**Overview**

Customer Behaviour And Analytics

Onboarding guide • Jan 2, 2025

Divider

# Abstract :

The **Shopping Dataset** contains data gathered from various consumer shopping sessions across multiple e-commerce platforms or retail environments. It encompasses detailed information on customer demographics, product preferences, browsing behavior, purchase histories, and transactional activities. Key features of the dataset include customer age, gender, ID, annual income, etc….. The dataset is valuable for studying consumer purchasing patterns, predicting future buying behaviors, optimizing product recommendations, and conducting market segmentation analysis. This data is typically used to develop recommendation systems, consumer behavior prediction models, and retail analytics applications. To address the dataset, various techniques such as data preprocessing, analysis, normalization, correlation assessment, and heatmap generation are employed. Additionally, data visualization methods are utilized to gain deeper insights and enhance interpretation. In the course of analyzing the dataset, various relationships between different variables were examined to understand their predictive power. Initially, the correlation between annual income and spending score was explored, with the model yielding an accuracy of 60%. This was followed by an investigation into the relationship between age and annual income, which also resulted in an accuracy of 60%. Similarly, the comparison between gender and annual income was assessed, and it too achieved an accuracy of 60%. These results suggest that, while the selected variables demonstrate some predictive value, the accuracy of the model remains moderate, indicating the need for further refinement or exploration of additional features to improve the model's performance. This dataset is utilized to predict the spending score in the context of consumer shopping behavior. By analyzing various features such as income, age, and gender, the goal is to develop a model that can accurately forecast an individual's spending tendencies, thereby offering valuable insights for targeted marketing strategies and personalized recommendations within the retail domain.

KEYWORDS :

We have few keywords in the shopping dataset

* Normalization
* Correlation
* Linear regression
* Heatmap
* Training, Testing
* Optimization

INTRODUCTION :

This dataset captures a wide range of customer attributes including Age, Gender, Spending Score, Annual Income, CustomerID, and more. Analyzing this data can help businesses make informed decisions, optimize product offerings, and enhance customer satisfaction. The Customer Shopping Preferences Dataset offers valuable consumer behavior and purchasing patterns insights. Understanding customer preferences and trends is critical for businesses to tailor their products, marketing strategies, and overall customer experience. I took a shopping data set to understand customer preferences and behaviors.

Analyzing this data can help businesses make informed decisions. The dataset stands as a valuable resource for businesses aiming to align their strategies with customer needs and preferences. Shopping datasets can be used to predict customer behavior and purchases using machine learning models, such as those that Identify potential customers, Analyze consumer behavior, Create personalized recommendations, Improve marketing strategies, Enhance customer experience, etc….This data is typically used to develop recommendation systems, consumer behavior prediction models, and retail analytics applications.

To address the dataset, various techniques such as data preprocessing, analysis, normalization, correlation assessment, and heatmap generation are employed. The shopping dataset provides a comprehensive look at customer behaviors and demographic characteristics, which can significantly enhance business decision-making. It includes crucial attributes such as CustomerID, Gender, Age, Annual Income, and Spending Score. These data points allow businesses to segment their customer base effectively and develop personalized marketing strategies. By understanding how each customer interacts with products and services, businesses can improve targeting and tailor promotions.

The dataset also provides valuable insights into purchasing patterns, helping businesses identify trends over time. Analyzing such data enables companies to refine their product offerings, pricing, and customer engagement tactics. With a deeper understanding of customer behavior, businesses can optimize sales and improve overall customer satisfaction. Additionally, this data helps businesses predict future trends and adapt to shifts in consumer preferences. By leveraging the shopping dataset, companies can create more efficient and customer-centric marketing campaigns. This ultimately leads to higher customer loyalty and increased revenue potential.

By leveraging attributes like Gender, Age, Annual Income, and Spending Score, businesses can craft personalized marketing strategies that address individual customer needs. For instance, a high-income customer may be more responsive to exclusive product offerings or early access to sales, while a younger customer may prefer promotions on tech gadgets or fashion. Personalized recommendations based on past purchases or browsing behavior can significantly enhance the shopping experience, leading to increased customer satisfaction and loyalty. Furthermore, targeted emails or advertisements tailored to specific customer segments can drive higher engagement, as the content is more relevant to the recipient. With data-driven insights, businesses can create more dynamic and customer-centric experiences that make customers feel understood and valued. Personalization not only boosts sales but also strengthens brand loyalty, as customers are more likely to return when they feel their preferences are being considered. Moreover, personalized marketing helps in building long-term relationships with customers, ensuring consistent sales and greater customer retention. By continually analyzing customer data, businesses can refine their strategies and stay ahead of trends. Ultimately, personalized marketing through data insights leads to improved performance and sustainable business growth.

RELATED PAPERS :

An innovative real-time scheduling framework is proposed to address a large-scale and multi-resource-constrained hybrid flow shop (HFS) scheduling problem. The framework utilizes the Dueling Double Deep Q Network (D3QN) algorithm to minimize the makespan. Firstly, HFS is transformed into a Manufacturing Petri Net (MPN) model [1], whose Markov state mechanisms are established. Secondly, the MPN simulation production process is repeatedly extrapolated and rewarded by a new mechanism, which generates a substantial amount of sample data. A multi-dimensional production information matrix is used as the MPN state input for the Q Value network, which is trained using the D3QN algorithm. Upon convergence to the optimal value function, the network model can be invoked in the online matching execution mechanism for optimal scheduling. In a 30×16 case, the average maximum completion time of this algorithm can be reduced to 293.09. The experimental results demonstrate that the network model trained by the D3QN algorithm under optimal hyperparameter settings has higher performance and response speed. In this paper[2] Deep neural networks (DNNs) are a robust and versatile machine learning technique that offers a wide range of applications across various domains. They constitute effective and approximate models that can replace realistic models. However, designing, training, and refining the model poses a significant challenge to achieve the desired outcome. To address this challenge, metaheuristics are employed for DNN structure optimization to improve their accuracy and efficiency. The present work proposes a Grey Wolf Optimization (GWO) technique to find an optimal or near-optimal DNN configuration to approximate a realistic FJSSP by considering makespan as a scheduling objective function. Different DNN hyper-parameters are adjusted according to the accuracy of the solutions, including the number of neurons in each hidden layer, activation functions for both hidden and output layers, and learning rate and momentum. For analyzing and evaluating DNN performances, an investigation is carried out considering the minimization of the maximum value of Relative Errors (REs). These latter are calculated between the DNN-predicted makespan values and their real corresponding values of the data validation set. The results obtained during the application of DNN to a case study representing a Flexible Job Shop Scheduling Problem (FJSSP) demonstrate the effectiveness and high precision of the model. The case study results confirm that the model can effectively handle FJSSP. This paper[3] presents a new generalization of the uniform dual-resource constrained flexible job shop scheduling problem (DRCFJSP) for automated flexible manufacturing systems, considering two multi-time constraints of setup time and transportation time. Design a mixed integer nonlinear programming model with multiple optimization objectives of minimizing maximum completion time, minimizing transportation time, and minimizing setup time. A hybrid multi-objective evolutionary algorithm (HMOEA) is proposed to solve the problem. Through simulation experiments on different data sets and algorithm comparisons, the effectiveness and feasibility of the proposed algorithm are verified. Finally, we designed thirteen processing cases under different conditions compared them with four different comparison algorithms and concluded that the HMOEA algorithm is superior in solving the DRCFJSP problem, which is of great significance to the use of algorithms for optimal design in actual production.In this digital era paper[4], the technology development is rapidly growing. The growth of digital technology also impacts the development of people’s lifestyles in daily life. One thing affected is the lifestyle of shopping and selling. The COVID-19 pandemic has caused all people to isolate at home for about 2 years. This is also very influential with a drastic change in lifestyle activities must be conducted online at home. As a result, the process of doing business and shopping is also conducted at home making online shopping grow rapidly. The development of online shopping is also a factor that drives sellers to do live-streaming shopping. This feature can help sellers in promoting products and helping customers to be able to shop more practically and interestingly. According to the development of live streaming shopping, it can be said that this feature is popular and demanded by many people in selling and shopping, especially on digital platforms because of its massive influence on both sellers and customers. This study aims to determine customer’s purchase intention in using live-streaming shopping. The research examines the relationship between Product Information Demonstration, Hedonic Value, Seller Reputation, Perceived Usefulness, Seller Interaction, and Product Promotion, on Purchase Intention. With 130 data from respondents collected from social media using Structural Equation Modeling (SEM) for processing data. In this study, we will see how the factors affect customer purchase intention in using live-streaming shopping.In this paper[5] study aims to determine the effect of using Mobile Food Ordering applications and Service Quality on Customer Loyalty through Customer Satisfaction at local coffee shops in Jakarta. The research is quantitative and correlational. Data collection was carried out through online questionnaires with 144 samples. Specifically, the sample consisted of 62.5% female and 37.5% male respondents, with the majority aged between 21-50 years old, providing a comprehensive view of the target consumer group for local coffee shops in Jakarta. This study shows that the use of Mobile Food Ordering Applications and Service Quality does not have a significant direct effect on Customer Loyalty. However, the use of Mobile Food Ordering Apps and Quality of Service shows an indirect influence on Customer Loyalty through Customer Satisfaction. The implications of this research suggest that enhancing customer satisfaction is crucial for increasing customer loyalty. This study fills the gap in understanding the indirect effects of service quality and mobile application usage on customer loyalty in the context of local coffee shops. The novelty of this research lies in its focus on the coffee shop industry in Jakarta, providing insights into how digital tools and service quality interact to influence consumer behavior. Environments.In this paper[6] it is frequently difficult to handle large-scale, high-dimensional data using traditional e-commerce data analysis methodologies, which makes it challenging to effectively capture user behavior and market trends. The capacity to precisely evaluate e-commerce data can be enhanced by the use of improved Bayesian algorithms, enabling more precise predictions of user behavior and market trends. This article gathers a lot of e-commerce data from e-commerce websites using web crawler technology, preprocesses it, and verifies its validity. It integrates normal distributions and dynamic modeling techniques based on Bayesian networks, and it uses a graph model to express the relationships between items like users, products, and transactions. The dataset was divided using 5-fold cross-validation, and the experimental findings demonstrated that the enhanced Bayesian algorithm had an average accuracy of 97.9% for predicting user behavior. The enhanced Bayesian algorithm's average user shopping cart conversion rate is 83.2%. The analytical performance of e-commerce data can be efficiently enhanced through the use of enhanced Bayesian algorithms. In this paper[7] in smart manufacturing, the job-shop scheduling problem (JSP) is a major obstacle that must be solved by the best possible sequencing of task operations. Dynamic job-shop environments require flexible scheduling systems that can adjust to changing conditions due to unpredictabilities like machine breakdowns. Traditional methods, which only provide the best answers when they are put into practice, are not adaptable enough to take into account shifting circumstances. Because of this limitation, temporal complexity has increased, highlighting the importance of sophisticated, flexible scheduling techniques in smart manufacturing. Several metaheuristic techniques, such as the well-known Ant Colony Optimization (ACO), are inspired by natural phenomena and are remarkably successful and efficient at solving extremely difficult (NP-hard) combinatorial optimization problems. This paper presents the implementation of an Ant Colony Optimization with Kalman Filter (ACO\_KF) model algorithm applied to solve the JSP. ACO\_KF is a combination of the recursive estimating algorithm for dynamic systems with the metaheuristic optimization algorithm inspired by ant foraging behavior to solve optimization problems. Our proposed approach aims to implement an ACO algorithm for solving a JSP and optimizing the makespan time by adjusting pheromone levels on paths. Also, the algorithm incorporates a Kalman filter to adaptively adjust pheromone levels according to recorded makespan times, to improve the convergence and efficiency of the ACO algorithm. Comparing the quality of the solutions to the most well-known outcomes from the most successful methods was necessary to evaluate the algorithm's performance on reference JSP. The solutions were obtained with remarkable efficiency and excellent quality.In this paper[8]Dorsaf Aribi 1; Olfa Belkahla Driss 2 and Hind El Haouzi 3.A massive number of studies have tackled the scheduling problem, but they mainly seek to solve the classic problem by reducing the real constraints of the environment like workers’ fatigue, which may lead to defective production, and the occurrence of unexpected events that make the initial scheduling obsolete. In this paper, we study the multi-objective dynamic flexible job shop-scheduling problem under workers’ fatigue constraints (DFJSP-WF) through three unexpected events: job insertion, machine breakdown, and job cancellation. First, a multi-objective model is established with objectives to minimize makespan and total weighted tardiness, earliness, and rejected parts due to workers’ errors, which depend on workers’ fatigue. Second, to deal with this model, a non-dominated sorting genetic algorithm II (NSGA-II) is adapted. Computational results are presented using three sets of well-known benchmark literature instances. In this paper[9] "Smart Shopping Cart Using Radio Frequency Identification" project revolutionizes the shopping experience through RFID technology, offering real-time cost updates to shoppers. This innovative system incorporates RFID readers, a microcontroller for instant cost computations, and an LCD screen for displaying the total expense. By tackling common issues associated with traditional shopping carts, like prolonged checkout procedures and the absence of live cost monitoring, our project significantly enhances the efficiency and convenience of shopping. Shoppers can now enjoy a seamless experience as the RFID technology effortlessly scans tagged products, allowing the microcontroller to promptly calculate the overall cost, which is then displayed on the LCD screen. This breakthrough not only streamlines the shopping process but also eliminates the need for manual price entries, reducing errors and enhancing overall customer satisfaction. The successful integration of these technologies marks a promising step toward the future of retail, where smart shopping carts pave the way for quicker, more accurate, and enjoyable shopping ventures. In this paper[10] A manufacturing company's production-related decision-making is to a large extent characterized by machine scheduling and support device operations management. All these industrial equipment types consume energy, often in the form of electricity. This electricity is more and more provided by renewable energy sources such as wind and solar power. The volatility of these power sources can lead to peak periods where feed-in management is required to stabilize a power grid. In this paper, we suggest increasing local industrial energy consumption in such periods to relieve the power grid. For this, we use models that are capable of synchronizing machine scheduling activities and supporting device charging operations with the availability of renewable energy. We then use a decentralized decision-making platform to coordinate the decision-making of various types of production-related equipment. By integrating a forecast for the occurrence of excessive renewable energy into this coordination platform, an opportunity is given to support environmentally oriented production decisions that avoid feed-in management actions in the power grid that surrounds the company. In a simulation study based on real-world data, we compare flow shop and job shop production environments, both under stochastically arriving jobs in such an energy-oriented setting. We furthermore introduce and examine the impact of machine-specific due date adjustment methods to achieve high processing rates and low job tardiness next to the energy-related goal. The presented approach is computationally analyzed concerning the trade-offs of these conflicting goals in both types of production environments.

METHODOLOGY :

Creating a methodology for analyzing or creating a shopping dataset involves several key steps, whether the goal is to gather, clean, and analyze the data or to use it for machine learning models. Below is a general methodology you can follow for a shopping dataset, which is typically structured to track consumer behavior, products, purchases, and other related aspects.

***DATA COLLECTION*** :

This raw shopping data is collected from the Kaggle platform.

***DATA PREPROCESSING*** *:*

Once data is collected, it must be cleaned and structured:

*Standardization and Normalization*: Scaling numerical data for consistency, is especially important when using machine learning algorithms.

In the context of data preprocessing, normalization is a crucial technique used to scale the numerical features of a dataset, ensuring that they are within a comparable range. One commonly employed method is Min-Max normalization, which scales the data to a specific range, typically between 0 and 1. This transformation is carried out by subtracting the minimum value of a feature from each data point and then dividing by the range (the difference between the maximum and minimum values). The formula for this normalization is:

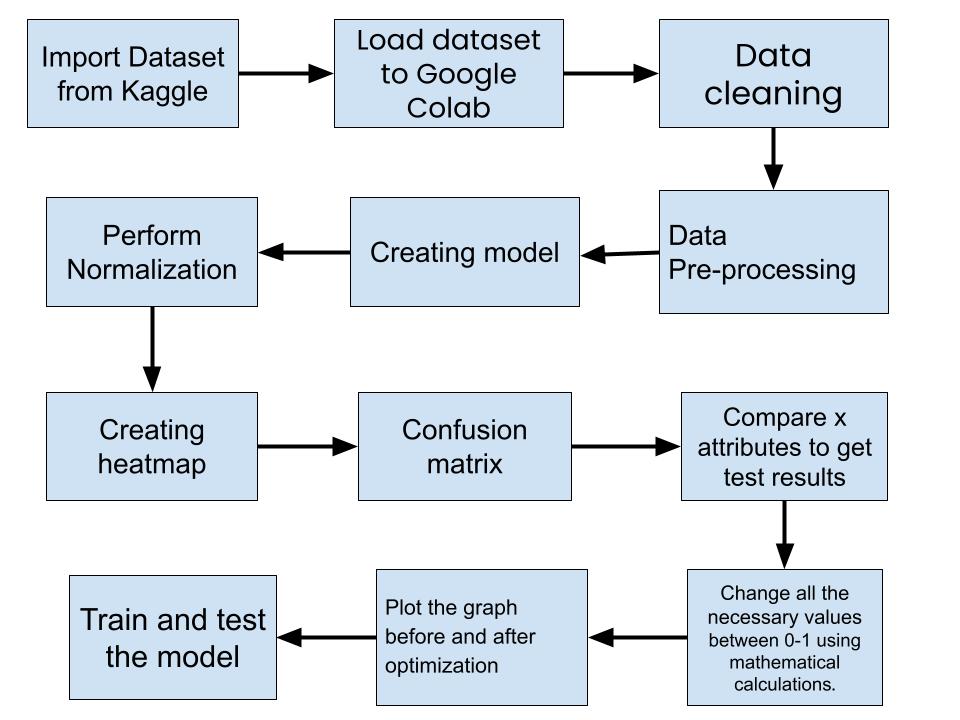
X′=X−Xmin/Xmax−Xmin

Where X is the normalized value, X is the original value, and Xmin ​ and Xmax​ are the minimum and maximum values of the feature, respectively.

Another essential technique used during the data analysis phase is correlation analysis, which is utilized to determine the relationships between the various numerical variables within the dataset. This is often visualized using a correlation heatmap, a graphical representation that illustrates the strength and direction of relationships between variables. In a correlation matrix, the values range from -1 (perfect negative correlation) to +1 (perfect positive correlation), with 0 indicating no linear relationship. The heatmap provides an intuitive way to visualize these correlations, aiding in the identification of highly correlated features, which can help in feature selection and model optimization.

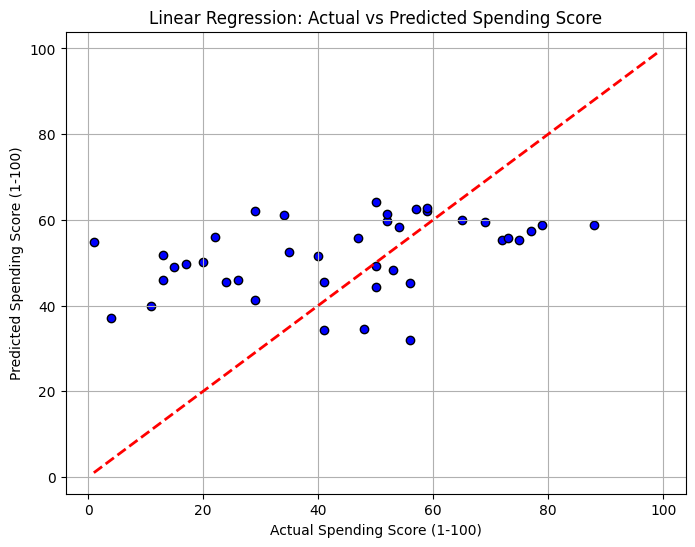
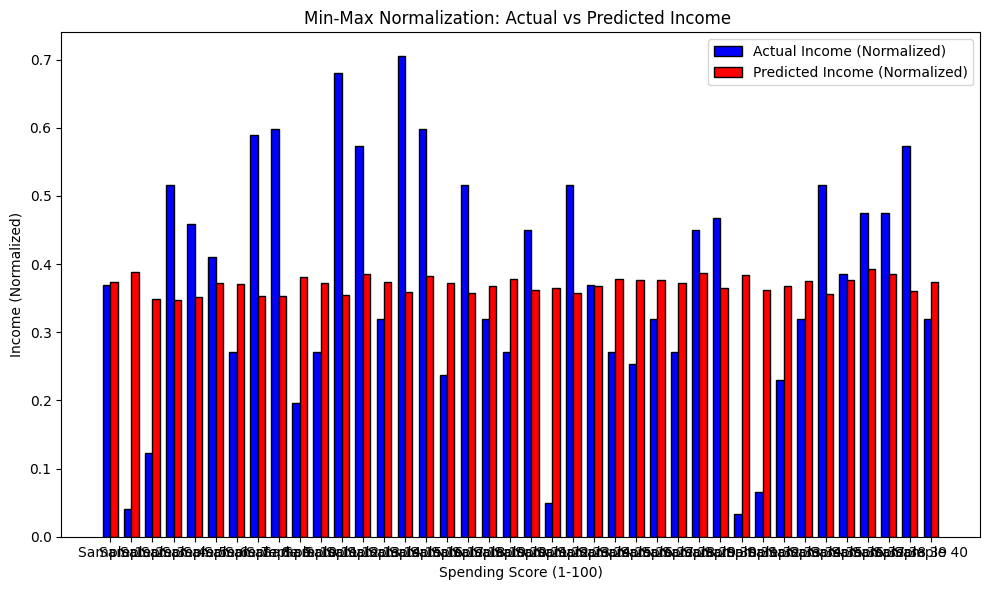
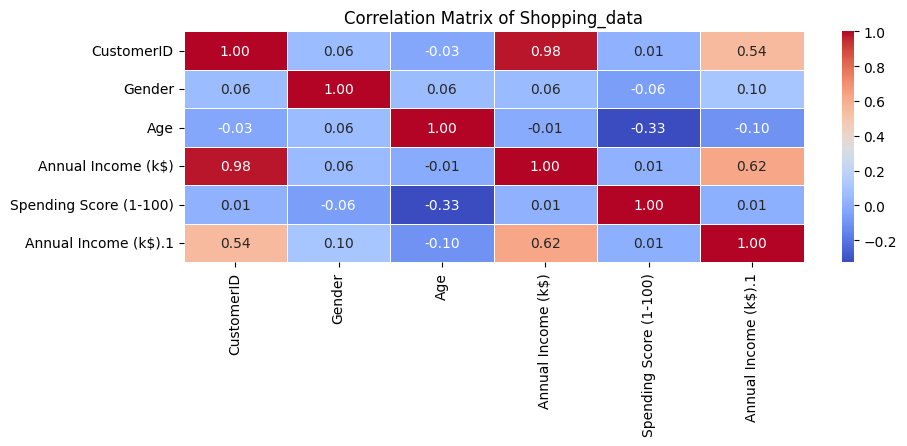
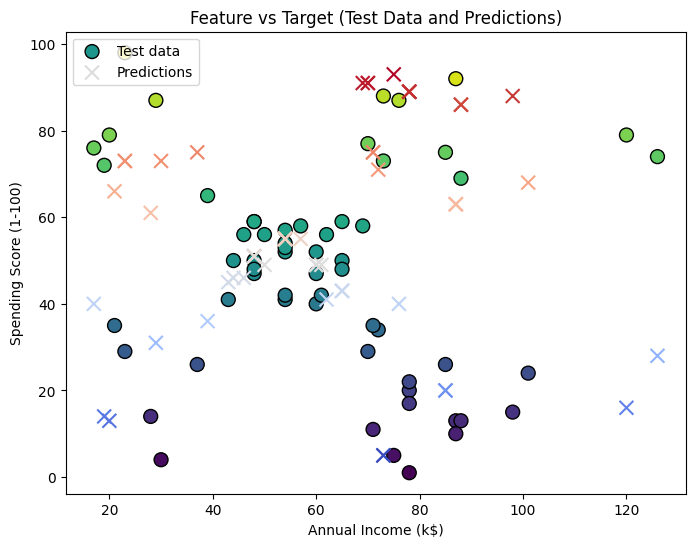
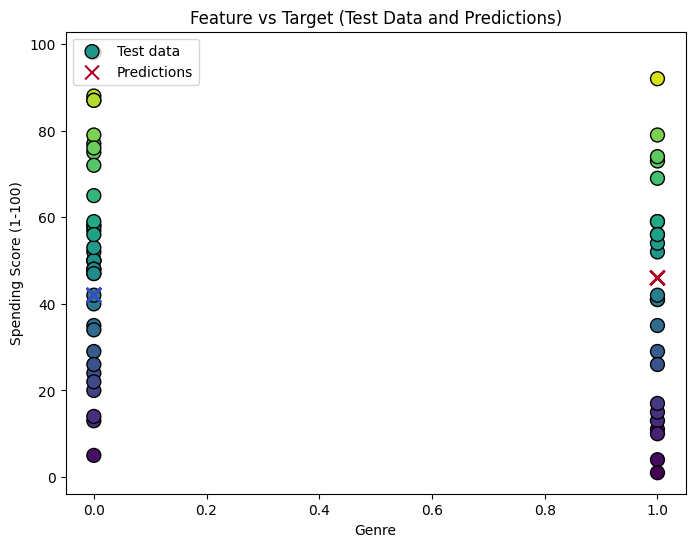
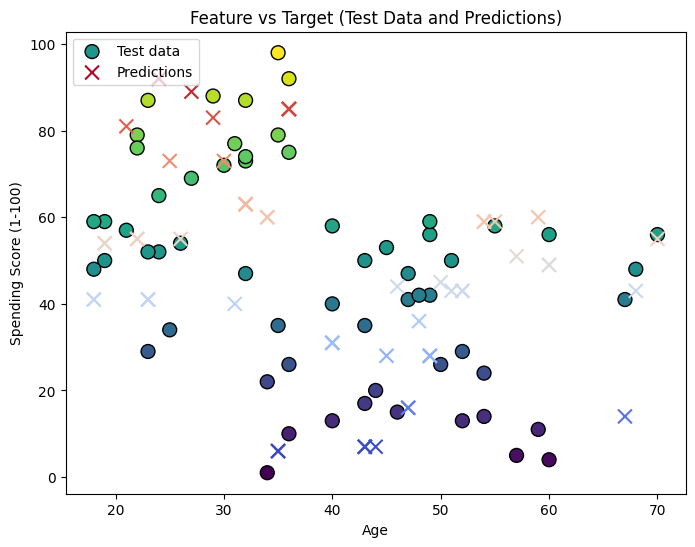
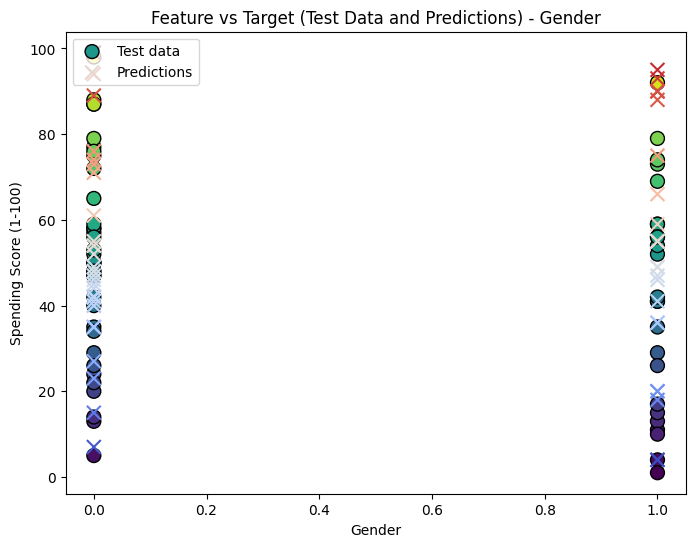
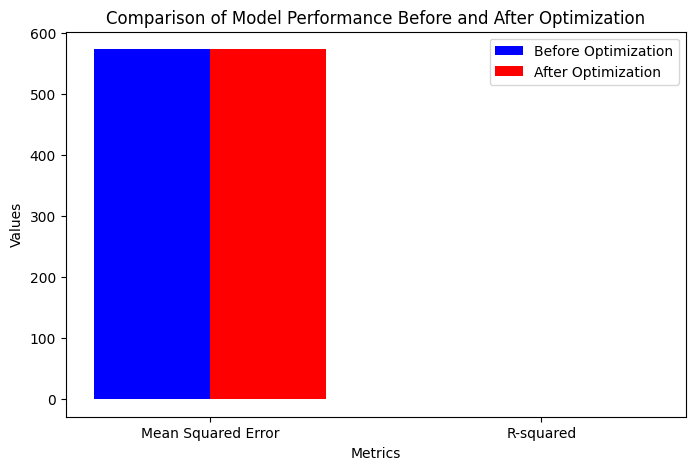
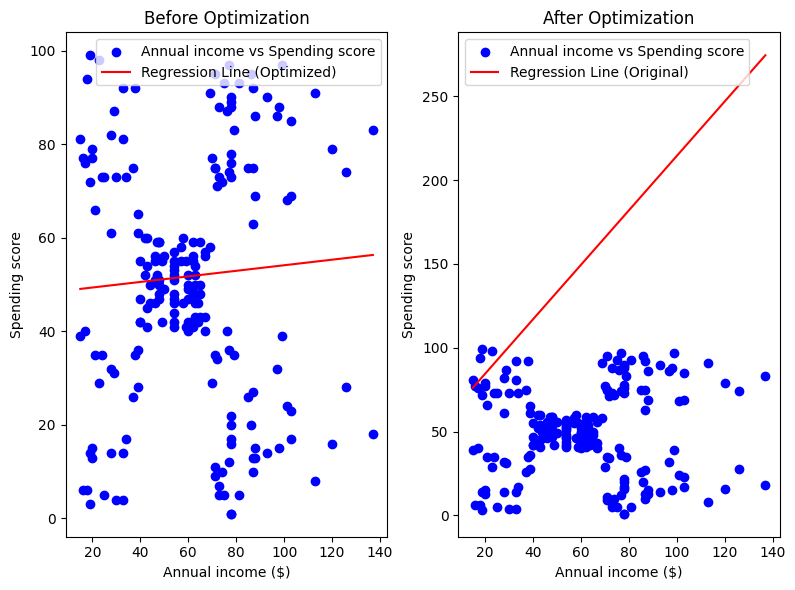
By applying Min-Max normalization and generating a correlation heatmap, we ensure that the dataset is appropriately scaled for analysis, and we gain valuable insights into the interrelationships between variables, facilitating more accurate modeling and interpretation.

*Categorical Encoding****:*** Converting categorical features like Male and Female into numerical representations.



REGRESSION ANALYSIS :

ALGORITHM :

* Step 1: Download the dataset from Kaggle.
* Step 2: Import the dataset from Google Drive to Google Colab.
* Step 3: Load the dataset into google colab.
* Step 4: Use some commands to perform various operations (eg: tail, ps.dropna, ps. size, ps.info, ps. describe, etc…)
* Step 5: Write a linear regression code based on a shopping data set using Matplotlib.
* Step 6: Write another linear graph using train\_test\_split to see the Actual and Predicted Spending Score.
* Step 7: Perform Min-Max Normalization to check Actual and Predicted Annual Income.
* Step 8: Change the character values into numerical values and create one CSV file of it.
* Step 9: Now write the code to get a correlation matrix and heatmap of the new CSV file.
* Step 10: Write a code to plot a graph of the Annual Income and Spending score of the customer using a confusion matrix.
* Step 11: Write a code to find the Spending Score based on the Customer’s Gender.
* Step 12: Write a code to find the Spending Score based on Customer’s Age.
* Step 12: Write a code to find the Spending Score based on all the dependent attributes like Customer’s Age, Gender, and Annual Income. 
* Step 13: Now change the annual income values between 0-1 using mathematical calculations .
* Step 14: Write the code for before and after optimization of the Spending Score for shopping.
* Step 15: Training and Testing the model helps us to get real-world performance.

PERFORMANCE EVALUTATION :

Evaluating the performance of a shopping dataset can be done from various perspectives, depending on the system's goals (e.g., recommendation accuracy, transaction efficiency, user engagement, etc.). The performance evaluation helps determine the effectiveness of the system and identify areas for improvement.

If the dataset is being used to power a recommendation engine, the primary performance metrics are:

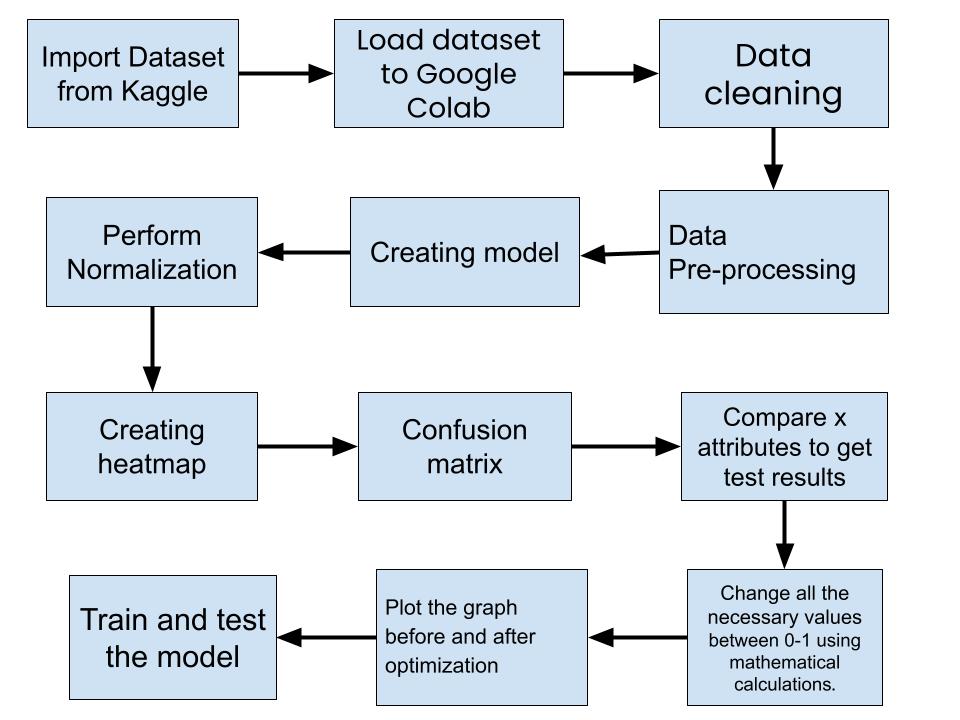
* **Accuracy Metrics:**
  + **Precision**: Measures the proportion of relevant items recommended to the user. Precision=True Positives/True Positives+False Positives.
  + **Root Mean Squared Error (RMSE)**: Measures the square root of the average squared differences between predicted and actual values.
  + **Correctness of User Information**: User profile data (e.g., Age, Customer ID, Gender, Annual Income, Spending score) must be accurate and up to 60% .

Based on my Shopping\_data set, Compared to the other age groups, the age group of people between 28 to 33is observed they be more interested in shopping. Irrespective of their annual income, their spending score is high compared to others.

To compare each person's annual income and spending score for the highest shopping rate, I would want to calculate a shopping rate for each individual based on their Annual Income, Age, and Gender*,* and then identify who has the highest shopping rate.

Females spending scores are higher when compared to Men. The Spending score percentage of Females is 56% whereas Men have 44%.

EXPERIMENT ILLUSTRATION :



RESULTS AND DISCUSSION:

The analysis of the Customer Shopping Preferences Dataset has provided valuable insights into consumer behavior and purchasing patterns, shedding light on the complex relationships between key attributes such as age, gender, annual income, and spending score. Through various analytical techniques including data preprocessing, normalization, correlation analysis, and regression modeling, this study has explored the factors that influence customer spending behavior.

The findings suggest that while there are predictive relationships between demographic factors and spending scores, the accuracy of these predictions remains moderate, indicating that additional features or more sophisticated models may be needed to enhance performance. Despite the moderate accuracy, the dataset proves to be a powerful tool for understanding customer preferences and can be effectively used to develop targeted marketing strategies, improve customer segmentation, and optimize product recommendations.

Moreover, the research reveals significant trends such as the higher spending tendencies of women compared to men and the notable shopping interest observed in the 28-33 age group, irrespective of income levels. This demographic insight is particularly useful for businesses looking to tailor their marketing efforts toward high-potential consumer segments.

Overall, the study underscores the importance of leveraging data-driven insights for more personalized, customer-centric approaches in the retail sector. By continuously refining models and incorporating additional features, businesses can further improve their strategies, enhance customer engagement, and drive long-term growth. The findings of this research highlight the potential of machine learning and data analytics in shaping the future of retail, offering businesses a competitive edge in an increasingly dynamic market.

CONCLUSION:

In conclusion, the analysis of the Customer Shopping Preferences Dataset reveals key insights into consumer behavior, highlighting demographic factors' influence on spending patterns. While predictive accuracy can be improved, these findings offer valuable guidance for businesses to refine marketing strategies, enhance customer segmentation, and drive growth through data-driven decision-making.

REFERENCES :

I took a shopping dataset from Kaggle and I preferred using the below pages to complete my project.

Wang, M., Liang, K., Xiao, D., & Hua, S. (2024). Data-driven real-time framework for hybrid flow shop scheduling based on dueling double deep Q network. *2024 5th International Seminar on Artificial Intelligence, Networking and Information Technology (AINIT)*, 1029-1033. <https://doi.org/10.1109/ainit61980.2024.10581542>

Mihoubi, B., Gaham, M., & Kouider, A. (2024). Grey wolf optimization-based deep neural network for a flexible job shop scheduling problem. *2024 2nd International Conference on Electrical Engineering and Automatic Control (ICEEAC)*, 1-5. <https://doi.org/10.1109/iceeac61226.2024.10576359>

Yang, L., Lv, Z., Du, B., & Guo, J. (2024). Multi-time constrained dual-resource flexible job shop scheduling based on multi-objective evolutionary algorithm. *2024 4th International Conference on Machine Learning and Intelligent Systems Engineering (MLISE)*, 12-15. <https://doi.org/10.1109/mlise62164.2024.10674321>

Jingga, F., & Santoso, J. R. (2024). Factors affecting the intention to buy customers in live streaming shopping. *2024 28th International Conference on Information Technology (IT)*, 1-5. <https://doi.org/10.1109/it61232.2024.10475762>

Sanny, L., Chans, M., & Andevta, R. (2024). Impact of mobile food ordering applications on consumer loyalty in Indonesian local coffee shops: A technological perspective. *2024 3rd International Conference on Creative Communication and Innovative Technology (ICCIT)*, 1-5. <https://doi.org/10.1109/iccit62134.2024.10701167>

Wang, L. (2024). Design of intelligent analysis method for e-commerce data based on improved Bayesian algorithm. *2024 Third International Conference on Distributed Computing and Electrical Circuits and Electronics (ICDCECE)*, 1-4. https://doi.org/10.1109/icdcece60827.2024.10548504

Kriouich, M., & Sarir, H. (2024). Integration of Kalman filter with ant colony optimization for solving the job shop scheduling problem. <https://doi.org/10.21203/rs.3.rs-4442314/v1>

Aribi, D., Belkahla Driss, O., & El Haouzi, H. (2023). Multi-objective optimization of the dynamic and flexible job shop scheduling problem under worker's fatigue constraints. *Proceedings of the 15th International Conference on Agents and Artificial Intelligence*. <https://doi.org/10.5220/0011663800003393>

Kesava, G. R., Reddy, V. K., Trivikram, S., & K, S. S. (2024). Smart shopping cart using radio frequency identification. *2024 IEEE International Students' Conference on Electrical, Electronics and Computer Science (SCEECS)*, 1-5. <https://doi.org/10.1109/sceecs61402.2024.10482043>

Scholz, S., & Meisel, F. (2023). Energy-aware coordination of manufacturing equipment in flow shop and job shop production environments with stochastic job arrival. <https://doi.org/10.2139/ssrn.4432290>