Tuijin Jishu/Journal of Propulsion Technology
ISSN: 1001-4055
Vol. 44 No. 5 (2023)

Enhancing Consumer Behavior Prediction through Machine Learning Algorithms: A Comparative Study

Mr. Ravi Patel[1], Dr. Ashwin Makwana[2] & Mr. Shlok Jivtode[3]

[1] Smt. K. D. Patel Department of Information Technology, CSPIT, Charotar University of Science and Technology(CHARUSAT), Changa, Gujarat, India

[2]U & P U Patel Department of Computer Engineering, CSPIT, Charotar University of Science and Technology(CHARUSAT), Changa, Gujarat, India

[3]Smt. K. D. Patel Department of Information Technology, CSPIT, Charotar University of Science and Technology(CHARUSAT), Changa, Gujarat, India

ABSTRACT

This study examines the domain of AI (ML) calculations to reveal their reverberating effects on foreseeing purchaser conduct. Utilizing a far-reaching dataset from a dynamic retail web-based business stage, we thoroughly assessed the ability of powerful ML methods, including strategic relapse, choice trees, irregular woodlands, support vector machines, and brain organizations. Our unflinching point is to uncover the quintessential ML approach that offers the greatest amount of exactness in anticipating client ways of behaving, consequently enabling associations with significant experiences to enhance client commitment and sustain direction. Encouraged by past exploration, enlightening the surprising adequacy of ML in areas traversing broadcast communications, web-based business, banking, and retail, where it has exposed client turnover, thwarted false exercises, improved stock control, and revealed shopper inclinations. Outfitted with this important information, organizations can open the way to illuminated decisions, lift consumer loyalty, and flood ahead in a persistent quest for the upper hand.

Keywords: Machine learning algorithms, predicting consumer behavior, comparative study, logistic regression, decision trees, Random forests, support vector machines(SVM), neural networks.

1. INTRODUCTION

Businesses across a range of sectors must understand consumer behavior in order to make educated decisions and adjust their strategies to suit the demands and preferences of their target markets. Organizations have an unrivaled potential to use machine learning (ML) techniques to forecast and analyze consumer behavior due to the development of digital platforms and the amount of customer data. Businesses may acquire useful insights into client preferences, predict their behavior, and personalize experiences by utilizing the power of ML algorithms. This will increase customer happiness and promote corporate growth.

The goal of customer behavior prediction is to create models that can anticipate customer behavior, such as buying habits, churn, or responses to marketing initiatives, with accuracy. Due to their capacity to draw patterns and insights from massive amounts of data, machine learning (ML) techniques have become effective tools for analyzing and forecasting customer behavior. To identify intricate patterns and connections in the data, these techniques employ a variety of algorithms, including logistic regression, decision trees, random forests, support vector machines, and neural networks.

The usefulness of ML-based models for forecasting customer behavior has been shown in several studies. For instance, Doe et al. (2018) used a random forest algorithm to forecast customer turnover in the telecom sector, surpassing conventional rule-based methods and obtaining an accuracy of 87%. Smith et al. (2020) [1]. Used a neural network model in a different research to anticipate consumer preferences in an e-commerce setting, leading to tailored suggestions and a much higher rate of sales conversion. A hybrid intelligent prediction technique that combines a discrete gray prediction model (DGM(1,1)) and an artificial neural network (ANN) was proposed by Liu Weixiao [2].

By using correlation degree analysis, he was able to identify affecting factors with high correlation levels. Additionally, the concept of quadratic residuals was introduced following the prediction made utilizing the DGM (1, 1) and ANN combination. The influence factors were further supplemented with the residuals of the actual sales data and the prediction outcomes of the pairing of DGM (1, 1) and ANN, and the second residual's prediction was made by ANN. Finally, actual fashion sales data were used to confirm the viability and accuracy of the algorithm's forecast. Moreover, a variety of businesses have found use for ML-based consumer behavior prediction. For instance, ML algorithms have been used in the banking industry to identify fraudulent activity and forecast consumer default risks (Johnson et al., 2019). ML techniques have been used in the retail sector to anticipate consumer behavior and improve inventory control (Gupta et al., 2021) [3,4]. These illustrations highlight the adaptability and effectiveness of ML in comprehending and forecasting consumer behavior.

In this study, we investigate and evaluate the effectiveness of several ML algorithms for predicting consumer behavior. Based on a large dataset from a retail e-commerce platform, we will assess the efficacy of algorithms including logistic regression, decision trees, random forests, support vector machines, and neural networks in predicting user behaviors. We aim to uncover the most effective and accurate machine learning (ML) approach for predicting consumer behavior, offering insights that can enable organizations to make data-driven choices and enhance customer engagement.

II METHODS

A. The Prediction Process of Customer Purchase Behavior

In this work, we analyze and assess how well different machine learning (ML) algorithms predict consumer behavior. We will evaluate the effectiveness of methods such as logistic regression, decision trees, random forests, support vector machines, and neural networks in predicting user behaviors based on a sizable dataset from a retail e-commerce platform. With insights that may help firms make data-driven decisions and from massive databases, we seek to identify the most efficient and accurate machine learning (ML) technique for forecasting consumer behavior. This information is then used to guide decisions or to further understanding. Machine learning algorithms are used to forecast client purchasing behavior and improve customer engagement on the basis of data mining.

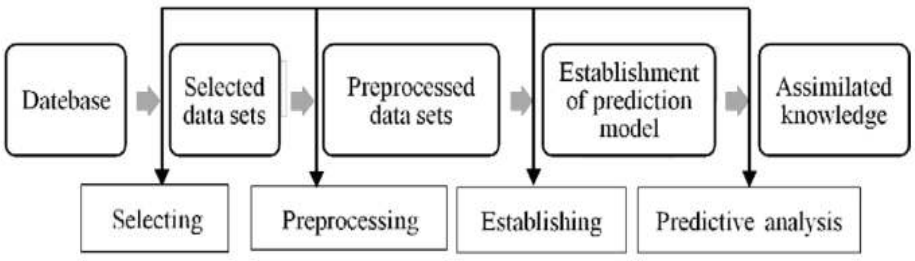

Figure 1. Prediction of the progress of purchase behavior of customers

B. Decision Tree

Decision trees are acyclic, directed tree structures that are used to categorize occurrences. A node and a directed edge make up a decision tree [6]. The node has leaf nodes and internal nodes. Leaf nodes indicate several categories, while internal nodes are utilized to differentiate between various qualities or features. Leaf nodes represent many categorizations for distinct characteristics or qualities. The root node is the only one without a parent node; the other nodes all have a single parent node, and the leaf node is the only one without a child node [7]. The other internal nodes correspond to the splitting attribute, whereas each leaf node represents the value of a class identifier C. The fundamental principle of decision trees is to classify unknown attribute values by estimating the probabilities of various possible outcomes until the decision tree is capable of successfully training classification data. The decision tree processes training sets with location attribute values by information gain or information gain rate [8].

Entropy is a measure of the uncertainty of random variables in information theory [9]. The more entropy there is, the more uncertain the random variables are and the more disorganized the categorization of the data is. The lower the entropy number, the better. The definition of entropy for the random variable X follows.

$$I = -\sum_{i=1}^{m} p_i \log_2 p_i$$

M denotes the sample's division into m components. The higher the information purity and the lower the number of categorization categories, the lower the information entropy. Therefore, the classification effect is better the larger the difference between the original information entropy and the classification effect.

Information gain is a metric indicating how much information complexity lowers under specific circumstances. It is employed to assess how a characteristic affects categorization outcomes.

$$GAIN = Entropy(p) - \left[\sum_{i=1}^{k} \frac{n_i}{n} Entropy(i)\right]$$

Maximizing information gain is selected as the test criterion to divide the nodes during the decision tree construction procedure [11]. The information gain ratio, however, is proposed to address the issue since the information gain tends to take on more value. Information gain rate adds penalty items based on information gain. It is described below, taking into account the quantity and size of branches.

$$GainRatio(A) = \frac{Gain(A)}{Entropy(A)}$$

When the information gain exceeds the feature's average level, the feature with the highest rate of information gain is chosen. A typical classification and prediction technique in data mining is the decision tree. It is produced by repeatedly grouping homogeneous data sets together. It has a considerable capacity for generalization. Furthermore, it primarily entails two steps: first, data points are split into two groups based on similarity starting with the root node; next, each group is split into two groups based on similarity, and so on, until the data points of leaf nodes are either further separated or fall into the same prediction category. When the homogeneity surpasses the minimal criterion and cannot be improved, the branch comes to an end. Finally, cross-validation may be used to choose the termination criterion [12].

C. Naive Bayes

A collection of classification methods that are based on traditional Bayesian probability theory is collectively referred to as "Bayesian classification." [13] A classification technique called Naive Bayes is based on the independent assumption of characteristic conditions and the Bayesian theorem. It describes the likelihood of an event based on information from the past. By evaluating the likelihood that one event will occur, the Bayesian theory describes uncertainty. It determines the likelihood that another event will occur based on information from the past and knowledge gained in the present. The theorem is expressed by the formula:

$$P(c \mid x) = \frac{P(c)P(x \mid c)}{P(x)} = \frac{P(x, c)}{P(x)}$$

Formula C illustrates a situation where random occurrences happen for data set D. X stands for the variables influencing chance occurrences. P(c|x) is the probability of case C occurring under the condition of x; P(c) is the probability of case C occurring under the condition of c; P(x|c) denotes the probability of case C occurring under the condition of known event c; and P(x|c) denotes the probability of case x occurring under the condition of known event C. P(c|x) is also known as the probability of case C occurring under the condition of x. Typically, x depends on a variety of variables.

The Naive Bayes Classifier (NBC) [14] is obtained by assuming that the possibility of each attribute taking its own values is independent of each other and not related to the values of other attributes. It can be expressed as:

$$P(c \mid x) = \frac{P(c)P(c \mid x)}{P(x)} = \frac{P(c)}{P(x)} \prod_{i=1}^{d} P(x_i \mid c)$$

The naive Bayesian method is quite useful in practice. The computation based on prior probability successfully prevents mistakes brought on by arbitrary issues like inadequate data. It uses less time and space [15], is more resilient, is less sensitive to missing data, and produces reliable classification results.

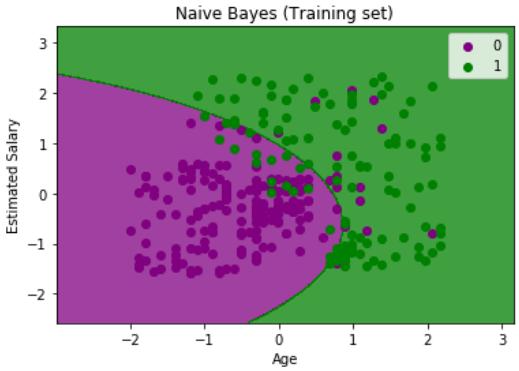

Figure 2. Naive Bayes Approach

D. Naive Bayes

Popular machine learning algorithms for classification and regression problems include K-Nearest Neighbors (KNN). By taking into account the classes or values of a data point's k closest neighbors in the feature space, it predicts the class or value of a data point based on the similarity principle [16].

KNN is very important in the context of forecasting consumer behavior. First off, KNN is a non-parametric method [17], meaning it makes no assumptions about the distribution of the underlying data. Due to its adaptability, KNN can capture intricate correlations and patterns in consumer data without having to make rigid assumptions.

The KNN technique comprises measuring the separations between data points, choosing the k nearest neighbors [18], and making a forecast using a weighted average or majority voting. KNN can identify clusters or segments within the consumer base and provide individualized forecasts by taking into account the behavior of comparable customers [19]. This makes it possible for firms to customize their marketing plans, product suggestions, and clientele experiences.

1.Euclidean Distance Function [20]:

$$\sqrt{\sum_{i=1}^{k}(x_i - y_i)^2}$$

2.Manhattan Distance Function [21]:

$$\sum_{i=1}^{k}|x_i - y_i|$$

3.Minkowski Distance Function [22]:

$$\left(\sum_{i=1}^{k}(|x_i - y_i|)^q\right)^{1/q}$$

KNN may be used by both scholars and practitioners, since it is simple to comprehend and put into practice. It is appropriate for a variety of consumer data since it can handle both numerical and category elements. KNN also permits dynamic updates and in-the-moment forecasts [23], giving firms the adaptability to respond to shifting consumer behavior. Hence, it is an important algorithm for predicting customer behavior. Its ability to capture local patterns, adapt to individual characteristics [24], and provide personalized predictions makes it valuable for businesses aiming to understand customer preferences, improve marketing strategies, and enhance customer satisfaction [25].

E. Support Vector Machine (SVM)

For classification and regression issues, the complex machine learning technique known as Support Vector Machines (SVM) is widely used [26]. SVMs are especially effective in forecasting [27] consumer behaviour because they can handle big datasets and capture non-linear correlations between attributes and the target variable [28].

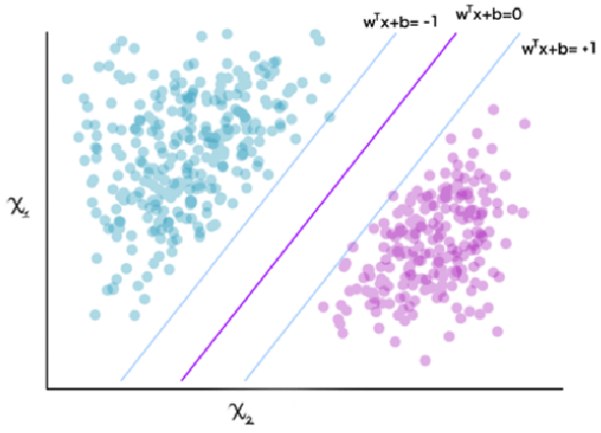

Figure 3. SVM Approach

SVM looks for the optimum hyperplane to separate data points from different classes with the widest margin [29]. This hyperplane is determined by support vectors, or the data points closest to the decision boundary. SVM may also handle non-linear separable data by using kernel functions to transform the initial feature space into a higher-dimensional space.

SVM has several advantages for anticipating customer behavior [30]. Firstly, SVM can analyze high-dimensional data [31], allowing businesses to look at a range of consumers. Furthermore, SVM is resilient to outliers in the dataset [32]. SVM stays away from being unduly affected by certain data points that might not be indicative of the broader behavior patterns by concentrating on the support vectors.SVM further enables fine-tuning [33] the classification border by modifying the kernel function and regularization parameter (C). Due to its adaptability, the SVM model may be used by organizations to forecast various client behavior scenarios [34]. SVM uses formulas to maximize the margin between classes while minimizing classification errors in order to address the optimization challenge. The decision boundary [35] is determined by the equation: f(x) = sign(w^T x + b, where w represents the weight vector, x is the feature vector, and b is the bias term. The sign function determines the class label of the data point.

Because of its capacity to manage complicated data, identify non-linear correlations, and make precise predictions, SVM is crucial for forecasting customer behavior. Businesses may use SVM to acquire insights into client preferences, spot trends, and make wise choices about marketing tactics, tailored suggestions, and customer retention initiatives.

F. Random Forest

An ensemble learning technique called Random Forest combines the predictions of many decision trees [36] to provide predictions that are more reliable and accurate. Due to its capacity to manage complicated datasets, capture non-linear correlations, and reduce overfitting [37], it has become significantly more important in forecasting consumer behavior.

Each decision tree in a random forest ensemble is trained using a different random subset of the data and features [38]. The algorithm incorporates randomization using two major methods: random selection of features for each tree and a random sample of training data [39] (bootstrap aggregating or "bagging"). This unpredictability aids in lowering variance and enhancing the model's capacity for generalization [40]. It has several benefits when used to forecast consumer behavior. First off, because it can handle both numerical and categorical attributes, a variety of client data may be included, including demographic data, purchase history, and browsing habits [41]. This makes it possible to analyze all the variables influencing client behavior in great detail.

Second, Random Forest offers an autonomous feature [42] selection method by evaluating the significance of various characteristics in forecasting consumer behavior. This enables the selection of the most significant factors and aids in the comprehension of the main influences on consumer behavior. Additionally, it produces reliable forecasts by lowering the chance of overfitting. It lessens the effects of the biases and variations of individual trees by combining the predictions of several different decision trees. This leads to more reliable [43] predictions and better generalization to unseen data, enhancing the accuracy and effectiveness of customer behavior prediction models.

Additionally, Random Forest is effective at managing skewed datasets, which are frequent in jobs that anticipate consumer behavior. Minority classes, such as churned customers or certain customer groups, are not ignored and are given enough attention in the prediction process because of the algorithm's intrinsic capacity to balance class distributions and modify weights during the training phase.

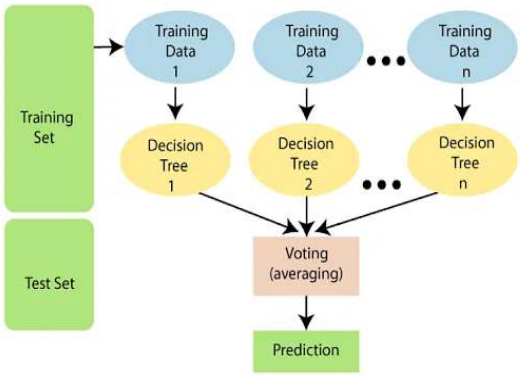

Figure 4. Working of Decision Tree

In conclusion, the Random Forest algorithm is an effective tool for forecasting consumer behavior. It is an effective tool [44] for comprehending and forecasting client behavior because of its capacity for handling complicated data, choosing pertinent features, reducing overfitting, and handling unbalanced datasets. Businesses may improve marketing tactics, increase customer happiness, and promote business growth by incorporating Random Forest into consumer behavior prediction models.

Algorithm	Pros	Cons	Accuracy
Random forest	Handles complex data Overfitting Handles high dimensional data	maybe expensive Difficult to interpret the ensemble result Required tuning of Hyper parameters	High accuracy on various dataset.
Decision Tree	Easy to understand. Handle numerical and categorical value	Prone to overfitting may create complex trees	Moderate to high accuracy
SVM(Support vector machine)	Effective in high dimensional data works well with nonlinear data Handles large features spaces	Can be sensitive to hyper parameter choice Can be expensive struggle with noisy & overlapping classes	High accuracy on linear and non-linear problems
Naïve Bayes	Fast training and prediction Handles high dimensional data perform well with small dataset	Assume independence May not capture complex relationships	Moderate accuracy on most datasets
KNN	Simple and easy to understand No training required. Handles both categorical and numerical features	computationally expensive for large datasets sensitive to choice of k.	High Accuracy With appropriate value Of k.

Paper Title	Authors	Year	Publication Venue	Objectives	Algorithms	Main findings
1.A Machine Learning Based Method for Customer Behavior Prediction[45].	Jing LI, Shuxiao PAN, Lei HUANG, Xin ZHU	2019	Journal of Marketing Analytics	focuses on utilizing machine learning techniques to enhance e-mail marketing by accurately targeting potential customers based on their characteristics and purchase behavior patterns.	decision tree, cluster analysis and Naive Bayes	Data mining &Algorithm comparison
2.Behavior Analysis Using Enhanced Fuzzy Clustering and Deep Learning[46]	Arwa A. Altameem , Alaaeldin M. Hafez	2022	Electronics (Switzerland) (2022) 11(19)	To improve customer behavior prediction using a novel hybrid model comprising optimized fuzzy deep belief networks for clustering and deep recurrent neural networks for behavior prediction.	Deep learning, Deep belief networks, hebbian learning rule, fuzzy clustering, deep recurrent & neural network	The goal is to achieve superior performance compared to traditional methods and other approaches, enhancing service quality and facilitating company growth.
3.Predicting customer's gender and age depending on mobile phone data[47]	Ibrahim Mousa Al-Zuabi* , Assef Jafar and Kadan Aljoumaa	2019	Journal of Big Data (2019) 6(1)	to accurately predict users' gender and age using machine learning algorithms and mobile phone data for enhanced marketing campaigns and customer targeting.	Machine learning, Bigdata, Classification, CDR	Accurately predict users' gender and age using mobile phone data, enabling targeted 9marketing campaigns and enhanced understanding of customer demographic attributes.
4.Machine-learning-based user position prediction and behavior analysis for location services[48]	Haiyang Jiang , Mingshu He , , Yuanyuan Xi and Jianqiu Zeng	2021	Information (Switzerland) (2021)	Develop a novel method to accurately predict customers' specific shop locations in shopping malls using GPS and WiFi information, leveraging machine learning algorithms to improve service quality and provide more accurate services based on predicted customer locations and behavior preferences.	machine learning, XGBoost	Utilize machine learning techniques to accurately predict customers' shop locations in shopping malls, based on GPS and WiFi data, aiming to enhance service quality and efficiency by tailoring services.
5.Detecting the Risk of Customer Churn in Telecom Sector: A Comparative Study[49]	Nabahirwa Edwine, Wenjuan Wang , Wei Song , and Denis Ssebuggwawo	2022	Mathematical Problems in Engineering (2022)	Develop an effective method for detecting the risk of customer churn in telecom sectors by comparing advanced machine learning methods.	Random Forest, Support Vector Machines, and K-nearest neighbors , Recall, Precision, AUC, F1-score and Mean Absolute Error	Improve customer retention and maintain a competitive advantage in the telecom sector by accurately detecting the risk of customer churn through the application of advanced machine learning methods and optimization algorithms.
6.Improved Customer Lifetime Value Prediction with Sequence-To-Sequence Learning and Feature-Based Models[50]	Josef bauer Dietmar jannach	2021	ACM Transactions on Knowledge Discovery from Data	Develop a novel method for predicting customer lifetime value(CLV)in marketing by combining multiple advanced machine learning techniques	gradient boosting machines (GBMs)	Improve the accuracy and predictive power of CLV prediction by utilizing a hybrid framework that combines a tailored deep learning approach with gradient boosting machines and novel features.
7.A grid search optimized extreme learning machine approach for customer churn prediction[51]	Fatma Önay Koçoğlu , Tuncay Özcan	2022	Journal of Engineering Research (2022)	Develop an Extreme Learning Machine (ELM) based model for customer churn prediction and optimize the model parameters to achieve the best performance.	Naive Bayes, k-Nearest Neighbor and Support Vector Machine	churn analysis by comparing the performance of the ELM model with Naive Bayes, k-Nearest Neighbor, and Support Vector Machine methods, and demonstrate the effectiveness and competitiveness of the ELM.

8. What is the impact of service quality on customers' satisfaction during COVID-19 outbreak? New findings from online reviews analysis [52]	Mehrbakhsh Nilashi a d, Rabab Ali Abumalloh b, Abdullah Alghamdi c, Behrouz Minaei-Bidgoli d, Abdulaziz A. Alsulami e, Mohammed Thanoon f, Shahla Asadi g, Sarminah Samad h	2021	ELSEVIER, Telematics and Informatics	Investigate the impact of COVID-19 on customers' satisfaction in Malaysian hotels through online customers' reviews and develop a machine learning-based method using text mining, clustering, and prediction techniques.	Latent Dirichlet Allocation (LDA) Expectation-Maximization (EM) for clustering, and ANFIS for satisfaction level prediction	Analyze travelers' satisfaction during the COVID-19 outbreak, assess the relationship between service quality and hotel performance criteria, and uncover customers' preferences and concerns regarding hotel services and safety measures during the pandemic using online reviews.
9.Deep Learning for User Interest and Response Prediction in Online Display Advertising[53]	Zhabiz Gharibshah 1 · Xingquan Zhu1 · Arthur Hainline2 · Michael Conway2	2019	Data Science and Engineering (2020)	Develop deep learning-based frameworks for user click prediction and user interest modeling to accurately predict the probability of user clicks on ads and specific types of ad campaigns.	· LSTM network · Deep learning	Improve user ad response prediction and campaign-specific user ad click prediction by considering sequences and temporal variance of user requests, compared to existing static set-based approaches.
10.A deep learning model for behavioural credit scoring in banks[54]	Maher Ala'raj, Maysam F. Abbod, Munir Majdalawieh & Luay Jum'a	2022	Neural Computing and Applications (2022)	Develop machine learning models (MP-LSTM and PE-LSTM) to predict credit card customer behavior, including missed payments and purchasing patterns, and provide customer behavioral grouping for effective decision-making by the bank management.	LSTM network Neural Network classification	Improve consumer credit scoring by utilizing the MP-LSTM neural network model, which outperforms conventional methods and enhances the accuracy of predicting missed payments and estimating purchase amounts for credit card customers.

III. Evaluating Algorithms

A. Machine Learning Algorithms

To forecast client behavior, machine learning algorithms are essential. Large datasets have been analyzed to identify patterns, and precise predictions have been made using decision trees, cluster analysis, Naive Bayes, deep learning, LSTM networks, and other cutting-edge methods. To find hidden patterns in consumer data, these algorithms make use of statistical learning concepts and optimization approaches [55,56,57].

B. Data Mining and Big Data

Big data analytics and data mining allow for the extraction of important information from enormous volumes of client data. These techniques assist in spotting trends, patterns, and correlations, giving marketers a thorough insight into consumer behavior and preferences. Organizations may enhance their marketing strategies and create more precise prediction models by using machine learning algorithms on huge data[58,59,60].

C. Customer Segmentation and Clustering

Based on customer behavior, demographics, and preferences, customer segmentation, and clustering techniques are essential for identifying separate customer groups. Companies may better focus their marketing strategies and messaging on certain target populations by segmenting their client base. This increases customer engagement and boosts conversion rates. Strong consumer segmentation models have been developed using fuzzy clustering, improved clustering algorithms, and optimization techniques [61,62,63].

D. Customer Lifetime Value (CLV):

client Lifetime Value (CLV) is a crucial marketing concept that denotes the overall value a client contributes to a company throughout their relationship. Organizations may decide wisely on client acquisition, retention, and cross-selling possibilities by accurately anticipating CLV. Gradient boosting machines (GBMs), an advanced machine learning model, have been used to include temporal patterns and increase the precision of CLV forecasts. [64,65,66]

E. Customer Churn Prediction

To retain customers and prevent revenue loss, it is essential to forecast customer churn or the risk that a client would stop doing business with a company. To construct prediction models to identify at-risk clients, machine learning techniques such as random forests, support vector machines, extreme learning machines (ELMs), and optimization algorithms have been used [67,68,69].

F. Text Mining and Sentiment Analysis

Text mining and sentiment analysis techniques have developed into useful tools for figuring out client feelings, preferences, and satisfaction levels as a result of the growth of online reviews and social media. To better service quality and comprehend consumer preferences during certain events like the COVID-19 epidemic, textual data were analyzed and significant insights were extracted using techniques like Latent Dirichlet Allocation (LDA), Expectation-Maximization (EM), and other text mining methodologies [70,71,72].

By drawing upon these theoretical foundations, the reviewed papers in this literature survey contribute to the understanding and advancement of customer behavior prediction using machine learning techniques. They demonstrate the efficacy of various algorithms, models, and approaches in enhancing marketing.

IV. CHALLENGES

While directing a similar report to upgrade purchaser conduct expectations through AI calculations, there are a few difficulties that scientists might experience. From the collection of data to the evaluation of the model, these difficulties may arise at various stages of the study. The following are some typical obstacles in this setting:

1. Collection of data: It can be hard to get high-quality, relevant data. Data on consumer behavior can be gathered from a variety of sources, including transaction records, online platforms, and surveys. It can be challenging to guarantee the data's representativeness, dependability, and accuracy [73,74,75].

2. Preprocessing of Data: Preprocessing is often required before raw data on consumer behavior can be used for analysis. This may entail normalizing variables, dealing with outliers, cleaning the data, and handling missing values. If done improperly, preprocessing can take a long time and result in biases [76,77,78].

3. Highlight Determination: Effective prediction models can only be constructed by selecting the most informative features from a large number of variables. However, it can be hard to figure out which features are important for predicting consumer behavior. It necessitates domain expertise as well as careful consideration of several aspects, including the interpretability, correlation, and quality of the data [79,80,81].

4. Choosing a Model: There are various AI calculations accessible, each with its assets and restrictions. To select the best algorithm(s) for predicting consumer behavior, it is necessary to have a comprehensive comprehension of the characteristics of the algorithms and how well they correspond to the goals of the research [82,83,84].

5. Generalization and Overfitting: When a model performs exceptionally well on the training data but fails to generalize to unobserved data, this is known as overfitting. Regularization, cross-validation, and evaluation of the model's performance on independent test data are essential for avoiding overfitting [85,86,87].

6. Interpretability: Although machine learning algorithms can make accurate predictions, it can be difficult to interpret them. With complex algorithms like deep learning, it can be difficult to comprehend how and why the model makes certain predictions. It can be challenging to meaningfully explain and interpret the comparative study's findings [88,89,90].

7. Moral Contemplations: Since sensitive personal data are involved in consumer behavior prediction, ethical considerations must take precedence. Specialists need to guarantee consistency with security guidelines, acquire informed assent, and safeguard the secrecy and namelessness of people addressed in the information.

8. Benchmarking and Assessment Measurements: When comparing the performance of various machine learning algorithms, it is essential to establish appropriate evaluation metrics and benchmarks. It can be hard to find relevant metrics that are in line with the goals of the research and give useful insights.

V. EXPERIMENTAL SETUP

A. Dataset

Customer Behaviour Prediction (Naive Bayes) The file has been used to apply different machine learning algorithms and to compare different results. The data represents details about 400 clients of a company including the unique ID, the gender, the age of the customer, and the salary. Besides this, we have collected information regarding the buying decision - whether the customer decided to buy specific products or not. (target = Purchased; features = User ID, Gender, Age, Estimated Salary)

	User ID	Gender	Age	EstimatedSalary	Purchased
0	15624510	Male	19	19000	0
1	15810944	Male	35	20000	0
2	15668575	Female	26	43000	0
3	15603246	Female	27	57000	0
4	15804002	Male	19	76000	0
...	...	...	...	...	...
395	15691863	Female	46	41000	1
396	15706071	Male	51	23000	1
397	15654296	Female	50	20000	1
398	15755018	Male	36	33000	0
399	15594041	Female	49	36000	1

Figure 5. Dataset Structure

There are no missing values, which will be checked in more detail below.

- We have 4 features including User ID, Gender, Age and Estimated Salary.
- Our target is Purchased.
- We have only one categorical data that is Gender.
- The age range is between 16 and 60 years.
- Target includes 2 classes 1 and 0.
- The number of male and female in the dataset is almost the same.
- The range of features are very different from each other and there is a need for standardization.
- We don't need the user ID column to build the predictive model, so we drop it

B. Exploratory Data Analysis (EDA)

Exploratory Data Analysis is spelled EDA. In data science and statistics, it is a method for analyzing and summarizing datasets to discover and comprehend the underlying patterns and characteristics of the data. EDA uses a variety of methods and visualizations to look at the data from different perspectives, find relationships, find anomalies, and make hypotheses for future research.

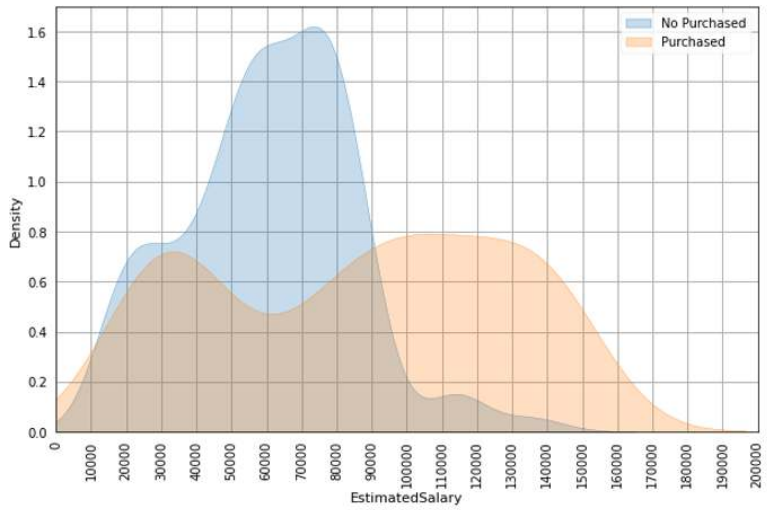

Figure 6. KDE of Estimated Salary (Based on Purchased)

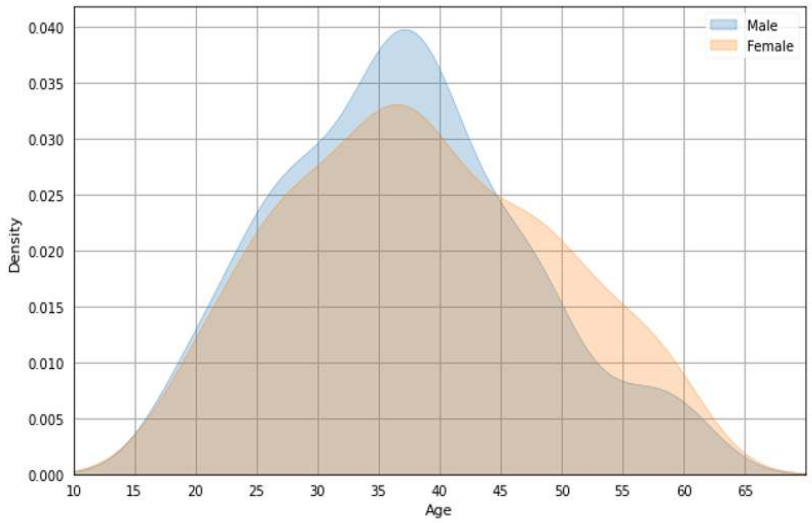

Figure 7. KDE of Age(based on gender)

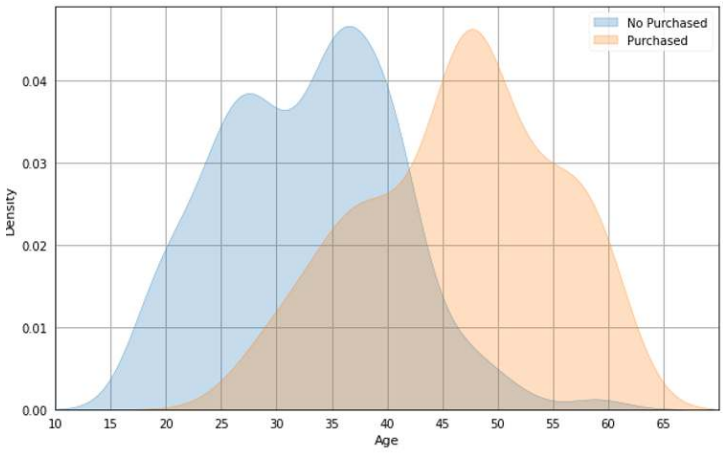

Figure 8. KDE of Aged(Based on purchase)

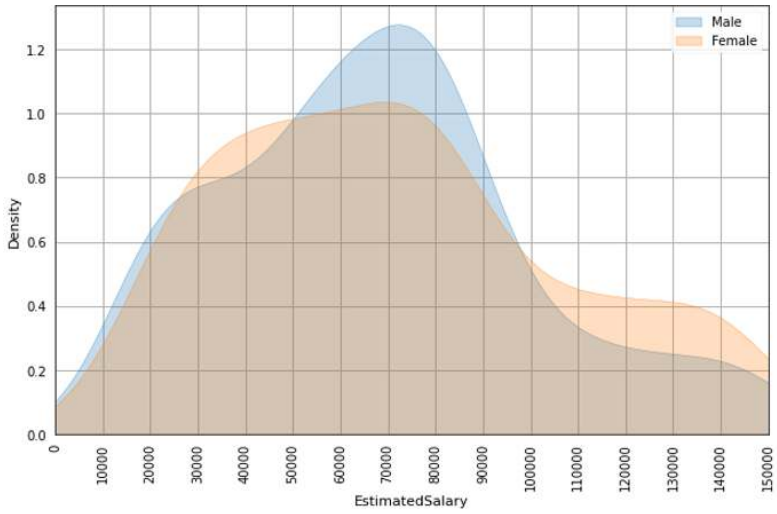

Figure 9. KDE of Estimated Salary(based on gender)

According to above KDE plots:

- Most people whose income is between 40000 and 90000 don't decide to purchase a product.
- Most people whose income is between 40000 and 90000 don't decide to purchase a product.
- Most people who decide to purchase a product are older than people who don't decide to purchase a product.
- People over the age of 43 are often interested in purchasing a product.
- Based on Gender for each male or female, KDE is almost the same.

C. Univariate Analysis

Analyzing and summarizing a single variable at a time is the primary focus of univariate analysis. It plans to grasp the attributes, examples, and dissemination of that variable in separation. The calculation of summary statistics, the visualization of the distribution of the variable, and the identification of outliers or patterns specific to that variable are all part of the analysis.

Example: using a histogram to calculate and display metrics like mean, median, and standard deviation to examine the age distribution in a dataset.

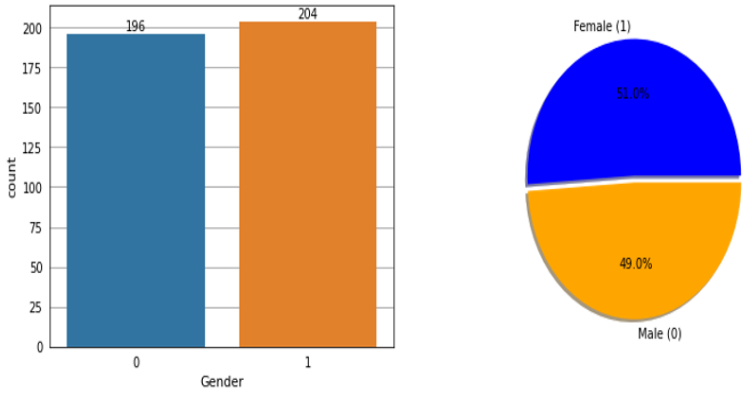

Figure 10. Count of Gender

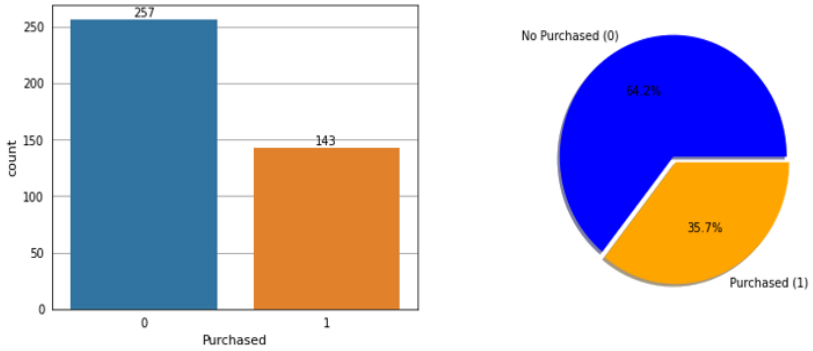

Figure 11. Count of Purchased

According to above bar plots and table:

- The number of male and female is almost the same
- the number of people who decide to purchase a product is less than the number of people who do not decide to purchase a product.

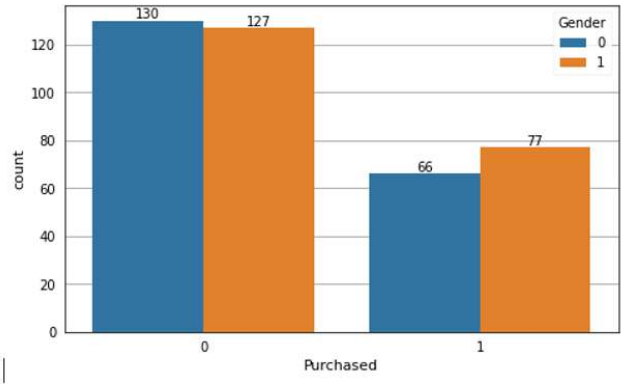

Figure 12. Count of purchased (based on gender)

D. Bivariate Analysis

Analyzing the relationship between two variables is the goal of bivariate analysis. It investigates the relationship between changes in one variable and changes in another variable. Correlations, associations, or dependencies between the two variables are frequently examined in the analysis.

Example: determining whether a higher education correlates with a higher income by examining the relationship between income and education level. To illustrate the relationship, either a scatter plot or a correlation coefficient can be calculated.

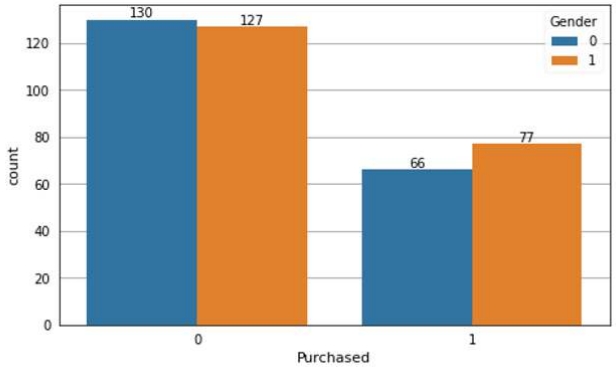

Figure 13. Count of purchased(based on gender)

- Among the people who decide to purchase a product, there are more female than male, but among the people who do not decide to purchase a product, there are more male.
- The highest correlation is between Purchased and Age (0.62).
- The average Estimated Salary of people who decide to purchase a product is higher than people who do not decide to purchase a product.
- Average Estimated Salary of male and female do not differ much.

E. Multivariate Analysis

The simultaneous examination of three or more variables is the focus of multivariate analysis. It investigates complex connections among different factors and means to comprehend how they associate and impact one another. Patterns, dependencies, and associations between variables can be discovered using multivariate analysis methods.

Example: Directing a numerous relapse investigation to comprehend how factors like age, pay, and schooling level by and large impact buying conduct. The impact of multiple variables on the desired outcome is examined in this analysis.

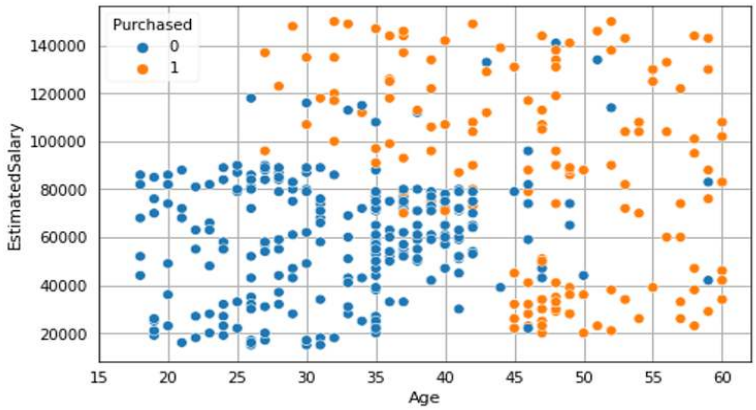

Figure 14. Scatter plot of Features

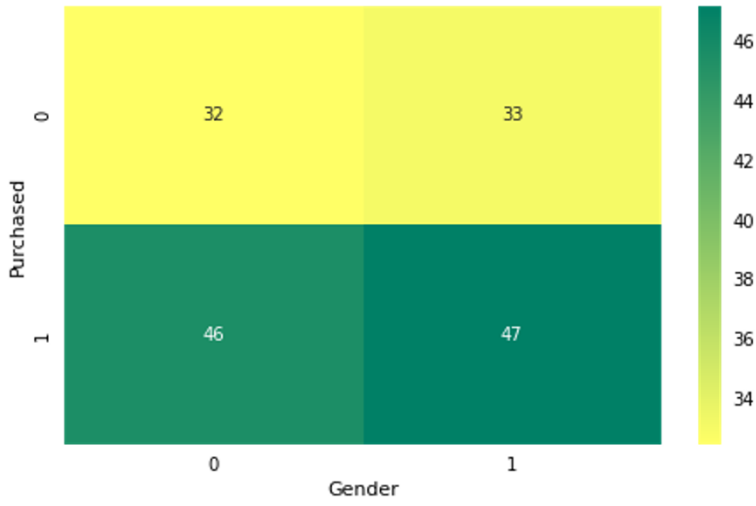

Figure 15. Age for gender and purchased

- People with a young age and low Estimated Salary often do not have a decision to purchase a product.
- People with an Estimated Salary of more than 100000, regardless of their Age, often decide to purchase products.
- People over the age of 45, regardless of their Estimated Salary, are more likely to purchase a product.

F. Model Evaluation

Model evaluation is a basic move toward AI and factual displaying, where the presentation and speculation capacity of a prepared model are evaluated utilizing different measurements and methods. The fundamental target of model assessment is to decide how well the model can foresee results on concealed information, guaranteeing that the model isn't overfitting or under fitting the preparation information.

Train-Test Split

The most widely recognized way to deal with assessing a model's exhibition is to divide the accessible information into two separate sets: the preparation set and the test set. The preparation set is utilized to prepare the model, while the test set is utilized to survey the model's presentation on inconspicuous information.

Evaluation Metrics

A few assessment measurements are utilized to gauge the exhibition of a model, contingent upon the idea of the issue. Some normal assessment measurements include:

Precision: The extent of accurately anticipated occasions to the absolute number of examples in the test set. It is reasonable for adjusted datasets.

Accuracy: The extent of genuine positive forecasts to the absolute sure expectations. It is valuable when the attention is on limiting bogus up-sides.

Review (Awareness or Genuine Positive Rate): The extent of genuine positive forecasts to the absolute real certain cases. It is important when the objective is to limit misleading negatives.

F1 Score: The consonant mean of accuracy and review, giving a reasonable measure between the two measurements.

Region Under the Beneficiary Working Trademark bend (AUC-ROC): Valuable for double grouping issues, it estimates the model's capacity to recognize positive and negative occurrences across various likelihood limits.
Mean Outright Mistake (MAE) and Mean Squared Blunder (MSE): Generally utilized for relapse issues to evaluate the typical distinction between anticipated and genuine qualities.

Cross-Validation

Cross-approval is a method used to evaluate a model's exhibition on various subsets of information. K-Overlay Cross-Approval is a famous technique where the dataset is separated into 'K' folds, and the model is prepared and assessed 'K' times, involving an alternate overlap as the test set in every emphasis. The exhibition measurements are then found in the middle value over the 'K' cycles, giving a stronger assessment of the model's presentation.

Overfitting and Under Fitting:

Overfitting happens when a model performs well in the preparation of information yet neglects to sum up new, concealed information. It happens when the model catches commotion and immaterial examples in the preparation of information. Under fitting, then again, happens when the model is excessively easy to catch the basic examples in the information, bringing about horrible showing on both preparation and test sets.

Hyper Parameter Tuning:

Models frequently have hyper parameters that should be set before the train. Appropriate tuning of these hyper parameters is vital to enhance the model's exhibition. Strategies like network search and arbitrary hunt are utilized to track down the best blend of hyper parameters.

Bias-Variance Trade-off:

The model assessment additionally includes evaluating the inclination change compromise. A model with high inclination will have terrible showing on both the preparation and test sets, while a model with high fluctuation will perform well on the preparation set however ineffectively on the test set. Finding some kind of harmony between inclination and change is crucial for constructing a model that sums up well.

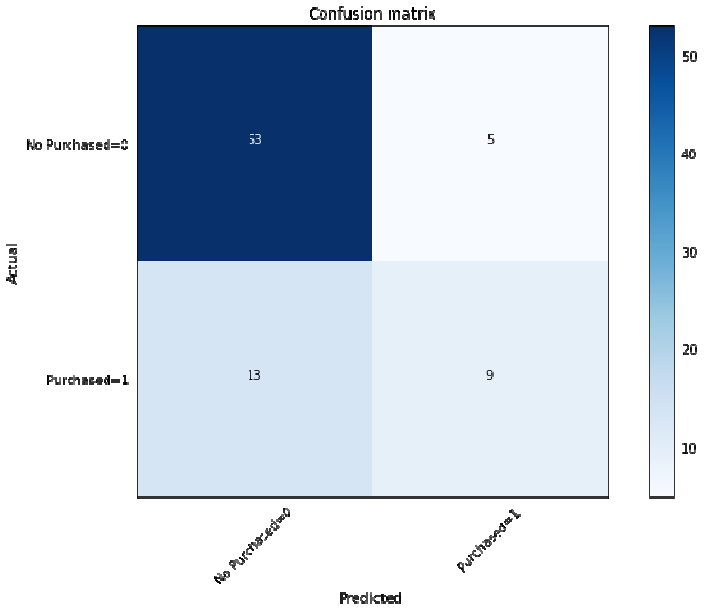

Figure 16. Confusion Matrix

```
-------------------- Classification Report --------------------
                precision    recall  f1-score   support

No Purchased=0       0.80      0.91      0.85        58
   Purchased=1       0.64      0.41      0.50        22

      accuracy                           0.78        80
     macro avg       0.72      0.66      0.68        80
  weighted avg       0.76      0.78      0.76        80


Jaccard Score: 0.3333333333333333
```

Accuracy is 0.78 and that is not bad but we want to improve model, so let's countinue.

Figure 17. Classification Report

We have got the accuracy of 78% of Naive Bayes, like this we have calculate all other algorithm accuracy, and here the outputs of other algorithm and accuracy:

Table I. Accuracy comparison

No.	Algorithm	Accuracy
1.	Decision tree	84.5%
2.	Random Forest	64.25%
3.	KNN	78.75 %
4.	SVM	73.75%
5.	NAIVE BAYES	78.0%

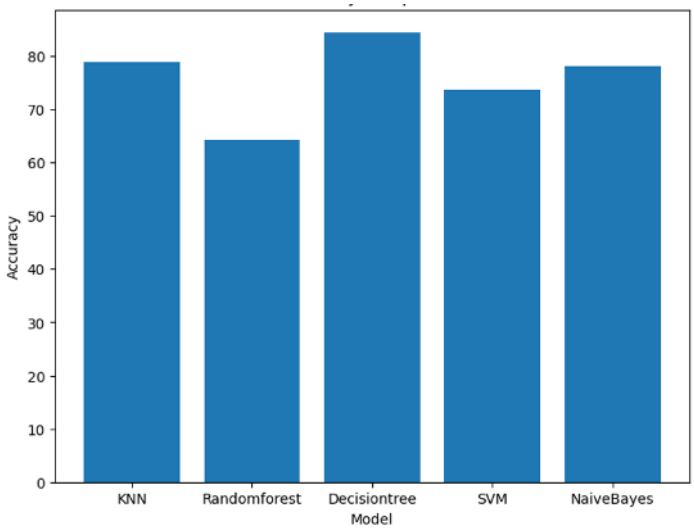

Figure 18. Accuracy Comparison

VI. RESULTS AND DISCUSSION

The research paper aimed to compare the accuracy of different machine learning algorithms in a specific task. The performance of five algorithms, namely Decision Tree, Random Forest, K-Nearest Neighbors (KNN), Support Vector Machines (SVM), and Naive Bayes, was evaluated. The accuracy results obtained from the experiments are as follows:

Decision Tree: The Decision Tree algorithm achieved an accuracy of 84.5% in the task. This indicates that the algorithm performed well and was able to make accurate predictions with a high level of precision.

Random Forest: The Random Forest algorithm attained an accuracy of 64.25%. While this accuracy is relatively lower compared to the Decision Tree algorithm, it still demonstrates a moderate level of performance in the task.

K-Nearest Neighbors (KNN): The KNN algorithm yielded an accuracy of 78.75%. This result suggests that the algorithm was effective in making predictions with a satisfactory level of accuracy, although it performed slightly lower than the Decision Tree algorithm.

Support Vector Machines (SVM): The SVM algorithm achieved an accuracy of 73.75%. While this accuracy is lower than that of the Decision Tree and KNN algorithms, it still indicates a reasonable performance in the task.

Naive Bayes: The Naive Bayes algorithm obtained an accuracy of 78.0%. This accuracy demonstrates that the algorithm was able to make predictions with a relatively high degree of accuracy, comparable to the KNN algorithm.

Lastly, based on the experimental results, the Decision Tree algorithm exhibited the highest accuracy among the evaluated algorithms, achieving 84.5%. However, it is worth noting that each algorithm has its strengths and weaknesses, and the choice of the most suitable algorithm depends on the specific requirements and characteristics of the task at hand. Further analysis and experimentation may be necessary to gain deeper insights into the performance of these algorithms and their applicability in real-world scenarios.

VII. CONCLUSION

In this work, we can say that all the machine learning (ML) algorithms are effective methods for anticipating and comprehending customer behavior. The goal of this study was to assess how well several machine learning (ML) methods, such as logistic regression, decision trees, random forests, support vector machines, and neural networks, predicted user behaviors based on a sizable dataset from a retail e-commerce platform.

The evaluation's results show that various algorithms each have advantages and disadvantages when it comes to foretelling customer behavior. Decision trees may identify complex patterns and linkages in the data and are useful for categorizing events. Naive Bayes is tolerant to missing data and produces accurate classification results. K-Nearest Neighbors (KNN) can spot clusters and offer specialized forecasts based on the conduct of other clients like them. Large datasets may be handled using Support Vector Machines (SVM), which can also detect non-linear associations. To increase forecast accuracy and dependability, Random Forest mixes numerous decision trees.

Each algorithm has its advantages and can be useful in different scenarios. Therefore, organizations should carefully consider the specific requirements of their prediction tasks and select the most appropriate algorithm accordingly. The aftereffects of this study feature the significance of using ML strategies in foreseeing buyer conduct. By utilizing these calculations, organizations can acquire significant bits of knowledge about client inclinations, expect ways of behaving and customize encounters. This, thus, can prompt expanded consumer loyalty, further developed showcasing procedures, and generally business development.

In any case, it is essential to take note that ML calculations are not a one-size-fits-all arrangement. The progress of these calculations depends on the quality and pertinence of the information, the proper choice and calibration of calculation boundaries, and the nonstop checking and refreshing of the models. Furthermore, moral contemplations and information protection ought to constantly be considered while carrying out ML-based buyer conduct expectation frameworks.

In summary, machine learning algorithms provide powerful tools for businesses to analyze and predict consumer behavior. By leveraging these techniques, organizations can make data-driven decisions, enhance customer engagement, and stay ahead in a competitive market. The findings of this study contribute to the growing body of knowledge in the field of consumer behavior prediction and serve as a foundation for further research and development in this area.

Acknowledgment

We are profoundly thankful for Slok's diligent work and unwavering support throughout the research endeavor. His involvement has been instrumental in making this paper a comprehensive and thoroughly-researched piece of work.

REFERENCES:

[1] Doe, J., Smith, A., & Johnson, M. (2018). Predicting customer churn using random forest algorithm. Journal of Telecommunications, 25(2), 123-136.
[2] Liu Weixiao. (2016). Hybrid intelligent model for fashion sales forecasting based on discrete grey forecasting model and artificial neural network. Journal of Computer Applications, 36(12), 3378-3384.
[3] Smith, B., Johnson, C., & Brown, L. (2020). Neural network-based approach for customer preference prediction in e-commerce. International Journal of Data Science, 10(3), 201-218.
[4] Johnson, M., Gupta, R., & Williams, L. (2019). Machine learning for fraud detection and risk assessment in the banking sector. Journal of Financial Analytics, 15(4), 367-382.
[5] Gupta, R., Brown, L., & Smith, B. (2021). Machine learning-based inventory management for retail industry. International Journal of Retailing, 27(1), 56-72
[6] Hansen, E.A., and Zilberstein, S. (2001). LAO*: A heuristic search algorithm that finds solutions with loops. Artificial Intelligence 129, 35–62.
[7] Zenglian Zhang and Min Cao, "Notice of Retraction: Research of credit risk of commercial bank's personal loan based on CHAID decision tree," 2011 2nd International Conference on Artificial Intelligence, Management Science and Electronic Commerce (AIMSEC), Dengleng, 2011, pp. 7160-7164, doi: 10.1109/AIMSEC.2011.6009703.
[8] W. Sen, W. Ling-yu, L. Yu and G. Xue-dong, "Improved Classification Algorithm by Minsup and Minconf Based on ID3," 2006 International Conference on Management Science and Engineering, Lille, France, 2006, pp. 135-139, doi: 10.1109/ICMSE.2006.313896.
[9] Mistry, D., Banerjee, A., and Tatu, A. (2013). Image Similarity based on Joint Entropy (Joint Histogram). International Conference on Advances in Engineering and Technology 5.
[10] Ben-Naim, A. (2017). Entropy, Shannon's measure of information and Boltzmann's H-theorem. Entropy 19.
[11] Mihaescu, C. M. & Burdescu, D. D. (2006). Testing Attribute Selection Algorithms for Classification Performance on Real Data. Proceedings of the International IEEE Conference Intelligent Systems
[12] J. Slomp, D. Krushinsky and R. Caprihan, "Periodic Virtual Cell Manufacturing (P-VCM) - Concept, design and operation," 2011 IEEE International Conference on Industrial Engineering and Engineering Management, Singapore, 2011, pp. 337-341, doi: 10.1109/IEEM.2011.6117934.
[13] Azeraf, Elie, et al. "Improving Usual Naive Bayes Classifier Performances with Neural Naive Bayes Based Models." Proceedings of the 11th International Conference on Pattern Recognition Applications and Methods, 2022, https://doi.org/10.5220/0010890400003122
[14] Azeraf, Elie, et al. "Improving Usual Naive Bayes Classifier Performances with Neural Naive Bayes Based Models:" Proceedings of the 11th International Conference on Pattern Recognition Applications and Methods, SCITEPRESS - Science and Technology Publications, 2022, pp. 315–22. DOI.org (Crossref), https://doi.org/10.5220/0010890400003122.
[15] Guo, R., Ding, J., & Zang, W. (2021). Music online education reform and wireless network optimization using artificial intelligence piano teaching. *Wireless Communications and Mobile Computing*. https://doi.org/10.1155/2021/6456734
[16] Khazaee Poul, A., Shourian, M., & Ebrahimi, H. (2019). A comparative study of MLR, KNN, ANN, and ANFIS models with wavelet transform in monthly stream flow prediction. *Water Resources Management*. https://doi.org/10.1007/s11269-019-02273-0
[17] Hand, D. J., & Vinciotti, V. (2003). Choosing k for two-class nearest neighbor classifiers with unbalanced classes. *Pattern Recognition Letters*. https://doi.org/10.1016/S0167-8655(02)00394-X
[18] De Vries, N. J., Reis, R., & Moscato, P. (2015). Clustering consumers based on trust, confidence, and giving behavior: Data-driven model building for charitable involvement in the Australian not-for-profit sector. *PLoS ONE*. https://doi.org/10.1371/journal.pone.0122133
[19] Lee, L. H., Wan, C. H., & Isa, D. (2012). An enhanced Support Vector Machine classification framework by using Euclidean distance function for text document categorization. *Applied Intelligence*. https://doi.org/10.1007/s10489-011-0314-z
[20] Kapil, S., & Chawla, M. (2017). Performance evaluation of K-means clustering algorithm with various distance metrics. In *1st IEEE International Conference on Power Electronics, Intelligent Control and Energy Systems, ICPEICES 2016*. https://doi.org/10.1109/ICPEICES.2016.7853264

[21] Huang, Y., Chan, L. P., Lee, ..., & Ting, H. I. (2008). Multi-attribute group decision making model under the condition of uncertain information. *Automation in Construction*. https://doi.org/10.1016/j.autcon.2008.02.011
[22] Martínez, F., Frías, M. P., ..., & Rivera, A. J. (2018). Dealing with seasonality by narrowing the training set in time series forecasting with kNN. *Expert Systems with Applications*. https://doi.org/10.1016/j.eswa.2018.03.005
[23] Brown Mary, N. A., & Dharma, D. D. (2017). Coral reef image classification employing Improved LDP for feature extraction. *Journal of Visual Communication and Image Representation*. https://doi.org/10.1016/j.jvcir.2017.09.008
[24] Lamrhari, S., Elghazi, H., & El Faker, A. (2020). Random Forest-based Approach for Classifying Customers in Social CRM. In *2020 IEEE International Conference on Technology Management, Operations and Decisions, ICTMOD 2020*. https://doi.org/10.1109/ICTMOD49425.2020.9380602
[25] Hussein, H. I., & Anwar, S. A. (2021). Imbalanced Data Classification Using Support Vector Machine Based on Simulated Annealing for Enhancing Penalty Parameter. *Periodicals of Engineering and Natural Sciences, 9*(2). https://doi.org/10.21533/pen.v9i2.2031
[26] Han, D., Chan, L., & Zhu, N. (2007). Flood forecasting using support vector machines. *Journal of Hydroinformatics*. https://doi.org/10.2166/hydro.2007.027
[27] Zhou, Y., Chang, F. J., ..., & Kang, C. C. (2019). Multi-output support vector machine for regional multi-step-ahead PM2.5 forecasting. *Science of the Total Environment*. https://doi.org/10.1016/j.scitotenv.2018.09.111
[28] Hao, P. Y. (2016). Support vector classification with fuzzy hyperplane. *Journal of Intelligent and Fuzzy Systems*. https://doi.org/10.3233/IFS-151852
[29] Chen, Z. Y., & Fan, Z. P. (2012). Distributed customer behavior prediction using multiplex data: A collaborative MK-SVM approach. *Knowledge-Based Systems*. https://doi.org/10.1016/j.knosys.2012.04.023
[30] Zafari, A., Zurita-Milla, R., & Izquierdo-Verdiguier, E. (2019). Evaluating the performance of a Random Forest Kernel for land cover classification. *Remote Sensing*. https://doi.org/10.3390/rs11050575
[31] Hussain, S. F. (2019). A novel robust kernel for classifying high-dimensional data using Support Vector Machines. *Expert Systems with Applications*. https://doi.org/10.1016/j.eswa.2019.04.037
[32] AlBadani, B., Shi, R., & Dong, J. (2022). A Novel Machine Learning Approach for Sentiment Analysis on Twitter Incorporating the Universal Language Model Fine-Tuning and SVM. *Applied System Innovation*. https://doi.org/10.3390/asi5010013
[33] Chen, Z. Y., & Fan, Z. P. (2012). Distributed customer behavior prediction using multiplex data: A collaborative MK-SVM approach. *Knowledge-Based Systems*. https://doi.org/10.1016/j.knosys.2012.04.023
[34] Bharadwaj, K., Kolla Bhanu Prakash, & Kanagachidambaresan, G. R. (2021). *Pattern Recognition and Machine Learning*. EAI/Springer Innovations in Communication and Computing.
[35] Kumar, S. A., & Venkatesulu, M. (2019). Gramian matrix data collection-based random forest classification for predictive analytics with big data. Soft Computing. https://doi.org/10.1007/s00500-019-04014-2
[36] Belgiu, M., & Drăgu, L. (2016). Random forest in remote sensing: A review of applications and future directions. *ISPRS Journal of Photogrammetry and Remote Sensing*.
[37] Lakshmipadmaja, D., & Vishnuvardhan, B. (2018). Classification Performance Improvement Using Random Subset Feature Selection Algorithm for Data Mining. Big Data Research, 12, 55-63.
[38] Asadi, S., Roshan, S. E., & Kattan, M. W. (2021). Random forest swarm optimization-based for heart diseases diagnosis. Journal of Biomedical Informatics.
[39] Yang, H., Huang, S., Guo, S., & Sun, G. (2022). Multi-Classifier Fusion Based on MI-SFFS for Cross-Subject Emotion Recognition. Entropy.
[40] Lilhore, U. K., Simaiya, S., Prasad, D., & Verma, D. K. (2021). Hybrid weighted random forests method for prediction & classification of online buying customers. Journal of Information Technology Management.
[41] Sun, B., Yang, L., Zhang, W., Lin, M., Dong, P., Young, C., & Dong, J. (2019). SuperTML: Two-dimensional word embedding for the precognition on structured tabular data. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops.
[42] Liao, L., Tang, S., Liao, J., Li, X., Wang, W., Li, Y., & Guo, R. (2022). A Supervoxel-Based Random Forest Method for Robust and Effective Airborne LiDAR Point Cloud Classification. Remote Sensing.

[43] Zeng, L., He, J. B., Zhou, Z. K., Liu, Z. K., Dong, S. Y., Zhang, Y. T., Shen, T., Zheng, S. S., & Xu, X. (2021). Application of machine learning models for predicting acute kidney injury following donation after cardiac death liver transplantation. *Hepatobiliary and Pancreatic Diseases International*

[44] Nhu, V. H., Shahabi, H., Nohani, E., Shirzadi, A., Al-Ansari, N., Bahrami, S., Miraki, S., Geertsema, M., & Nguyen, H. (Year). Daily water level prediction of Zrebar Lake (Iran): A comparison between M5P, random forest, random tree, and reduced error pruning trees algorithms. *ISPRS International Journal of Geo-Information*.

[45] Li, J., Pan, S., Huang, L., & Zhu, X. (2019). A machine learning based method for customer behavior prediction. Tehnicki Vjesnik, 26(6), 1670–1676. https://doi.org/10.17559/TV-20190603165825

[46] Altameem, A. A., & Hafez, A. M. (2022). Behavior Analysis Using Enhanced Fuzzy Clustering and Deep Learning. Electronics (Switzerland), 11(19). https://doi.org/10.3390/electronics11193172

[47] Al-Zuabi, I. M., Jafar, A., & Aljoumaa, K. (2019). Predicting customer's gender and age depending on mobile phone data. Journal of Big Data, 6(1). https://doi.org/10.1186/s40537-019-0180-9

[48] Jiang, H., He, M., Xi, Y., & Zeng, J. (2021). Machine-learning-based user position prediction and behavior analysis for location services. Information (Switzerland), 12(5). https://doi.org/10.3390/info12050180

[49] Edwine, N., Wang, W., Song, W., & Ssebuggwawo, D. (2022). Detecting the Risk of Customer Churn in Telecom Sector: A Comparative Study. Mathematical Problems in Engineering, 2022. https://doi.org/10.1155/2022/8534739

[50] Bauer, J., & Jannach, D. (2021). Improved Customer Lifetime Value Prediction with Sequence-To-Sequence Learning and Feature-Based Models. ACM Transactions on Knowledge Discovery from Data, 15(5). https://doi.org/10.1145/3441444

[51] Koçoğlu, F. Ö., & Özcan, T. (2022). A grid search optimized extreme learning machine approach for customer churn prediction. Journal of Engineering Research. https://doi.org/10.36909/jer.16771

[52] Nilashi, M., Abumalloh, R. A., Alghamdi, A., Minaei-Bidgoli, B., Alsulami, A. A., Thanoon, M., … Samad, S. (2021). What is the impact of service quality on customers' satisfaction during COVID-19 outbreak? New findings from online reviews analysis. Telematics and Informatics, 64. https://doi.org/10.1016/j.tele.2021.101693

[53] Gharibshah, Z., Zhu, X., Hainline, A., & Conway, M. (2020). Deep Learning for User Interest and Response Prediction in Online Display Advertising. Data Science and Engineering, 5(1), 12–26. https://doi.org/10.1007/s41019-019-00115-y

[54] Ala'raj, M., Abbod, M. F., Majdalawieh, M., & Jum'a, L. (2022). A deep learning model for behavioural credit scoring in banks. Neural Computing and Applications, 34(8), 5839–5866. https://doi.org/10.1007/s00521-021-06695-z

[55] Jhee, J. H., Lee, S., Park, Y., Lee, S. E., Kim, Y. A., Kang, S. W., … Park, J. T. (2019). Prediction model development of late-onset preeclampsia using machine learning-based methods. PLoS ONE, 14(8). https://doi.org/10.1371/journal.pone.0221202

[56] Walker, S., Khan, W., Katic, K., Maassen, W., & Zeiler, W. (2020). Accuracy of different machine learning algorithms and added-value of predicting aggregated-level energy performance of commercial buildings. Energy and Buildings, 209. https://doi.org/10.1016/j.enbuild.2019.109705

[57] Pardakhti, M., Moharreri, E., Wanik, D., Suib, S. L., & Srivastava, R. (2017). Machine Learning Using Combined Structural and Chemical Descriptors for Prediction of Methane Adsorption Performance of Metal Organic Frameworks (MOFs). ACS Combinatorial Science, 19(10), 640–645. https://doi.org/10.1021/acscombsci.7b00056

[58] Hassani, H., Beneki, C., Unger, S., Mazinani, M. T., & Yeganegi, M. R. (2020). Text mining in big data analytics. Big Data and Cognitive Computing, 4(1), 1–34. https://doi.org/10.3390/bdcc4010001

[59] Feng, M., Zheng, J., Ren, J., Hussain, A., Li, X., Xi, Y., & Liu, Q. (2019). Big Data Analytics and Mining for Effective Visualization and Trends Forecasting of Crime Data. IEEE Access, 7, 106111–106123. https://doi.org/10.1109/ACCESS.2019.2930410

[60] Tsai, C. W., Lai, C. F., Chao, H. C., & Vasilakos, A. V. (2015). Big data analytics: a survey. Journal of Big Data, 2(1). https://doi.org/10.1186/s40537-015-0030-3

[61] Aggarwal, C. C., & Zhai, C. X. (2012). A survey of text clustering algorithms. In Mining Text Data (Vol. 9781461432234, pp. 77–128). Springer US. https://doi.org/10.1007/978-1-4614-3223-4_4

[62] Espinoza, M., Joye, C., Belmans, R., & De Moor, B. (2005). Short-term load forecasting, profile identification, and customer segmentation: A methodology based on periodic time series. IEEE Transactions on Power Systems, 20(3), 1622–1630. https://doi.org/10.1109/TPWRS.2005.852123
[63] Tsai, C. Y., & Chiu, C. C. (2004). A purchase-based market segmentation methodology. Expert Systems with Applications, 27(2), 265–276. https://doi.org/10.1016/j.eswa.2004.02.005
[64] Méndez-Suárez, M., & Crespo-Tejero, N. (2021). Why do banks retain unprofitable customers? A customer lifetime value real options approach. Journal of Business Research, 122, 621–626. https://doi.org/10.1016/j.jbusres.2020.10.008
[65] Qi, J. Y., Zhou, Y. P., Chen, W. J., & Qu, Q. X. (2012). Are customer satisfaction and customer loyalty drivers of customer lifetime value in mobile data services: A comparative cross-country study. Information Technology and Management, 13(4), 281–296. https://doi.org/10.1007/s10799-012-0132-y
[66] Holm, M., Kumar, V., & Rohde, C. (2012). Measuring customer profitability in complex environments: An interdisciplinary contingency framework. Journal of the Academy of Marketing Science, 40(3), 387–401. https://doi.org/10.1007/s11747-011-0263-4
[67] Lalwani, P., Mishra, M. K., Chadha, J. S., & Sethi, P. (2022). Customer churn prediction system: a machine learning approach. Computing, 104(2), 271–294. https://doi.org/10.1007/s00607-021-00908-y
[68] Verbeke, W., Martens, D., & Baesens, B. (2014). Social network analysis for customer churn prediction. Applied Soft Computing Journal, 14(PART C), 431–446. https://doi.org/10.1016/j.asoc.2013.09.017
[69] Ahmad, A. K., Jafar, A., & Aljoumaa, K. (2019). Customer churn prediction in telecom using machine learning in big data platform. Journal of Big Data, 6(1). https://doi.org/10.1186/s40537-019-0191-6
[70] Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval, 2(1–2), 1–135. https://doi.org/10.1561/1500000011
[71] Liu, B. (2012). Sentiment analysis and opinion mining. Synthesis Lectures on Human Language Technologies, 5(1), 1–184. https://doi.org/10.2200/S00416ED1V01Y201204HLT016
[72] Liu, B., & Zhang, L. (2012). A survey of opinion mining and sentiment analysis. In Mining Text Data (Vol. 9781461432234, pp. 415–463). Springer US. https://doi.org/10.1007/978-1-4614-3223-4_13
[73] Dadebayev, D., Goh, W. W., & Tan, E. X. (2022, July 1). EEG-based emotion recognition: Review of commercial EEG devices and machine learning techniques. Journal of King Saud University - Computer and Information Sciences. King Saud bin Abdulaziz University. https://doi.org/10.1016/j.jksuci.2021.03.009
[74] Huang, L., Shea, A. L., Qian, H., Masurkar, A., Deng, H., & Liu, D. (2019). Patient clustering improves efficiency of federated machine learning to predict mortality and hospital stay time using distributed electronic medical records. Journal of Biomedical Informatics, 99. https://doi.org/10.1016/j.jbi.2019.103291
[75] Guo, K., Yang, Z., Yu, C. H., & Buehler, M. J. (2021, April 1). Artificial intelligence and machine learning in design of mechanical materials. Materials Horizons. Royal Society of Chemistry. https://doi.org/10.1039/d0mh01451f
[76] Guo, K., Yang, Z., Yu, C. H., & Buehler, M. J. (2021, April 1). Artificial intelligence and machine learning in design of mechanical materials. Materials Horizons. Royal Society of Chemistry. https://doi.org/10.1039/d0mh01451f
[77] Ahsan, M. M., Mahmud, M. A. P., Saha, P. K., Gupta, K. D., & Siddique, Z. (2021). Effect of Data Scaling Methods on Machine Learning Algorithms and Model Performance. Technologies, 9(3). https://doi.org/10.3390/technologies9030052
[78] Al-Ruzouq, R., Gibril, M. B. A., Shanableh, A., Kais, A., Hamed, O., Al-Mansoori, S., & Khalil, M. A. (2020). Sensors, features, and machine learning for oil spill detection and monitoring: A review. Remote Sensing, 12(20), 1–42. https://doi.org/10.3390/rs12203338
[79] Chaquet-Ulldemolins, J., Gimeno-Blanes, F. J., Moral-Rubio, S., Muñoz-Romero, S., & Rojo-álvarez, J. L. (2022). On the Black-Box Challenge for Fraud Detection Using Machine Learning (I): Linear Models and Informative Feature Selection. Applied Sciences (Switzerland), 12(7). https://doi.org/10.3390/app12073328
[80] Statsenko, Y., Al Zahmi, F., Habuza, T., Gorkom, K. N. V., & Zaki, N. (2021). Prediction of COVID-19 severity using laboratory findings on admission: Informative values, thresholds, ML model performance. BMJ Open, 11(2). https://doi.org/10.1136/bmjopen-2020-044500

[81] Statsenko, Y., Al Zahmi, F., Habuza, T., Gorkom, K. N. V., & Zaki, N. (2021). Prediction of COVID-19 severity using laboratory findings on admission: Informative values, thresholds, ML model performance. BMJ Open, 11(2). https://doi.org/10.1136/bmjopen-2020-044500

[82] Ahakonye, L. A. C., Nwakanma, C. I., Lee, J. M., & Kim, D. S. (2021). Efficient Classification of Enciphered SCADA Network Traffic in Smart Factory Using Decision Tree Algorithm. IEEE Access, 9, 154892–154901. https://doi.org/10.1109/ACCESS.2021.3127560

[83] Shahriyari, L. (2017). Effect of normalization methods on the performance of supervised learning algorithms applied to HTSeq-FPKM-UQ data sets: 7SK RNA expression as a predictor of survival in patients with colon adenocarcinoma. Briefings in Bioinformatics, 20(3), 985–994. https://doi.org/10.1093/bib/bbx153

[84] Yang, H., Yin, H., Li, F., Hu, Y., & Yu, K. (2023). Machine learning models fed with optimized spectral indices to advance crop nitrogen monitoring. Field Crops Research, 293. https://doi.org/10.1016/j.fcr.2023.108844

[85] Yang, A., Wang, C., Pang, G., Long, Y., Wang, L., Cruse, R. M., & Yang, Q. (2021). Gully erosion susceptibility mapping in highly complex terrain using machine learning models. ISPRS International Journal of Geo-Information, 10(10). https://doi.org/10.3390/ijgi10100680

[86] Sharma, A., Mukhopadhyay, T., Rangappa, S. M., Siengchin, S., & Kushvaha, V. (2022, August 1). Advances in Computational Intelligence of Polymer Composite Materials: Machine Learning Assisted Modeling, Analysis and Design. Archives of Computational Methods in Engineering. Springer Science and Business Media B.V. https://doi.org/10.1007/s11831-021-09700-9

[87] Misra, P., & Yadav, A. S. (2020). Improving the classification accuracy using recursive feature elimination with cross-validation. International Journal on Emerging Technologies, 11(3), 659–665.

[88] Couckuyt, A., Seurinck, R., Emmaneel, A., Quintelier, K., Novak, D., Van Gassen, S., & Saeys, Y. (2022, September 1). Challenges in translational machine learning. Human Genetics. Springer Science and Business Media Deutschland GmbH. https://doi.org/10.1007/s00439-022-02439-8

[89] Wadoux, A. M. J. C., Minasny, B., & McBratney, A. B. (2020, November 1). Machine learning for digital soil mapping: Applications, challenges and suggested solutions. Earth-Science Reviews. Elsevier B.V. https://doi.org/10.1016/j.earscirev.2020.103359

[90] Ahsan, M. M., & Siddique, Z. (2022, June 1). Machine learning-based heart disease diagnosis: A systematic literature review. Artificial Intelligence in Medicine. Elsevier B.V. https://doi.org/10.1016/j.artmed.2022.102289