The Consumer Behavior and Shopping Habits: A Deep Dive into Customer Choices & Web Application

Md. Baker
Electrical and Computer Engineering
North South University
Dhaka, Bangladesh
md.baker@northsouth.edu

Syed Niamul Kazbe Rayian
Electrical and Computer Engineering
North South University
Dhaka, Bangladesh
syed.rayian@northsouth.edu

Fatema Akter Rimi
Electrical and Computer Engineering
North South University
Dhaka, Bangladesh
fatema.akter02@northsouth.edu

***Abstract*—This research paper explores the prediction of consumer purchase behavior using a combination of public and independently collected datasets and applied various machine learning models. The study implements and compares the performance of eight classifiers: K-Nearest Neighbors (KNN), Decision Tree, Random Forest, Logistic Regression, Naive Bayes, Support Vector Machine (SVM), AdaBoost, and Gradient Boosting. We utilized hyperparameter optimization techniques like GridSearchCV to enhance model performance. The dataset includes factors from the Theory of Planned Behavior (TPB): Attitude (ATTB), Social Norm (SN), Perceived Behavior (PB) and Perceived Behavioral Control (PBC). Our results indicate that AdaBoost achieves the highest accuracy at 90%, while Decision Tree exhibits the lowest accuracy at 81%. The findings underscore the significance of personalized marketing strategies and the influential role of social media in shaping consumer behavior. A Django web application was developed to predict consumer purchase behavior as High or Low, based on these optimized models. Furthermore, we applied Fuzzy Association Rule mining to interpret the model and Explainable AI techniques, specifically LIME (Local Interpretable Model-agnostic Explanations), to provide transparency and interpretability to our model's predictions.**

Keywords— Consumer Purchase Behavior, AdaBoost, Explainable AI (XAI), Local Interpretable Model-agnostic Explanations (LIME), Fuzzy Association Rule Mining, Social Media Marketing, E-commerce, Theory of Planned Behavior (TPB), Hyperparameter Optimization, GridSearchCV,

I. Introduction

The advent of social media marketing represents a monumental development in the history of commerce, fundamentally transforming traditional marketing methodologies and ushering marketers into a new era. Over the last decade, this technical revolution has re-centered consumers within the business world, equipping marketers with an innovative toolkit to engage and integrate them into brands in unprecedented ways. Social media marketing enables a dynamic interaction between businesses and consumers, fostering deeper connections and brand loyalty through personalized and real-time communication. Understanding the influence of social media on purchase behavior (PB) decision-making is crucial for marketers. In the field of e-commerce, instead of classical market segmentation, personalization based on machine learning (ML) is an effective and innovative method of users' classification that applies approaches, providing a more efficient forecast of consumers' specifics and preferences [5]. For example, when organizations launch new products or services, early adoption by a small group of consumers often paves the way for broader acceptance. The growing number of social media users has attracted marketers. Marketers have recognized that social media marketing is an important part of their marketing communication strategies. Also, social media helps organizations communicate with their customers [1].

This project is about consumer behavior and shopping habits, which encompass the myriad ways in which individuals make decisions about purchasing products and services, influenced by a variety of psychological, social, and economic factors. Understanding these behaviors is crucial for businesses aiming to tailor their marketing strategies effectively. Moreover, by using social media, consumers have the power to influence other buyers through reviews of products or services used [2]. Consumers are also influenced by other psychosocial characteristics like income, purchase motivation, company presentation, company or brand's presence on social networks, demographic variables (age, sex, disposable income, etc.), workplace method of payment, type of store (online or physical), etc. Consumers are influenced by internal factors such as personal preferences, attitudes, and perceptions, as well as external factors including social influences, cultural trends, and economic conditions. The decision-making process typically involves several stages: recognition of need or desire, information search, evaluation of alternatives, purchase decision, and post-purchase behavior. In recent years, the rise of e-commerce and digital marketing has significantly altered shopping habits. Consumers now have access to vast amounts of information online, enabling them to compare products, read reviews, and make informed decisions more easily than ever before. Additionally, social media platforms have become powerful tools for shaping consumer opinions and driving purchase decisions through targeted advertising and influencer endorsements.

Social media marketing leverages various platforms to promote businesses, engage with consumers, and stimulate business growth. As of April 2023, there were 4.48 billion social media users globally, representing nearly 60% of the world's population [3]. This vast user base underscores the widespread appeal and utility of social media platforms, which offer convenience and valuable services to end users. The burgeoning popularity of social media is closely linked to its ability to connect people, provide entertainment, and facilitate information sharing, making it an indispensable

tool in modern life [3]. Furthermore, the e-commerce industry is poised to generate approximately $3.65 trillion in revenue by 2023, reflecting a compound annual growth rate of 11.22%. This growth trajectory is expected to continue, with market volume projected to reach $5.58 trillion by 2027. These figures highlight not only the escalating use of social media but also the increasing dependence of consumers on these platforms for their shopping needs. By tapping into this potential, e-commerce businesses can enhance their visibility, foster customer relationships, and drive substantial growth [3]. The integration of social media marketing strategies is thus becoming essential for businesses aiming to thrive in the digital economy, as it allows them to effectively connect with consumers, influence purchasing decisions, and sustain competitive advantage in a rapidly evolving market landscape.

This article [3] introduces a groundbreaking approach that merges the theory of planned behavior (TPB) with advanced machine learning techniques to develop precise predictive models for consumer purchase behavior. The study delves into three core elements of TPB—attitude, social norms, and perceived behavioral control—to uncover the primary factors influencing online purchasing decisions. Employing eight machine learning algorithms, including K-nearest neighbor, Decision Tree, Random Forest, Logistic Regression, Naive Bayes, Support Vector Machine, AdaBoost, and Gradient Boosting, the research aims to forecast consumer purchasing patterns with high accuracy. Among these, gradient boosting emerged as the most effective model. To further demystify this model's predictions, explainable AI (XAI) techniques, specifically LIME (local interpretable model-agnostic explanations), were utilized, providing transparency into the decision-making process of the gradient boosting model. This integration of TPB and machine learning offers valuable insights into consumer behavior on social media platforms, facilitating a more efficient allocation of marketing resources. By identifying customers with stronger purchasing inclinations, managers can strategically direct their budgets to maximize returns. The study assessed the predictive power of seven combinations of TPB factors in determining purchase behavior, highlighting the importance of factors such as ATTD1, ATTD3, and PBC4. The XAI analysis revealed that lower Likert scores on these variables are associated with reduced motivation for online purchases, offering a nuanced understanding of consumer tendencies [3]. This unique contribution not only advances the field of consumer behavior analysis but also provides practical applications for enhancing marketing strategies through targeted investments, thereby optimizing the impact of marketing efforts in the digital age.

This research [4] investigates five key dimensions of social network marketing—entertainment, customization, interaction, word of mouth, and trends—that can influence consumer purchase behavior (CPB), using the Facebook Marketplace in Hungary as a case study. The study employs a combination of structural equation modeling (SEM) and unsupervised machine learning approaches, specifically hierarchical cluster analysis (HCA) and K-means algorithms, to analyze the data. The findings of this research provide robust evidence supporting all five hypotheses, demonstrating that each factor significantly impacts CPB. The first hypothesis (H1) confirms that entertainment positively influences consumer purchase behavior, highlighting the role of engaging content in attracting and retaining customers. The second hypothesis establishes that customization, or the tailoring of content and advertisements to individual preferences, enhances consumer purchasing decisions on the platform. Interaction or communication, addressed in the third hypothesis, also shows a positive relationship with CPB, underscoring the importance of active engagement between sellers and buyers. The fourth hypothesis confirms that word of mouth, including recommendations and reviews, is a powerful driver of purchase behavior, reflecting the value of social proof and peer influence [4]. Lastly, the study verifies the positive effect of trends and influencer marketing, as posited in the fifth hypothesis, illustrating how contemporary and influential content can sway consumer decisions. This comprehensive analysis not only reinforces the significance of these five dimensions in social network marketing but also provides practical insights for marketers aiming to optimize their strategies on platforms like Facebook Marketplace. By leveraging these factors, businesses can better influence consumer behavior, enhance user engagement, and ultimately drive sales in the digital marketplace [4].

This study [5] investigates the interplay between fundamental personality dimensions and users' online shopping behaviors, utilizing the Ten-Item Personality Inventory (TIPI) test to construct individual consumer profiles. By applying machine learning models such as decision trees and random forests, the research effectively predicts consumer behavior in e-commerce settings based on distinct personality traits. The dataset encompasses responses in three languages—English, German, and Bulgarian—collected through a comprehensive survey that includes sections on online store features, user preferences, personality characteristics, risk averseness, and demographic analysis. The findings highlight the potential of personalized user interfaces, tailored to specific user groups, to enhance customer experiences and drive business success across various sectors, including strategic management, marketing, and e-commerce [5]. This user-centric approach, which combines personality insights with predictive techniques, shows promise for refining decision-making processes and boosting overall customer satisfaction in online retail environments. Additionally, the study's conceptual framework offers scholars and industry experts a systematic method for analyzing and interpreting consumer behavior. The use of the TIPI test allows for the identification of correlations between personality traits and preferences for e-commerce website functionalities, providing detailed insights into how specific personality characteristics influence purchasing decisions. By evaluating forecast accuracy using the mean absolute percentage error (MAPE), the research adds a layer of quantitative rigor, ensuring the reliability and validity of its findings [5]. This comprehensive approach underscores the significance of integrating psychological insights with technological advancements to foster more personalized and effective e-commerce strategies.

II. Proposed System

In this part, we are going to describe our dataset, preprocessing system, and different models that we have used in our work.

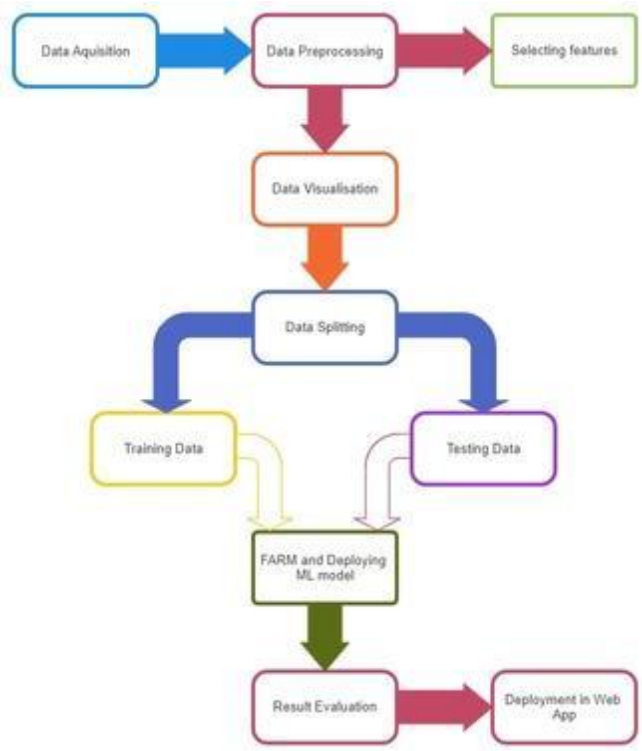

Fig. 1: Working sequences of the proposed system.

In Fig.1, we show the workflow of our project. In this project, we used both a public dataset and our own collected data to apply our machine learning model. Initially, we acquired the data and proceeded with data preprocessing, including selecting relevant features. Next, we visualized the data to understand its distribution and patterns. We then split the data into training and testing sets, with 80% for training and 20% for testing. The training data was used to train the machine learning models, while the testing data was reserved for evaluation. We applied Fuzzy Association Rule Mining to identify important patterns and optimize the model before deployment. Finally, we evaluated the results and deployed the model in a web application for practical use. Our results and the comparative analysis of different machine learning models, including their Accuracy, Precision, Recall & F1-score, are discussed in the Result and Discussion section.

A. Dataset

The initial dataset used in this study was derived from the research conducted by Zhou et al. [6]. This dataset comprised 219 complete responses collected through a meticulously structured questionnaire. The rationale behind using a survey-based data collection method was rooted in its efficiency and ability to gather targeted responses to specific questions. Surveys are particularly advantageous in collecting data from a large sample size, offering valuable insights into participants' perspectives while ensuring anonymity, thereby mitigating common method bias.

The distribution of the questionnaires was facilitated through Google Forms, targeting social media users residing in Malaysia. Zhou et al. employed a snowball sampling methodology, which is a non-probabilistic technique useful for gathering data from individuals with specific traits relevant to the study. Initial participants were recruited and then encouraged to enlist further subjects, ensuring the inclusion of active internet users. The survey link was distributed via WeChat, a popular social media platform, to maximize reach among Malaysian online users.

The questionnaire was divided into two sections: Section A, consisting of 10 questions related to demographic characteristics and basic social media usage, and Section B, containing 16 questions aimed at understanding the perception of factors influencing purchase intentions and decisions. Responses in Section B were collected using a 5-point Likert scale, ranging from "Strongly Disagree" (1) to "Strongly Agree" (5).

B. Data Integration

In addition to the data collected by Zhou et al., we conducted an independent data collection process to augment the dataset. Our survey followed a similar structure and methodology to ensure consistency and comparability. A total of 92 additional responses were collected, bringing the overall dataset to 311 participants.

Our data collection targeted the same demographic group: social media users residing in Bangladesh. We employed the same snowball sampling technique, distributing the survey via Google Forms and using social media platforms, including Facebook, to reach potential respondents. This approach ensured that the extended dataset maintained the same characteristics and demographic focus as the original dataset.

C. Survey Design

Both the original and additional surveys were designed to gather comprehensive data on the factors influencing online purchase behavior. The structured questionnaire ensured that responses were consistent and relevant to the study's objectives. The survey design incorporated multiple strategies to minimize common method bias, such as ensuring participant anonymity and crafting concise, precise questions to avoid ambiguity.

Common method bias was mitigated through several strategies, including ensuring anonymity and designing a concise and clear questionnaire. Non-response bias was assessed by comparing early and late respondents using a paired samples t-test in SPSS (version 29). This analysis showed minimal non-response bias, confirming the reliability of the dataset. The results of our study suggest that of the 16 variables examined of selected TPB factors.

TABLE I. SELECTED TPB FACTORS AND DEFINITIONS WITH RESPECTIVE FEATURE SET

TPB Factors	Definitions	Feature Set
Attitude (ATTD)	Attitude pertains to the extent to which an individual possesses a positive or negative assessment of the behavior under consideration.	ATTD1: Social media advertisements can assist me to learn about the existence of the product. ATTD2: Compared to other advertising platforms, social media advertisements can more easily get my attention. ATTD3: I will look for more production-related information if prominent keywords like promotion and discount are used on social media. ATTD4: I have previously engaged in the acquisition of a product that came to my attention via social media.
Social Norms (SN)	Social norms refer to the societal expectations and moral beliefs that exert social pressures on individuals, influencing their behavior and actions.	SN1: My family influence my purchasing decision towards social media marketing. SN2: People around me believe that I should purchase on social media. SN3: It makes me happy if many individuals use social media to make purchases. SN4: My friends encourage me to purchase through social media.
Perceived Behavioral Control (PBC)	Perceived behavioral control pertains to an individual's subjective assessment of the level of ease or difficulty associated with executing a particular behavior.	PBC1: Frequent advertisement of a product on social media led me to purchase it. PBC2: I will use social media as a purchasing reference channel in the future. PBC3: I will recommend my friends to use social media as a reference when deciding what to buy in the future. PBC4: I will recommend my family to use social media as a reference when deciding what to buy in the future.
Purchase Behavior (PB)	Purchase behavior refers to a person's willingness to buy a product.	PB1: I am willing to purchase a social media-marketed product. PB2: There is a high likelihood that I would purchase a product due to social media's influence. PB3: I am readily influenced by social media advertisements and subsequently engage in purchasing behavior. PB4: As a result of social media's influence, I had the experience of purchasing a product.

D. Data visualization

Figure 2,3,4,5 provides comprehensive insights into various demographic and behavioral aspects of the surveyed individuals, specifically focusing on gender distribution, attention to social media advertisements (ADSM), purchasing behavior influenced by social media, and age distribution. Figure 6,7,8,9 constructs Attitude, Social Norm, Perceived Behavioural Control, and Purchase Behavior. These charts illustrate the distribution of responses across various categories.

Fig 2, illustrates the gender distribution among the respondents. It shows that 50.5% of the participants are male, while 49.5% are female. This nearly equal distribution indicates a balanced representation of both genders in the survey, with a slight majority of males.

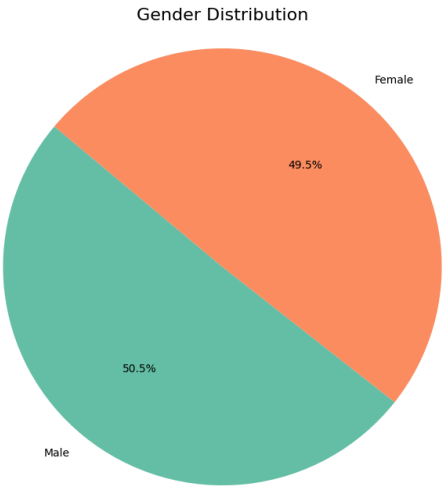

Fig. 2: Gender Distribution

Each slice of the age distribution pie chart is labeled with the corresponding percentage in fig 3, providing a clear visual comparison of the size of each age group within the overall dataset. The chart effectively shows that the majority of respondents are between 23-28 years old, followed by those in the 17-22 years old range. The remaining age groups are comparatively smaller, with the "Others" category being the least represented.

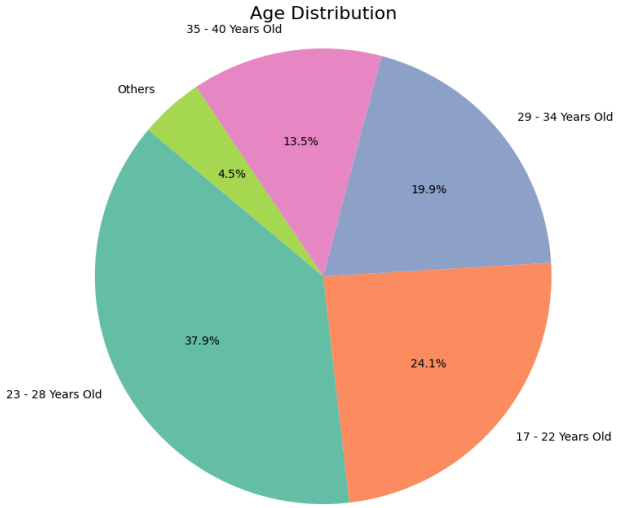

Fig. 3: Age Distribution

Fig 4, focuses on how many respondents pay attention to advertisements on social media. It reveals that 64% of the participants do pay attention to these ads, whereas 36% do not. This significant majority indicates that social media is an effective platform for advertising, capturing the interest of a large portion of users.

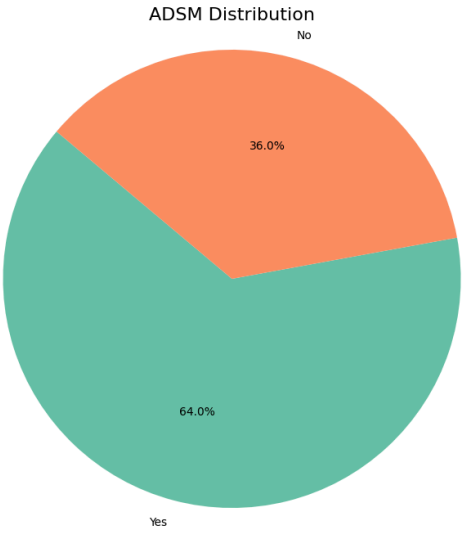

Fig. 4: Attention to Advertisements on Social Media (ADSM)

Fig 5, depicts the respondents' purchasing behavior influenced by social media. An overwhelming 90.7% of the participants have made purchases due to social media influence, while only 9.3% have not. This highlights the powerful impact social media has on consumer purchasing decisions, making it a critical tool for marketers.

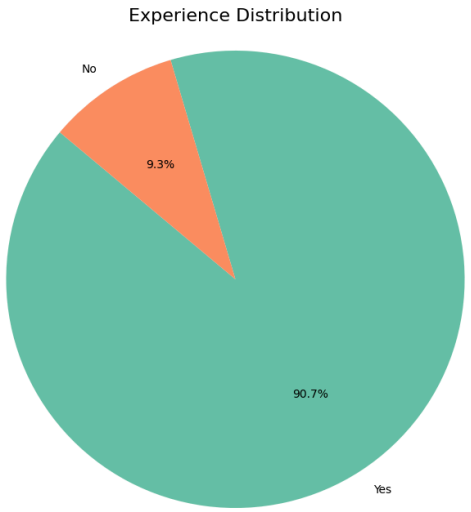

Fig. 5: Experience of purchasing due to social media influence

Fig 6, for the Attitude construct, the majority of responses for each statement generally fall within the "Agree" and "Neutral" categories, indicating a broadly positive or neutral attitude among respondents. Specifically, ATTD1 shows that 37.9% of respondents agree, while ATTD2, ATTD3, and ATTD4 have 30.2%, 34.1%, and 36.0% in the "Agree" category, respectively. The distribution for "Strongly Agree" also remains significant across all statements.

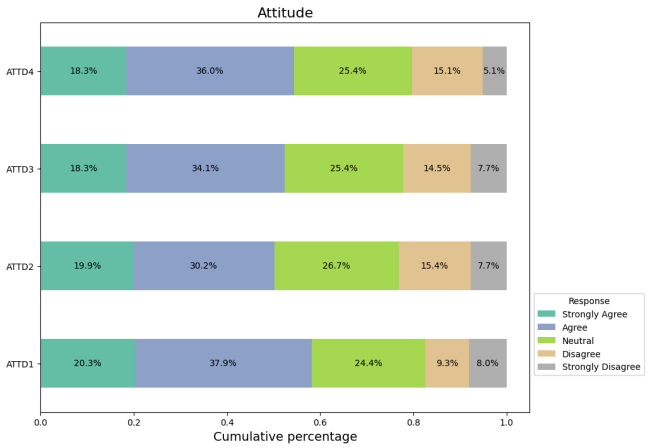

Fig. 6: Responses from attitude factors.

In fig 7, the Social Norm construct, responses are more varied and evenly spread among "Agree," "Neutral," and "Disagree" categories. For instance, SN1 and SN2 have substantial proportions in the "Neutral" category (28.0% and 28.9%, respectively), while SN3 and SN4 show notable agreement levels (25.1% and 28.0%, respectively). This suggests diverse perceptions of social norms among the respondents.

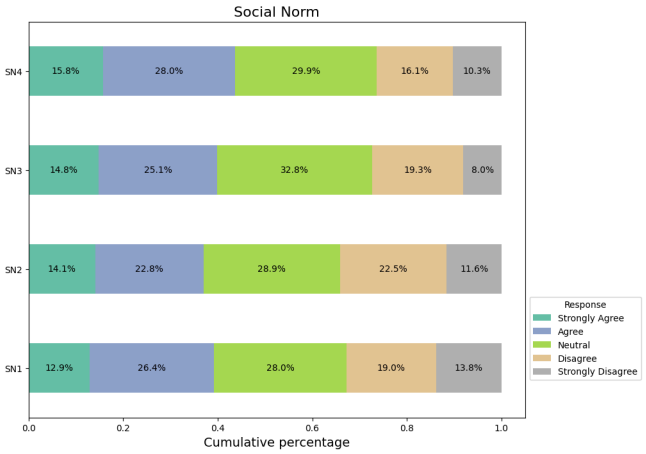

Fig. 7: Responses from social norm factors.

In fig 8, the Perceived Behavioural Control (PBC) construct follows a similar pattern to Attitude, with responses distributed across "Agree," "Neutral," and "Disagree" categories, indicating mixed perceptions of control over behaviors. Notably, PBC1 and PBC2 have significant "Agree" responses (26.4% and 32.2%, respectively), while PBC3 and PBC4 show high "Neutral" responses (29.3% and 30.5%, respectively).

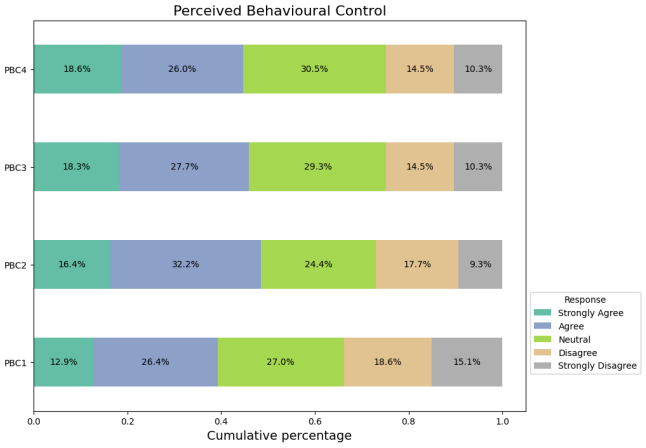

Fig. 8: Responses from PBC factors.

Fig 9, illustrates the distribution of responses across four different purchase behaviors (PB1, PB2, PB3, PB4). Each horizontal bar is segmented into five categories: Strongly Agree, Agree, Neutral, Disagree, and Strongly Disagree. The length of each segment represents the percentage of respondents falling into that category. For example, for PB1, the largest segments are Neutral (34.4%) and Agree (28.6%), indicating that most respondents either agreed or remained neutral about this purchase behavior. In contrast, PB4 has a higher proportion of Strongly Agree responses (22.2%) compared to the other behaviors. This chart helps to quickly compare how respondents feel about each purchase behavior, showing variations in agreement and neutrality as well as disagreement. The accompanying legend provides a color-coded key to interpret the different response categories.

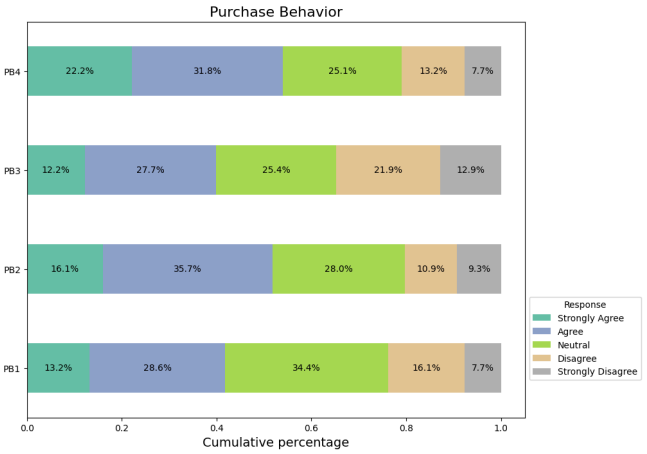

Fig. 9: Responses from PB factors.

Fig 10, illustrates the frequency of responses in five categories: Neutral, Disagree, Strongly Disagree, Agree, and Strongly Agree. The Neutral category has the highest frequency, with nearly 90 responses, indicating that many respondents neither strongly agree nor disagree with the statements. The Agree category follows with around 80 responses, and Strongly Agree has approximately 60 responses, suggesting a positive sentiment among many respondents. The Disagree category has about 50 responses, showing some level of disagreement, while Strongly Disagree is the least frequent category with fewer than 30 responses, indicating limited strong opposition. Overall, the histogram highlights a trend towards neutrality and agreement among the respondents.

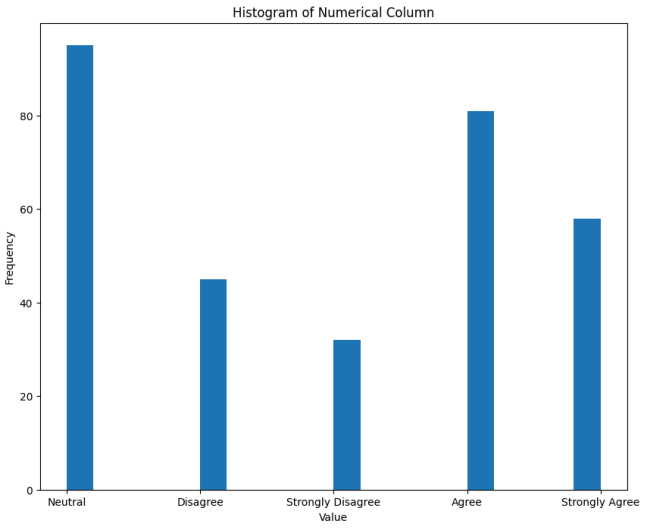

Fig. 10: Frequency of responses

E. Dataset Preprocessing

The dataset is balanced with no null or missing and duplicate values. In our dataset, we use eight different models with different models having their barriers. For example, the decision tree model cannot work on null values. That's why we need to pre-process our dataset.

Data preprocessing involved converting categorical values into numerical ones using label encoding (as detailed in Table II). Following the Theory of Planned Behavior (TPB), features related to attitude, social norms, and perceived behavioral control were retained.

TABLE II. CONVERTING CATEGORICAL VALUES INTO NUMERICAL ONES USING LABEL ENCODING

Categorical Value	Numerical Value
Strongly Disagree	1
Disagree	2
Neutral	3
Agree	4
Strongly Agree	5

Purchase behavior (PB) features were merged into a single column (PB-inf) and classified into LOW and HIGH categories using Fuzzy Association Rule mining, which identified nuanced patterns and enhanced the categorization within the data.

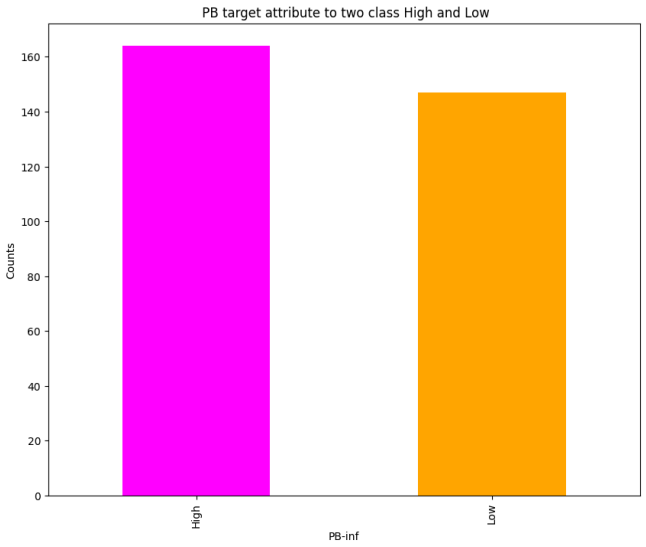

Fig. 11: PB target attribute to two class High and Low

We also used Minmaxscler as feature scaling. At first, we used feature scaling, and then we placed our data into the train and tested it with a ratio of 8:2 (The train set consisted of 80% of the data. The test set was the remaining 20% of the data.)

F. Preparing Machine Learning Algorithms

1. GridSearchCV: Grid search is a parameter tuning technique that systematically builds and examines a model for every set of algorithm parameters in a grid. It provides the optimal model parameters by fine-tuning the hyperparameter. It improves a model's performance but takes a long time because it looks through every possible combination of the specified grid components..

2. Explainable AI: Explainable artificial intelligence is a technique that is helpful for people who are not experts in machine learning or need help understanding the output of algorithms[9]. This process is normally called "Explainable AI."It is mainly used to explain any AI model. In our work, we applied LIME, which is very user-friendly.

3. LIME: "Local Interpretable Model-agnostic Explanations" is known as "LIME". The LIME approach uses a locally approximable interpretable model to faithfully explain the predictions of any classifier or regressor. In our work, we used it to understand our models.

4. Fuzzy Association Rule: (Fuzzy Association Rule), Similar to the methods for traditional association rule mining, it employs fuzzy logic to mine the rules from uncertain / vague data. Unlike traditional methods Fuzzy Association Rule will give way for discovery of rules more flexible to data that fails to have clear boundaries between the categories. It classifies elements on the basis of belonging to different sets and identifies the association for each pair of elements in a dataset by assigning a membership degree to elements. Frequent pattern mining algorithm (Fuzzy Association Rule) is a significant tool within the data mining toolbox for applications like retail market basket analysis, medical diagnosis, and decision-making systems; it allows the discovery of hidden patterns and relationships among uncertain data.

G. Machine Learning Models

1. Decision tree: A decision tree which is a widely used machine learning technique because it can classify dependent variables into different classes. It scans the data and identifies the most important independent factors, considering several possible variables to make predictions [7]. A decision tree is a supervised classification method. That is a simple structure where evaluation results are displayed in terminal nodes, and checks for one or more functions are displayed in non-terminal nodes. Previous research on human action detection using decision tree classifiers has produced encouraging results.

2. Random Forest: The Random Forest ensemble understanding technique creates many decision trees during training. A subset of the bagging technique called the RF performs better when noise and poor discriminating data are not affected by parameter initialization overfitting[8]. Breiman's study shows that the RF's generalization error converges as the number of trees rises, a feature lacking in most other classifiers[9]. With its excellent accuracy, ability to handle big datasets with plenty of features, and ability to provide feature-importance insights, Random Forest is very useful for classification and regression applications.

3. Logistic Regression: This fundamental machine-learning technique is used in binary classification [10]. Despite its name, its purpose is to calculate the probability that an instance will fall into a given class. The program simulates this chance using the logistic function, which adjusts the input data linearly. Logistic regression is a versatile tool that may be applied in many different settings because of its effectiveness, interpretability, and simplicity. Predictions are based on a threshold, and coefficients are optimized by maximum likelihood estimation. In many machine-learning contexts, the fundamental logistic regression method is utilized in finance and medicine to predict binary outcomes.

4. AdaBoost: AdaBoost, which stands for Adaptive Boosting, is an ensemble learning algorithm for classification and regression [11]. AdaBoost is a learning algorithm that combines weak learners (such as SVC) to form a strong classifier. It is especially effective with weak learners or models that are better than random chance. It constructs a series of weak learners to correct errors and iteratively focuses on misclassified instances, assigning higher weights to them. A weighted average of these learners is used to make the final prediction, resulting in a robust and accurate model.

5. KNN: K-Nearest Neighbors (KNN) is a simple machine-learning algorithm for classification and regression [12]. It can generate predictions by comparing newly collected data points to labeled data points in the training set. The algorithm determines the K nearest neighbors using weighted averaging or majority voting, computes distances, and forecasts the class label for classification or regression value. Depending on the parameter K selection, KNN is an instance-based, non-parametric model.

6. SVM (Support Vector Machine): SVM (Support Vector Machine) is a strong and robust supervised machine learning algorithm used for classification tasks, but can also be used for regression problems. SVM tries to find a hyperplane which is the optimal at separating different classes. SVM increases the accuracy and robustness of classification by increasing the margin between the data points of different classes. The method transforms the data using the kernel trick into a higher dimension where a linear separator is more effective. Useful in high-dimensional spaces and when the number of dimensions is greater than the number of samples. Although SVMs can be costly to computers in terms of both CPU cycles and memory (both during training and testing), their ability to reduce high-level decision-making to small and simple decision trees makes even the most computationally intensive problems solvable.

7. Naive Bayes: Naive Bayes is a machine learning algorithm that leverages Bayes' theorem to make probabilistic predictions, commonly utilized for classification purposes. It operates under the assumption that features are independent when given the class label. This simplification enables the development of quick and effective models, even when dealing with limited data. While the independence assumption is seldom accurate in practical situations, Naive Bayes delivers impressive performance, particularly in tasks like text classification (e.g., spam detection and sentiment analysis). The algorithm computes the likelihood of each class based on the feature values and assigns the class with the highest likelihood as the prediction. Naive Bayes models are straightforward, speedy, and demand only a small training dataset, making them well-suited for scenarios requiring real-time predictions and emphasis on interpretability.

8. Gradient Boosting: Gradient Boosting is an ensemble machine learning technique used for both classification and regression tasks. It constructs models in a step-by-step fashion by combining predictions from multiple weak learners, often decision trees, to develop a powerful predictive model. Each subsequent tree is trained to rectify the errors of its predecessors by minimizing a loss function, hence progressively enhancing the model's precision with each step. Gradient Boosting accommodates various loss functions, covering those relevant to classification (such as log-loss) and regression (like mean squared error). Renowned for its exceptional accuracy and capability to handle intricate data structures, Gradient Boosting finds widespread application in areas like fraud detection, ranking systems, and predictive analytics. The technique's adaptability and performance have established it

as a favored option in machine learning competitions and practical scenarios.

III. Results and Discussion

This section of the paper discusses the results of all different Machine Learning algorithms for Consumer Behavior and Shopping Habits. We have applied eight Machine Learning models in the dataset "Consumer Behavior and Shopping Habits". In our dataset, there are 27 columns and 311 rows of attributes. We split the data into train (80%) and test (20%) data. In 27 columns, we used one column as our target variable column, which has 11 classes and a multiclass problem, and the other 16 columns are features or predictors. We evaluate the performance metrics from the model, including accuracy, Recall, F1 score, and precision. These matrices can be calculated through the following equations:

$$Accuracy = \frac{correct\ predictions}{all\ predictions} = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

$$Precision\_c = \frac{TP\_c}{TP\_c + FP\_c} \quad (2)$$

$$Recall\_c = \frac{TP\_c}{TP\_c + FN\_c} \quad (3)$$

$$F1\ Score\_c = \frac{2\times(Precision\_c \times Recall\_c)}{Precision\_c + Recall\_c} \quad (4)$$

A. Table

TABLE III. HYPERPARAMETER VALUES' RANGES FOR VARIOUS ML MODELS

Model	**Hyperparameter Value Range**	**Optimized value**
KNN	{'n_neighbors': np.arange(1, 11)}	{'n_neighbors': 5}
Decision Tree	{'max_depth': np.arange(1, 11), 'criterion': ['gini', 'entropy']}	{'criterion': 'gini', 'max_depth': 3}
Random Forest	{'n_estimators': [10, 50, 100, 200], 'max_depth': np.arange(1, 11), 'criterion': ['gini', 'entropy']}	{'criterion': 'entropy', 'max_depth': 4, 'n_estimators': 200}
Logistic Regression	{'C': np.logspace(-4, 4, 9), 'solver': ['newton-cg', 'lbfgs', 'liblinear', 'sag', 'saga']}	{'C': 0.01, 'solver': 'sag'}
Naive Bayes	{}	{}
SVM	{'C': np.logspace(-3, 3, 7), 'kernel': ['linear', 'poly', 'rbf', 'sigmoid'], 'gamma': ['scale', 'auto']}	{'C': 0.01, 'gamma': 'scale', 'kernel': 'poly'}
AdaBoost	{'n_estimators': [10, 50, 100, 200], 'learning_rate': [0.001, 0.01, 0.1, 1]}	{'learning_rate': 0.1, 'n_estimators': 200}
Gradient Boosting	{'n_estimators': [10, 50, 100, 200], 'learning_rate': [0.001, 0.01, 0.1, 1], 'max_depth': np.arange(1, 11)}	{'learning_rate': 0.01, 'max_depth': 3, 'n_estimators': 200}

Table III illustrates the hyperparameter values' ranges for all the ML models.

TABLE IV. PRECISION METRICS OF VARIOUS CLASSIFIERS FOR OUTPUT WITH OPTIMIZED HYPERPARAMETERS

Model	ATTB, SN, PBC		ATTB		SN		PBC		PBC, ATTB		ATTB, SN		PBC, SN	
Classifier	HIGH	LOW	HIGH	LOW	HIGH	LOW	HIGH	LOW	HIGH	LOW	HIGH	LOW	HIGH	LOW
KNN	0.88	0.90	0.88	0.90	0.79	0.69	0.79	0.71	0.87	0.81	0.85	0.83	0.81	0.69
Decision Tree	0.78	0.85	0.84	0.92	0.95	0.66	0.78	0.81	0.84	0.92	0.84	0.81	0.74	0.72
Random Forest	0.88	0.87	0.87	0.81	0.76	0.63	0.78	0.74	0.84	0.81	0.88	0.84	0.75	0.66
Logistic Regression	0.96	0.83	0.96	0.83	0.77	0.72	0.87	0.79	0.90	0.84	0.90	0.84	0.84	0.81
Naive Bayes	0.93	0.85	0.93	0.85	0.77	0.70	0.90	0.84	0.90	0.84	0.87	0.81	0.87	0.79
SVM	0.93	0.82	0.93	0.82	0.84	0.68	0.83	0.74	0.93	0.85	0.88	0.87	0.93	0.78
AdaBoost	0.91	0.90	0.90	0.79	0.74	0.72	0.84	0.78	0.94	0.88	0.90	0.84	0.81	0.77

Gradient Boosting	0.92	0.76	0.94	0.88	0.71	0.69	0.84	0.78	0.91	0.87	0.88	0.84	0.81	0.75

Table IV. illustrates the performance matrices of various classifiers we used for parameters. As per this table, AdaBoost outperforms all the machine learning models with the highest accuracy of 0.90 and Decision Tree has the lowest accuracy of 0.81.

TABLE V. RECALL METRICS OF VARIOUS CLASSIFIERS FOR OUTPUT WITH OPTIMIZED HYPERPARAMETERS

Model	ATTB, SN, PBC		ATTB		SN		PBC		PBC, ATTB		ATTB, SN		PBC, SN	
Classifier	HIGH	LOW	HIGH	LOW	HIGH	LOW	HIGH	LOW	HIGH	LOW	HIGH	LOW	HIGH	LOW
KNN	0.91	0.87	0.91	0.87	0.67	0.80	0.70	0.80	0.82	0.87	0.85	0.83	0.67	0.83
Decision Tree	0.88	0.73	0.94	0.80	0.55	0.97	0.85	0.73	0.94	0.80	0.82	0.83	0.76	0.70
Random Forest	0.88	0.87	0.82	0.87	0.58	0.80	0.76	0.77	0.82	0.83	0.85	0.87	0.64	0.77
Logistic Regression	0.82	0.97	0.82	0.97	0.73	0.77	0.79	0.87	0.85	0.90	0.85	0.90	0.82	0.83
Naive Bayes	0.85	0.93	0.85	0.93	0.70	0.77	0.85	0.90	0.85	0.90	0.82	0.87	0.79	0.87
SVM	0.82	0.93	0.82	0.93	0.64	0.87	0.73	0.83	0.85	0.93	0.88	0.87	0.76	0.93
AdaBoost	0.91	0.90	0.79	0.90	0.76	0.70	0.79	0.83	0.88	0.93	0.85	0.90	0.79	0.80
Gradient Boosting	0.73	0.93	0.88	0.93	0.73	0.67	0.79	0.83	0.88	0.90	0.85	0.87	0.76	0.80

Table V, we see performance metrics of various classifiers with optimized hyperparameters. Here eight models are used. As per this table, AdaBoost outperforms all the machine learning models with the highest accuracy of 0.90 and Decision Tree has the lowest accuracy of 0.81.

TABLE VI. F1-SCORE OF VARIOUS CLASSIFIERS FOR OUTPUT WITH OPTIMIZED HYPERPARAMETERS

Model	ATTB, SN, PBC		ATTB		SN		PBC		PBC, ATTB		ATTB, SN		PBC, SN	
Classifier	HIGH	LOW	HIGH	LOW	HIGH	LOW	HIGH	LOW	HIGH	LOW	HIGH	LOW	HIGH	LOW
KNN	0.90	0.88	0.90	0.88	0.72	0.74	0.74	0.75	0.84	0.84	0.85	0.83	0.73	0.76
Decision Tree	0.83	0.79	0.89	0.86	0.69	0.78	0.81	0.77	0.89	0.86	0.83	0.82	0.75	0.71
Random Forest	0.88	0.87	0.84	0.84	0.66	0.71	0.77	0.75	0.83	0.83	0.86	0.85	0.69	0.71
Logistic Regression	0.89	0.89	0.89	0.89	0.75	0.74	0.83	0.83	0.88	0.87	0.88	0.87	0.83	0.82
Naive Bayes	0.89	0.89	0.89	0.89	0.73	0.73	0.88	0.87	0.88	0.87	0.84	0.84	0.83	0.83
SVM	0.87	0.87	0.87	0.87	0.72	0.76	0.77	0.78	0.89	0.89	0.88	0.87	0.83	0.85
AdaBoost	0.91	0.90	0.84	0.84	0.75	0.71	0.81	0.81	0.91	0.90	0.88	0.87	0.80	0.79
Gradient Boosting	0.81	0.84	0.91	0.90	0.72	0.68	0.81	0.81	0.89	0.89	0.86	0.85	0.78	0.77

Table VI, we see the F1-score of various classifiers with optimized hyperparameters. Here eight models are used. As per this table, AdaBoost outperforms all the machine learning models with the highest accuracy of 0.90 and Gradient Boost has the lowest accuracy of 0.81.

TABLE VII. ACCURACY OF VARIOUS CLASSIFIERS FOR OUTPUT WITH OPTIMIZED HYPERPARAMETERS

Model	ATTB, SN, PBC	ATTB	SN	PBC	PBC, ATTB	ATTB, SN	PBC, SN
KNN	0.89	0.89	0.73	0.75	0.84	0.84	0.75
Decision Tree	0.81	0.87	0.75	0.79	0.87	0.83	0.73
Random Forest	0.87	0.84	0.68	0.76	0.83	0.86	0.70
Logistic Regression	0.89	0.89	0.75	0.83	0.87	0.87	0.83
Naive Bayes	0.89	0.89	0.73	0.87	0.87	0.84	0.83
SVM	0.87	0.87	0.75	0.78	0.89	0.87	0.84
AdaBoost	0.90	0.84	0.73	0.81	0.90	0.87	0.79
Gradient Boosting	0.83	0.90	0.70	0.81	0.89	0.86	0.78

Table VII, we see performance metrics of various classifiers with optimized hyperparameters. Here eight models are used. As per this table, AdaBoost outperforms all the machine learning models with the highest accuracy of 0.90 and Decision Tree has the lowest accuracy of 0.81.

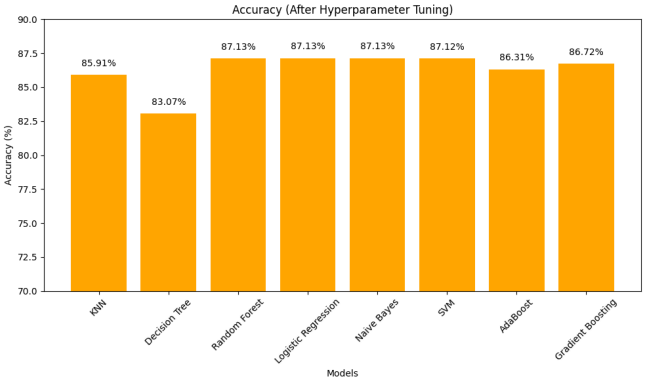

Fig. 12: Accuracy of all ML models with optimized hyperparameters using GridSearchCV or RandomSearchCV.

Fig. 12, shows the accuracy of the ML models in the form of a bar graph with optimized hyperparameters using GridSearchCV or RandomSearchCV. It shows that Random Forest, Logistic regression and Naive Bayes has the best accuracy of 87.13%% for the dataset and Decision Tree has the lowest accuracy of 83.07%.

B. Confusion Matrix

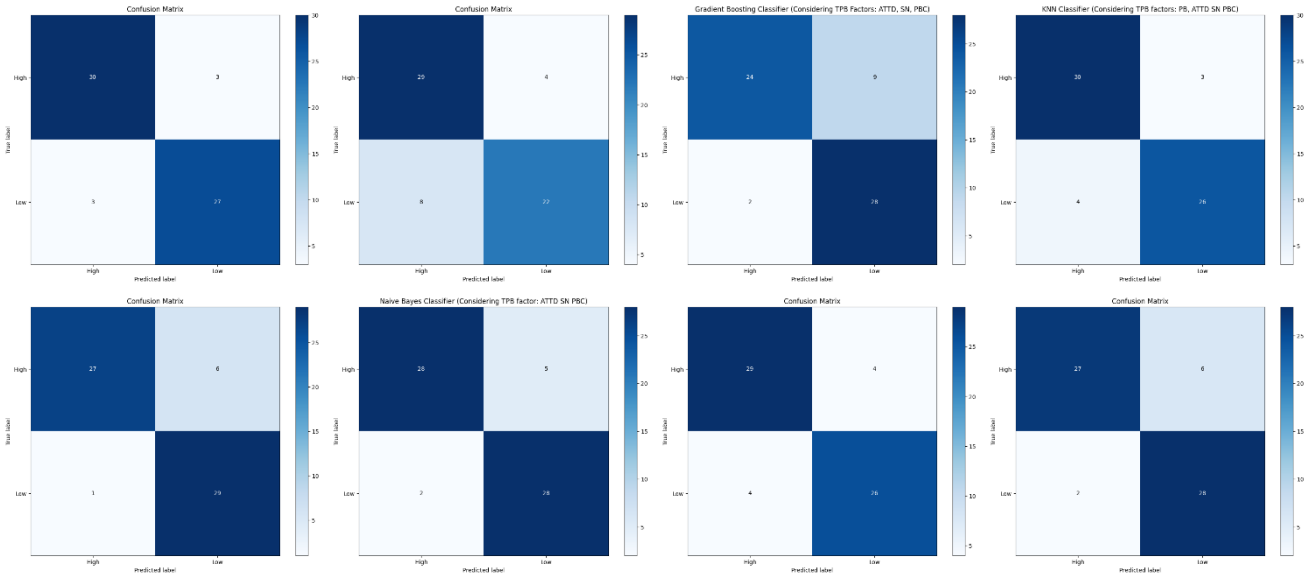

Fig. 13: Confusion Matrices of all ML models with optimized hyperparameters using GridSearchCV or RandomSearchCV.

Fig 13, shows the confusion matrices for best-performing models in terms of the theory of planned behavior factors' possible combinations. Fig 8(a) from the confusion matrices performed better than others. Therefore, AdaBoosting with combined Attitude, social Norm and Perceived behavioral Control factors performed the most promising. It was able to classify 33 HIGH classes and 27 LOW classes correctly. However, 3 HIGH classes were classified as Low and 3 LOW classes as HIGH. Hence, 6 classes were miss classified, still the lowest among other models.

C. Receiver operating characteristic and Precision recall curves

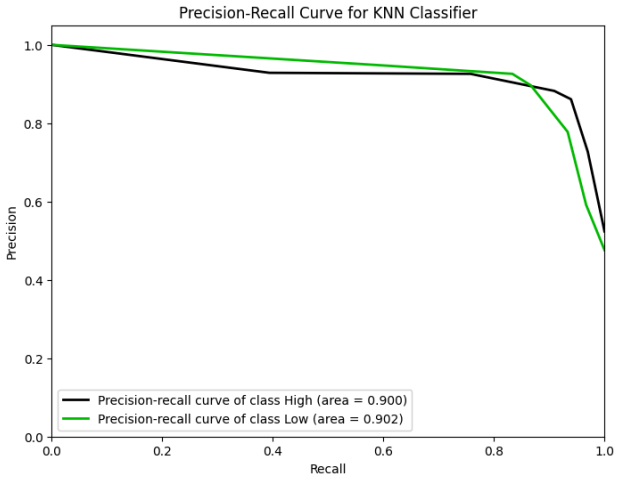

Fig. 14: Precision-Recall Curve of KNN model with optimized hyperparameters using GridSearchCV or RandomSearchCV.

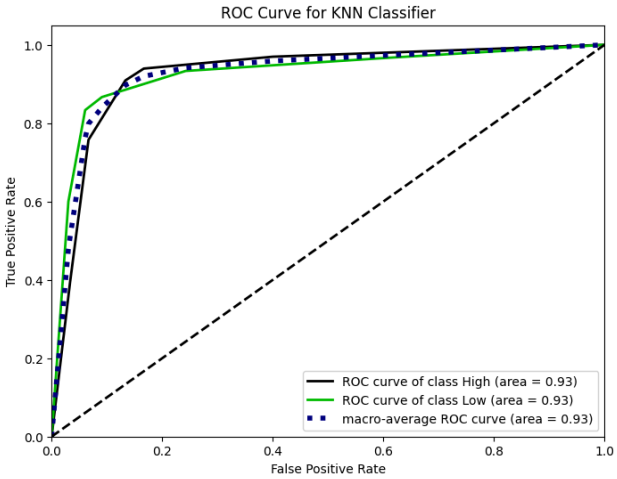

Fig. 15: ROC Curve of KNN model with optimized hyperparameters using GridSearchCV or RandomSearchCV.

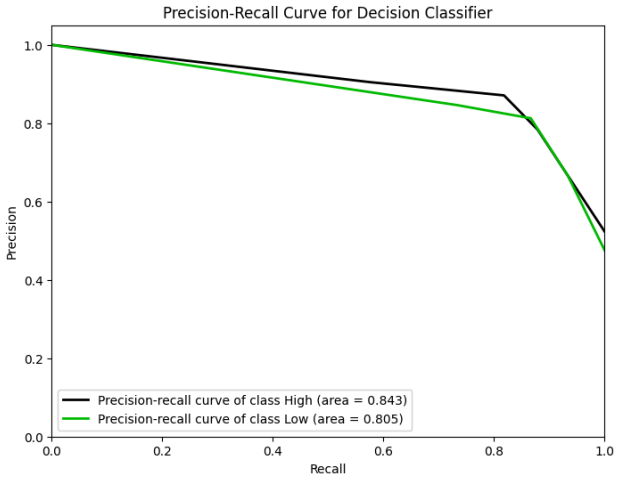

Fig. 16: Precision-Recall Curve of Decision Tree model with optimized hyperparameters using GridSearchCV or RandomSearchCV.

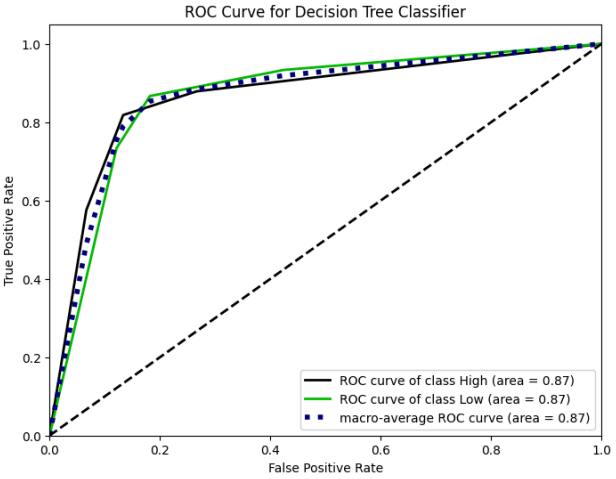

Fig. 17: ROC Curve of Decision Tree model with optimized hyperparameters using GridSearchCV or RandomSearchCV.

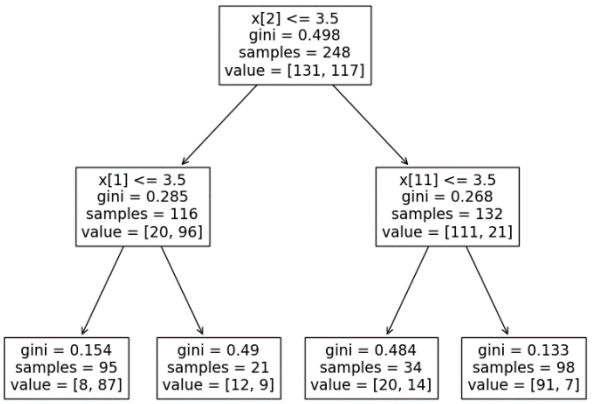

Fig. 18: Tree Diagram of Decision Tree model with optimized hyperparameters using GridSearchCV or RandomSearchCV.

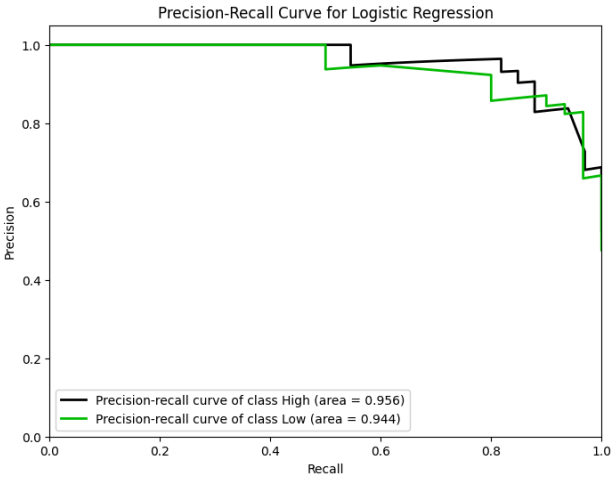

Fig. 19: Precision-Recall Curve of Logistic Regression model with optimized hyperparameters using GridSearchCV or RandomSearchCV.

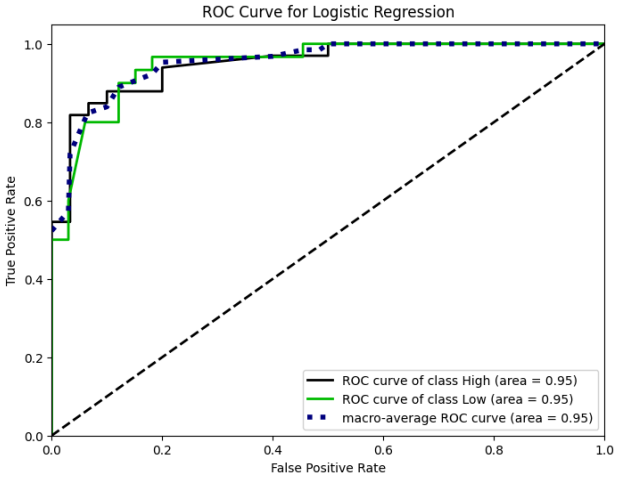

Fig. 20: ROC Curve of Logistic Regression model with optimized hyperparameters using GridSearchCV or RandomSearchCV.

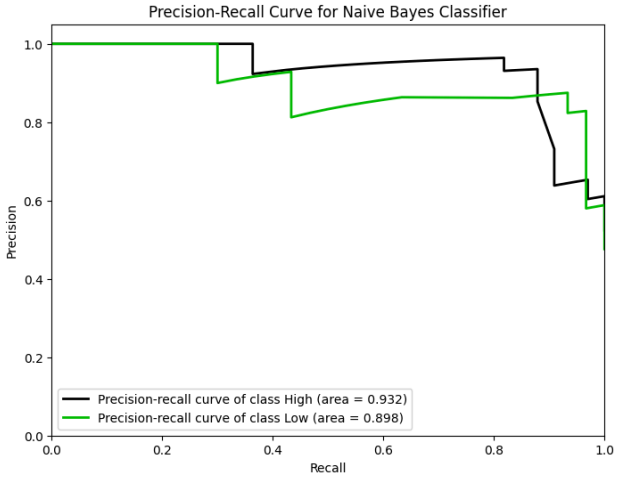

Fig. 21: Precision-Recall Curve of Naïve Bayes model with optimized hyperparameters using GridSearchCV or RandomSearchCV.

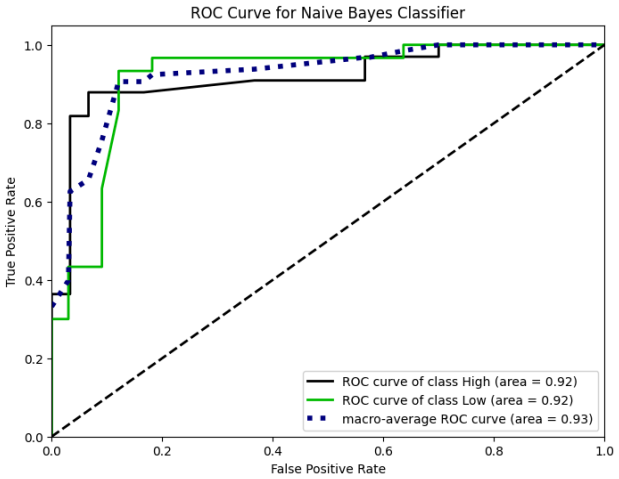

Fig. 22: ROC Curve of Naïve Bayes model with optimized hyperparameters using GridSearchCV or RandomSearchCV.

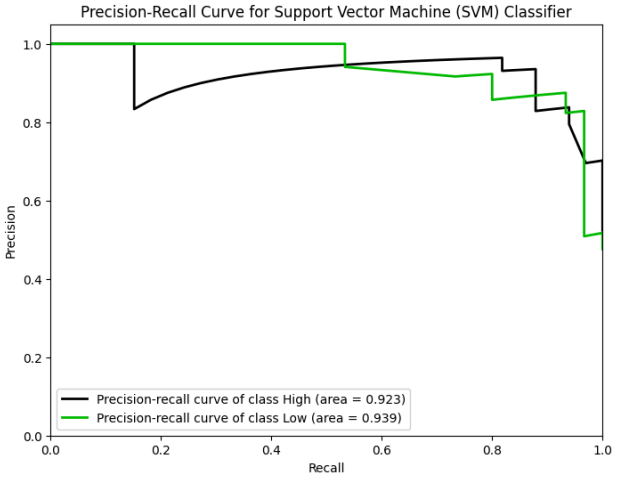

Fig. 23: Precision-Recall Curve of Support Vector Machine model with optimized hyperparameters using GridSearchCV or RandomSearchCV.

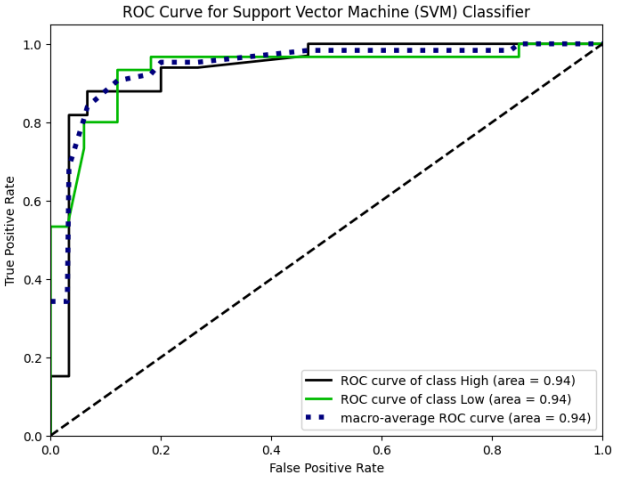

Fig. 24: ROC Curve of Support Vector Machine model with optimized hyperparameters using GridSearchCV or RandomSearchCV.

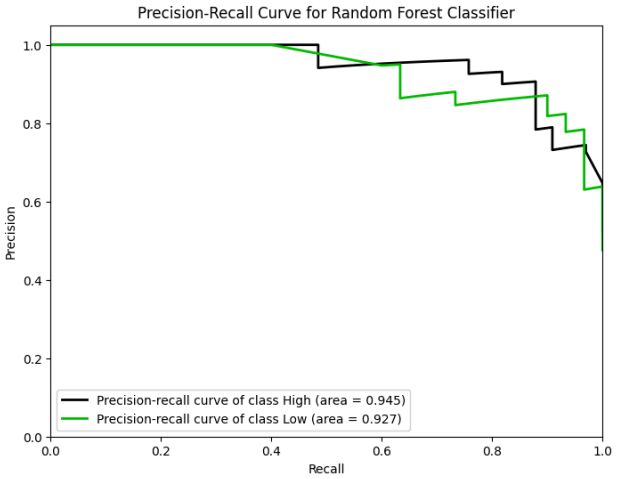

Fig. 25: Precision-Recall Curve of Random Forest model with optimized hyperparameters using GridSearchCV or RandomSearchCV.

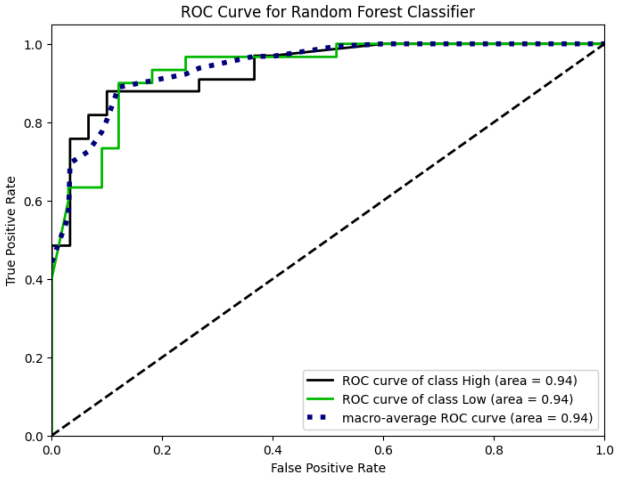

Fig. 26: ROC Curve of Random Forest model with optimized hyperparameters using GridSearchCV or RandomSearchCV.

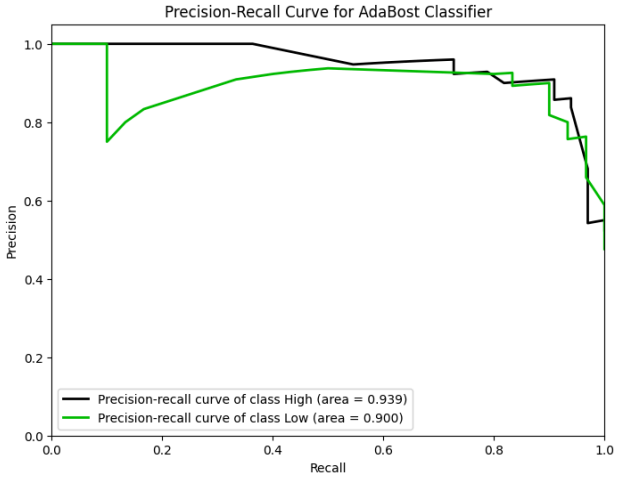

Fig. 27: Precision-Recall Curve of AdaBoost model with optimized hyperparameters using GridSearchCV or RandomSearchCV.

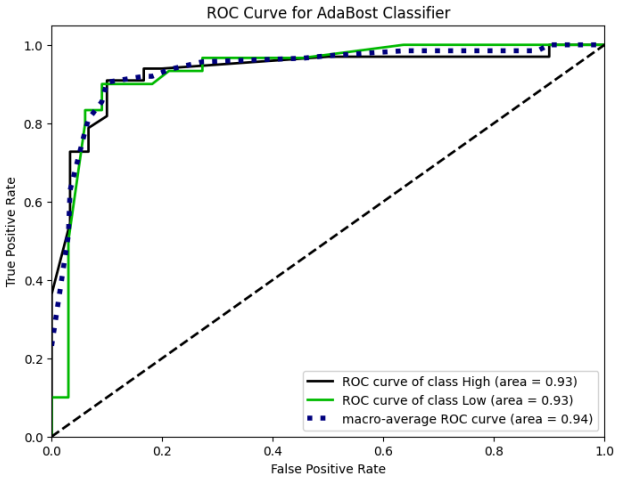

Fig. 28: ROC Curve of AdaBoost model with optimized hyperparameters using GridSearchCV or RandomSearchCV.

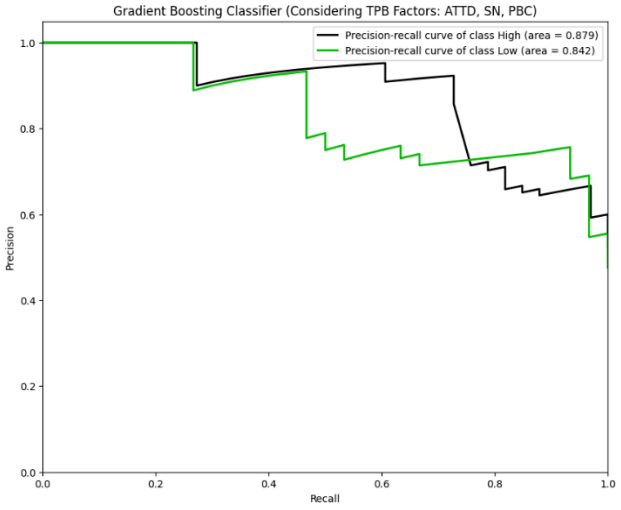

Fig. 29: Precision-Recall Curve of Gradient Boosting model with optimized hyperparameters using GridSearchCV or RandomSearchCV

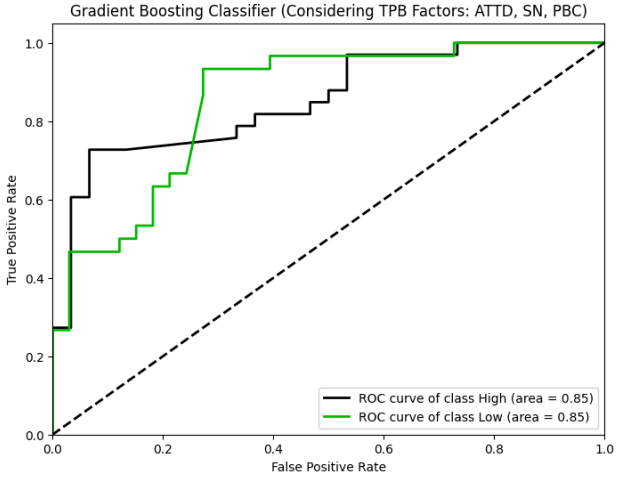

Fig. 30: ROC Curve of Gradient Boosting model with optimized hyperparameters using GridSearchCV or RandomSearchCV.

D. Explainable AI (LIME):

Fig. 31: Explaterpretation of purchase behavior prediction using explainable AI (LIME) for Adaboost Model

The provided visualization of LIME explainable AI explains the model's prediction for a specific instance, showing a 67% probability for the "High" class and a 33% probability for the "Low" class. The middle section lists features and their contributions to this prediction. Features such as ATTD3 <= 3.00, PBC1 <= 2.00, ATTD4 <= 3.00, PBC4 <= 2.00, and SN3 <= 2.00 have the highest positive contributions, significantly increasing the likelihood of the "High" prediction. Conversely, features like ATTD2 <= 3.00

and SN4 <= 2.00 have slight negative contributions, but their impact is minimal. The right section lists the actual values for these features, confirming they meet the conditions favoring the "High" class. Overall, the most influential features, particularly ATTD3 and PBC1, drive the model towards predicting the "High" class.

By analyzing the contributions, we can understand which features and their specific conditions most influence the model's decision for this particular instance. This helps in interpreting the model's behavior and ensuring its predictions align with domain knowledge and expectations.

IV. Comparative Analysis

A. *Accuracy Comparison*

We evaluated the accuracy of various classifiers with optimized hyperparameters, revealing significant differences in performance. AdaBoost emerged as the most accurate model, achieving an accuracy of 90%, while the Decision Tree had the lowest accuracy at 81%. This demonstrates the effectiveness of boosting techniques in handling complex classification tasks by combining weak learners to form a strong classifier.

B. *Precision, Recall, and F1 Score Analysis*

Precision, recall, and F1 scores were calculated to provide a more comprehensive evaluation of model performance. Logistic Regression, Naive Bayes, and SVM displayed high precision and recall values, indicating their robustness in correctly identifying both high and low consumer purchase behaviors. Specifically, the Logistic Regression model achieved a precision and recall of 97% for the 'low' class, while SVM reached a precision and recall of 93% for the 'high' class. The F1 scores across these models were consistently high, with AdaBoost also performing well, particularly with an F1 score of 90% for the 'high' class.

C. *Runtime Comparison*

The runtime performance of each classifier was also compared, highlighting the computational efficiency of the models. Simple models like KNN and Logistic Regression had shorter training times compared to more complex models like Gradient Boosting and AdaBoost. Despite its longer runtime, AdaBoost's superior accuracy and F1 score justify its computational cost, making it a viable choice for practical applications where prediction accuracy is critic

V. Integration of the Model & Create a Web-Application

A Django-based web application was developed to provide an accessible platform for predicting consumer purchase behavior. The application allows users to input relevant features (ATTB, SN, PBC, PB) and receive predictions on whether the consumer behavior is classified as High or Low. The backend leverages the trained AdaBoost model, given its highest accuracy and balanced performance metrics, to deliver reliable predictions.

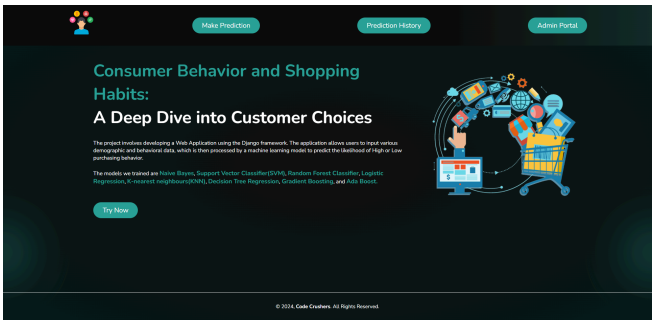

Fig. 32: Home page

The user-friendly interface allows for seamless navigation and interaction, with a main menu that includes options for "Make Prediction," "History," and "Admin Portal." In the "Make Prediction" section, users can input data and get immediate predictions. The "History" section stores previous predictions, which can be used as new data to update and enhance the dataset, while the "Admin Portal" provides administrative access to manage and maintain the application.

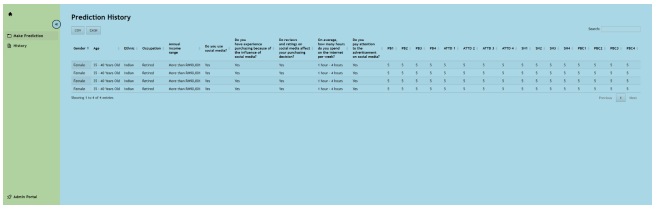

Fig. 35: History

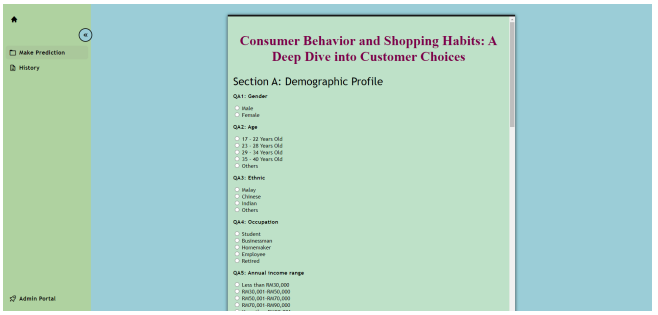

Fig. 33:Get user input data and get immediate predictions

Fig. 34: Prediction

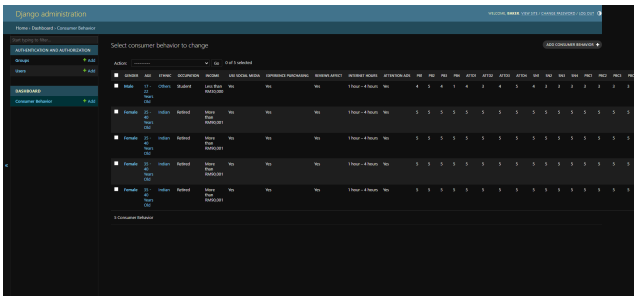

Fig. 36: Admin Portal

The backend leverages a trained AdaBoost model due to its highest accuracy and balanced performance metrics, ensuring reliable predictions that aid businesses in making data-driven decisions by understanding and anticipating consumer behavior trends effectively. Models such as Naive Bayes, Support Vector Classifier (SVM), Random Forest Classifier, Logistic Regression, K-Nearest Neighbours (KNN), Decision Tree Regression, and Gradient Boosting were also trained for comparison.

The integration of Local Interpretable Model-agnostic Explanations (LIME) into the application ensures that users can obtain explanations for each prediction, enhancing the transparency and usability of the tool[13]. The application includes a "History" feature that stores past predictions, allowing this historical data to be leveraged to continuously update and improve the dataset, ensuring the model remains accurate and relevant over time.

Designed to be accessible, the web application provides a reliable platform for businesses of all sizes to leverage consumer behavior insights[13]. Users can easily try the application by clicking on the "Try Now" button available on the main interface. This application aims to assist businesses in making data-driven decisions by understanding consumer behavior trends, anticipating future purchase behaviors, and enhancing strategic planning and marketing efforts. By providing accurate predictions and clear explanations, the tool empowers businesses to harness the power of machine learning for consumer behavior analysis, ultimately driving better business outcomes.

VI. Conclusions

This study successfully demonstrates the applicability of various machine learning models in predicting consumer purchase behavior. AdaBoost was identified as the best-performing model, achieving the highest accuracy of 90%. The comparative analysis highlights the strengths and weaknesses of each model in terms of accuracy, precision, recall, F1 score, and runtime. The developed Django web application showcases the practical implementation of these models, offering a valuable tool for businesses to predict and analyze consumer behavior. The integration of Explainable AI techniques, specifically LIME, adds a layer of transparency, making the predictions more interpretable and actionable.

Future work can focus on several avenues to enhance the findings of this study. Firstly, incorporating a broader range of features, including real-time data from social media and transaction history, could improve the accuracy of predictions. Additionally, exploring other advanced machine learning techniques, such as deep learning and reinforcement learning, may yield better performance. Another promising direction is the integration of sentiment analysis to gauge consumer emotions and their impact on purchase behavior. Furthermore, conducting similar studies across different demographic groups and regions could provide more generalized insights. Finally, continuous improvement of the web application with more interactive features and real-time data processing capabilities will make it an even more powerful tool for businesses.

References

[1] Prasath, P., & Yoganathen, A. (2018). Influence of social media marketing on consumer buying decision making process. *SLIS Student research journal*, *1*(1), 1-12.

[2] Ioanăs, E., & Stoica, I. (2014). Social media and its impact on consumers behavior. *International Journal of Economic Practices and Theories*, *4*(2), 295-303.

[3] Md. Shawmoon Azad, Shadman Sakib Khan, R. Hossain, R. Rahman, and S. Momen, "Predictive modeling of consumer purchase behavior on social media: Integrating theory of planned behavior and machine learning for actionable insights," *PloS one*, vol. 18, no. 12, pp. e0296336–e0296336, Dec. 2023, doi: https://doi.org/10.1371/journal.pone.0296336.

[4] Ebrahimi, P.; Basirat, M.; Yousefi, A.; Nekmahmud, M.; Gholampour, A.; Fekete-Farkas, M. Social Networks Marketing and Consumer Purchase Behavior: The Combination of SEM and Unsupervised Machine Learning Approaches. Big Data Cogn. Comput. 2022, 6, 35. https://doi.org/10.3390/bdcc6020035

[5] Rumen Ketipov, Angelova, V., Lyubka Doukovska, & Schnalle, R. (2023). Predicting User Behavior in e-Commerce Using Machine Learning. Cybernetics and Information Technologies, 23(3), 89–101. https://doi.org/10.2478/cait-2023-0026

[6] Zhou Y, Loi AMW, Tan GWH, Lo PS, Lim W. The survey dataset of The Influence of theory of planned behaviour on purchase behaviour on social media. Data in Brief. 2022;42:108239. pmid:35592771

[7] M. T. Huyut and Z. Huyut, "Effect of ferritin, INR, and D-dimer immunological parameters levels as predictors of COVID-19 mortality: A strong prediction with the decision trees," Heliyon, vol. 9, 2023.

[8] N. E. J. Asha, E. U. Islam and R. Khan, "Low-Cost Heart Rate Sensor and Mental Stress Detection Using Machine Learning," International Conference on Trends in Electronics and Informatics, pp. 1369-1374, 2021.

[9] I. Tasin, T. U. Nabil, S. Islam andR Khan, "Diabetes prediction using machine learning and explainable AI techniques", Healthcare Technology Letters, vol. 10, 2023.

[10] F. Hidayat and T. M. S. Astsauri, "Applied random forest for parameter sensitivity of low salinity water Injection (LSWI) implementation on carbonate reservoir," Alexandria Engineering Journal, vol. 61, pp. 2408-2417, 2022.

[11] Kalagotla, S. K., Gangashetty, S. V., & Giridhar, K. (2021). A novel stacking technique for prediction of diabetes. Computers in Biology and Medicine, 135, 104554. https://doi.org/10.1016/j.compbiomed.2021.104554

[12] Zhang Z. Introduction to machine learning: k-nearest neighbors. Annals of translational medicine. 2016;4(11). pmid:27386492

[13] Automated stroke Prediction using Machine Learning: An explainable and exploratory study with a web application for early intervention. (2023). IEEE Journals & Magazine | IEEE Xplore. https://ieeexplore.ieee.org/document/10130159