INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD)
© 2026 IJNRD | Volume 11, Issue 5, May 2026 | ISSN: 2456-4184 | IJNRD.ORG

PREDICTION OF CUSTOMERS PURCHASE BEHAVIOUR IN A MALL USING MACHINE LEARNING TECHNIQUE

[1]Onima Tigga, [2]Mili Dutta, [3]Jaya Pal
[1]Assistant Professor, [2]Assistant Professor, [3]Assistant Professor
[1]Department of Comp. Sc. & Engg,
[1]Birla Institute of Technology, Mesra, Ranchi, India

Abstract: Understanding consumer buying behavior is important for organizations in order to improve targeting consumers, maximize marketing efforts, and increase overall sales success. In order to identify whether a consumer is tending to make a purchase or not, the Naïve Bayes, Random Forest, Gradient Boosting algorithms are used in this paper's predictive modelling technique. In this work, we used a comprehensive set of data made up of consumer behavior to forecast consumer purchasing behavior using these techniques. The Gradient Boosting technique is found better than the other techniques, and it is used to create a model for prediction because it can handle complicated connections and detect non-linear patterns. The technique includes a number of steps, such as feature engineering, model training, and performance evaluation. On the labelled dataset, the Gradient Boosting classifier is trained, with prediction accuracy being optimized. On the other hand accuracy achieved by Gradient Boosting, Naïve Bayes and Random Forest are significantly improved and they are 80%, 68%, and 78% respectively.

IndexTerms **– Deep Learning, Naïve Bayes, Random Forest, Gradient Boosting.**

I. INTRODUCTION

According to Porter and Millar (1985), companies should add customer value to their products or services in order to maintain a competitive advantage. These days, one of the most popular topics is consumer behaviour predicting. It helps solve problems in a wide range of areas, starting with individual purchase decision forecasting. A wide range of intricate aspects, such as the socio-demographic characteristics of the area and the interactions between various types of nearby services, have an impact on their client flow[1].

The complex task of predicting consumer spending behaviour necessitates the integration of cutting-edge analytics methods with a thorough comprehension of market dynamics and consumer psychology. Businesses may drive growth, improve competitiveness, and cultivate long-term customer connections by utilising predictive analytics to obtain important insights into customer preferences and behaviours. The precision and depth of spending forecasts are anticipated to increase as technology develops further, giving companies the advantage to stay ahead of the curve in a market that is becoming more and more competitive. Understanding and forecasting consumer purchasing trends is now essential for organisations looking to prosper in today's dynamic and fiercely competitive business environment. Forecasting customer behaviour, especially when it comes to making purchases, is extremely valuable to businesses that want to properly adjust their tactics, allocate resources as efficiently as possible, and cultivate enduring connections with their clients [2]. Advancements in machine learning and analytics have provided organisations with the ability to predict future spending patterns with unprecedented precision and to delve deeply into the nuances of customer purchasing behaviour. The field of customer purchasing behaviour spending pattern prediction involves using data-driven techniques to identify the underlying factors impacting consumers' spending patterns. Businesses can gain important insights into the preferences, inclinations, and spending propensities of their customers by leveraging vast amounts of data, including transaction histories, demographic profiles, internet interactions, and market trends. These insights provide the foundation for creating predictive models that explain historical behaviour and forecast future behaviour, enabling businesses to proactively modify their tactics in response to changing customer needs [3]. This introduction seeks to clarify the approaches used in predictive analysis, highlight the revolutionary effects of using such insights on organisations across several sectors, and examine the importance of spending pattern prediction in understanding client purchasing behaviour. The ability to predict customer spending patterns holds immense potential for driving growth, fostering customer loyalty, and maintaining a competitive edge in today's dynamic marketplace. Applications for this ability range from customised marketing campaigns and targeted promotions to inventory optimisation and customer experience enhancement. An effective framework for evaluating enormous volumes of data and deriving insightful conclusions is offered by machine learning, a branch of artificial intelligence. Businesses can obtain a more profound comprehension of client spending behaviour by utilising algorithms that can recognise complex patterns and correlations within data. In order to forecast several aspects of buying behaviour, such as the likelihood that a consumer will make a purchase, how much they are likely to spend, and the variables affecting their decisions, this study aims to leverage the potential of machine learning. Finally, by enabling firms to keep ahead of market trends, increase operational efficiency, and foster enduring client loyalty, the study's findings may contribute to strategic decision-making processes[8].

The main contribution of this paper is as follows: The dataset is first normalized using the z-score normalization technique, and then the attributes are reduced using correlation. Then, with these datasets, we apply machine learning techniques such as Naïve Bayes (NB), Random Forest (RF), and Gradient Boosting (GB). In these machine learning approaches, performance measurements like as accuracy, precision, recall, and F1-score are calculated and compared.

The workflow diagram of the research paper is depicted in Figure 1 as follows:

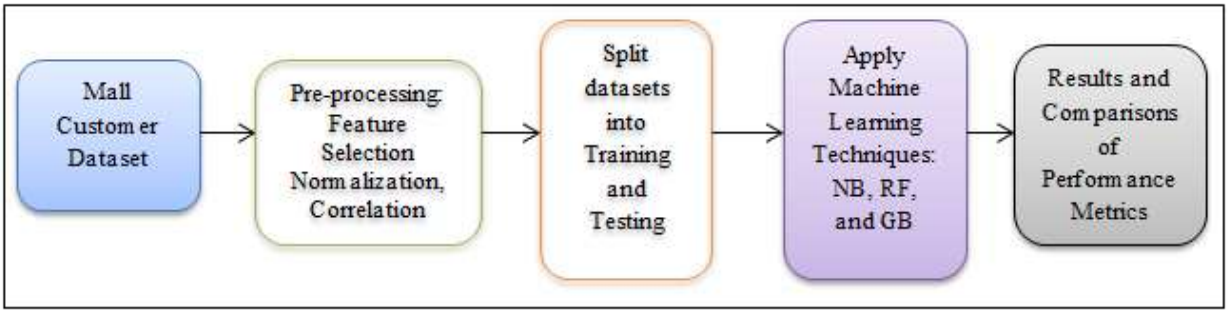

Figure 1: Workflow diagram of the research paper.

The arrangement of this research work is as follows: Explanation of the literature review is incorporated in Section II. The methodologies are defined in Section III. The experiment which includes dataset and evaluation metrics is specified in Section IV. The research work concludes in Section V.

II. LITERATURE REVIEW

A fundamental component of marketing strategy has always been comprehending and forecasting consumer purchasing behaviour. Machine learning (ML) algorithms have given organisations new tools and methods to analyse large volumes of data and derive insights that can be put to use. Numerous research have looked into the use of ML algorithms to forecast the purchasing behaviour of customers in recent years, providing insightful information about consumer preferences, trends, and decision-making processes[2].

Various ML algorithms' predictive power over consumer buying behaviour has been the subject of several research. To analyse the effectiveness of several algorithms in forecasting customer attrition and purchase behaviour in e-commerce contexts, Gupta et al. (2019) examined the efficacy of decision trees, random forests, and support vector machines. They found that ensemble approaches, such random forests, are more effective at producing accurate predictions[3].

When developing predictive models for consumer purchasing behaviour, feature engineering is essential. The extraction of temporal, geographical, and contextual features is one of the novel techniques to feature engineering that researchers have studied. In order to effectively anticipate future purchase behaviour, Zheng et al. (2020) suggested a unique feature engineering approach based on consumer trajectory data and utilising recurrent neural networks (RNNs)[4].

To improve customer engagement and satisfaction, personalised recommendation systems have made extensive use of machine learning algorithms. Using deep learning methods to identify complex patterns in customers' purchase histories and preferences, Li et al. (2018) created a collaborative filtering-based recommendation system for e-commerce platforms. Their research showed a considerable increase in user satisfaction and suggestion accuracy [5].

Social media platforms offer a rich source of data for understanding consumer preferences and sentiments. Artificial intelligence (ML) systems, in particular natural language processing (NLP) methods, have been used to evaluate social media content and forecast consumer purchasing patterns. In order to forecast the success of movie releases, Wang et al. (2019) performed sentiment analysis on Twitter data, illuminating the potential of ML algorithms in predicting consumer behaviour based on social media signals. Additionally, to optimise price choices and increase profitability, machine learning algorithms have been applied to revenue management and dynamic pricing methods. Using machine learning (ML) algorithms to predict demand and modify prices in real-time[6]. Wang et al. (2019) created a dynamic pricing model for ride-sharing services. According to their research, revenue performance and customer happiness can be enhanced by using machine learning (ML) techniques. Researchers have underlined the significance of resolving ethical and privacy problems, even if ML algorithms have the potential to forecast client purchase behaviour with great power. The ethical ramifications of algorithmic bias, data privacy, and the responsible use of consumer data in predictive modelling have all been studied [6]. According to Acquisti et al. (2021), maintaining consumer trust and regulatory compliance in ML-based systems requires ensuring openness and fairness [7].

III. METHODOLOGY

3.1 Dataset

This research work made use of one dataset from the Kaggle [9]. The Mall Customers dataset applied in this research contains 200 data instances with 4 attributes and a class identifying the Spending Score. The Mall Customers dataset is a collection of instances who visited a mall. This dataset includes various attributes such as customer ID, gender, age, annual income, and spending score. It is commonly used for customer segmentation, which is a marketing strategy to group customers based on their

behaviour patterns to enhance sales, revenue, and customer satisfaction. A score assigned by the mall based on customer behavior and spending nature, where 2 represents the highest spending when spending score is greater than or equal to 66, 1 represents the medium spending when spending score is greater than or equal to 26, and 0 represents the lowest spending. The dataset's detailed descriptions are presented in Table 1.

Table 1. Datasets Description

Name of Datasets	Instances	Attributes
Mall Customers	200	4

The attribute description of the Mall Customers dataset is given in Table 2 as follows:

Table 2. Attribute Description of Mall Customers dataset

Attribute Name	Description
Customer ID	Customer ID, continuous from 1 to 200
Gender	Gender, continuous from 0 to 1(female/male)
Age	Age, continuous from 18 to 70
Annual Income	Annual Income, continuous from 15K$ to 137K$
Spending Score	Spending Score, continuous from 1 to 100

The Correlation among the attributes of the given dataset is shown in the Figure 2 as follows:

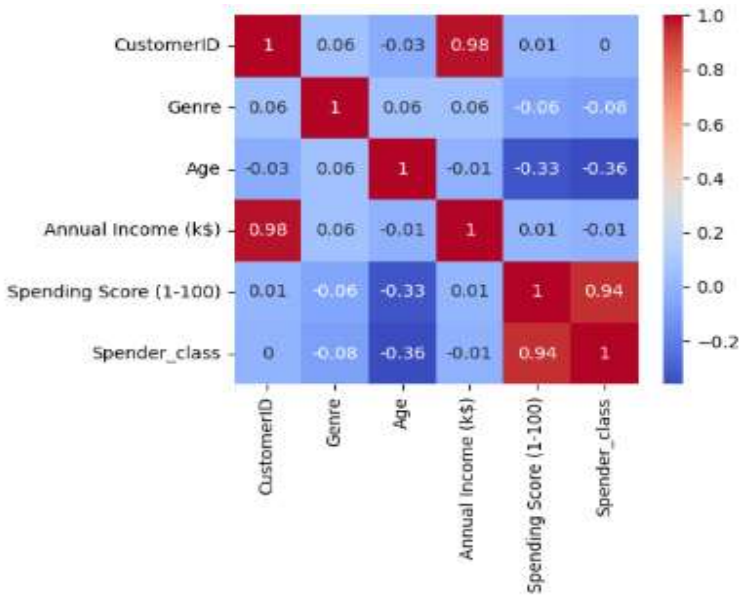

Figure 2. Correlation among the attributes of the dataset.

3.2 Machine Learning Techniques

In this section, various Machine Learning (ML) Techniques, Performance metrics, and Feature Selection are explained as shown in the following sub-sections.

3.2.1 Naïve Bayes

Uncertainty presents challenges in numerous classification tasks. A measure of confidence must be associated with each prediction of class labels. Probability theory accomplishes this by providing a method for quantifying and managing data uncertainty. A model for probabilistic classification is the Naive Bayes (NB) classifier. The classification process involves figuring out how likely a data instance will have a certain class label y based on its set of attribute values x[10, 11]. The posterior probability is articulated through the Equation (3.1) as follows:

$$P(y/x)=\frac{P(x/y)P(y)}{P(x)} \quad (3.1)$$

Where, P(x/y) represents the conditional probability of the attributes. P(y/x) quantifies the probability of observing x given the distribution of instances associated with y. P(y) denotes the prior probability associated with class labels. The following Equation (3.2) determines the value of P(x):

$$P(x)=\sum_j P(x/y_j)\,P(y_j) \tag{3.2}$$

This classifier uses a feature vector $X=(x_1, x_2,...., x_n)$with n dimensions to represent each data sample. Here we look at the m classes$(C_1, C_2,..., C_m)$. X is an unidentified data instance classifier chosen from the class with the highest posterior probability. For every $1 \le k \le m$, the NB classifier assigns an unknown sample X to class C_1, if and only if P (C_1 / X) > P (C_k / X) [13].

3.2.2 Random Forest

Random Forest (RF) is an ML technique that uses ensemble learning to create a classifier by dividing a dataset into sub-datasets. The ensemble models are constructed using these datasets. RF builds many decision trees without pruning, creating a forest that has a predicted value for a specific data sample. The majority voting of the trees determines the predicted value [13]. A Decision Tree model serves as the foundation for the RF machine learning (ML) model. RF models, derived from the idea of Decision Trees, generate a large number of trees ('n') that greatly improve prediction accuracy compared to a single tree. It is achieved by randomly selecting a subset of trees from the training set without replacement. Decision Trees typically have a tree-like configuration with a primary node, the root or decision node, positioned at the top. The Random Forest approach generates M trees based on decision-making and compares them with a single tree to improve prediction accuracy without replacement. Without pruning, we create multiple Decision Trees to construct a forest that forecasts the value for sample data.

One type of ensemble classifier is the Random Forest Classifier, which requires the setting of three important components before training: node size, cardinality of trees, and data points of RF. Next, we use a divider to solve regression and classification problems.

To construct each DT using data points from the training datasets also referred to as bootstrap samples, and then combine these DTs with Random Forest. It uses one-third of the training sample as test data, also known as a sample from the bag, and then returns it. The trees for each decision the majority will vote on the most common categorical features, weighting the trees to determine the predicted category. The RF technique predicts a result based on the result of several Decision Trees. The accuracy of the result increases while increasing the number of trees [14].

3.2.3 Gradient Boosting

Gradient boosting is a machine learning technique that combines multiple weak prediction models into a single ensemble. These weak models are typically decision trees, which are trained sequentially to minimize errors and improve accuracy. By combining multiple decision tree regressors or decision tree classifiers, gradient boosting can effectively capture complex relationships between features[15].

One of the key benefits of gradient boosting is its ability to iteratively minimize the loss function, resulting in improved predictive accuracy. However, one must be conscious of overfitting, which occurs when a model becomes too specialized to the training data and fails to generalize well to new instances. To mitigate this risk, practitioners must carefully tune hyperparameters, monitor model performance during training and employ techniques like regularization, pruning or early stopping. By understanding these challenges and taking steps to address them, practitioners can successfully harness the power of gradient boosting—including the use of regression trees—to develop accurate and robust prediction models for various applications [16].

Mean Squared Error (MSE) is one loss function used to evaluate how well a machine learning model's predictions match actual data. MSE calculates the average of the squared differences between the predicted and observed values.

After each tree is trained its predictions are shrunk by multiplying them with the learning rate η which ranges from 0 to 1. This prevents overfitting by ensuring each tree has a smaller impact on the final model [17]. Once all trees are trained predictions are made by summing the contributions of all the trees. The final prediction is given by the formula (3.3):

$$y_{\text{pred}}=y_1+\eta \cdot r_1+\eta \cdot r_2+ \cdot +\eta \cdot r_n \tag{3.3}$$

Where r_1, r_2, ….. ,r_n are the errors predicted by each tree.

3.3 Feature Selection Approaches

In the field of ML, processing of high-dimensional data is a stimulating challenge. Dimensionality Reduction is important as it enhances data comprehension, reduces computing costs, and prevents models from becoming overly complex or simplistic by refining them.

i) ***Pearson Correlation:*** Pearson's Correlation is one way of assessing the relationship between a feature and an outcome variable; it may also be used to choose features. Pearson's Correlation functions as a technique to examine the association between a feature and the response variable, and it can be employed for feature selection. This methodology is employed to discern the interconnections among the attributes present within a dataset. The Correlation (r)

value lies between [-1,1], where -1 shows negatively correlated, +1 shows positively correlated and 0 shows no Correlation as shown in Equations (3.4) & (3.5).

The Pearson Correlation Coefficient (r) is shown as

$$r = \frac{\text{cov}(x,y)}{\sqrt{\text{var}(x)\text{var}(y)}} \quad (3.4)$$

Where cov & var are Covariance and Variance respectively. Also, r can be evaluated as follows:

$$r = \frac{\sum_{i=1}^{m}(x_i-\overline{x})(y_i-\overline{y})}{\sqrt{\sum_{i=1}^{m}(x_i-\overline{x})^2 \sum_{i=1}^{m}(y_i-\overline{y})^2}} \quad (3.5)$$

Where $\overline{x}$ = mean of independent data, $\overline{y}$ = mean of dependent data, m = number of data points, x_i = individual independent variable, y_i = individual dependent variable. Correlation (r) represents the linear relation between the dependent and independent variables. To overcome the risk of Overfitting, simple Correlation Coefficient (r) is used for nonlinear preprocessing [10].

3.4 Normalization

Z-Score Approach: Standardization takes place after the pre-processing phase. The given Equation demonstrates the procedure of standardizing data via Z-Score Standardization. This method employs a scaling range of -1 to 1 for feature values. This study employs the characteristics of a Normal Distribution, defined by Equation (3.6) [12].

$$z = \frac{x-\mu}{\sigma} \quad (3.6)$$

Where μ is a Mean having a value of 0 and a Standard Deviation σhaving 1 for each attribute.

3.5 Performance Metrics

To validate how well the machine learning algorithms perform for classification methods and a comparison with NB, RF, and Gradient Boosting techniques for (100%) of data, for the Mall Customers datasets, we have used performance metrics such as: accuracy, precision, recall, and F1-score.

3.6 Confusion Matrix

It is a square matrix that is n x n, where n is a class label, employed for assessing the performance of the ML model, given in Figure 3. The diagonal values in the matrix determine the accurate predictions. In this matrix, columns denote actual values and rows shows predicted values and vice-versa [12].

		Actual class: True	Actual class: False
Predicted class	True	T_p	F_p
	False	F_n	T_n

Figure 3: Confusion Matrix

Where, T_p (True Positive): In this table of confusion matrix, it gives the actual class's value as true, and as well as the predicted class's is also true. T_n (True Negative): In this case, it is found that the actual class's value is false, and it also shows the predicted value to be false. F_p (False Positive): In this case, it is said that the actual class's value is false but the predicted class shows it as true. F_n (False Negative): In this case, the value of actual class is true but the value of predicted class is false.

The formulae used for performance evaluation are as follows:

i. Accuracy: It evaluates the correct predictions of the model for the overall dataset. It is a good performance measure which is given in the Equation (3.7), but not as good for the imbalanced data [12].

$$\text{Accuracy} = \frac{T_p+T_n}{T_p+T_n+F_p+F_n} \quad (3.7)$$

ii. Precision: It calculates the number of the classes which is correctly predicted out of the total true values. This can be determined by using the following Equation (3.8) [14].

$$\text{Precision} = \frac{T_p}{T_p+T_n} \quad (3.8)$$

iii. Recall: Recall is the number of the actual classes to be able to predict correctly with the model, given in Equation (3.9).

$$\text{Recall}=\frac{T_p}{T_p+F_n} \tag{3.9}$$

iv. F1-Score: It is defined by the following Equation (3.10) [12].

$$F1-\text{Score}=2*\frac{\text{Recall*Precision}}{\text{Recall+Precision}} \tag{3.10}$$

v. *MSE:* MSE is calculated using the Equation (3.11), where n represents the number of tuples [13, 14].

$$\text{MSE}=\frac{1}{n}\sum_{i=1}^{n}(P_i-A_i)^2 \tag{3.11}$$

IV. RESULTS AND EXPERIMENTS

In this research work, we have used three ML Techniques: Naïve Bayes, Random Forest, and Gradient Boosting for Mall Customers dataset. The performance analyses of the above techniques are depicted in Table 3.

Table 3. Performance Measures

Machine Learning Techniques	Performance Measures				
	Accuracy	**Precision**	**Recall**	**F1-Score**	**MSE**
Naïve Bayes	0.68	0.78	0.68	0.66	0.475
Random Forest	0.78	0.77	0.78	0.77	**0.375**
Gradient Boosting	**0.80**	**0.83**	**0.80**	**0.80**	0.425

In Table 3, Gradient Boosting gives Accuracy, Precision, Recall, and F1-score as 0.80, 0.83, 0.80, and 0.80 respectively for dataset Mall Customers which are better than NB and RF as shown in bold. The pictorial representation of the accuracy of three ML techniques Naïve Bayes, Random Forest, and Gradient Boosting for Mall Customers dataset are shown in Figure 4.

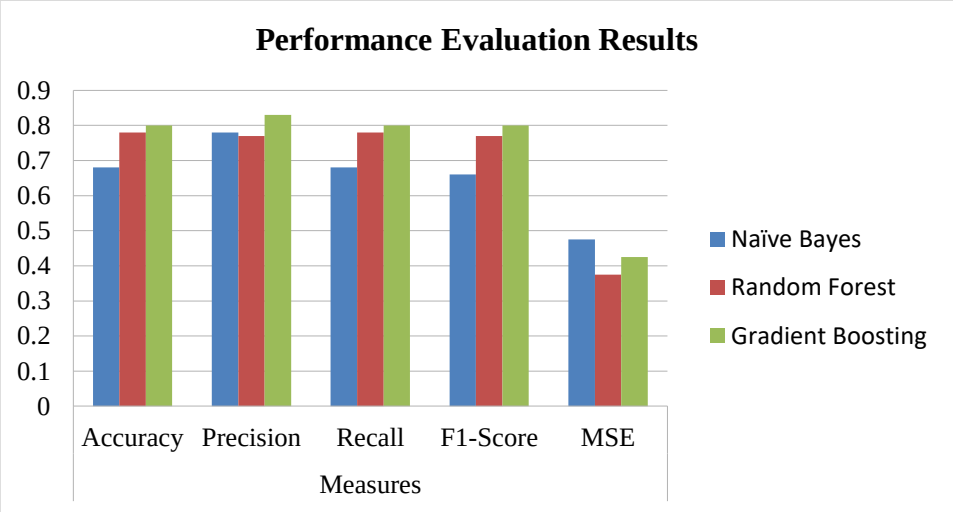

Figure 4. Performance Measurements of NB, RF, and GB.

The Confusion Matrix for the Gradient Boosting is shown in Figure 5 as below:

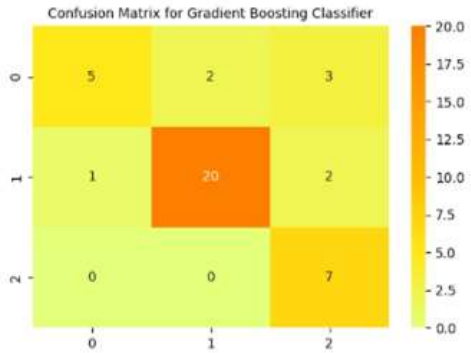

Figure 5. Confusion matrix of the Gradient Boosting

V. CONCLUSION AND FUTURE WORK

The knowledge collected from this study can help organizations create more successful marketing plans, individualized deals, and consumer retention programs. In conclusion, this study has paved the way for future research and development. Future efforts in this area show excellent promises for improving the estimation of customer purchase behavior and aiding informed business choices with more data, the improvement of features, and the development of different modelling methodologies. In this research work, several classifiers like Naïve Bayes (NB), Random Forest (RF), and Gradient Boosting classifiers have been applied on the mall dataset. Overall performance of the gradient boosting method is significantly improved compared to other classifiers and state-of-the-art methods.

Although both our model's accuracy are not as high as we are anticipating, it is crucial to highlight that there is still a lot of space for future enhancements and further development. Despite this, our study gave useful insight into the dataset and probable factors affecting purchasing decisions.

There are various directions for future research that can be taken in order to increase the model's capacity for prediction. The addition of supplemental characteristics, including data on consumer browsing habits, demographics, or past purchases of the customer, may significantly enhance model performance. Future efforts in this area show excellent promises for improving the estimation of customer purchase behavior and aiding informed business choices with more data, the improvement of features, and the development of different modelling methodologies.

REFERENCES

[1] Porter, M.E. and Millar, V.E. (1985) How Information Gives You Competitive Advantage. Harvard Business Review, 63, 149-160.
[2] Bharadwaj, S. G., Varadarajan, P. R., & Fahy, J. 1993. Sustainable Competitive Advantage in Service Industries: A Conceptual Model and Research Propositions. Journal of Marketing, 57(4), 83–99. https://doi.org/10.2307/125222.
[3] Gupta, A., Jaiswal, S., & Pandey, M. 2019. A comparative study of classification algorithms for customer purchase behavior analysis, International Journal of Engineering and Advanced Technology, 8(5), 2911-2917.
[4] Zheng, Y., Lin, X., & Han, J. 2020. Trajectory-based purchase prediction with deep learning, Proceedings of the AAAI Conference on Artificial Intelligence, 34(7), 11652-11659.
[5] Li, Y., Zhang, Y., & Li, H. 2018. Deep learning for personalized product recommendation. *Proceedings of the ACM International Conference on Information and Knowledge Management, 187-196.
[6] Wang, Z., Wang, X., & Li, Y. 2019. Sentiment analysis for movie prediction: A hybrid approach combining CNN and LSTM. Expert systems with Applications,132,292-302
[7] Acquisti, A., Taylor, C., & Wagman, L. 2021. The economics of privacy. Journal of Economic Literature,59(2), 442-492
[8] Vaganov,D.,Funkner,A.,Kovalchuk,S.,Guleva,V.,Bochenina,K.,2018. Forecasting purchase categories with transition graphs using financial and social data in : International Conference on Social Informatics, Springer, pp. 439-454. Patterson P. G., 2007, Demographic correlates of loyality in a service context, Journal of Service Marketing.
[9] https://www.kaggle.com/datasets/kandij/mall-customers
[10] Khilari N., Hadawale P., Shaikh H., Kolase S., 2021. Analysis of Machine Learning Algorithm to Predict Wine Quality, International Research Journal of Engineering and Technology (IRJET). Volume: 08 Issue 12, DOI: 10.32628/IJSRSET229235
[11] Bhardwaj P., Tiwari P., Olejar K., Parr W., Kulasiri D., 2022, A Machine Learning application in wine quality prediction, Machine Learning with Applications, Vol 8, 100261.
[12] Tigga O., Pal J. and Mustafi D. 2023, Performance Analysis of Machine Learning Algorithms for Data Classification, International Conference on Machine Intelligence with Applications (ICMIA 2023), Birla Institute of Technology, Mesra , Off Campus Lalpur, Ranchi, India,07-08, December 2023, https://doi.org/10.1063/5.0214183
[13]Carpita M., Goli S., 2023. Categorical Classifiers in multiclass classification with imbalanced datasets. WILEY, https://doi.org/10.1002/sam.11624 .

[14] Tigga O., Pal J., Mustafi D., 2023. A Comparative Study of Multiple Linear Regression and KNNs using Machine Learning, Fifth IEEE International Conference on Electrical, Computer and Communication Technologies, **INSPEC Accession Number:** 23456658, IEEE Xplore, https://**doi: 10.**1109/ICECCT56650.2023.10179713.
[15] Murari T., Chandra Sekhara Rao M. V. P., 2024. Evaluating the Performance of Xgboost and Gradient Boost Models with Feature Extraction in FMCG Demand Forecasting: A Feature-Enriched Comparative Study, Journal of Theoretical and Applied Information Technology, Vol. 102. No 9, ISSN: 1992-8645.
[16] Abdullah-All-Tanvir, Ali K., Muzahidul Islam, Salekul I., Swakkhar S. 2023. A gradient boosting classifier for purchase intention prediction of online shoppers, Heliyon 9. e15163, https://doi.org/10.1016/j.heliyon.2023.e15163.
[17] Arushi S., Renuka D. S. M. 2023. Predicting Online Customer Purchase using Gradient Boost Classifier, International Journal for Research in Applied Science & Engineering Technology (IJRASET) ISSN: 2321-9653; IC Value: 4598; Volume 11 Issue VI - Available at www.ijraset.com