learning model in terms of performance measure metrics computations based on the fscmrmr feature selection method after applying the proposed predicting customer purchase behavior.

TABLE 4. The mean values generated by the machine learning models based on fscmrmr method

Machine Learning Model	Accuracy	F1-Score	Precision	AUC	Sensitivity	MCC
Decision Tree	0.708	0.707	0.706	0.758	0.708	0.401
k-Nearest Neighbor	0.641	0.638	0.637	0.682	0.641	0.260
Support Vector Machine	0.635	0.636	0.644	0.698	0.635	0.273
Artificial Neural Networks	0.753	0.751	0.752	0.805	0.753	0.493
Random Forest	0.773	0.771	0.772	0.829	0.773	0.534
Logistic Regression	0.714	0.712	0.712	0.778	0.714	0.413
Naive Bayes	0.745	0.745	0.745	0.813	0.745	0.479

Figure 4 presents the machine learning models performance based on FSCMRMR filter-based method after applying the proposed predicting customer purchase behavior.

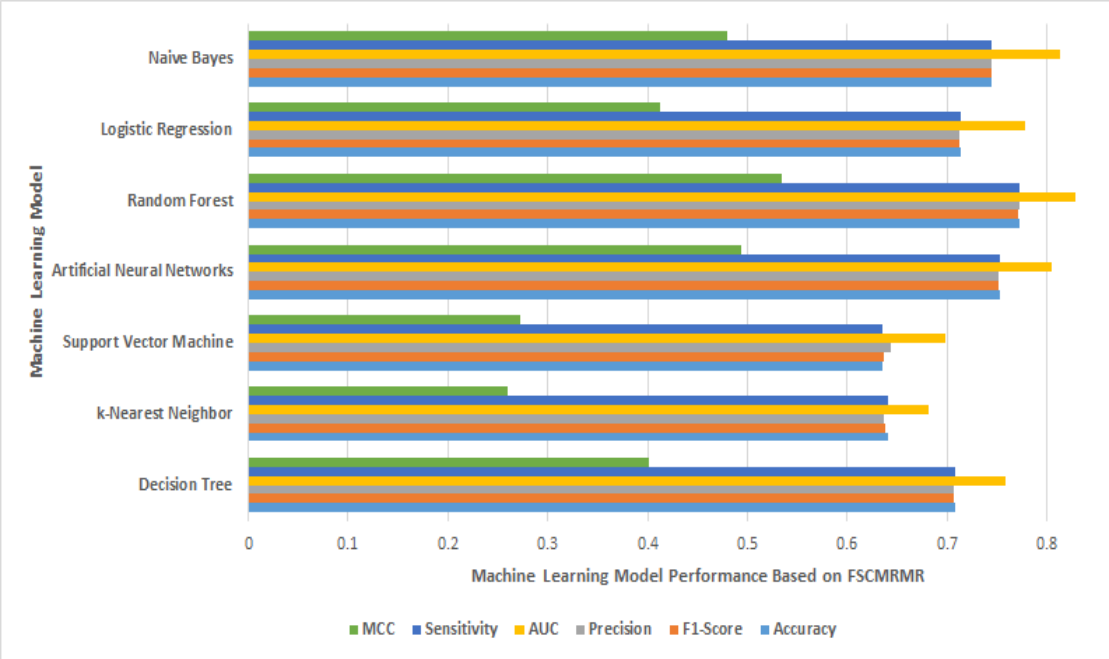

Fig. 4. The machine learning algorithms performance based on FSCMRMR method

Figures 5 and 6 show the confusion matrix results of the kNN, SVM, and RF machine learning models using the fscchi2 and fscmrmr filter-based methods respectively.

Fig. 5. The confusion matrix of kNN, SVM, and RF machine learning models using the fscchi2 filter-based method.

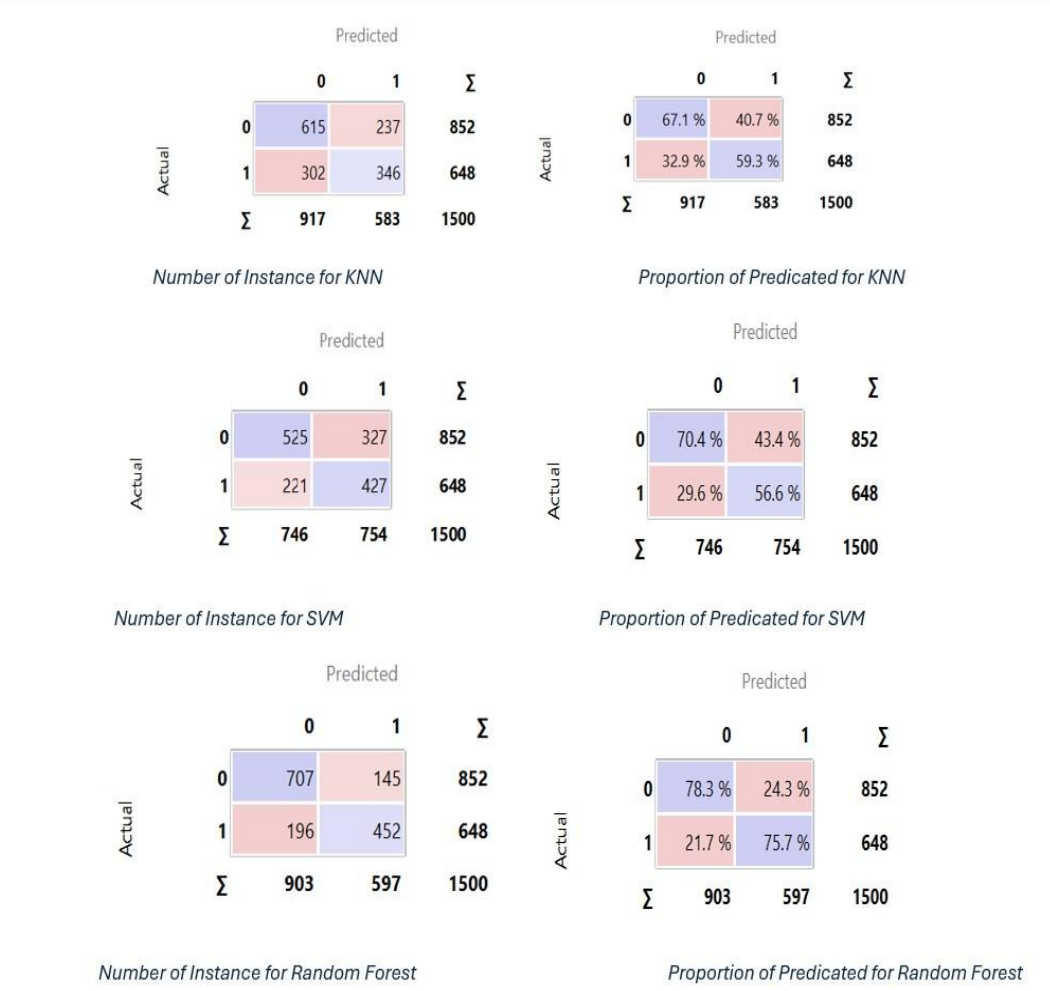

Fig. 6. The confusion matrix of kNN, SVM, and RF machine learning models using the fscmrmr filter-based method.

Figures 7 and 8 show the confusion matrix results of the Naive Bayes, Decision Tree, Artificial Neural Network, and Linear Regression machine learning algorithms using the fscchi2 and fscmrmr filter-based methods respectively.

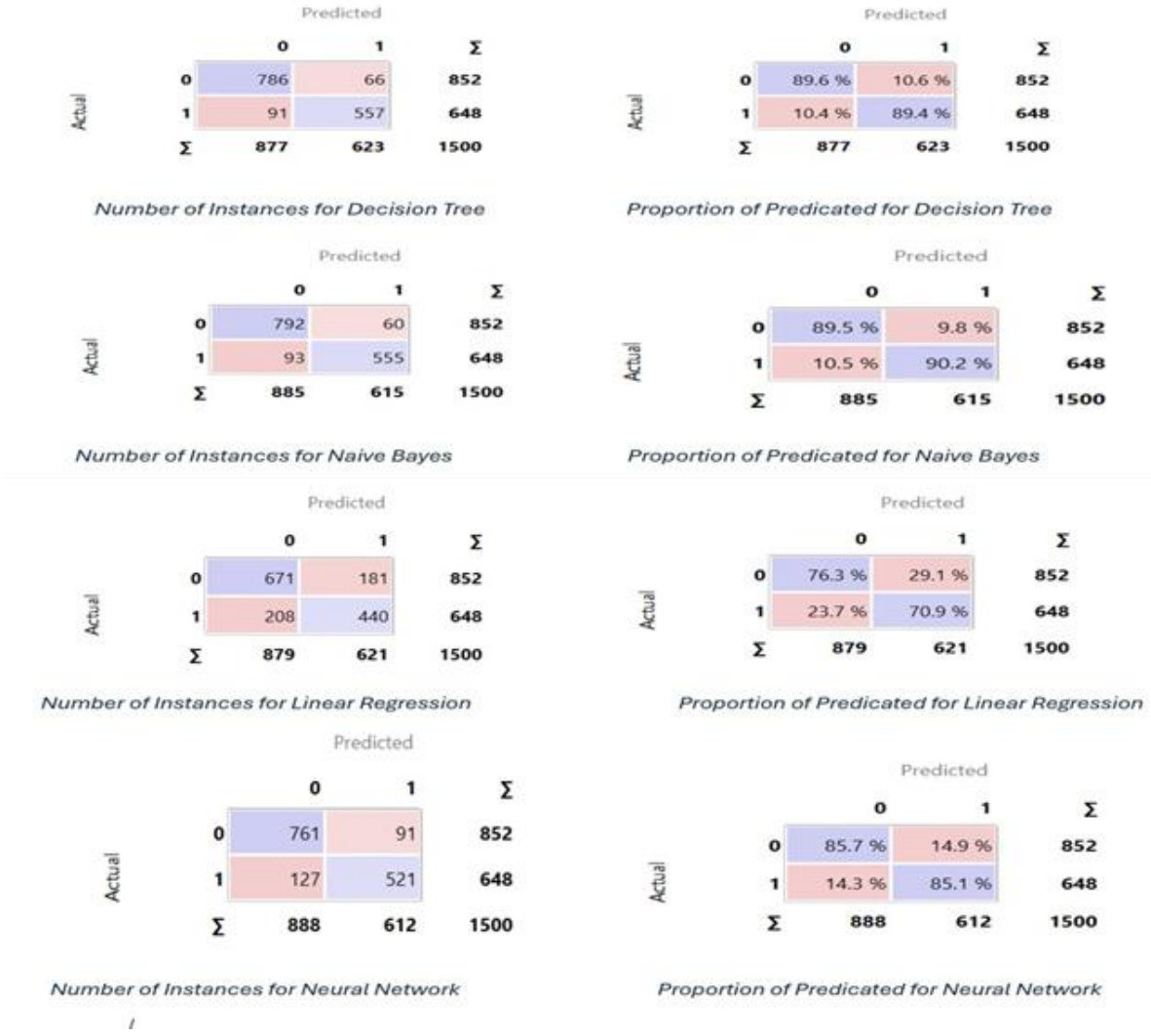

Fig. 7. The confusion matrix results generated by Naive Bayes, Decision Tree, Artificial Neural Network, and Linear Regression machine learning algorithms using the fscchi2 filter-based method.

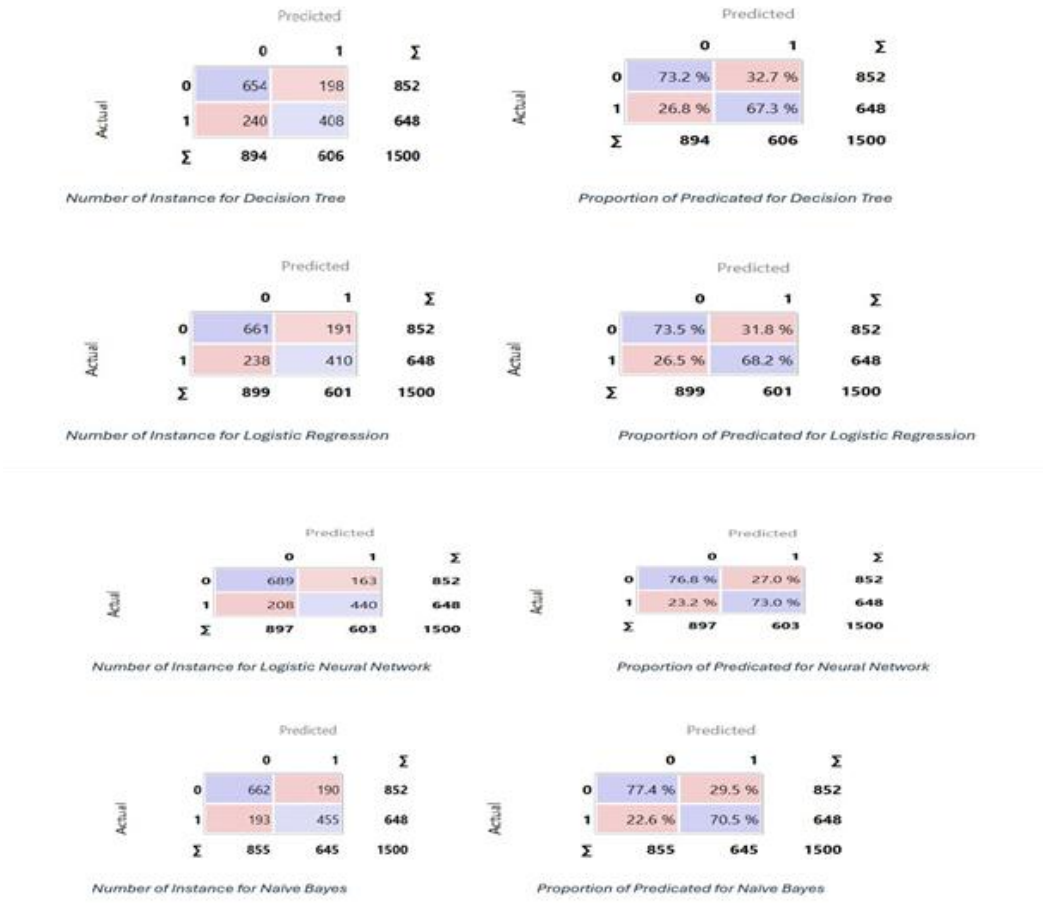

Fig. 8. The confusion matrix results are generated by Naive Bayes, Decision Tree, Artificial Neural Network, and Linear Regression machine learning algorithms using the fscmrmr filter-based method.

The confusion matrix of each model indicates how many correct predicated customer purchase behaviors were found or missed. This detailed information is crucial as it offers numerous important advantages in business intelligence [38] such as Customer Retention [39], Optimized Pricing Strategies [40], Product Development [41], and Enhanced Customer Experience [42].

Since fscchi2 explicitly evaluates each feature's statistical association with the target class, it may produce more discriminative features for classification tasks than fscmrmr. This is demonstrated by the experiment results shown in figures 3 and 4. Additionally, because fscchi2 only works on one feature at a time, it is computationally more efficient and avoids noise from redundant features. Only a few features are helpful to the goal, but it functions best when there are many features. Fscmrmr may perform better if the data contains continuous or highly connected variables, whereas fscchi2 works better when characteristics are categorical or non-negative and the aim is discrete.

Several performance analyses were conducted to evaluate the machine learning algorithms and establish predictions built on their output. We applied the proposed prediction method described in algorithm 1 on a new dataset to predict the new customers' purchase behavior. Predictions' results can be used to determine if the customers will make a purchase based on their input features. Figure 9 describes the process of applying the proposed method to generate predictions about new instances that have not yet been observed by using Random Forest machine learning model.

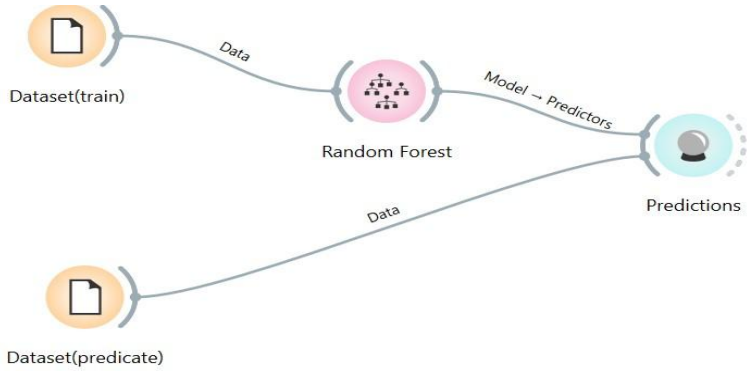

Figure 9: Prediction process of Random Forest algorithm

Figure 10 shows the predication results of the Random Forest method on a new dataset.

	Random Forest (1)	PurchaseStatus	Age	Gender	AnnualIncome	umberOfPurchase	ProductCategory	meSpentOnWebsi	LoyaltyProgram	DiscountsAvailed
1	0.26 : 0.74 → 1	1	40	1	66120.3	8	0	30.5686	0	5
2	0.85 : 0.15 → 0	0	20	1	23579.8	4	2	38.2401	0	5
3	0.10 : 0.90 → 1	1	27	1	127821	11	2	31.6332	1	0
4	0.08 : 0.92 → 1	1	24	1	137799	19	3	46.1671	0	4
5	0.57 : 0.43 → 0	1	31	1	99301	19	1	19.8236	0	0
6	0.89 : 0.11 → 0	0	66	1	37758.1	14	4	17.8275	0	2
7	0.00 : 1.00 → 1	1	39	1	126883	16	3	42.0854	1	4
8	1.00 : 0.00 → 0	0	64	1	39707.4	13	2	17.1903	1	0
9	0.97 : 0.03 → 0	0	43	0	102797	20	1	6.02347	0	3
10	0.00 : 1.00 → 1	1	20	1	63854.9	16	0	38.5725	0	5
11	0.20 : 0.80 → 1	1	66	1	66200	11	1	1.41553	1	5
12	0.11 : 0.89 → 1	1	70	1	83556.7	11	2	42.174	0	4
13	0.05 : 0.95 → 1	1	54	0	114467	9	2	17.626	1	5
14	0.97 : 0.03 → 0	0	64	1	31880.9	17	0	22.7535	1	1
15	0.05 : 0.95 → 1	1	19	1	107486	7	1	28.2603	1	4
16	0.84 : 0.16 → 0	0	70	1	67049.6	2	4	21.8951	0	0
17	0.61 : 0.40 → 0	1	51	1	129174	13	4	21.8088	0	0
18	0.58 : 0.42 → 0	0	18	1	128374	0	2	16.7683	0	4
19	1.00 : 0.00 → 0	0	57	0	71740.7	12	2	40.6967	0	2
20	0.07 : 0.93 → 1	1	20	0	121499	12	3	21.2401	1	0
21	0.73 : 0.27 → 0	0	59	0	92768.2	3	2	1.03702	0	5
22	0.10 : 0.90 → 1	1	19	0	116945	17	3	5.95379	0	5

Fig. 10. Predication results of Random Forest method

The results of this research show that the Naive Bayes, Random Forest, and Decision trees models appeared as the most efficient machine learning algorithms for customer purchase behavior prediction. High AUC and MCC values indicate the robustness of these algorithms in classifying the customer purchase status accurately, i.e. the probability of the customer buying (0: No, 1: Yes). The SVM and artificial neural network models also performed adequately but were less effective compared to the leading model, Random Forest. On the other hand, Logistic Regression model provided moderate results, while kNN under-performed due to its sensitivity to the choice of kernel or the high dimensional of the data.

The proposed method is evaluated and validated against current methods in literature and shows superiority in accuracy. Various machine learning models are explored to effectively classify customer purchase status. Our analysis highlighted the unique strengths of each model and their applicability to the dataset's characteristics.

4. Conclusions and Future Work

This study highlighted the use of machine learning algorithms to significantly anticipate client purchasing behavior. Several machine learning models, Decision Trees, k-Nearest Neighbor, Artificial Neural Networks, Random Forest, Logistic Regression, Naive Bayes, and Support Vector Machine, were applied on the Orange data mining tool, able to predict future consumer behavior, and discover important purchase trends by examining transactional, demographic, and behavioral data. The results show that enhancing prediction accuracy requires careful consideration of feature selection, data preparation, and model tuning. Furthermore, incorporating temporal and behavioral data greatly improves model performance and provides more in-depth understanding of how consumers make decisions. The research reveals that machine learning methods, such as Decision Tree and Random Forest, are effective tools for organizations looking to enhance client retention, optimize inventories, and customize marketing campaigns. As part of future work, we plan to extend this work by finding out which stacking-based meta learning and ensemble techniques, such as Random Forest with Support Vector Machine and Artificial Neural Networks with Naïve Bayes, would accomplish higher prediction results. Furthermore, employing several benchmark datasets from various industries, such as retail, banking, and subscription services, would increase the proposed approach truthfulness.

References

[1] Adewusi, A.O., Okoli, U.I., Adaga, E., Olorunsogo, T., Asuzu, O.F. and Daraojimba, D.O., 2024. Business intelligence in the era of big data: A review of analytical tools and competitive advantage. Computer Science & IT Research Journal, 5(2), pp.415-431.

[2] Adeniran, I.A., Efunniyi, C.P., Osundare, O.S. and Abhulimen, A.O., 2024. Transforming marketing strategies with data analytics: A study on customer behavior and personalization. International Journal of Management & Entrepreneurship Research, 6(8), pp.41-51.

[3] Tang, J., Wu, X. and Tang, L., 2024, September. Automobile Customer Behavior Analysis and Service Optimization Based on Big Data and AI Technology. In 2024 14th International Conference on Information Technology in Medicine and Education (ITME) (pp. 1146-1150). IEEE.

[4] Liu, D., Huang, H., Zhang, H., Luo, X. and Fan, Z., 2024. Enhancing customer behavior prediction in e-commerce: A comparative analysis of machine learning and deep learning models. Applied and Computational Engineering, 55, pp.181-195.

[5] Charanasomboon, T., and Viyanon, W. A comparative study of repeat buyer prediction: Kaggle acquired value shopper case study. In *Proceedings of the 2nd International Conference on Information Science and Systems* (New York, NY, USA, 2019), ICISS '19, Association for Computing Machinery, p. 306–310.

[6] Orogun, A. and Onyekwelu, B., 2019. Predicting consumer behaviour in digital market: a machine learning approach.

[7] Adesina, A.A., Iyelolu, T.V. and Paul, P.O., 2024. Leveraging predictive analytics for strategic decision-making: Enhancing business performance through data-driven insights. World Journal of Advanced Research and Reviews, 22(3), pp.1927-1934.

[8] Al-Otaibi, Y.D., 2024. Enhancing e-Commerce Strategies: A Deep Learning Framework for Customer Behavior Prediction. *Engineering, Technology & Applied Science Research*, *14*(4), pp.15656-15664.
[9] Akter, M.S. and Islam, R., Big Data Analytics and Predictive Analysis In Enhancing Customer Relationship Management (CRM): A Systematic Review Of Techniques.
[10] Sun, C., Adamopoulos, P., Ghose, A. and Luo, X., 2022. Predicting stages in omnichannel path to eCommerce: A deep learning model. Information Systems Research, 33(2), pp.429-445.
[11] Kishore, M.K., Shaik, S. and Panda, B.S., 2024. Exploration of the Sentiment-Driven Forecasting Models for Predicting Consumer ECommerce Patterns on Social Media. IJIRT, 10(12).
[12] Velásquez, J.D., 2025. An analysis of trends, challenges, and opportunities in retail analytics. International Journal of Market Research, p.14707853251315585.
[13] Pramiarsih, E.E., 2024. Consumer behavior in the digital era. International Journal of Financial Economics, 1(3), pp.662-674.
[14] Naz, H. and Kashif, M., 2025. Artificial intelligence and predictive marketing: an ethical framework from managers' perspective. Spanish Journal of Marketing-ESIC, 29(1), pp.22-45.
[15] Jangid, M. and Kumar, R., 2024. Deep learning approaches to address cold start and long tail challenges in recommendation systems: a systematic review. Multimedia Tools and Applications, pp.1-33.
[16] Abou El Houda, Z., Brik, B. and Khoukhi, L., 2022. "why should i trust your ids?": An explainable deep learning framework for intrusion detection systems in internet of things networks. IEEE Open Journal of the Communications Society, 3, pp.1164-1176.
[17] Beyerer, J., Hagmanns, R. and Stadler, D., 2024. Pattern recognition: introduction, features, classifiers and principles. In Pattern Recognition. De Gruyter Oldenbourg.
[18] Customer shopping behavior analysis using rfid and machine learning models. *Information (Switzerland) 14*, 10 (2023).
[19] Singh, A., Thakur, N., and Sharma, A. A review of supervised machine learning algorithms. In *2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom)* (2016).
[20] Naeem, S., Ali, A., Anam, S., and Ahmed, M. An unsupervised machine learning algorithm: Comprehensive review. *IJCDS Journal 13* (04 2023), 911–921.
[21] https://www.kaggle.com/code/chandanarprasad/p5-predict-customer-purchase-behavior. Accessed: May 15th, 2025.
[22] El Khattabi, M.Z., El Jai, M., Lahmadi, Y., Oughdir, L. and Rahhali, M., 2024. Understanding the interplay between metrics, normalization forms, and data distribution in K-means clustering: A comparative simulation study. Arabian Journal for Science and Engineering, 49(3), pp.2987-3007.
[23] Yaghoubi, E., Khamees, A. and Vakili, A.H., 2024. A systematic review and meta-analysis of artificial neural network, machine learning, deep learning, and ensemble learning approaches in field of geotechnical engineering. *Neural Computing and Applications*, *36*(21), pp.12655-12699.
[24] Khan, T.A., Sadiq, R., Shahid, Z., Alam, M.M. and Su'ud, M.B.M., 2024. Sentiment analysis using support vector machine and random forest. Journal of Informatics and Web Engineering, 3(1), pp.67-75.
[25] Salman, H.A., Kalakech, A. and Steiti, A., 2024. Random forest algorithm overview. *Babylonian Journal of Machine Learning*, *2024*, pp.69-79.

[26] Halder, R.K., Uddin, M.N., Uddin, M.A., Aryal, S. and Khraisat, A., 2024. Enhancing K-nearest neighbor algorithm: a comprehensive review and performance analysis of modifications. Journal of Big Data, 11(1), p.113.
[27] Singh, H.P., Singh, N., Mishra, A., Sen, S.K., Swarnkar, M. and Pandey, D., 2024, February. Logistic Regression based Sentiment Analysis System: Rectify. In *2024 IEEE International Conference on Big Data & Machine Learning (ICBDML)* (pp. 186-191). IEEE.
[28] Mienye, I.D. and Jere, N., 2024. A survey of decision trees: Concepts, algorithms, and applications. IEEE access.
[29] Sanchez-Franco,´ M. J., Navarro-Garc´ıa, A., and Rondan-´ Cataluna, F. J.˜ A naive Bayes strategy for classifying customer satisfaction: A study based on online reviews of hospitality services. *Journal of Business Research 101*, C (2019), 499–506.
[30] Zhou, Y., and Bouadjenek, S. A. M. R. A comprehensive review of handling missing data: Exploring special missing mechanisms. *arXiv:2404.04905v1 [stat.ME] 7 Apr 2024*.
[31] Barbieri, M.C., Grisci, B.I. and Dorn, M., 2024. Analysis and comparison of feature selection methods towards performance and stability. *Expert Systems with Applications*, p.123667.
[32] Das, A., 2024. New methods to compute the generalized chi-square distribution. *arXiv preprint arXiv:2404.05062*.
[33] Wang, Y., Li, X. and Ruiz, R., 2022. Feature selection with maximal relevance and minimal supervised redundancy. *IEEE transactions on Cybernetics*, *53*(2), pp.707-717.
[34] Popchev, I. and Orozova, D., 2023. Algorithms for Machine Learning with Orange System. *International Journal of Online & Biomedical Engineering*, *19*(4).
[35] Sathyanarayanan, S. and Tantri, B.R., 2024. Confusion matrix-based performance evaluation metrics. African Journal of Biomedical Research, pp.4023-4031.
[36] Takona, J.P., 2024. Research design: qualitative, quantitative, and mixed methods' approaches. Quality & Quantity, 58(1), pp.1011-1013.
[37] Obi, J.C., 2023. A comparative study of several classification metrics and their performances on data. World Journal of Advanced Engineering Technology and Sciences, 8(1), pp.308-314.
[38] Adewusi, A.O., Okoli, U.I., Adaga, E., Olorunsogo, T., Asuzu, O.F. and Daraojimba, D.O., 2024. Business intelligence in the era of big data: A review of analytical tools and competitive advantage. *Computer Science & IT Research Journal*, *5*(2), pp.415-431.
[39] Adeniran, I.A., Efunniyi, C.P., Osundare, O.S., Abhulimen, A.O. and OneAdvanced, U., 2024. Implementing machine learning techniques for customer retention and churn prediction in telecommunications. *Computer Science & IT Research Journal*, *5*(8).
[40] Chowdhury, M.S., Shak, M.S., Devi, S., Miah, M.R., Al Mamun, A., Ahmed, E., Hera, S.A.S., Mahmud, F. and Mozumder, M.S.A., 2024. Optimizing E-Commerce Pricing Strategies: A Comparative Analysis of Machine Learning Models for Predicting Customer Satisfaction. The American Journal of Engineering and Technology, 6(09), pp.6-17.
[41] Cooper, R.G. and McCausland, T., 2024. AI and new product development. *Research-Technology Management*, *67*(1), pp.70-75.
[42] Bhuiyan, M.S., 2024. The role of AI-Enhanced personalization in customer experiences. *Journal of Computer Science and Technology Studies*, *6*(1), pp.162-169.