International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 07 Issue: 06 | June - 2023 SJIF Rating: 8.176 ISSN: 2582-3930

GROCERY DATA ANALYSIS USING MACHINE LEARNING

Prof. Soni Kendre[1]

Pratik Umap[2], Shubham Salunkhe[3], Atharv Khurase[4], Yash Lakhotiya[5]

Computer Engineering Department, ISBM College of Engineering Nande, Pune-412115, India Savitribai Phule Pune University

Abstract: **Grocery retailers generate vast amounts of data from various sources, including sales transactions, customer loyalty programs, inventory management systems, and online platforms. This abundance of data presents a valuable opportunity to extract actionable insights and optimize business operations. In recent years, machine learning techniques have gained significant attention for their ability to analyzed large and complex datasets. This abstract provides an overview of the application of machine learning in grocery data analysis and highlights its benefits and challenges. The objective of grocery data analysis using machine learning is to uncover patterns, trends, and relationships within the data that can drive decision-making, enhance operational efficiency, and improve customer experience. We also implemented a voice-based searching feature that allows users to search for products using voice commands. The results of this project were presented using Power BI in real-time, enabling users to view changes. ML model predicted the user's next order items that have high probability user can order.**

Key Words: Grocery Data Analysis, Machine Learning, Power BI, Voice-based searching

I. INTRODUCTION

In the modern era of retail, the grocery industry has witnessed a massive influx of data from various sources such as point-of-sale systems, e-commerce platforms, customer loyalty programs, and social media. This data deluge provides a rich opportunity for grocery retailers to harness the power of machine learning techniques to extract actionable insights, enhance operational efficiency, and gain a competitive advantage in the market. Grocery data analysis using machine learning refers to the application of advanced algorithms and statistical models to analyze and interpret large volumes of grocery-related data. Machine learning algorithms can learn from historical data patterns, make predictions, identify trends, and uncover hidden relationships that may not be apparent through traditional analytical approaches. This enables retailers to make data-driven decisions, optimize processes, and deliver a personalized shopping experience to their customers.

The prime goal of this research paper is to analyze grocery data and determines the insight by applying advanced machine learning techniques and increase the user experience.

II. PROBLEM DEFINITION

Ecommerce is a very rich problem. A lot of rich insights and ideas can be extracted. Also, new contributions can be added.

The main objective of the project is to make Grocery retailers to understand the current customer's behavior and to predict future customers' purchasing products. Leveraging customer transaction data can help in understanding customers' purchasing behavior, offering right bundles and promotions, assortment planning and inventory management to retain customers, improve sales and extend their relationship with customers.

The objective is to extract meaningful insights from the available grocery data to assist in decision-making, improve operational efficiency, and enhance the overall customer experience.

Machine learning is an effective tool that can be used to examine data on grocery data and spot trends that are difficult to spot with the unaided eye.

III. LITERATURE SURVEY

The paper is written by Shruthi Gurudath she wrote the model paper of Market Basket Analysis & Recommendation System Using Association Rules Market Basket Analysis is a key method known and utilized by substantial retailers to reveal relationships between products, like bread, butter, etc. It works by searching for a mix of products that happen together every now and then in exchange. To give it another perspective, it enables retailers to recognize connections between things that individuals purchase. With the continuous growth of information technology, massive amounts of data are collected and stored by enterprises. It is very important for enterprises to transform this data into useful information and knowledge for decision making in dynamic markets. This value-added information discovered from Market Basket Analysis can be used to support decision making.

The paper is written by Shuvechchha Kunwar she wrote the model paper of K-Means Clustering for Instacart Recommendations This paper is inspired by the extensive use of Recommendation Systems in this digital era. It draws concepts from Machine Learning and Data Science to develop a recommendation model employing Instacart's User Dataset. It aims to utilize the concept of collaborative filtering which predicts relevant products based on the behavior patterns of similar users. K-Means Clustering is used to split customers into distinct groups depending on their attributes. The predictions are made for each cluster of users based on the cluster's collective purchase pattern.

The paper is written by Lingling Zhang he wrote the model paper of purchasing behavior analysis using binary classification over recent decades, the retail market of organic products has constantly been growing worldwide in response to the evolving consumers' demand for food quality, freshness, environment concern and health. However, despite high-profit margins, the growth of organic business is hampered by fragile supply chain, high operating cost and poor alignment between retailer supply and consumers' demand. Thus, it is important to have a better understanding of organic purchasing behavior of consumers and further optimize the organic distribution system accordingly. This research analyzed the purchasing behaviors features of consumers ordered.

IV. ASSUMPTIONS AND DEPENDENCIES

Here are some of the assumptions and dependencies that we have to consider while creating an analysis model using machine learning:

Sufficient Historical Data Availability: It is assumed that an adequate amount of historical transactional data, including customer purchase history and product details, is available for training the ML model. Sufficient data is necessary to capture user behavior patterns and generate accurate recommendations.

Relevance of Historical Data: It is assumed that past user behavior is a relevant indicator of future behavior. The ML model assumes that user preferences and purchase patterns remain consistent over time and can be used to predict future purchasing decisions.

User Item Interaction: The ML model assumes that user-item interactions, such as purchases or views, reflect the user's interest and preference for specific products. It assumes that these interactions can be used to infer the user's preferences and make accurate recommendations.

Data Accuracy and Quality: The accuracy and quality of the collected data are assumed to be reliable. It is assumed that the data has been properly cleaned, normalized, and preprocessed to ensure the ML model's effectiveness and accuracy.

Evaluation data: The evaluation data should be separate from the training data. This data is used to evaluate the accuracy of the model.

Deployment: The model should be deployed in a way that is accessible to users. The model should be easy to use and interpret.

V. Design

A data-flow diagram is a visual representation of how data moves through a process. Using a flowchart, specific operations based on the data can be depicted. Dataflow diagrams can be displayed using a variety of notations. A process must contain at least one of the endpoints (source and/or destination) for each data flow. Another data-flow diagram that divides a process into sub-processes can be used to represent a process in more detail. The structured analysis modeling tools include the dataflow diagram.

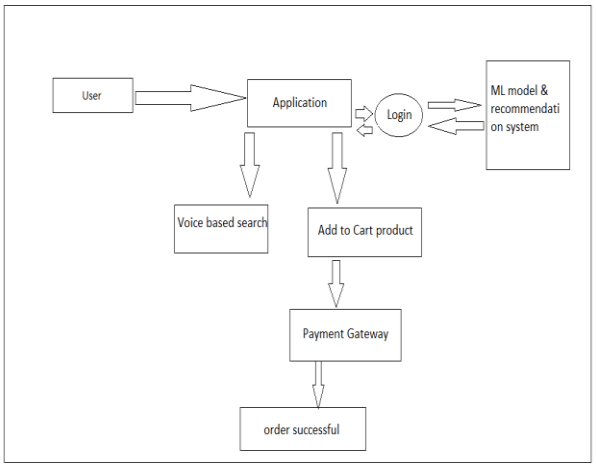

Figure 1: System model

Data is collected from a variety of sources, The data is cleaned, pre-processed, and engineered to create features that are used to train a machine-learning model. The model is evaluated and deployed so that users can use it

VI. PROPOSED MODEL

PLANNING AND ANALYSIS:
We have read various Grocery analysis reports and documents and found that product, gender, seasons, and most order product are some of the basic factors that are involved in grocery data analysis. These factors will help us to build an accurate model that can help to predict grocery products and prevent them from happening. product, gender, seasons, and most order product are some of the basic factors that are involved in grocery data analysis. By taking these factors into account, we can build an accurate model that can help to predict grocery product that can recommend to individual customers.

DATA GATHERING:
We will use a variety of sources to gather data for machine learning, including Kaggle. Kaggle is a great resource for finding high-quality data sets, and it makes it easy to import data into our model. We have found a data set on Kaggle that contains 200 thousand customers' records. This data set includes a variety of factors that led to the recommendation This data set is high quality and well-curated, and we believe that it will be very helpful in building an accurate model that can help to predict item set for recommendation.

DATA PRE-PROCESSING:
A key stage of the data mining process is data processing. It is the process of preparing raw data for analysis by cleaning, formatting, and other changes. The reliability, uniformity, and completeness of the data can be improved via data processing. This may result in simpler and more accurate data analysis outcomes. The data set we have gathered lacks proper formatting and has missing values. This indicates that before using the data for analysis, processing is required. Among the things we must accomplish are the following: Data cleaning involves removing errors and differences from the data. Putting the data into a format that is compatible with our data mining tools is known as formatting the data.

DATA TRANSFOMATION:
This entails putting the data in a format that is better suited for analysis. We will have a tidy, formatted, and transformed data set that is suitable for analysis once we have completed the data pre-processing. As a result, our results from data analysis will be simpler and more accurate

DATA VISUALIZATION:
The visual representation of data is known as data visualization. Data patterns and trends that would be challenging to spot without them are now visible to us. Finding hidden patterns in data sets with the aid of data visualization can aid in the selection of features. For example, let's say we have a data set of grocery. We can use data visualization to see how the different features of the grocery are related to each other. For example, we can see if there is a relationship between the age of the customer and the other data. Once we have found hidden insights in the data set, we can use this information to select features that are important for predicting grocery recommendation. For example, if we find that the most order product is a significant predictor of grocery data, we can include this feature in our model. Finding hidden insights in data sets can be done with the help of the potent tool known as data visualization. We can choose features that are crucial for estimating grocery severity by using data visualization.
The histogram represents the most order product from customers.

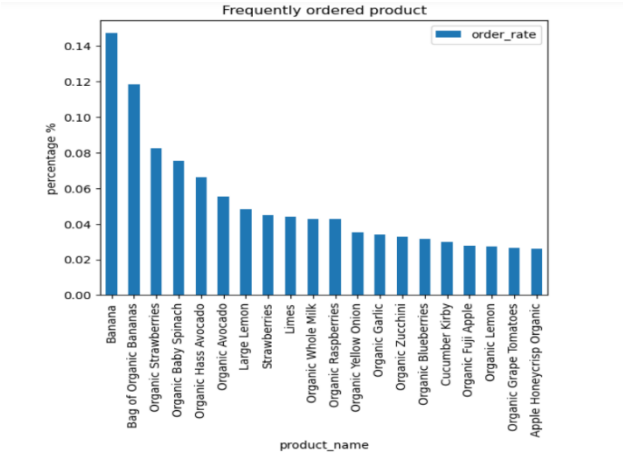

Figure 2: most order product

The histogram represents the frequently order product department.

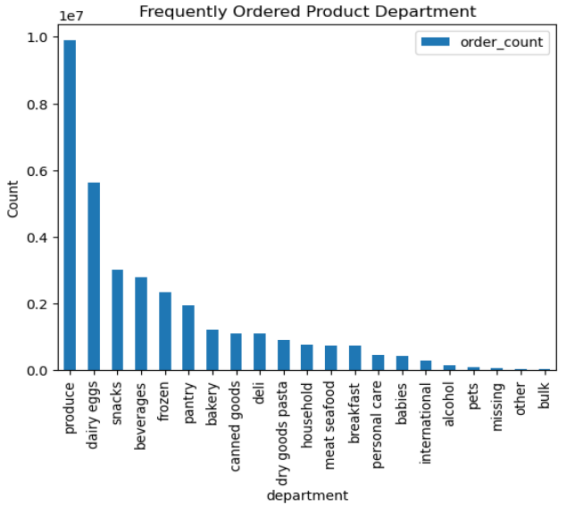

Figure 3: Frequently order product department

The pie-chart represents the Contribution of each Dept in sales

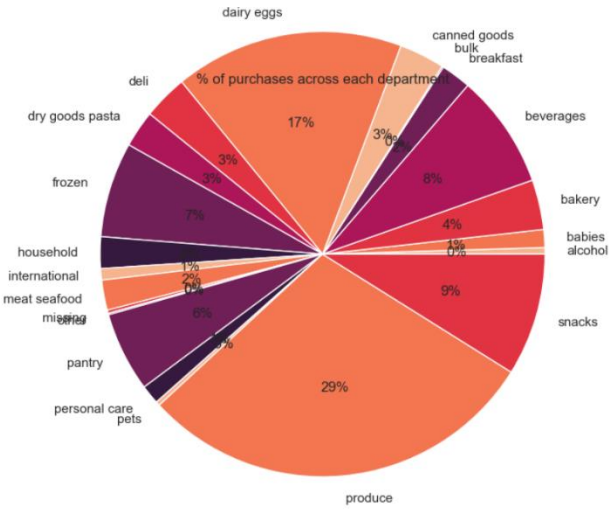

Figure 4: Contribution of each Dept in sale

FEATURE SELECTION:

I have tried with my own features and some are taken from other resources, blogs and Kaggle discussions. These features are calculated only based on prior data did not include train and test data

- User_product_ratio
- Day_of_week_reordered ratio
- Hour_of_day_reordered ratio
- Day_since_prior_order_reordered_ratio
- Product_Day_of_week reordered ratio
- Product_hour_of_day_reordered ratio
- User_hour_of_day
- User_day_of_week
- Day since prior order for a particular product
- How many times user purchased the product

We believe that these features are the most important for predicting grocery recommendation. We will use these features to build a machine learning model that can help to predict recommendation item set.

MODEL SELECTION:
We have selected machine learning algorithm that is used for building a model. We have used a Light GBM model for prediction.
We believe that these models are the most suitable for our data set and will be able to provide us with accurate predictions. We will then tune the hyper parameters of the selected model to further improve its accuracy.

MODEL TRAINING:
We will split the data set into train and test sets. We will then train the model on the train set and evaluate the model on the test set. We will then tune the hyper parameters of the selected model to further improve its accuracy. Finally, we will save the pickle file of the model, which we will use in the front end to predict item set that is used for recommendation.

MODEL EVALUATION:
Model evaluation is the process measure the accuracy of the model. It involves using the evaluation data to measure the accuracy of the model. Many different methods can be used to evaluate a machine learning model, some of the most common methods include accuracy, precision, recall, F1 score. The best method for evaluating a machine learning model will depend on the specific application.

MODEL DEPLOYMENT:
The Flask library was used to deploy and defined routes the model. Making web applications for machine learning models is simple with the help of React Js. We used the pickle used for storing the predicted data. By entering customer's details and receiving a recommendation.

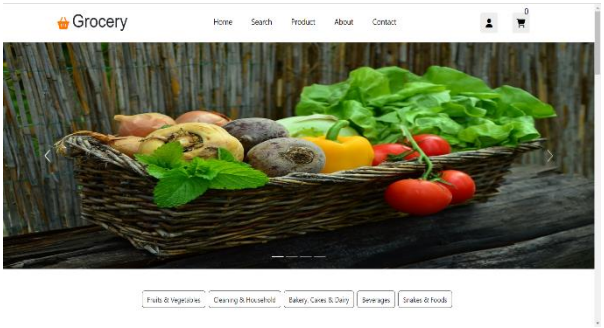

Figure: A Screenshot of The Home Page

The visual page of the application recommended products to each user based on its past history.

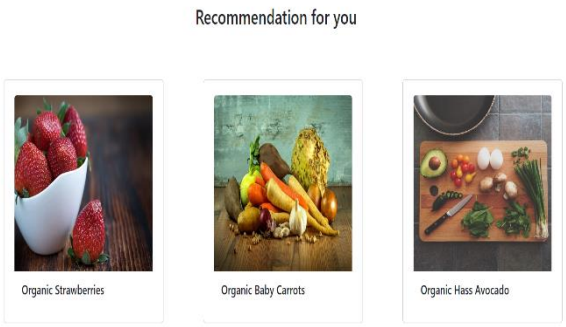

Figure: A Screenshot of Recommendation page

VOICE BASED SEARCHING PRODUCT:
In this project we have added another feature that is voice-based searching product where user can search product based on voice and typing also. we used react-speech-recognition hook that will be provided the voice-based searching properties.

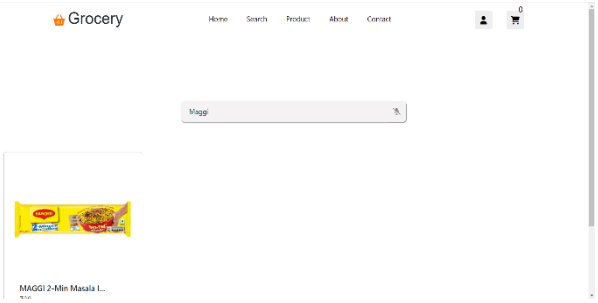

Figure: A Screenshot of searching page

REALTIME ANALYSIS USING POWER BI:

Real-time analysis with Power BI involves using the Power BI platform to analyze and visualize real-time data generated by the system. Here we build dashboard it is used for many purposes like improve business, planning marketing schema to increase sale.
In dashboard we show the multiple insight that are drawn from the data set like top selling product from each department, contribution of each department in sale etc.

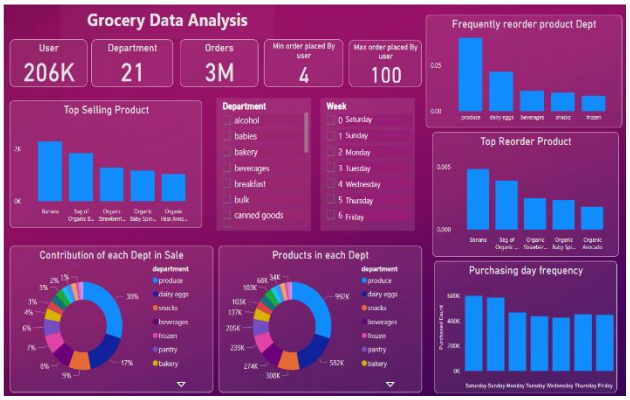

Figure: A Screenshot Power BI dashboard

The visual page of the application allows users to visualize the data. The data can be visualized in a variety of ways, including charts, graphs, and maps. The visual page allows users to explore the data and identify trends.

SOFTWARE AND HARDWARE REQUIREMENTS

Software Requirements*:*
Jupyter notebook.
VS code.
Node Js
Power BI
Any operating system (Windows/Linux).
Other required Python libraries (NumPy, Pandas, Matplotlib, etc.)

Hardware Requirements:
RAM: 8 GB (minimum requirement).
Hard Disk: 100 GB working space (minimum requirement).
Processor: Intel Core i5 or higher, or AMD Ryzen 5 or higher
CPU with at least 4 cores (8 cores or more is recommended)

VII. OTHER SPECIFICATION

A. ADVANTAGES

1. Improved User Experience: By leveraging machine learning and recommendation systems, the project provides personalized recommendations to users. This enhances the user experience by suggesting relevant products based on their preferences and purchase history, leading to increased customer satisfaction and engagement.
2. Increased Sales and Revenue: The accurate product recommendations generated by the machine learning model can lead to an increase in cross-selling and upselling opportunities. By suggesting complementary or related items, users are more likely to make additional purchases, resulting in increased sales and revenue for the business.
3. Efficient and Convenient Shopping: The voice-based searching feature adds convenience to the shopping experience. Users can easily search for products by using voice commands, eliminating the need for manual typing and navigation. This saves time and effort, making the shopping process more efficient and user-friendly.
4. Real-Time Data Analysis: The integration of Power BI for real-time data analysis provides valuable insights into various aspects of the grocery business. It enables the identification of popular products, analysis of sales patterns, and understanding customer preferences. This information can be utilized for strategic decision-making, inventory management, and marketing strategies.
5. Enhanced Operational Efficiency: By analyzing real-time data, the project allows the identification of low-demand or never-ordered products. This information can be used to optimize inventory management, reduce wastage, and streamline procurement processes. It also helps in identifying the departments or product categories that contribute most to sales, enabling the business to focus resources and efforts accordingly.

B. LIMITATIONS

1. Cold Start Problem: The ML model's effectiveness heavily relies on the availability of sufficient user data. For new users or items with limited data, accurate recommendations may be challenging to provide due to the "cold start" problem.
2. Data Sparsity: Sparse data, where users have limited purchase or interaction history, can affect the accuracy of recommendations. Additional techniques, such as incorporating auxiliary data sources or utilizing hybrid models, can mitigate this limitation.
3. Privacy and Data Security: Collecting and analyzing user data raises concerns about privacy and data security. Ensuring robust data protection measures and obtaining user consent are essential to address these concerns and

maintain user trust.

4. Voice Recognition Accuracy: The voice-based search feature's performance is contingent on the accuracy of the underlying voice recognition system. Limitations in voice recognition technology may result in errors or inaccurate search results.
5. Hardware Dependencies: The voice-based search feature relies on appropriate hardware, such as microphones or voice-enabled devices. Compatibility issues or limitations in hardware capabilities may impact the user experience.

C. CHALLENGES

Data Quality and Availability: Obtaining high-quality and relevant data for training the machine learning models can be challenging. Inaccurate or incomplete data can impact the performance of the models and the accuracy of recommendations. Additionally, ensuring the availability of real-time data for analysis can require efficient data collection and processing mechanisms.

Model Training and Optimization: Developing and fine-tuning machine learning models can be a complex task. Selecting the appropriate algorithms, optimizing hyperparameters, and addressing overfitting or underfitting are challenges that need to be tackled. Achieving the right balance between model accuracy and computational efficiency is crucial, especially when dealing with large-scale datasets.

Scalability and Performance: As the project involves real-time analysis and recommendation systems, handling a large volume of data and user requests can be demanding. Ensuring the system's scalability and performance is essential to provide a seamless user experience. Optimizing algorithms and leveraging technologies like distributed computing or cloud resources may be necessary.

Integration of Multiple Components: Integrating different components such as the machine learning model, recommendation engine, voice-based searching, and real-time analysis can be challenging. Ensuring smooth communication and synchronization between these components, handling data flow, and managing dependencies requires careful planning and development.

User Experience and Usability: Designing an intuitive and user-friendly interface is crucial for user adoption and satisfaction. Incorporating voice-based searching while maintaining a seamless and accurate user experience can pose design and implementation challenges. Balancing the system's capabilities with ease of use and ensuring a smooth transition between different functionalities are important considerations.

D . FUTURE PROSPECTS

Refining the Machine Learning Model: Continuously refining and optimizing the machine learning model used for predicting user purchases can lead to more accurate and reliable recommendations. This involves exploring advanced algorithms, incorporating more diverse data sources, and fine-tuning model parameters.

Enhancing Recommendation System: Improving the recommendation system by incorporating additional factors such as user feedback, ratings, and reviews can further personalize and enhance the recommendations. Implementing advanced recommendation algorithms like collaborative filtering or hybrid approaches can also be explored.

Advanced Voice Recognition: Investing in advanced voice recognition technologies, such as natural language processing and speech-to-text conversion, can enhance the voice-based searching feature. This includes improving the accuracy and speed of voice recognition, handling various accents and languages, and enabling more natural and intuitive voice commands.

Enhancing Data Analysis and Visualization: Continuously enhancing the real-time grocery data analysis using Power BI can provide more comprehensive insights into sales patterns, customer preferences, and inventory management. Developing interactive dashboards, incorporating advanced analytics techniques, and exploring data mining approaches can enable more robust and actionable insights

VIII. CONCLUSION

In conclusion, our project demonstrated the potential of machine learning in analysing grocery store data to uncover useful insights. The insights gained can help improve the store's performance by providing information on customer behaviour, popular products, and inventory levels. Additionally, the implementation of a voice-based searching feature adds a new dimension to the user experience, allowing customers to interact with the store in a new and innovative way. Overall, this project highlights the importance of data analysis and the potential of machine learning in the grocery store industry.

In summary, this project successfully developed a machine learning model and recommendation system for predicting user purchases and introduced a voice-based searching feature. Real-time grocery data analysis using Power BI provided valuable insights into sales patterns. The project achieved its objectives of improving user experience, enhancing recommendations, and providing data-driven insights. However, limitations exist in the accuracy of predictions and assumptions made in the

project. Overall, this project has contributed to personalized recommendations, improved user experience, and informed decision-making in the grocery industry.

IX. ACKNOWLEDGMENT

We would like to take this opportunity to thank all the people who were part of this seminar in numerous ways, people who gave unending support right from the initial stage.

In particular, we wish to thank Prof. Soni Kendre as an internal project guide who gave their co-operation timely and precious guidance without which this project would not have been a success. We thank them for reviewing the entire project with painstaking efforts and more of his, unbanning ability to spot the mistakes.

We would like to thank our H.O.D Prof. B. B. Gite for his continuous encouragement, support and guidance at each and every stage of the project.

And last but not least we would like to thank all my friends who were associated with me and helped us in preparing our project. The project named “Grocery Data Analysis Using ML” would not have been possible without the extensive support of people who were directly or indirectly involved in its successful execution.

X. REFERENCES

Julander. Basket Analysis: A New Way of Analyzing Scanner Data. International Journal of Retail and Distribution Management, V (7), pp 10-18

S. Erpolat, “Comparison of Apriori and FP-Growth Algorithms on Determination of Association Rules in Authorized Automobile Service Centres,” Anadolu Univ. J. Soc. Sci., vol. 12, no. 2, pp. 137– 146, 2012.

Beliakov, G., et al. (2019). Forecasting Sales in the Retail Industry Using Machine Learning Techniques. In Proceedings of the 12th International Conference on Soft Computing and Pattern Recognition (pp. 65-70).

Datta, S., et al. (2020). Customer Segmentation in the Grocery Retail Industry Using Machine Learning Approaches. In Proceedings of the 10th International Conference on Management and Information Technology (pp. 12-19).

Jain, A., et al. (2018). Optimization of Inventory Management in the Grocery Industry using Machine Learning. International Journal of Computer Science and Information Technology, 10(3), 63-69.

Li, Q., et al. (2019). Pricing Optimization in Grocery Retail Using Machine Learning: A Case Study. In Proceedings of the 3rd International Conference on Machine Learning and Soft Computing (pp. 123-132).

Pappu, R., et al. (2020). Machine Learning for Shelf Space Optimization in Grocery Stores. In Proceedings of the 14th International Conference on Intelligent Systems and Knowledge Engineering (pp. 187-193).