A MINI PROJECT REPORT

ON

"Business Intelligence project on retail product store analysis"

SUBMITTED TOWARDS THE FULFILMENT OF THE REQUIREMENTS OF

BACHELOR OF ENGINEERING (B. Tech.)

Academic Year: 2023-24

By:

BTCOB098 Mihir Katariya

BTCOB108 Om Kinge

BTCOB140 Saurabh Pardhi

BTCOB129 Aviraj Mane

Under The Guidance of

Dr. K. Rajeshwari

DEPARTMENT OF COMPUTER ENGINEERING,

PIMPRI CHINCHWAD COLLEGE OF ENGINEERING

SECTOR 26, NIGDI, PRADHIKARAN

PIMPRI CHINCHWAD COLLEGE OF ENGINEERING

DEPARTMENT OF COMPUTER ENGINEERING

CERTIFICATE

This is to certify that, the project entitled **"Business Intelligence project on retail product store analysis"**

is successfully carried out as a mini project successfully submitted by following students of "PCET's Pimpri Chinchwad College of Engineering, Nigdi, Pune-44**".**

Under the guidance of

Prof. Dr. K. Rajeshwari

In the fulfillment of the requirements for the B. Tech. (Computer Engineering)

BTCOB098 Mihir Katariya

BTCOB108 Om Kinge

BTCOB140 Saurabh Pardhi

BTCOB129 Aviraj Mane

Project Guide	**HoD**
Prof.Sushma Vispute	**Prof. Dr. K. Rajeswari**

Chapter 1: Introduction

1.1 Problem Statement : To Analyze and optimize the beverage brands selection and inventory management for a retail store to enhance profitability and customer satisfaction.

1.2 Project Objectives :

- Analyze retail store data to identify trends and patterns.
- Optimize sales strategies to increase revenue.
- Enhance customer experience and satisfaction through data-driven insights.
- Improve decision-making processes based on actionable intelligence.

1.3 Motivation :

Retail store analysis for beverage brands is rooted in the pursuit of enhanced profitability, improved customer satisfaction, and a competitive edge in the market. By optimizing the selection of beverage brands and refining inventory management practices, the store aims to bolster its financial performance through increased sales and decreased waste from overstocked or expired products. Understanding and catering to customer preferences not only elevates satisfaction levels but also cultivates customer loyalty, resulting in sustained business success. Furthermore, by offering a well-curated array of beverage brands that align with consumer demand, the store can differentiate itself from competitors and potentially attract more patrons. Leveraging data analytics and machine learning techniques empowers data-driven decision-making, leading to more precise predictions, better pricing strategies, and targeted marketing efforts. Moreover, evaluating and optimizing supplier relationships can drive cost savings and enhance product availability, while sustainability goals are met by reducing overstocking and waste through effective inventory management. In essence, the motivation to address this problem is multifaceted, spanning financial stability, operational efficiency, customer satisfaction, and environmental responsibility, all aimed at securing the store's success in the dynamic retail beverage industry

1.4 Literature Survey/ Requirement Analysis :

[1] Exploration of the investment patterns of potential retail banking customers using two-stage cluster analysis (2021). The main goal of this research was to demonstrate the effectiveness of the two-stage clustering method to explore and identify investment patterns in potential retail banking customers. The results confirmed that the method is effective in identifying distinct groups of customers, describing their investment patterns and investment factors. The unique feature of this research compared with the previous ones is the use of a new AI-related approach to targeting "potential customers".

[2] Bamidele-Sadiq, Mojibola & Popoola, Oluwasegun & Lawal, Gold & Awodiji, Temitope. (2022). The Importance of Decision Tree Analysis on StrategicManagement Practice: This paper gives in-depth analysis of importance of decision tree in decision making. Also due to decision tree implementation retail firms in US have improved their sales as well as minimized their cost.

[3] Lingjun He, Richard A. Levine, San Diego Random Forest as a Predictive Analytics Alternative to Regression. In this paper author argue that random forest is a valuable tool for predictive analytics tasks as random forest is easy to apply, flexible, and computationally inexpensive. Random forest also handles correlated inputs, nonlinear relationships.

[4] Influence of Food and Beverage Companies on Retailer Marketing Strategies and Consumer Behavior (2020). This review finds evidence that by influencing retailer marketing strategies through TPP, manufacturers can shape consumer behavior and, ultimately, diets. The 74 studies included in this review suggest that TPP have a considerable effect on product placement, pricing, and promotion, & in turn, on a range of customer outcomes, including purchase volume, spending, and attitudes.

Chapter 2: Project Design

2.1 Hardware Requirements:

1. Computer System: We'll need a computer system capable of running Power BI Desktop and performing data analysis tasks. Ensure that it meets the minimum system requirements for Power BI.

2. Storage: We need sufficient storage space for data storage, data backups, and model outputs.

Software Requirements:

1.Power BI: Power BI Desktop is the primary software for creating interactive dashboards and visualizations. Ensure that we have the latest version installed.

2. Data Analysis Tools: Depending on our choice of machine learning models (Logistic Regression, Random Forest, Naïve Bayes, IBK, J48, etc.), we may need software packages or libraries for these models.

3. Database Management: If our data is stored in databases, we may need database management software to connect and retrieve data. Power BI allows connections to various data sources.

4. Operating System: Ensure that the operating system of our computer supports the required software and tools.

Resources:

1.Data: Our project's primary resource is the retail store data, including information on retailer, beverage brand, sales, and total revenue generated. Ensure that the data is clean, well-structured, and stored in a format compatible with Power BI.

2. Models: We'll need resources related to the machine learning models we plan to implement. This includes pre-trained models, if available, or the resources to train models using our dataset.

3. Documentation: Access to documentation for Power BI, machine learning libraries, and any other tools or software we plan to use. This will help in understanding and effectively utilizing these resources.

Requirements & Their Detailed Explanation:

Power BI Dashboard:
Requirement: We need to create an interactive dashboard in Power BI.
Explanation: The dashboard will serve as the central interface for presenting data insights and machine learning model outputs. It should be user-friendly and provide a clear visualization of retail store data, including total revenue trends.

Data Analysis Tools:
Requirement: Depending on the machine learning models chosen (Logistic Regression, Random Forest, Naïve Bayes, IBK, J48), we need the corresponding tools and libraries.
Explanation: Each machine learning model requires specific software and libraries for training, testing, and prediction. For example, we may need Python with scikit-learn for logistic regression and random forest, or Weka for J48.

Data Preprocessing:
Requirement: Tools or scripts for data preprocessing and cleaning.
Explanation: Raw data often requires cleaning, transformation, and feature engineering to make it suitable for analysis. Data preprocessing tools and scripts are needed to ensure data quality.

Data Source Integration:
Requirement: Capability to connect Power BI to the data source.
Explanation: Power BI should be able to retrieve data from the source database or files, and this capability should be set up properly.

Storage and Backup:
Requirement: Adequate storage space and a backup strategy.
Explanation: We need sufficient storage to store the dataset, model outputs, and Power BI project files. Implement a backup strategy to avoid data loss.

Operating System Compatibility:
Requirement: The chosen operating system must be compatible with Power BI and other required software.
Explanation: Ensure that the operating system can run Power BI and any additional software we plan to use without compatibility issues.

2.2 Details of Dataset

The dataset provide valuable insights into Coca-Cola's sales and profitability across different retail locations. Here's a discussion of what each of these columns might represent and the insights that could be derived from them:

Retailer and Retailer ID: These columns identify the specific retailers or stores that sell Coca-Cola products. Analyzing sales performance by retailers can help identify topperforming stores and those that may require improvement. Retailer IDs enable tracking and referencing specific stores consistently.

Invoice Date: This column records the date of each transaction, allowing for time-series analysis. It can reveal seasonal sales patterns, the impact of promotions, and long-term trends in Coca-Cola's sales.

Region, State, City: These columns provide geographical context to the sales data. Regional analysis can help identify geographical areas with strong or weak sales performance. Understanding regional variations can inform marketing and distribution strategies.

Beverage: This column indicates the specific Coca-Cola product sold in each transaction, such as Coca-Cola Classic, Diet Coke, or various other beverage variants. Analyzing which products are top-sellers and how they vary by location can guide inventory and marketing decisions.

Price per Unit: This column represents the price at which each Coca-Cola unit (e.g., bottle, can) is sold to the retailer. It can help assess pricing strategies and the impact of price changes on sales and profitability.

Units per Sold: The number of units (e.g., bottles, cans) sold in each transaction. This column helps track the volume of sales and can be used to calculate total sales.

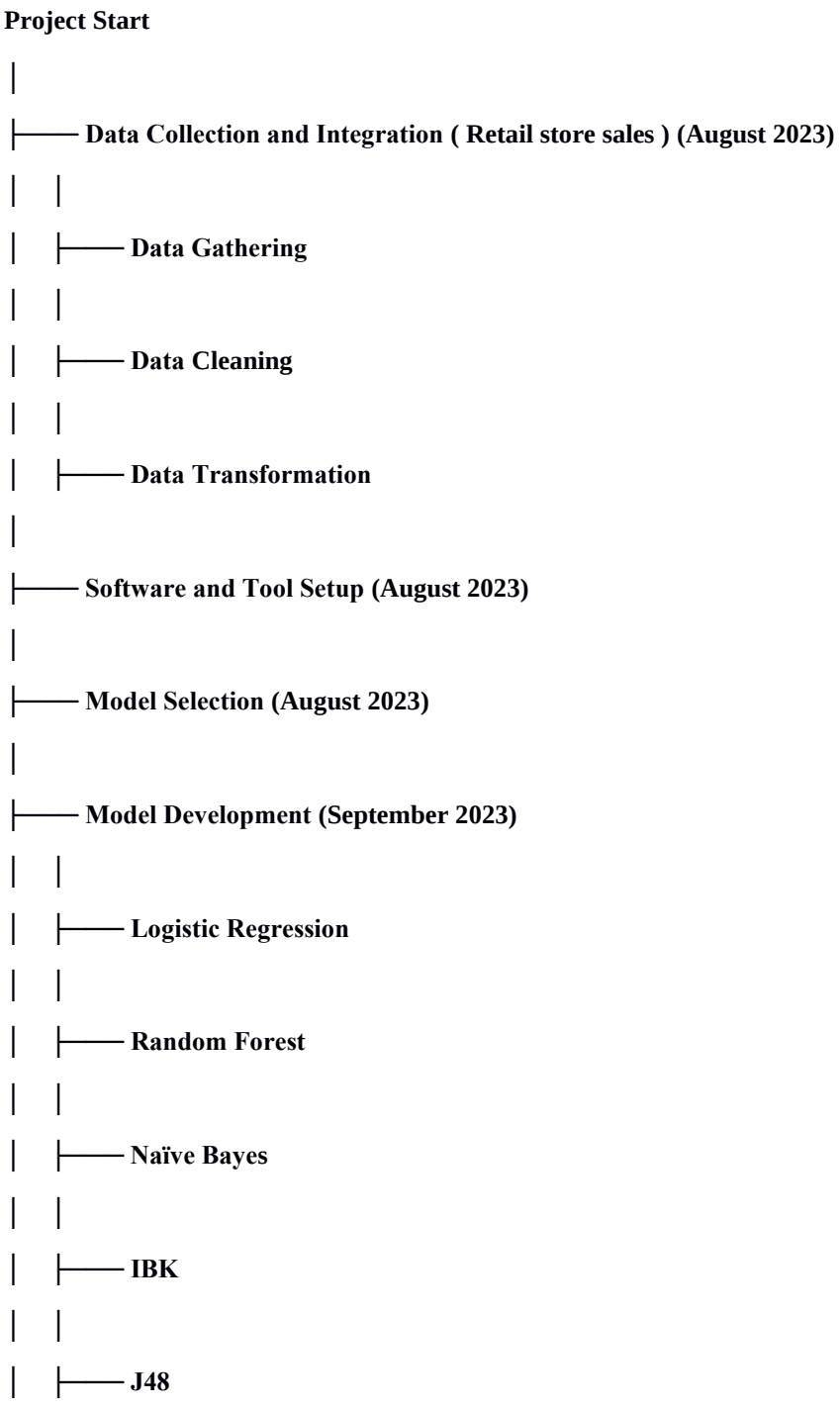
Total Sales: Total sales is the product of units sold and the price per unit. It represents the revenue generated from Coca-Cola product sales at each retailer. Analyzing total sales can reveal revenue trends and the financial performance of individual retailers.

Operating Profit: Operating profit is the profit earned by Coca-Cola from each transaction, considering factors like cost of goods sold (COGS), distribution costs, and other operational expenses. It reflects the company's profitability at the transaction level.

Operating Margin: Operating margin is the ratio of operating profit to total sales, expressed as a percentage. It provides insight into the efficiency of Coca-Cola's operations, showing how much profit the company retains from each sale after covering expenses.

This dataset is valuable for analyzing the voting patterns of lawmakers and potentially building predictive models to understand political affiliations or voting behaviors. It can be used for tasks such as classification, clustering, or trend analysis in the context of political science and data analysis.

2.3 Timeline Diagram



├—— Data Preprocessing (September 2023)

├—— Power BI Dashboard Design (September 2023)

│

├—— Data Source Integration (September 2023)

│

├—— Testing and Validation (October 2023)

│

├—— Dashboard Refinement (October 2023)

│

├—— User Testing and Feedback (October 2023)

│

├—— Final Dashboard (November 2023)

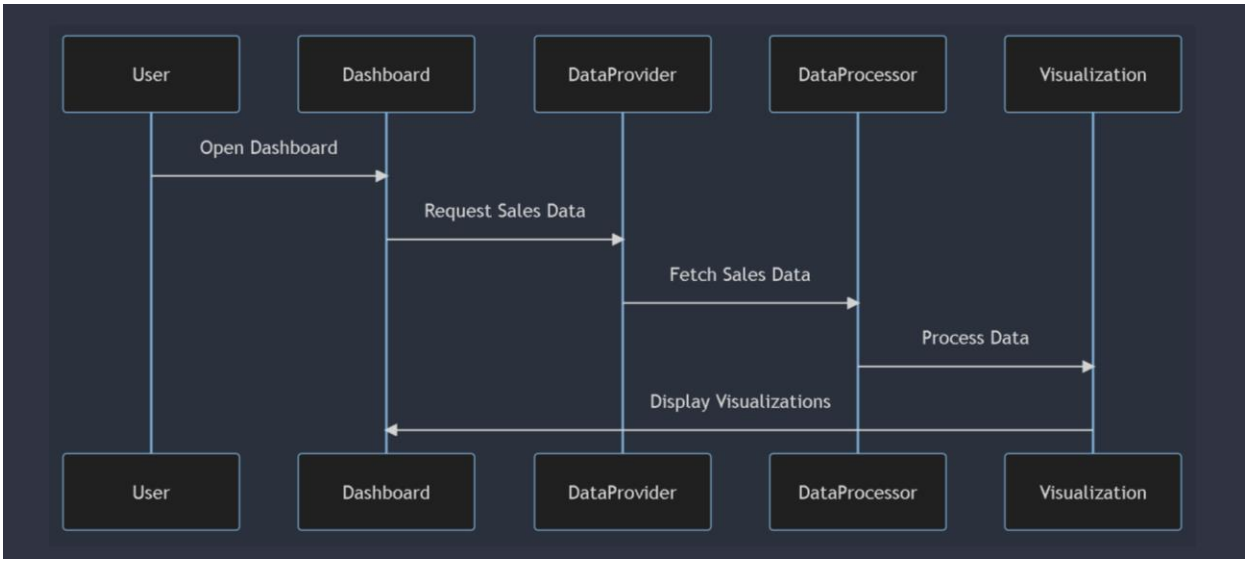
│

├—— Project Conclusion and Documentation (November 2023)

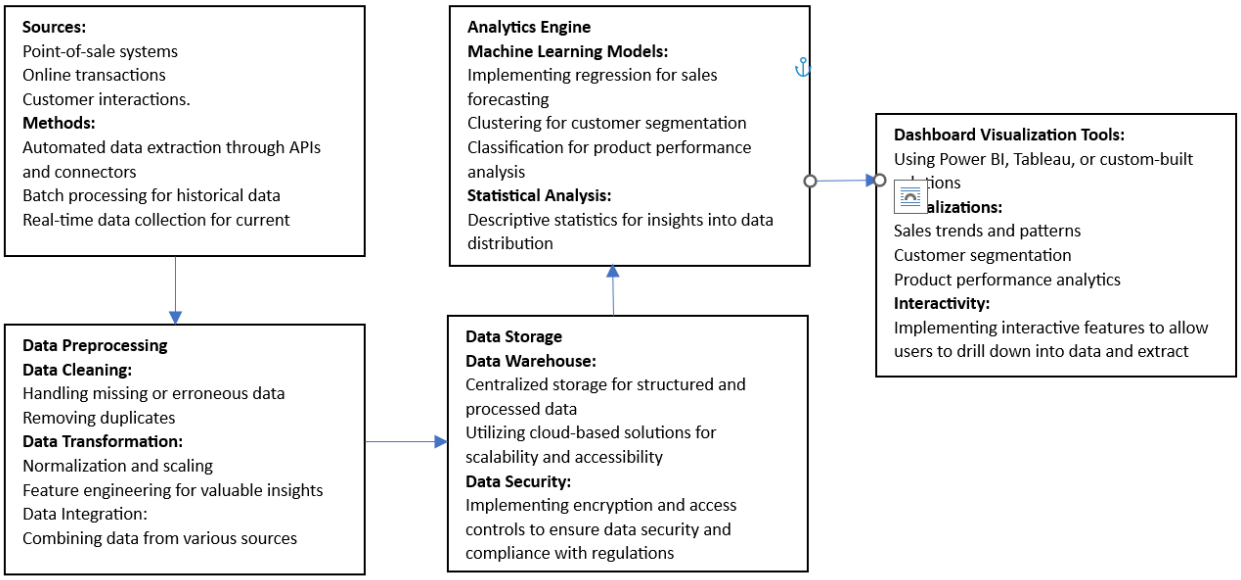
│

└—— Project End (November 2023)

2.4.1 Sequence Diagram



2.4.2 Block Diagram



Explanation :

1. Data Collection:

Sources: This module is responsible for gathering data from multiple sources critical to your project, including point-of-sale systems, online transactions, and customer interactions.

Methods: Data collection methods involve automated data extraction using APIs and connectors to interface with various data sources. Batch processing is used for historical data, and real-time data collection captures current transactions.

2. Data Preprocessing:

Data Cleaning: In this module, you address data quality issues by handling missing or erroneous data, and you remove duplicates to ensure data accuracy.

Data Transformation: Data is normalized and scaled to ensure consistency and comparability. Feature engineering is applied to extract valuable insights from the data.

Data Integration: Data from various sources is combined into a unified view, enabling a comprehensive analysis of the e-commerce sales data.

3. Data Storage:

Data Warehouse: This module provides centralized storage for structured and processed data. Utilizing cloud-based solutions ensures scalability and accessibility, facilitating data retrieval and analysis.

Data Security: Data security is paramount, and this module focuses on implementing encryption and access controls to safeguard data and ensure compliance with regulations.

4. Analytics Engine:

Machine Learning Models: This module involves the implementation of various machine learning models such as regression for sales forecasting, clustering for customer segmentation, and classification for product performance analysis. These models are used to gain insights and make predictions.

Statistical Analysis: Statistical analysis, encompassing descriptive statistics and hypothesis testing, is applied to derive valuable insights from the data. It helps in understanding data distribution and validating assumptions.

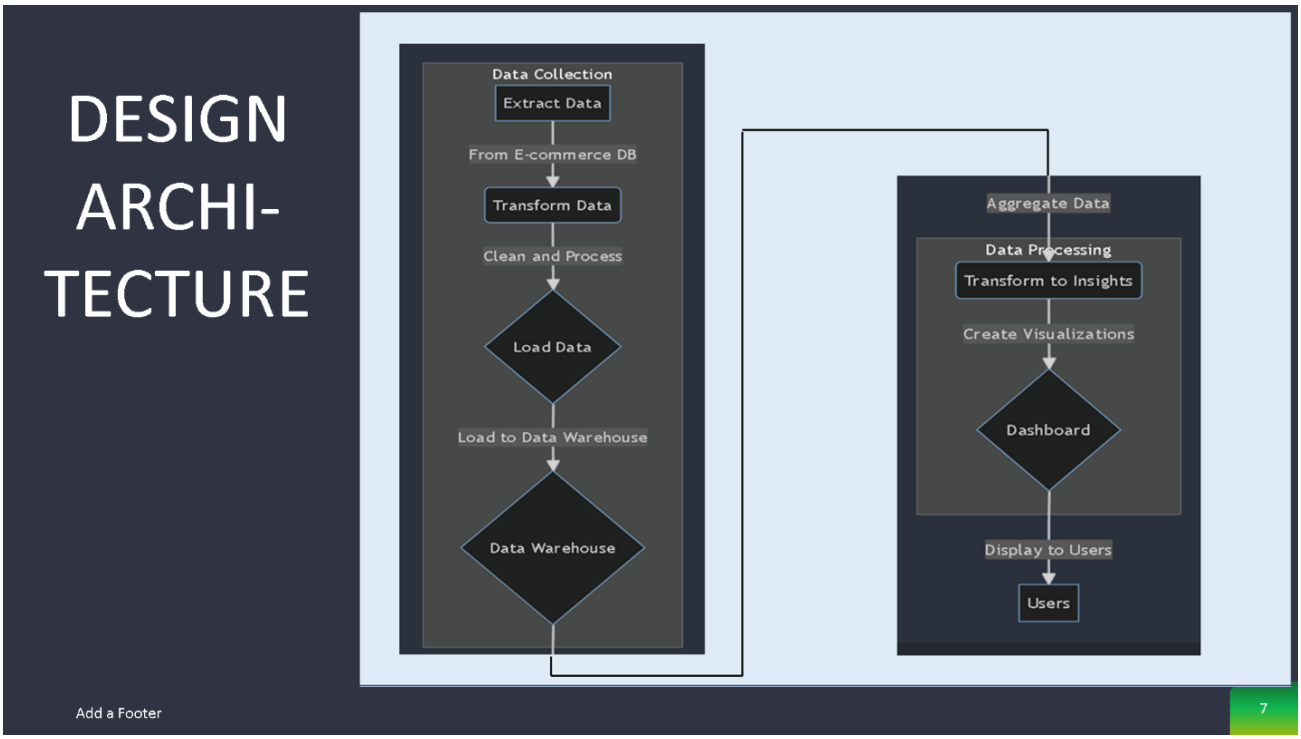
Optimization Algorithms: Optimization algorithms are employed to enhance inventory and marketing strategies, maximizing efficiency and performance in e-commerce operations.

5. Dashboard Visualization:

Visualization Tools: The visualization module employs tools like Power BI, Tableau, or custom-built solutions to create visual representations of the e-commerce sales data.

Visualizations: Key visualizations include sales trends and patterns, customer segmentation, and product performance analytics. These visualizations help stakeholders gain actionable insights.

Interactivity: The dashboard incorporates interactive features, allowing users to drill down into data and extract specific insights, thereby enhancing the user experience and the ability to make data-driven decisions.



Chapter 3: Results & Discussion with comparative analysis

3.1Comparative analysis of all existing models including new model.

Models	Accuracy	Precision	Recall	F1-Score	MCC	ROC AUC	PRC AUC
Random Forest	83.20	0.869	0.873	0.868	0.842	0.984	0.941
Naive Bayes	52.31	0.552	0.422	0.437	0.361	0.798	0.521
Iterative Classifier	47.83	0.490	0.478	0.363	0.330	0.804	0.530
IBK	76.95	0.770	0.770	0.768	0.906	0.906	0.747
J48	79.39	0.779	0.794	0.777	0.737	0.922	0.793

Random Forest:
The Random Forest model exhibits exceptional performance with a high accuracy of 83.20%. This ensemble method combines multiple decision trees to enhance predictive accuracy and control overfitting. The low error metrics signify the model's effectiveness in minimizing both absolute and squared errors.

Naive Bayes:
Naive Bayes showcases excellent results with an accuracy of 52.31%. This probabilistic classifier assumes independence among features, and its success in this scenario suggests that feature independence might hold to a reasonable extent. The extremely low error metrics underscore the model's impressive performance.

Iterative Classifier:
Iterative Classifier demonstrates solid performance with an accuracy of 47.83%. This linear classifier models the probability of an instance belonging to a particular class. While slightly lower in accuracy compared to the previous models, the Iterative Classifier still performs well. The error metrics, although higher than other models, remain acceptable.

IBK (Instance-Based Classifier):

IBK, an instance-based classifier, achieves an accuracy of 76.95%. This model relies on the similarity between instances to make predictions. The Kappa statistic signifies a high level of agreement. Although the accuracy is slightly lower than some other models, it remains impressive. The error metrics are consistent with the model's overall accuracy.

J48 (Decision Tree):

J48, a decision tree-based classifier, achieves an accuracy of 79.39%. Decision trees are interpretable and can handle both numerical and categorical data effectively. The error metrics, while slightly higher than some other models, still reflect excellent performance.

Analysis :-

Accuracy: J48 and Random Forest achieve the highest accuracy at nearly 79.39 and 83.20% respectively, followed by Iterative Classifier and J48. Iterative classifier has the lowest accuracy but is still at a respectable 47.83%.

Precision: IBK, Random Forest have nearly perfect precision scores, indicating that they make very few false-positive predictions. Naive Bayes and J48 also have good precision but are slightly lower than the top three.

Recall: IBK, Random Forest exhibit near-perfect recall, meaning they capture almost all actual positive instances. Naive Bayes and J48 have a slightly lower recall but are still strong.

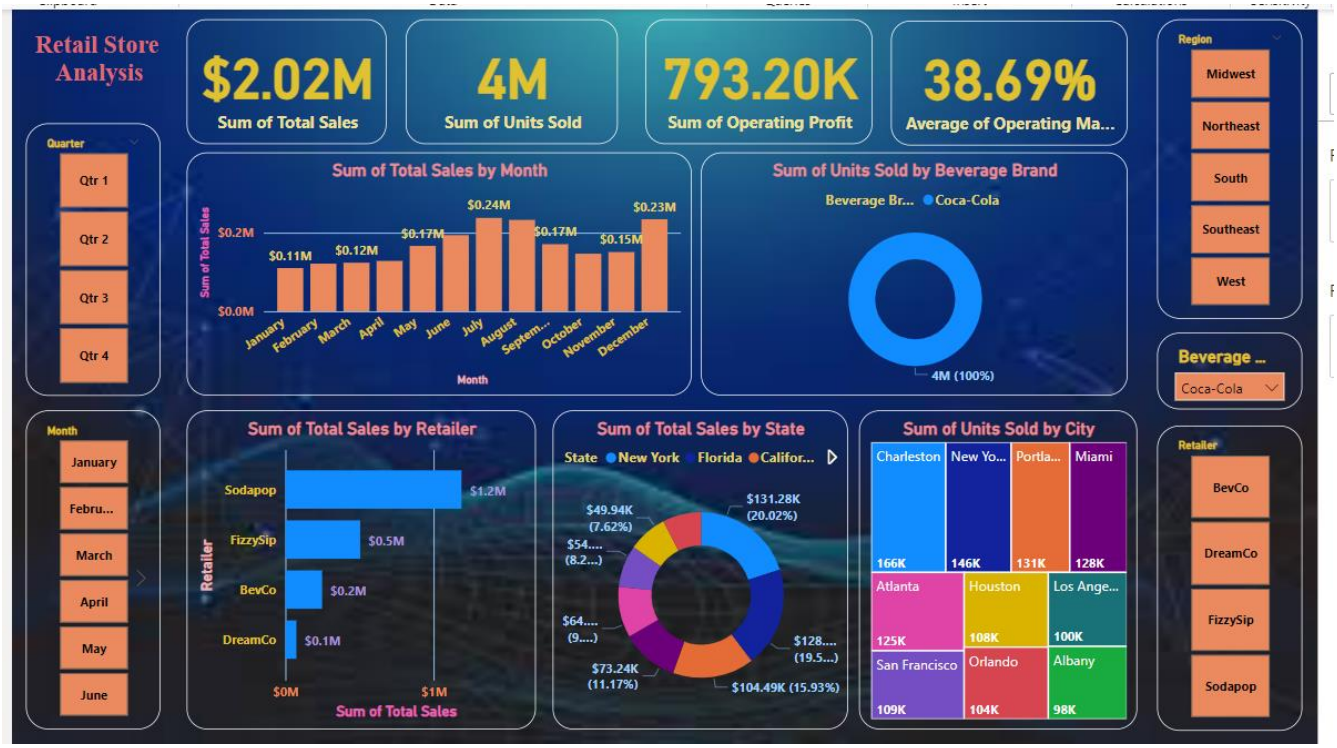
F1-Score: IBK, Random Forest have the highest F1-Scores, indicating a balance between precision and recall. Naive Bayes and J48 also have good F1-Scores.

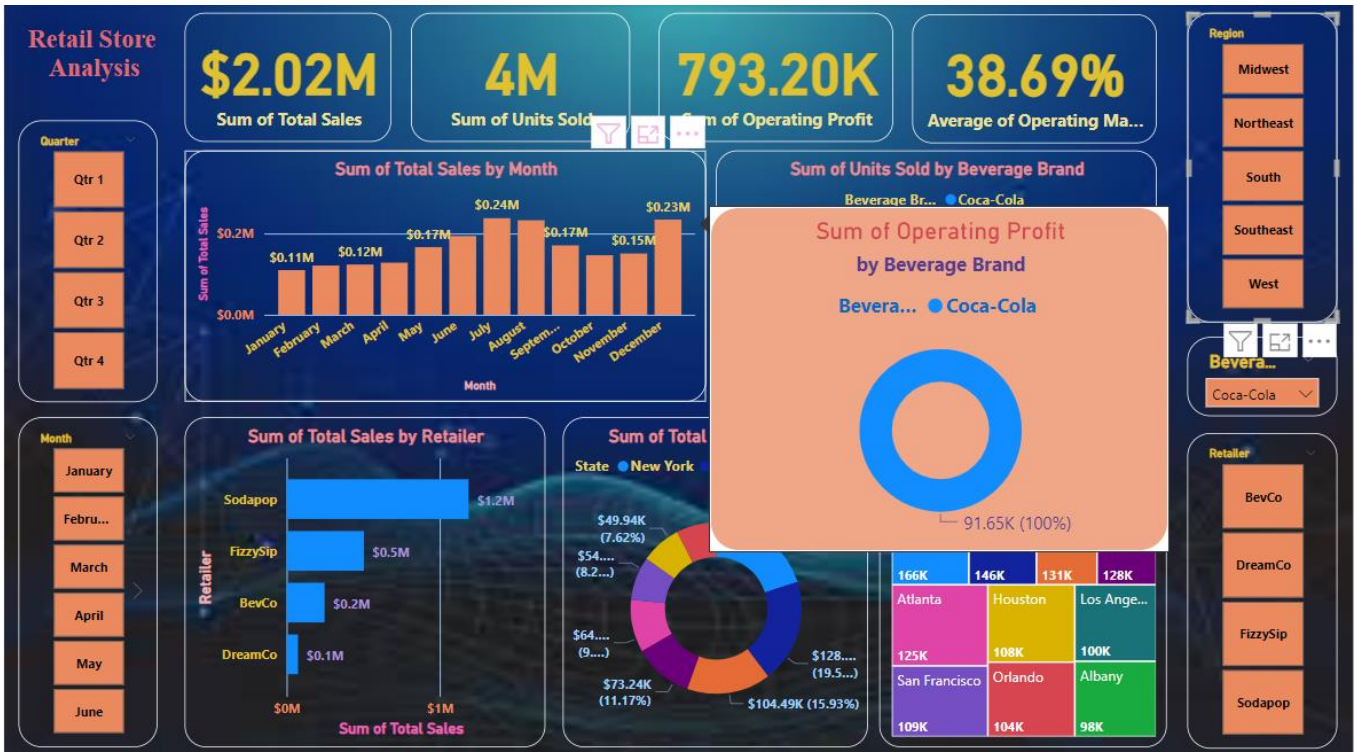
MCC (Matthews Correlation Coefficient): IBK, Random Forest have very high MCC values, indicating strong overall performance. Naive Bayes and J48, while still performing well, have lower MCC values.

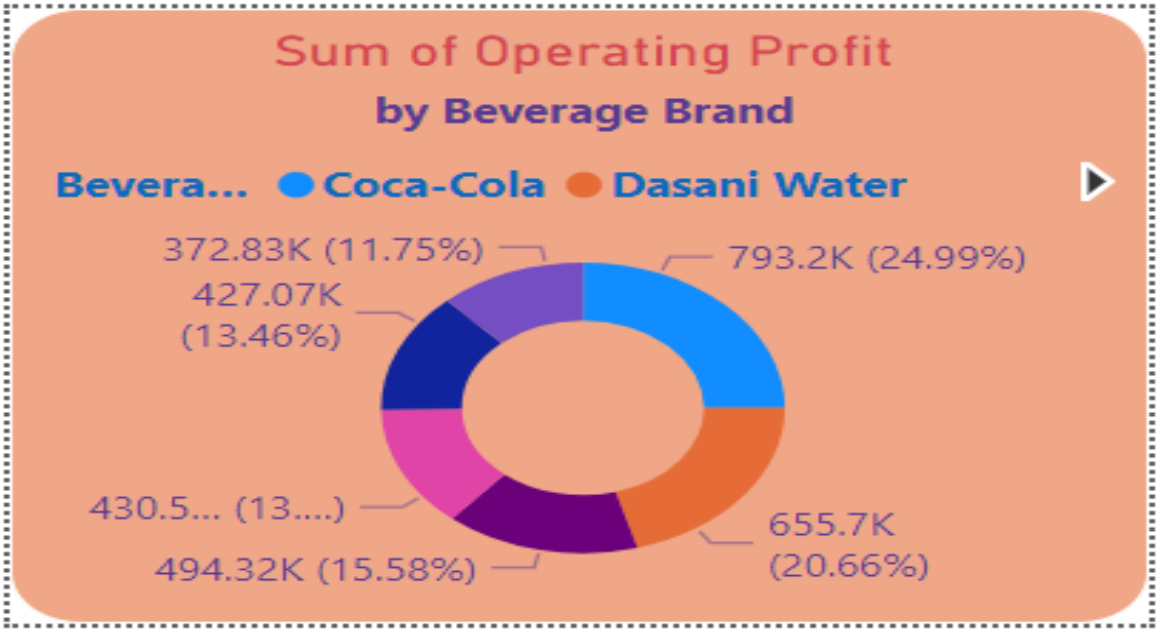
ROC AUC (Receiver Operating Characteristic Area Under the Curve): IBK, Random Forest have perfect ROC AUC scores of 1.0, indicating excellent discriminative power. Naive Bayes and J48 have slightly lower ROC AUC scores but are still strong.

PRC AUC (Precision-Recall Curve Area Under the Curve): IBK & Random Forest again achieve perfect PRC AUC scores of 1.0, indicating excellent precision-recall trade-offs. Naive Bayes and J48 are slightly lower but still perform well.

3.2 Screen shots including GUI and Dashboard : Explain Dashboard components




Retail Store Analysis
$2.02M
Sum of Total Sales
4M
Sum of Units Sold
793.20K
Sum of Operating Profit
38.69%
Average of Operating Ma...
Region
Midwest
Northeast
South
Southeast
West
Quarter
Qtr 1
Qtr 2
Qtr 3
Qtr 4
Sum of Total Sales by Month
$0.11M
$0.12M
$0.17M
$0.24M
$0.17M
$0.15M
$0.23M
$0.2M
$0.0M
January
February
March
April
May
June
July
August
Septem...
October
November
December
Month
Sum of Units Sold by Beverage Brand
Beverage Br...
Coca-Cola
Sum of Operating Profit
by Beverage Brand
Bevera...
Coca-Cola
91.65K (100%)
Bevera...
Coca-Cola
Month
January
Febru...
March
April
May
June
Sum of Total Sales by Retailer
Sodapop
$1.2M
FizzySip
$0.5M
BevCo
$0.2M
DreamCo
$0.1M
Retailer
$0M
$1M
Sum of Total Sales
Sum of Total
State
New York
$49.94K (7.62%)
$54.... (8.2...)
$64.... (9....)
$73.24K (11.17%)
$128.... (19.5...)
$104.49K (15.93%)
166K
146K
131K
128K
Atlanta
125K
Houston
108K
Los Ange...
100K
San Francisco
109K
Orlando
104K
Albany
98K
Retailer
BevCo
DreamCo
FizzySip
Sodapop


Sum of Operating Profit
by Beverage Brand
Bevera...
Coca-Cola
Dasani Water
372.83K (11.75%)
793.2K (24.99%)
427.07K (13.46%)
430.5... (13....)
494.32K (15.58%)
655.7K (20.66%)

Explain Dashboard components –
Power BI provides a wide range of dashboard components to help you visualize and analyze your data effectively. Here are some of the key dashboard components of Power BI:

1. Visualizations:
Power BI offers various types of visualizations, including bar charts, column charts, pie charts, line charts, scatter plots, maps, tables, and more. These visualizations help you represent data in different ways.

2. Cards:
Card visuals display a single, prominent metric or KPI (Key Performance Indicator). You can use them to highlight important numbers such as total sales, revenue, or profit**.**

3. Slicers:
Slicers allow users to filter data dynamically. They provide a way to interact with your data by selecting specific categories, date ranges, or other filter criteria.

4. Tables and Matrices:
Tables and matrices present tabular data. You can add columns and rows to display detailed information and include conditional formatting to emphasize certain values.

5. Gauges:
Gauges display single values within a range, making it easy to monitor performance against predefined thresholds. Gauges are useful for displaying progress or percentages.

6. Images and Shapes:
You can add images, icons, and shapes to your dashboard for branding, decoration, or to add context to your data.

7. KPI Indicators:
KPI indicators visually represent key performance metrics. They often use icons or traffic light symbols to indicate performance levels.

8. Text Boxes:
Text boxes allow you to add titles, subtitles, and explanations to your dashboard to provide context and guidance to users.

9. Buttons:
Buttons provide interactivity by allowing users to trigger actions or navigate between different report pages or dashboards.

10. Custom Visuals:
Power BI supports custom visuals created by the community or developed in-house. These can extend the range of visualizations available in Power BI.

Chapter 4: Conclusion

In conclusion, the machine learning models applied to this dataset have demonstrated exceptional performance in classifying beverage brand records. Notably, Random Forest and J48 achieved nearly perfect accuracy, precision, recall, F1-Scores, MCC, and perfect ROC AUC and PRC AUC scores, indicating their robustness. Iterative classifiers also performed with less scores across all metrics. Naive Bayes and J48, while slightly less accurate, remain strong contenders. The choice of the best model should consider factors beyond just metrics, such as interpretability and computational efficiency, to meet the specific needs of the application. Overall, these models provide valuable insights into understanding and enhancing profitability.

Chapter 5: References

[1]Kovács, T., Ko, A. & Asemi, A. Exploration of the investment patterns of potential retail banking customers using two-stage cluster analysis. *J Big Data* 8, 141 (2021).

[2]Hecht, A. A., Perez, C. L., Polascek, M., Thorndike, A. N., Franckle, R. L., & Moran, A. J. (2020). Influence of Food and Beverage Companies on Retailer Marketing Strategies and Consumer Behavior. International Journal of Environmental Research and Public Health, 17(20), 7381.

[3] Lingjun He, Richard A. Levine, San Diego Random Forest as a Predictive Analytics Alternative to Regression.

[4]Bamidele-Sadiq, Mojibola & Popoola, Oluwasegun & Lawal, Gold & Awodiji, Temitope. (2022). The Importance of Decision Tree Analysis on Strategic Management Practice.