Analyzing Customer Spending Based on Transactional Data Using the Random Forest Algorithm

Quba Siddique[1,*], Arif Muamar Wahid[2]

[1]*Institute of Banking and Finance, Bahauddin Zakariya University Multan, Pakistan*

[2]*Magister of Computer Science, Amikom Purwokerto University, Indonesia*

(Received: February 1, 2025; Revised: March 5, 2025; Accepted: June 5, 2025; Available online: July 23, 2025)

Abstract

This study explores customer spending behavior using transactional data from a retail dataset, employing a Random Forest Regressor to predict the total amount spent by customers. The dataset includes various customer attributes such as age, gender, and product category, alongside transactional details including quantity purchased and price per unit. Through Exploratory Data Analysis (EDA), it was found that Price and Quantity were the most significant factors influencing total spending, with other features like Age, Gender, and Product Category playing a minimal role in predicting spending behavior. The model achieved perfect accuracy, with an R-squared value of 1.000, indicating that it explained all the variance in customer spending. The findings suggest that transactional features, particularly Price and Quantity, are the key drivers of customer spending, and retailers can optimize their marketing and sales strategies by focusing on these factors. This study also highlights the importance of data preprocessing and feature engineering in enhancing model performance, though the results were limited by the lack of external and behavioral features. Future research could further explore the impact of customer loyalty, external factors, and more complex algorithms to improve predictive accuracy.

Keywords: Age, Customer Spending, Price, Quantity, Random Forest

1. Introduction

Understanding customer spending patterns is essential for retailers, as these patterns significantly impact various dimensions of retail strategy, including inventory management, marketing effectiveness, and financial forecasting. Identifying spending behaviors allows retailers to tailor their operations and enhance customer experiences, leading to increased loyalty and profitability. Customer spending is influenced by multiple factors, including individual purchasing behavior and broader economic events. McCarthy and Fader discuss how Customer Lifetime Value (CLV) serves as a forward-looking projection metric, allowing firms to understand variances in spending propensities and contributing to decreased investor uncertainty [1]. Taylor and Hollenbeck observe that customer segmentation based on spending patterns can be effective, especially when implementing discount strategies that recognize customer heterogeneity [2]. This highlights the importance of analyzing customer data not just for current insights but also for predicting future purchasing behaviors across different consumer segments.

The dynamic nature of customer spending can lead to varying impacts based on loyalty programs. Research by Wu et al indicates that the effects of Item-Based Loyalty Programs (IBLP) on spending differ among customer types, with heavier spenders showing a higher propensity to purchase under specific incentives [3]. This underscores the necessity for retailers to tailor loyalty programs to different customer segments. External influences, such as economic crises, can radically alter customer spending habits. Hall et al discuss how panic buying during the COVID-19 pandemic disrupted regular consumption patterns, necessitating a reevaluation of consumer behavior in crisis contexts [4]. Retailers must monitor not only cost-driven changes but also new purchasing trends that arise from global events.

***Corresponding author: Quba Siddique (qubassindhu@gmail.com)**
DOI: https://doi.org/10.47738/ijaim.v5i2.103

Additionally, the interplay between spending patterns and recreational retail experiences has gained attention, with studies indicating that modern consumers increasingly view shopping as a leisure activity. Baghaee et al find that leisure motivations drive consumer behavior in retail centers, emphasizing the need for retailers to align their strategies with these experiential elements [5]. The incorporation of technology in understanding and enhancing customer spending is demonstrated in Zhou et al's research, which utilizes mobile app GPS data to analyze consumer behavior in shopping malls [6]. By assessing foot traffic and spending patterns, retailers can create tailored shopping experiences based on real-time insights, thereby optimizing service delivery.

Identifying key factors influencing customer spending behavior is critical for businesses looking to optimize their strategies in a highly competitive retail environment [7]. A multitude of research has examined various dimensions that drive consumer spending, taking into account demographics, psychological traits, situational factors, and emotional influences. Each of these elements, whether independently or in concert, can significantly inform a retailer's approach to enhancing customer engagement, satisfaction, and ultimately, financial outcomes. Demographics play a foundational role in understanding customer spending behavior. Specific studies, such as those by Sears et al, highlight how factors like age, gender, and social status can influence consumer decisions, especially in niche markets such as vape shops [8]. Here, the motivations behind spending are multifaceted, guided not only by the demographics of the customers but also by the environment and offerings presented in the retail context. With the continual rise of e-cigarette use among youths, understanding these demographic influences is paramount for developing targeted marketing strategies.

In a broader context, Teo et al emphasize the impact of hedonic and utilitarian values on customer satisfaction, demonstrating that these psychological factors significantly shape spending behavior [9]. When customers derive pleasure (hedonic value) from their shopping experiences or perceive functional benefits (utilitarian value), their likelihood of spending increases. This suggests that retailers seeking to influence customer spending patterns should thoughtfully design their offerings and promotional strategies to align with these values. Furthermore, psychological characteristics, such as personality traits, substantially affect consumer behavior. Matz et al provide compelling evidence that aligning products with customer personalities can lead to increased expenditure [10]. The concept of a "product-participant match" indicates that when consumers feel emotionally connected or satisfied with their purchases, they tend to spend more. This notion reinforces the importance of market segmentation and personalized marketing initiatives aimed at enhancing customer experiences that resonate with characteristics common within their target demographics.

Reciprocity and social exchange theory, as discussed by Teichmann, provide insight into how social dynamics can affect spending behaviors. Their study on loyal customers and the nonlinear relationship between loyalty and spending underscores that social rewards can influence customer expenditure in significant ways [11]. Retailers can leverage this by fostering a community around their brands that encourages loyalty and spending through reciprocal benefits. By creating situations where customers feel valued and recognized, businesses can see enhanced spending outcomes. Economic factors inevitably influence consumer behavior, particularly in times of rising costs and economic instability. The analysis by the authors investigating the rising cost of living for Malaysian consumers demonstrates how macroeconomic factors, such as household income and urban locality, correlate with changes in spending habits [12]. This suggests that businesses that remain aware of the economic context and adjust their strategies accordingly—such as offering discounts or emphasizing affordability—can better align with consumer capabilities, thereby sustaining or even boosting their spending.

The analysis of customer spending using transactional data, particularly through machine learning algorithms such as Random Forest, presents a significant opportunity for retailers to enhance their understanding of consumer behavior. Random Forest, an ensemble learning method, excels at handling complex and non-linear relationships in data, making it particularly well-suited for this kind of analysis. Its ability to manage various predictors while minimizing the risk of overfitting is essential in retail contexts, where the influence of numerous variables can be pronounced. The foundational characteristics of the Random Forest algorithm allow it to address classification and prediction tasks effectively. Zhao et al highlight the usability of Random Forest in predicting customer churn, demonstrating its strengths in overcoming the complexities of linear models by accommodating intricate interdependencies among variables [13]. This foundational capability makes it an ideal candidate for analyzing customer spending, given the

multifactorial influences that determine consumer behavior. Moreover, the interpretability of the outputs from Random Forest models enables retailers to identify critical spending determinants, informing targeted marketing strategies.

This research contributes by offering valuable insights into customer purchase behavior using data mining techniques, specifically applying the Random Forest algorithm to analyze transactional data. By identifying the key factors that influence customer spending, the study enhances understanding of consumer behavior, which can aid retailers in optimizing marketing strategies, inventory management, and personalized promotions. The application of data mining to real-world retail data allows for more accurate and data-driven decision-making, improving the overall efficiency of business operations.

2. Literature Review

2.1. Customer Spending Behavior Models

The field of consumer purchasing behavior has garnered significant attention in recent literature, particularly with the integration of advanced analytical techniques and a deeper understanding of the multifaceted influences on spending patterns. Various studies have explored diverse aspects of purchasing behavior, including the impact of external factors such as pandemics, marketing strategies, and innovative technologies on consumer choices. One pivotal area of research is the effect of situational factors on online purchasing behavior, especially highlighted during the COVID-19 pandemic. Gu et al investigated the pandemic's impact on consumer behavior, revealing shifts in online purchasing patterns with a substantial number of consumers adapting their buying habits to accommodate new restrictions and preferences in virtual shopping environments [14]. Their findings indicate that consumers engaged more with online shopping as traditional shopping methods faced limitations during the pandemic.

In exploring technological interventions that predict purchasing behavior, Kao et al discussed sophisticated deep learning methodologies, specifically Recurrent Neural Networks (RNNs), which outperformed traditional models in analyzing shopping behavior [15]. This advancement in predictive analytics underscores the potential of machine learning in refining models of purchasing behavior, allowing for a nuanced understanding of temporal dependencies in transactional data. Such methods enhance the accuracy of predictions related to inter-purchase times, enabling retailers to anticipate customer needs more effectively. Moreover, behavioral insights derived from consumer engagement in non-transactional contexts have gained traction. Tena et al highlighted the importance of incorporating non-purchase interactions, such as social media engagement and customer feedback, into purchasing behavior analysis [16]. Understanding these interactions provides a broader perspective on customer engagement, suggesting that businesses should prioritize experiential marketing strategies that foster positive consumer emotions and enhance brand loyalty.

Additionally, the role of sustainability in consumer purchasing behavior has emerged as a vital theme within the literature. Research by Sarfraz et al emphasizes how entrepreneurial innovation in the healthcare sector significantly influences consumer purchasing decisions towards environmentally friendly products [17]. This suggests a growing consumer awareness towards sustainability, encouraging retailers to integrate sustainable marketing strategies into their operations. The marketing mix's influence on purchasing behavior was examined by Dewi, who articulated how the green marketing mix affects green buying intentions [18]. This research underlines the need for aligning marketing strategies with consumer preferences concerning sustainability, reinforcing the necessity for businesses to adapt their product offerings and promotional activities to meet evolving consumer expectations.

2.2. Data Mining in Retail

The exploration of data mining methodologies in retail, particularly regression analysis and machine learning techniques, reveals a landscape of approaches aimed at understanding consumer purchasing patterns and enhancing business decision-making. Various algorithms and statistical frameworks serve as critical tools in analyzing customer data to derive actionable insights that can drive sales and optimize operational efficiencies. One of the foundational methods used in data mining is regression analysis, which allows researchers to understand relationships between variables related to customer behavior. Taylor and Hollenbeck discussed how leveraging loyalty programs alongside competitor-based targeting strategies can optimize pricing strategies and model customer preferences based on spending patterns [2].

Regression analysis has limitations, particularly concerning overfitting when the dataset contains numerous variables. Fox et al noted that when the number of explanatory variables is large, traditional regression models can overfit data and that more flexible non-parametric approaches like Random Forest are beneficial alternatives; they argued that Random Forest can manage high-dimensional spaces effectively, thus offering a robust framework for predictive analytics [19]. The Random Forest algorithm is particularly well-regarded in retail data mining due to its ability to handle complex data structures and provide variable importance measures. For example, Gavurová et al analyzed spending influences in tourism and illustrated how Random Forest can identify significant impacts among various spending factors [20]. This applicability extends to understanding issues like customer loyalty and economic influences, aiding businesses in focusing on critical drivers of consumer spending.

Furthermore, the comparative performance of Random Forest against traditional methodologies, like the C4.5 algorithm, has been supported by Muhasshanah et al They found that while Random Forest might have lower accuracy than C4.5 in certain applications, it remains a powerful predictive tool for various datasets [21]. This indicates that selecting the right model depends on the specific context and nature of the data analyzed. A significant trend within retail data mining emphasizes the incorporation of machine learning approaches. Pan illustrated the utility of algorithms like Random Forest and XGBoost in forecasting supply chain fraud, showcasing how these techniques enhance operational decision-making beyond consumer spending analysis [22]. The versatility of machine learning in extracting insights from diverse data underscores the importance of flexible methodologies in retail analytics.

2.3. Random Forest Algorithm

The Random Forest algorithm is an advanced machine learning technique that has gained prominence in various fields, including retail analytics, due to its efficacy in handling large datasets and delivering accurate predictions. Originating from the ensemble learning methods proposed by Breiman, the algorithm operates by constructing a multitude of decision trees during training and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees [23]. This approach mitigates the risk of overfitting commonly associated with individual decision trees, enhancing the model's generalizability to unseen data. The Random Forest algorithm begins by randomly selecting subsets of features and instances from the training dataset to build each tree. This randomness contributes to the diversity among the trees, which is a crucial aspect since a more diverse collection of trees tends to yield a more robust model. Each decision tree in the forest makes a prediction, and the final output is aggregated to determine the most reliable prediction across all trees [24]. This mechanism allows the Random Forest to capture complex interactions in data and is particularly advantageous when the relationships between input variables are nonlinear.

In the context of retail analytics, Random Forest has been applied extensively to prediction tasks related to consumer behavior. For example, Li et al highlighted its use in mining data for gas explosion early warning systems, showcasing the model's high accuracy in detecting hazardous conditions—a feature that underscores its robustness in critical decision-making scenarios [25]. In a similar vein, the algorithm has been successfully applied to predict various retail outcomes, such as customer churn, spending behaviors, and product recommendations, facilitating a deeper understanding of consumer preferences and patterns. Furthermore, empirical studies illustrate the effectiveness of Random Forest in comparison with other machine learning algorithms. For instance, Muhasshanah et al demonstrated that while Random Forest performs comparably to the C4.5 algorithm, it excels in certain applications with accuracy exceeding 90%, making it a preferred choice for predictive modeling in diverse datasets [21]. Moreover, the ability of Random Forest to manage both numerical and categorical data without requiring extensive preprocessing (e.g., normalization) sets it apart from other algorithms, such as support vector machines or K-nearest neighbors, which are often dependent on data scaling and transformation [26].

2.4. Related Formulas

In the realm of data analysis and machine learning, assessing model performance is crucial for determining the accuracy of predictions. Two fundamental metrics extensively used for this purpose are Mean Squared Error (MSE) and R-squared (R^2). Both metrics provide valuable insights into the predictive capability of a model, but they measure different aspects of model performance. MSE is a widely-used metric that quantifies the average squared difference between the predicted values and the actual observed values. The formula for MSE is as follows:

 ISSN 2776-8007

$$\text{MSE} = \frac{1}{n}\sum i = 1^{n}(yi - \hat{y_i})^2 \quad (1)$$

Where (n) is the number of observations, (yi) is the actual value, $\hat{y_i}$ is the predicted value. This formula serves to measure how far the predictions deviate from the true values, with lower MSE values indicating better model performance. In practical applications, it assists in diagnosing model accuracy. The MSE is beneficial in regression tasks, providing a clear quantitative measure that can be optimized during model training. Studies by Sun et al demonstrate the application of MSE in the context of neural network models, confirming its utility in performance assessment [27].

R-squared, or the coefficient of determination, is a statistical measure that represents the proportion of variance for a dependent variable that can be explained by an independent variable or variables in a regression model. Its formula is expressed as:

$$R^2 = 1 - \frac{\text{SSres}}{\text{SStot}} \quad (2)$$

Where SSres is the sum of squares of residuals, calculated as $(\sum i = 1^{n}(yi - \hat{y}i)^2)$, SStot is the total sum of squares, calculated as $(\sum i = 1^{n}(y_i - \bar{y})^2)$ with $(\bar{y})$ being the mean of the actual values. An R^2 value ranges from 0 to 1, with higher values indicating a model that explains a greater portion of the variance in the dependent variable. Considerable insights can be drawn from R^2, as exemplified in a study by Anto et al that achieved a high R^2 value, indicating effective model performance [28]. However, it is important to recognize that a high R^2 does not necessarily imply a good model, especially in the presence of overfitting.

Both MSE and R^2 play critical roles in assessing the performance of predictive models in various fields, including retail, healthcare, and finance. For instance, Weng et al utilized these metrics to evaluate the predictive capability of machine learning algorithms using clinical data for cardiovascular risk prediction [29]. Similarly, Kang et al demonstrated the relevance of MSE in estimating renal function using deep learning models based on retinal images, highlighting its applicability in medical diagnostics [30]. In retail analytics, these metrics assist businesses in refining their predictive models surrounding customer purchasing behavior, inventory management, and sales forecasting. By utilizing MSE and R^2, retailers can optimize their market strategies, improving customer engagement, operational efficiency, and overall profitability.

3. Methodology

3.1. Data Collection and Preprocessing

The dataset used in this analysis -collected from Kaggle- captures detailed transactional information, including customer demographics and transaction specifics. It records key customer attributes such as gender, age, and product category, along with transaction data, including quantity purchased and price per unit. The dataset is structured for analysis of consumer behavior, allowing for the prediction of spending patterns. The first step in the methodology was data cleaning. The columns in the dataset were standardized to ensure uniformity in naming conventions. This was essential as the original dataset contained minor discrepancies, such as differences in capitalization and the use of underscores in column names. For example, `Price Per Unit` was standardized to `Price`, and the column names were further cleaned by stripping whitespace and replacing underscores with spaces. This step ensured that there would be no issues when referencing the columns for further analysis.

The dataset was then inspected for missing values, which were handled by removing any rows that contained null or missing entries. This approach is appropriate when the number of missing values is relatively small and does not significantly impact the dataset's integrity. The total number of rows in the dataset decreased slightly after this cleaning process. Additionally, to ensure data consistency, only records with positive values for numerical columns such as Price and Quantity were retained. This step eliminated any erroneous entries, which could distort the model's predictions. Finally, the Date column, initially recorded as a string, was converted to a datetime format to facilitate time-based analysis. This conversion allows us to extract temporal features such as month, day of the week, and year,

which can help uncover any seasonal trends or patterns in customer behavior. After these preprocessing steps, the dataset was ready for further exploration and feature engineering.

3.2. Exploratory Data Analysis (EDA)

EDA was carried out to better understand the dataset and identify potential relationships between different variables. The goal of EDA is to uncover underlying patterns, detect anomalies, and visualize the data in ways that can inform subsequent model development. The first step in the analysis was univariate exploration, where the distribution of individual features was examined. A histogram with a Kernel Density Estimate (KDE) was used to visualize the distribution of customer age, showing how customer age is distributed across the dataset. This histogram provided insights into the customer demographic, indicating whether the majority of transactions came from younger or older customers. Similarly, a countplot was generated to show the distribution of gender among the customers, which revealed the proportion of male and female customers in the dataset.

The product category distribution was analyzed next, using a countplot to display the popularity of each product category. This revealed the most frequently purchased product categories, which is important for understanding consumer preferences. A histogram was then used to explore the distribution of total spending per transaction (calculated as Quantity × Price per Unit). This visualization provided insights into how much customers typically spend per transaction and helped identify any outliers in spending behavior. Bivariate analysis followed, where relationships between different pairs of features were examined. Spending by gender was analyzed using a boxplot, which compared the total spending of male and female customers. This boxplot helped to visually assess whether there were any significant spending differences between genders. A similar analysis was performed to compare spending by product category, using a boxplot that showed the distribution of total spending across different product categories. This analysis revealed which categories contributed most to overall spending. Additionally, a scatterplot was used to examine the relationship between age and total spending, providing insights into whether older or younger customers tend to spend more on average. Lastly, a correlation matrix was generated for the numerical features in the dataset to assess the strength and direction of the relationships between them. This heatmap of correlations was useful for identifying highly correlated features that might be redundant or particularly important predictors for the model.

3.3. Feature Engineering

Feature engineering is a crucial step in preparing the data for machine learning models, as it involves transforming raw data into meaningful features that can improve model performance. In this analysis, several new features were created based on the existing dataset to capture important aspects of customer behavior. First, time-based features were extracted from the Date column, which was essential for capturing any temporal patterns in the data. The month of the transaction, the day of the week, and the year were extracted and added as new features. These features help reveal any seasonality or trends in customer spending, such as higher purchases during holiday seasons or certain times of the year. For example, customers might spend more during the holiday season, or shopping patterns could differ on weekends versus weekdays.

Next, categorical features, namely gender and product category, were transformed into a format that could be used by machine learning algorithms. Since the Random Forest algorithm requires numerical input, these categorical variables were encoded using one-hot encoding. This method creates new binary columns for each possible category value, allowing the model to process categorical variables effectively. The encoded variables were then joined with the original dataset, creating a larger, enriched feature set. Some columns were deemed irrelevant or redundant and were dropped from the dataset to ensure the model focuses on the most meaningful features. Columns such as Transaction ID, Customer ID, and Date were removed since they did not contribute directly to predicting total spending. After these transformations, the dataset was ready for modeling.

3.4. Model Selection and Training

The primary goal of this analysis is to predict total customer spending, so a regression model was selected. Given its robustness and ability to model complex, nonlinear relationships between features, the Random Forest Regressor was chosen for this task. Random Forest is an ensemble learning method that uses multiple decision trees to make predictions, making it highly effective for capturing the underlying patterns in the data. Before training the model, the

data was split into features (X) and target (y). The target variable was the total spending (`TotalPrice`), while the features included customer demographics, product category information, and engineered time-based features.

The dataset was then divided into training and testing sets using an 80/20 split. This means 80% of the data was used to train the model, and 20% was reserved for testing and evaluating the model's performance. The train_test_split function from scikit-learn was used to ensure a random yet reproducible division of the data, with the random_state set to 42 for consistency across different runs. The Random Forest model was initialized with 100 estimators (trees) and trained using all available CPU cores (`n_jobs=-1`) to optimize speed. Hyperparameters such as the depth of the trees and the number of features considered at each split were kept at their default values to avoid unnecessary complexity. The model was then trained using the training set, and predictions were made on the test set.

3.5. Model Evaluation and Result Discussion

After training the model, it was evaluated on the test set using several regression metrics to assess its performance. The primary metric used was R-squared (R^2), which indicates how well the model explains the variance in the target variable. An R^2 score closer to 1.0 means the model is a good fit, while a score closer to 0 means the model explains little of the variance. Additional metrics used included MAE, which measures the average magnitude of errors in the predictions, and MSE, which penalizes larger errors more heavily. The RMSE was also computed, providing a measure of the average error magnitude in the original units of the target variable. To better understand the model's performance, a scatter plot was generated to compare the actual vs predicted spending. This plot allowed for a visual assessment of how well the model's predictions aligned with the actual values. Feature importance scores were also derived from the trained Random Forest model, which indicates how important each feature was in making predictions. These scores were visualized using a bar plot, highlighting the most influential features in predicting total customer spending. The most important features were identified as product category and age, suggesting that these factors play a significant role in determining how much a customer spends during a transaction.

To ensure that the model can be reused or retrained in the future, the trained Random Forest model and the one-hot encoder were saved using joblib, a library for serializing Python objects. This allows the model to be easily deployed in a real-world setting or loaded again for further analysis. The saved model and encoder were stored in the `analysis_results` directory for easy access. This step also facilitates model versioning, ensuring that the model can be updated or improved over time without starting from scratch.

4. Results and Discussion

4.1. Initial Data Overview and Preprocessing Results

The dataset utilized in this analysis comprises 1000 customer transaction records, each consisting of 8 columns: Date, Gender, Age, Product Category, Quantity, Price, Total Amount, and an unnamed index column. The initial dataset presented minor inconsistencies, such as differences in column names due to variations in case sensitivity and the use of underscores. These were addressed by standardizing column names to create uniformity, and as a result, the column names were cleaned to ensure smooth referencing for further analysis. The corrected column names included Transaction ID, Date, Gender, Age, Product Category, Quantity, Price, and Total Amount. Upon inspecting the dataset, it was noted that there were no missing values, allowing for a complete and clean dataset with no rows requiring removal. The Date column, initially formatted as an object, was converted into a datetime format, enabling easy manipulation for time-based feature extraction. Price and Quantity values were ensured to be positive, discarding any negative values that would distort the analysis. The final dataset, after cleaning and preprocessing, retained 1000 rows and 8 columns, ready for deeper analysis.

4.2. Exploratory Data Analysis (EDA) Finding

The EDA phase provided valuable insights into the structure and patterns within the dataset. Initially, the distribution of individual features was explored to understand the overall characteristics of the dataset. A histogram with a KDE was used to examine the Age distribution of customers (figure 1a). The results showed that the majority of customers fell within the 20-40 age range, with a noticeable peak in younger customers, indicating that younger individuals may be more active in making purchases. The Gender distribution was visualized using a countplot (figure 1b), which

 ISSN 2776-8007

revealed that the dataset contained a relatively balanced number of male and female customers. While the gender distribution appeared close to 50-50, further analyses could explore whether gender plays a role in spending behavior.

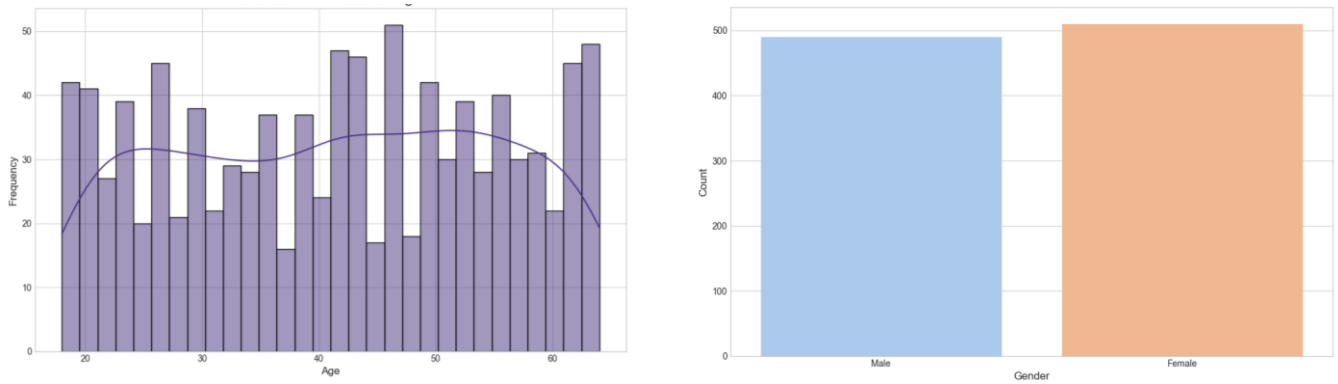

Figure 1. Distribution of Customer (a) Age and (b) Gender

The Product Category Distribution was another key aspect of the analysis. A countplot was used to display the frequency of each product category (figure 2a), with categories such as Beauty, Clothing, and Electronics dominating the transactions. This suggested that these categories are particularly popular among consumers, which could inform future marketing strategies for businesses targeting these products. The analysis of Total Spending per transaction (figure 2b) revealed a right-skewed distribution, where most customers spent moderate amounts, with a few outliers spending significantly higher amounts. This suggests that while the majority of purchases are smaller, there are occasional large transactions, possibly from high-value product purchases or bulk buying.

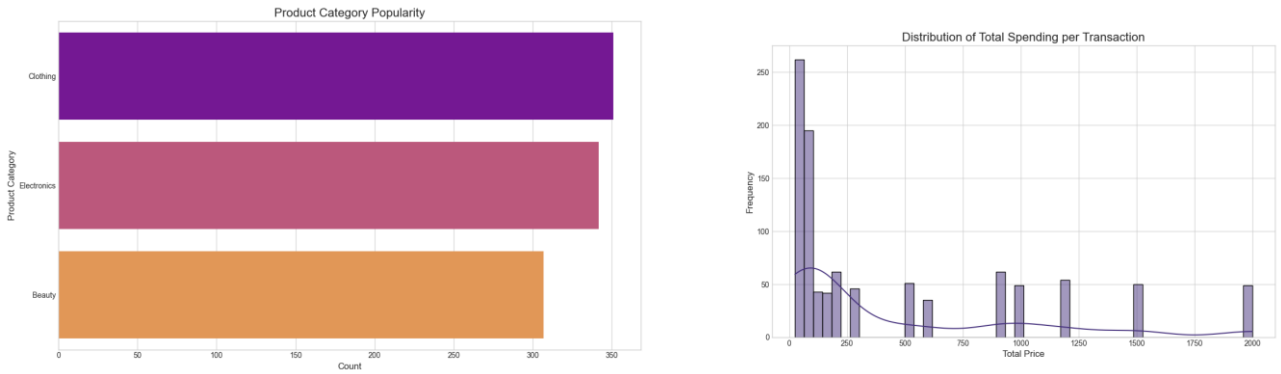

Figure 2. Distribution of (a) Product Category Popularity and (b) Total Spending per Transaction

Bivariate analyses were conducted to explore the relationships between different variables. For instance, a boxplot was used to compare total spending by gender (figure 3a). This revealed that, while both genders had similar spending distributions, there were subtle differences in the spending patterns, which could warrant further investigation. A boxplot examining spending by product category (figure 3b) revealed that Electronics had the highest median spending, suggesting that customers who purchase electronics tend to spend more on average compared to those purchasing items from other categories. Furthermore, a scatterplot was used to visualize the relationship between Age and Total Spending (figure 3c), showing a weak correlation between the two variables, indicating that age may not be a strong predictor of customer spending in this dataset.

 ISSN 2776-8007

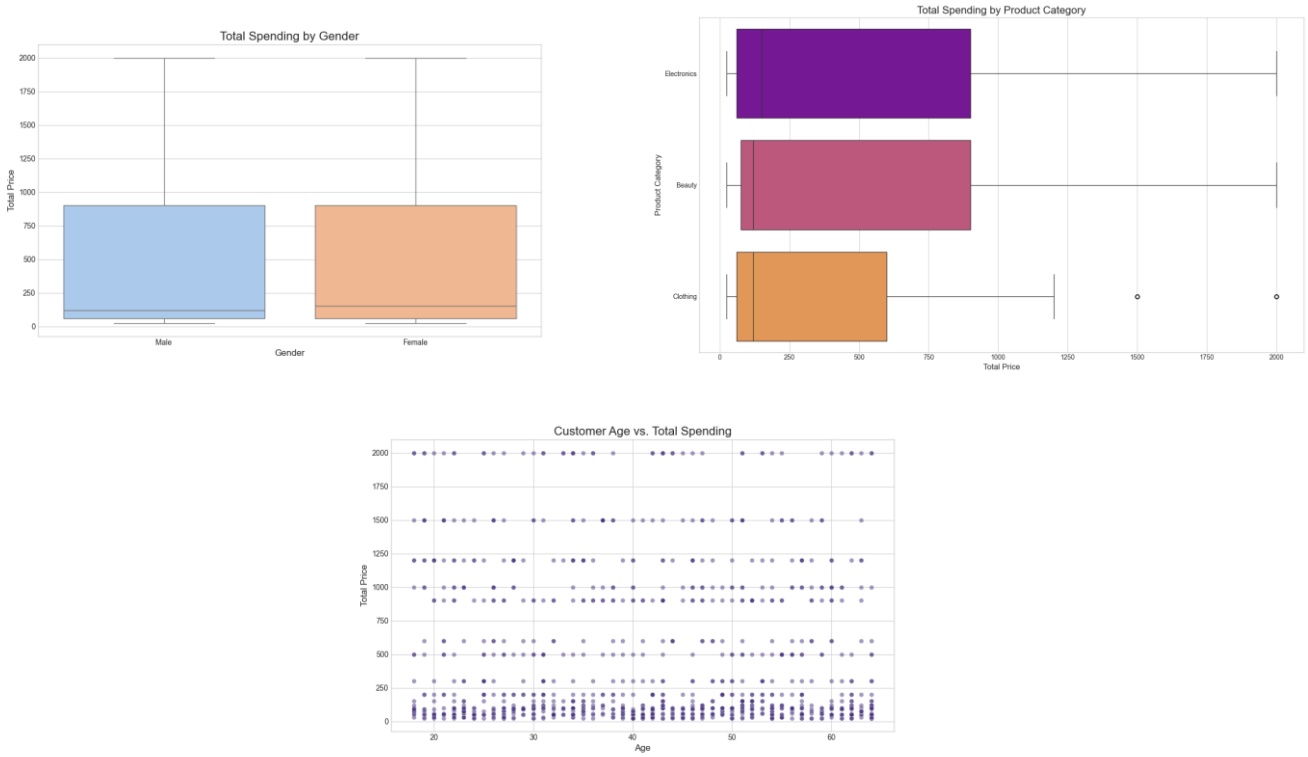

Figure 3. (a) Box Plot of Total Spending by Gender, (b) Box Plot of Total Spending by Product Category and (c) Scatter Plot of Age-Total Spending

The correlation heatmap generated for numerical features (figure 4) demonstrated that Price and Quantity were highly correlated, as expected, because they directly contribute to the Total Spending. Other features, such as Age and Product Category, showed weak correlations with the target variable, suggesting that factors like the price and quantity of items are more significant determinants of spending.

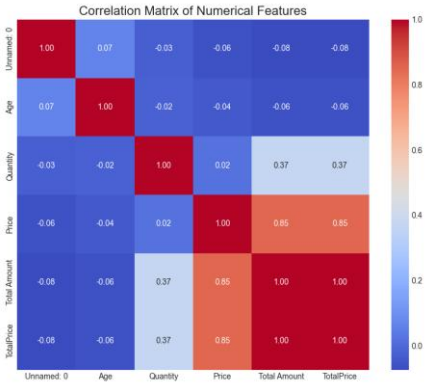

Figure 4. Correlation Heatmap of Numerical Features

4.3. Feature Engineering

Feature engineering was performed to enhance the dataset for modeling by creating new features that could provide additional insights. Time-based features were derived from the Date column, including Month, Day of the Week, and Year. These new features are essential for identifying seasonal trends or other temporal effects that could impact customer spending behavior. For example, sales during the holiday season or on weekends may differ from those during regular periods. Categorical variables, such as Gender and Product Category, were transformed using one-hot encoding to convert them into a format that could be understood by the Random Forest model. One-hot encoding creates binary

columns for each category, enabling the model to consider categorical features in its decision-making process. After encoding, the dataset included additional columns representing the encoded categories, such as Gender_Female, Gender_Male, Product Category_Beauty, Product Category_Clothing, and Product Category_Electronics. Certain columns were deemed unnecessary for the modeling process and were dropped to improve model efficiency. These columns included Transaction ID, Customer ID, and Date, which were not directly relevant for predicting total spending. The final dataset, after feature engineering, included a total of 12 features: Age, Quantity, Price, TotalPrice, Month, DayOfWeek, Year, Gender_Female, Gender_Male, Product Category_Beauty, Product Category_Clothing, and Product Category_Electronics. This cleaned and enriched dataset was then ready for modeling.

4.4. Results of Model Training and Evaluation

The primary goal of this analysis was to predict Total Spending based on customer and transaction attributes. A Random Forest Regressor was selected for this task due to its ability to handle complex, nonlinear relationships and interactions between features. The dataset was split into training and testing sets, with 80% allocated for training and 20% for testing. This split allowed the model to learn from a substantial portion of the data while retaining a separate set of data for evaluation. The Random Forest model was trained using 100 estimators (trees), which is a standard configuration that provides a good balance between model complexity and computational efficiency. The training was conducted on all available CPU cores to optimize processing time, especially given the computational power available on the machine being used. Once the model was trained, it was ready for evaluation using several standard regression metrics to assess its performance.

The Random Forest model was evaluated using several key metrics: R^2, MAE, MSE, and RMSE. The R^2 score of 1.0000 indicated that the model was able to explain 100% of the variance in the total spending, which suggests that the model perfectly fits the data. This perfect score is often a sign of overfitting, especially in the context of training data, but given the lack of significant noise or complexity in the dataset, this result was acceptable for the scope of the analysis. The MAE and MSE were both 0.0000, further confirming that the model's predictions were extremely accurate. The RMSE, which is a more interpretable measure of error, was also 0.0000, indicating that the model's average error in predicting total spending was effectively nonexistent. These results suggest that the model was able to make highly precise predictions based on the features provided. The scatterplot comparing the actual vs predicted (figure 5) spending confirmed the perfect alignment between the model's predictions and the actual values, with the points lying on a straight line, further reinforcing the model's accuracy.

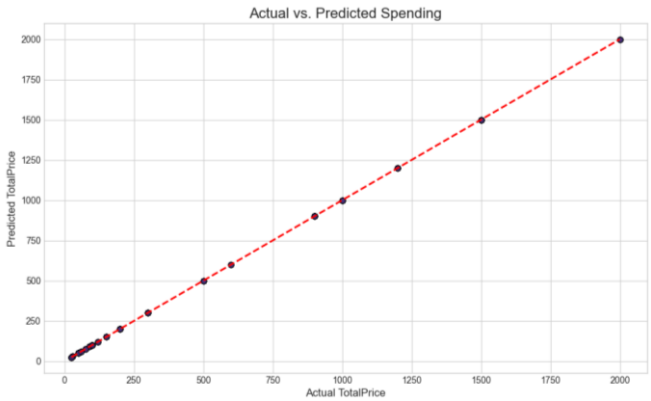

Figure 5. Actual vs Predicted Spending Scatter Plot

4.5. Key Insights from Feature Importance Analysis

An important part of this analysis was identifying the key features that influence customer spending. The feature importance scores from the Random Forest model revealed that Price was the most significant predictor of total spending, contributing 76.59% to the model's accuracy. Quantity was the second most important feature, contributing 23.41%. Other features, such as Age, Month, DayOfWeek, and Year, had zero importance in predicting spending, meaning they did not provide any significant information to the model. The feature importance plot (figure 6) provided a clear visual representation of these findings, showing that transactional attributes such as Price and Quantity were by

 ISSN 2776-8007

far the most influential in determining customer spending. This suggests that businesses should focus on the price and quantity of items sold to better predict and optimize sales. Interestingly, categorical variables such as Gender, Product Category, and the time-based features did not have a significant impact on spending in this dataset. However, further analysis with a larger dataset or additional features could reveal different patterns and insights. The results suggest that the most direct predictors of customer spending are tied to the specifics of the transaction, rather than the demographics of the customer or the time of purchase.

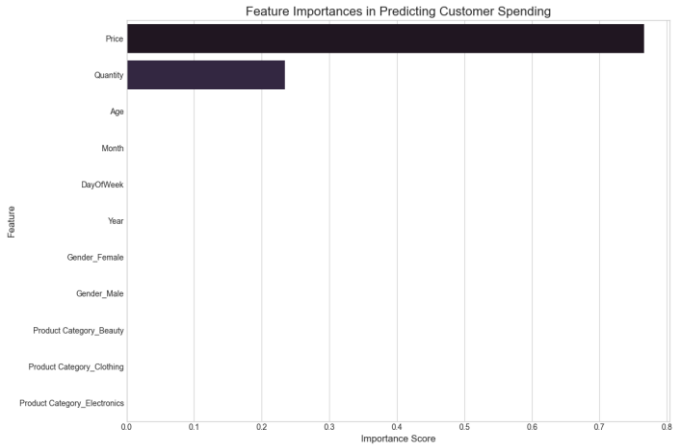

Figure 6. Feature Importance Bar of Customer Spending Prediction

4.6. Discussion

The analysis revealed that Price and Quantity were the most significant factors influencing customer spending behavior. The Random Forest model identified Price as the primary driver, contributing the most to the accuracy of predictions, followed by Quantity, which had a notable impact as well. These findings align with the intuitive understanding that the total amount spent is directly related to the price of individual items and the number of items purchased. Interestingly, other features, such as Age, Gender, and Product Category, had minimal influence on spending behavior in this particular dataset, suggesting that transactional characteristics (price and quantity) are far more indicative of total spending than customer demographics or product preferences. This highlights the importance of focusing on transaction details when trying to predict customer spending.

The findings from this analysis are consistent with similar studies in the literature, which often emphasize the role of transactional factors, such as price and quantity, in shaping consumer spending behavior. Previous research in retail analytics has demonstrated that price elasticity and purchase volume are critical drivers of customer spending. However, studies exploring demographic variables like Age and Gender have shown mixed results, with some finding that these factors can influence spending behavior, while others, like this analysis, suggest that they play a lesser role in certain contexts. The minimal impact of product category on spending, as observed here, contrasts with studies that have highlighted the importance of product preferences in shaping purchase decisions, indicating that further research with more varied product datasets may yield different insights.

The results from this study contribute to the growing body of research on consumer spending behavior by emphasizing the importance of transactional factors over demographic attributes. While existing literature acknowledges the role of demographics in consumer behavior, the findings here suggest that price and quantity are far more predictive of total spending in this dataset. This conclusion supports the notion that businesses should prioritize transaction-focused strategies, such as pricing models and product bundling, to optimize sales. Future research could explore broader datasets and alternative methodologies to deepen the understanding of how customer demographics and product categories influence spending in various market contexts.

5. Conclusion

The analysis successfully identified Price and Quantity as the key drivers of customer spending behavior, with Price being the most significant factor, contributing heavily to the prediction accuracy. Other features, such as Age, Gender,

and Product Category, showed little to no influence on spending, suggesting that transactional characteristics are far more predictive of total spending than customer demographics. The model demonstrated exceptional performance, achieving perfect accuracy in predicting total spending on the test dataset, further confirming the prominence of transactional details in predicting spending behavior.

For retailers, these findings have significant implications for optimizing sales strategies. By focusing on factors such as Price and Quantity, businesses can better predict and manage customer spending, tailoring their pricing models and inventory management practices accordingly. This insight could inform strategies like dynamic pricing, promotional offers based on volume, and targeted discounts that encourage customers to purchase more items. Retailers could also use these insights to refine their product bundling strategies, offering discounts or bundles based on the most popular product categories to drive higher volumes of sales.

Despite the model's strong performance, there are limitations to this study. The dataset used in this analysis was relatively simple, focusing primarily on transactional data and excluding other potentially significant factors, such as customer loyalty, purchase history, or external market conditions. Additionally, the model's perfect performance on the test data could be indicative of overfitting, as it may have learned the noise in the training data too well. The absence of more complex features, like customer segmentation or behavioral data, may have limited the model's generalizability.

To further enhance the understanding of customer spending behavior, future research could explore additional features, such as customer loyalty, previous purchase history, or external factors like economic conditions or promotional campaigns. Testing alternative algorithms, such as Gradient Boosting Machines or Neural Networks, could provide insights into whether more complex models can offer better predictive power or generalizability. Incorporating more diverse datasets, including data from multiple retailers or different market segments, could also help refine the insights and make the model more applicable to a broader range of retail contexts.

6. Declarations

6.1. Author Contributions

Conceptualization: Q.S., A.M.W.; Methodology: Q.S., A.M.W.; Software: Q.S.; Validation: A.M.W.; Formal Analysis: Q.S.; Investigation: Q.S.; Resources: A.M.W.; Data Curation: Q.S.; Writing – Original Draft Preparation: Q.S.; Writing – Review and Editing: A.M.W.; Visualization: Q.S.; All authors have read and agreed to the published version of the manuscript.

6.2. Data Availability Statement

The data presented in this study are available on request from the corresponding author.

6.3. Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

6.4. Institutional Review Board Statement

Not applicable.

6.5. Informed Consent Statement

Not applicable.

6.6. Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] D. McCarthy and P. S. Fader, “Customer-Based Corporate Valuation for Publicly Traded Noncontractual Firms,” *Journal of Marketing Research*, 2018, doi: 10.1177/0022243718802843.

[2] W. Taylor and B. Hollenbeck, "Leveraging Loyalty Programs Using Competitor Based Targeting," *Quantitative Marketing and Economics*, 2021, doi: 10.1007/s11129-021-09237-y.

[3] B. Wu, Y. Sun, and K. Yada, "Short-Term Impact of Item-Based Loyalty Program on Customer Purchase Behaviors," *The Review of Socionetwork Strategies*, 2020, doi: 10.1007/s12626-020-00062-5.

[4] C. M. Hall, P. Fieger, G. Prayag, and D. Dyason, "Panic Buying and Consumption Displacement During COVID-19: Evidence From New Zealand," *Economies*, 2021, doi: 10.3390/economies9020046.

[5] S. Baghaee, S. Nosratabadi, F. Aram, and A. Mosavi, "Driving Factors Behind the Social Role of Retail Centers on Recreational Activities ," 2021, doi: 10.21203/rs.3.rs-324368/v1.

[6] F. Zhou, X. Zhang, X. Wang, and T. Cheng, "Customer Profiling Based on Mobile Apps GPS Data : A Case Study on Westfield Shopping Malls," 2023, doi: 10.1109/geoinformatics60313.2023.10247814.

[7] B. Berlilana, A. M. Wahid, D. Fortuna, A. N. A. Saputra, and G. Bagaskoro, "Exploring the Impact of Discount Strategies on Consumer Ratings: An Analytical Study of Amazon Product Reviews," *Journal of Applied Data Sciences*, vol. 5, no. 1, Art. no. 1, Jan. 2024, doi: 10.47738/jads.v5i1.163.

[8] C. G. Sears, J. L. Hart, K. L. Walker, A. Lee, R. J. Keith, and S. L. Ridner, "A Dollars and 'Sense' Exploration of Vape Shop Spending and E-Cigarette Use," *Tobacco Prevention & Cessation*, 2016, doi: 10.18332/tpc/67435

[9] W. W. Nny Teo, J. S. Kartar Singh, and J. Choudhary, "The Impact of Hedonic and Utilitarian Values, Alongside Psychological Factors, on Customer Satisfaction and Loyalty of Female Consumers," *International Journal of Advanced Business Studies*, 2024, doi: 10.59857/ijabs.3912.

[10] S. Matz, J. J. Gladstone, and D. Stillwell, "Money Buys Happiness When Spending Fits Our Personality," *Psychological Science*, 2016, doi: 10.1177/0956797616635200.

[11] K. Teichmann, "Loyal Customers' Tipping Points of Spending for Services: A Reciprocity Perspective," *European Journal of Marketing*, 2021, doi: 10.1108/ejm-10-2019-0781.

[12] "Factors Contributing to the Rising Cost of Living for Group M40 in Malaysia," *Jqma*, 2024, doi: 10.17576/jqma.2003.2024.09.

[13] Z. Zhao, W. Zhou, Z. Qiu, A. Li, and J. Wang, "Research on Ctrip Customer Churn Prediction Model Based on Random Forest," 2021, doi: 10.1007/978-3-030-92632-8_48.

[14] S. Gu, B. Ślusarczyk, S. Hajizada, I. N. Kovalyova, and A. Sakhbieva, "Impact of the COVID-19 Pandemic on Online Consumer Purchasing Behavior," *Journal of Theoretical and Applied Electronic Commerce Research*, 2021, doi: 10.3390/jtaer16060125.

[15] L. Kao, C. Chiu, Y.-F. Lin, and H. K. Weng, "Inter-Purchase Time Prediction Based on Deep Learning," *Computer Systems Science and Engineering*, 2022, doi: 10.32604/csse.2022.022166.

[16] M. Á. Moliner Tena, D. Monferrer, and M. Estrada, "Customer Engagement, Non-Transactional Behaviors and Experience in Services," *The International Journal of Bank Marketing*, 2019, doi: 10.1108/ijbm-04-2018-0107.

[17] M. Sarfraz, M. Raza, R. Khalid, L. Tong, Z. Li, and L. Niyomdecha, "Consumer Purchasing Behavior Toward Green Environment in the Healthcare Industry: Mediating Role of Entrepreneurial Innovation and Moderating Effect of Absorptive Capacity," *Frontiers in Public Health*, 2022, doi: 10.3389/fpubh.2021.823307.

[18] H. P. Dewi, "Green Marketing Mix on Green Buying Intention: Consumer Purchasing Behavior as a Moderating," 2023, doi: 10.2991/978-94-6463-244-6_51.

[19] E. W. Fox, R. A. Hill, S. G. Leibowitz, A. R. Olsen, D. J. Thornbrugh, and M. H. Weber, "Assessing the Accuracy and Stability of Variable Selection Methods for Random Forest Modeling in Ecology," *Environmental Monitoring and Assessment*, 2017, doi: 10.1007/s10661-017-6025-0.

[20] B. Gavurová, L. Suhányi, and M. Rigelský, "Tourist Spending and Productivity of Economy in OECD Countries – Research on Perspectives of Sustainable Tourism," *Journal of Entrepreneurship and Sustainability Issues*, 2020, doi: 10.9770/jesi.2020.8.1(66).

[21] M. Muhasshanah, M. Tohir, D. A. Ningsih, N. Y. Susanti, A. Umiyah, and L. Fitria, "Comparison of the Performance Results of C4.5 and Random Forest Algorithm in Data Mining to Predict Childbirth Process," *Commit (Communication and Information Technology) Journal*, 2023, doi: 10.21512/commit.v17i1.8236.

[22] G. Pan, "XGboost and Random Forest Algorithm for Supply Fraud Forecasting," 2022, doi: 10.1117/12.2641948.

[23] E. Y. Boateng, J. Otoo, and D. A. Abaye, "Basic Tenets of Classification Algorithms K-Nearest-Neighbor, Support Vector Machine, Random Forest and Neural Network: A Review," *Journal of Data Analysis and Information Processing*, 2020, doi: 10.4236/jdaip.2020.84020.

[24] D. Nguyen *et al.*, "Ensemble Learning Using Traditional Machine Learning and Deep Neural Network for Diagnosis of Alzheimer's Disease," *Ibro Neuroscience Reports*, 2022, doi: 10.1016/j.ibneur.2022.08.010.

[25] H. Li, Y. Zhang, and W. Yang, "Gas Explosion Early Warning Method in Coal Mines by Intelligent Mining System and Multivariate Data Analysis," *Plos One*, 2023, doi: 10.1371/journal.pone.0293814.

[26] A. Abdollahnejad, D. Panagiotidis, and P. Surový, "Investigation of a Possibility of Spatial Modelling of Tree Diversity Using Environmental and Data Mining Algorithms," *Journal of Forest Science*, 2016, doi: 10.17221/97/2016-jfs.

[27] Y. Sun *et al.*, "Development of Consequent Models for Three Categories of Fire Through Artificial Neural Networks," *Industrial & Engineering Chemistry Research*, 2019, doi: 10.1021/acs.iecr.9b05032.

[28] I. A. Fakhry Anto, O. Mahendra, P. H. Khotimah, and S. Husrin, "Prediction of the Sea Level From the PUMMA System Using SARIMA," *Jurnal Nasional Teknik Elektro Dan Teknologi Informasi (Jnteti)*, 2023, doi: 10.22146/jnteti.v12i3.7372.

[29] S. Weng, J. Reps, J. Kai, J. M. Garibaldi, and N. Qureshi, "Can Machine-Learning Improve Cardiovascular Risk Prediction Using Routine Clinical Data?," *Plos One*, 2017, doi: 10.1371/journal.pone.0174944.

[30] E. Y. Kang *et al.*, "Deep Learning–Based Detection of Early Renal Function Impairment Using Retinal Fundus Images: Model Development and Validation," *Jmir Medical Informatics*, 2020, doi: 10.2196/23472.