The 7th Sense For Data-Driven Decision Mastery

Raja Brundha A
Assistant Professor
Department of Artificial Intelligence and Data Science,
Sri Sai Ram Engineering College,
Chennai, Tamil Nadu, India.
rajabrundha.ai@sairam.edu.in

Abubacker S
Student
Department of Artificial Intelligence and Data Science,
Sri Sai Ram Engineering College,
Chennai, Tamil Nadu, India.
nav.abubacker@gmail.com

Deva Praveen K
Student
Department of Artificial Intelligence and Data Science,
Sri Sai Ram Engineering College,
Chennai, Tamil Nadu, India.
devapraveen20@gmail.com

Bastin A
Student
Department of Artificial Intelligence and Data Science,
Sri Sai Ram Engineering College,
Chennai, Tamil Nadu, India.
biobastin2005@gmail.com

Vetri Thirumagan S
Student
Department of Artificial Intelligence and Data Science,
Sri Sai Ram Engineering College,
Chennai, Tamil Nadu, India.
vetrithirumagans@gmail.com

***Abstract*— Our project automates the manual data analytics pipeline, transforming it into a seamless, AI-powered process designed to revolutionize decision-making and drive strategic innovation. By connecting to a client's database, it automates data preprocessing, cleaning, and imputation, ensuring accurate data types through sophisticated feature-template similarity calculations. The system generates a comprehensive abstracted text file, detailing metadata, feature descriptions, missing values, and statistical summaries, alongside advanced analyses such as RFM, customer behavior clustering, and churn prediction. Leveraging cutting-edge generative AI with retrieval-augmented generation (RAG) and few-shot prompting, the project converts text files into embeddings, indexed in a vector database for insightful, similarity-based content retrieval. This allows decision-makers to quickly access vital insights, interpret complex data, and make informed decisions, ultimately driving business success. By providing a fast, accurate, and thorough analysis, our project empowers decision-makers to stay ahead of trends, transforming how data is used in strategic planning, and setting a new global standard in the competitive business world.**

Keywords—Retail Data Analytics, Retrieval augmented generation kRAG), Large Language Models (LLMs), Natural Language Processing, Few-Shot Learning, Business Intelligence, Analytics Agent, Statistical Patterns, Data-Driven Decision Marking, Retail trends , Business growth.

I. INTRODUCTION

In today's data-driven world, businesses are inundated with vast amounts of data that hold the potential to unlock significant insights. However, the manual process of analyzing this data is often time-consuming, error-prone, and inefficient. As companies strive to remain competitive, the need for an automated, reliable, and intelligent data analytics system has never been more critical.

Traditional data analysis methods often require extensive human intervention, from data cleaning and preprocessing to complex statistical analysis and visualization. These steps, while essential, can lead to inconsistencies and delays, impeding timely decision-making. The limitations of manual processes underscore the necessity for a more streamlined, automated solution that can handle large datasets with precision and speed.

Our project addresses these challenges by developing an advanced automated data analytics pipeline. This system is designed to seamlessly connect to client databases, fetch relevant data, and perform comprehensive preprocessing tasks, including cleaning, imputation, and ensuring correct data types. By automating these foundational steps, the system significantly reduces the potential for human error and accelerates the overall analysis process.

Beyond preprocessing, the system also generates detailed abstracts of the data, encompassing metadata, feature descriptions, and statistical summaries. These abstracts include rigorous statistical tests to identify normal distributions, outliers, and influential points. By providing this level of detail, the system ensures that decision-makers have access to a complete and accurate picture of the data at their disposal.

One of the standout features of our project is the integration of generative AI techniques, specifically retrieval-augmented generation (RAG) and few-shot prompting. These technologies enable the system to convert abstracted data files into embeddings, which are then indexed in a vector database. This allows for efficient, similarity-based content retrieval, facilitating more insightful and actionable analysis.

In addition to its analytical capabilities, the system incorporates a question-answering chatbox that interacts with decision-makers, helping to clarify requirements and refine analysis outputs. This interactive component ensures that the analysis is not only thorough but also aligned with the specific needs and goals of the business, enhancing its relevance and utility.

By automating and enhancing the data analysis process, our project empowers decision-makers with faster, more accurate, and more comprehensive insights. This not only accelerates strategic planning and implementation but also sets a new standard for how businesses leverage data in a rapidly evolving and competitive global marketplace.Related Works

II. LITERATURE SURVEY

The evolving landscape of retail, fashion, and data analysis has seen significant advances in machine learning, large language models (LLMs), and data-driven decision-making. This literature survey examines eight key papers, each contributing uniquely to the field, ranging from retail

forecasting to anomaly detection and customer segmentation. These studies offer insights into the potential of automated systems to revolutionize industries through advanced analytics and artificial intelligence. This narrative explores the comparative importance of these papers and their contributions to current research.

The work by Shuangyu Pang (2022) lays a foundational understanding of retail sales forecasting using machine learning, primarily focusing on Walmart's sales data. The study's emphasis on Random Forest Regression as the optimal model for sales prediction highlights the importance of feature importance and model accuracy. Though limited by its data scope (Walmart), the study provides a structured methodology for handling large retail datasets, making it a critical contribution to retail analytics.

When compared with *İnanç Kabasakal's* (2020) work on customer segmentation using the RFM model, Pang's paper complements this approach by offering a sales forecasting framework that businesses could use to optimize their marketing strategies developed from segmentation results. While Kabasakal's study directly addresses customer behavior and segmentation, Pang's approach helps retailers adjust inventory and staffing to meet forecasted demand, ensuring that the results of segmentation analyses can be acted upon with precision. [1]

Shifting from retail forecasting to automated fashion analysis, Yujuan Ding et al. (2024) introduce the GPT-FAR system, which utilizes LLMs to generate detailed fashion reports from catwalk data. The novel use of GPT-4V for garment classification, collective analysis, and report generation marks a significant leap in automating subjective tasks typically reserved for human experts.

FashionReGen expands on the more straightforward predictive models explored in Pang's work by incorporating the generative capabilities of LLMs, extending the scope of analysis beyond numerical predictions into the realm of creativity and subjective interpretation. While Kabasakal's RFM analysis remains grounded in data segmentation, FashionReGen opens new avenues for retail analytics by providing actionable insights into fashion trends, an area of growing importance in the fast-paced fashion industry.[2]

Exploring deeper into the intersection of LLMs and data retrieval, Yunfan Gao et al. (2023) provide a comprehensive survey on Retrieval-Augmented Generation (RAG), a paradigm that enhances LLMs by retrieving and incorporating relevant external data during text generation. This framework is crucial in advancing the capabilities of models like GPT-FAR, as it ensures that generated content is not only coherent but also accurate and contextually relevant.[8]

The ***Gao et al.*** study offers an expanded understanding of how information retrieval and augmentation can be leveraged in models like those discussed by Ding et al, allowing for a more robust fashion report generation system. Compared to Pang's work, which remains static in its predictive model, the RAG framework offers a dynamic approach to enhance LLM output by making real-time data adjustments.[3]

The exploration of whether GPT-4 can function as a data analyst, as proposed by Liying Cheng et al. (2023), adds a thought-provoking dimension to the discussion of LLMs in retail and business analytics. Cheng's work demonstrates that while LLMs like GPT-4 can outperform entry-level human data analysts, there are limitations in the comprehensiveness and depth of analysis when compared to senior analysts.

This paper finds its greatest relevance when juxtaposed with Pang's study, which relies on the static analysis of sales data. The research by Cheng et al raises the question of whether LLMs could ultimately replace traditional machine learning models in tasks like those performed by Pang, especially when real-time adaptability and detailed insights are required.[4]

Dawit Dibekulu Alem's (2020) overview of data analysis serves as a foundational paper in this survey. Alem stresses the importance of data analysis in research, providing a broad overview of the techniques used to extract insights from datasets. Though less specialized than the other papers, Alem's work underlines the essential methodologies that are critical across all studies, whether it be retail sales forecasting, customer segmentation, or LLM-based fashion report generation.

Compared to the advanced methodologies of Pang and Ding, Alem's contribution lies in its accessibility and emphasis on fundamental techniques, offering a grounding perspective that can be built upon with more complex models and systems.[5]

In the field of anomaly detection, this survey highlights the importance of identifying irregularities in data, which can lead to actionable insights across various sectors. While Sudeep B Chandramana's (2017) paper on retail analytics underscores the application of anomaly detection in fraud prevention, it provides a broader business perspective that ties into the work of Pang and Kabasakal by illustrating how anomalies in sales and customer behaviour can be used to adjust marketing strategies and prevent financial losses.[6].

Each of the eight papers contributes uniquely to the fields of retail analytics, fashion report generation, and data analysis. *Pang's sales forecasting model* provides a robust framework for retail decision-making, while Ding et al.'s *FashionReGen* introduces the potential of LLMs in subjective areas like fashion reporting. *Gao et al.'s work on RAG* enhances the potential of LLMs in dynamic contexts, and *Cheng et al.* critically evaluates the role of AI in data analysis, challenging the dominance of traditional human analysts.[7]. Finally, *Alem, Kabasakal, and Chandramana* provide crucial perspectives on foundational data analysis techniques, customer segmentation, and anomaly detection, respectively, tying the entire field together in a comprehensive narrative of automation, machine learning, and AI.

a) METHODOLOGY

This project employs a multi-step approach to automate the data analysis process from start to finish. It begins with connecting to the client's database, where relevant data tables are fetched for analysis. Automated preprocessing tasks follow, including data cleaning, imputation, and ensuring correct data types through feature-template similarity calculations. An abstracted text file is then generated, detailing metadata, feature descriptions, and comprehensive statistical summaries, alongside key statistical tests. The system identifies outliers and conducts advanced analyses, such as RFM and customer behavior clustering, to provide deeper insights. Leveraging generative AI, particularly retrieval-augmented generation (RAG) and few-shot prompting, the abstracted data is converted into embeddings, which are indexed in a vector database for efficient similarity-based retrieval. Finally, a question-answering chatbox engages with decision-makers, offering precise insights and strategic recommendations tailored to their specific needs, thus transforming the way businesses leverage data for informed decision-making.

Figure 3.1: Agent flow (9)

A. Data Collection and Preprocessing:

a) Data Collection:

In the First Step, data collection, accessing necessary data from the company's Data Warehouse is essential. It's preferable to rely on the business' own data for automated knowledge discovery, ensuring valuable insights for clients. By leveraging internal data sources, the process becomes more tailored and effective in uncovering pertinent information.

Dataset: Retail-Shopping Transaction Dataset from Kaggle

https://www.kaggle.com/datasets/zeesolver/consumer-behavior-and-shopping-habits-dataset?select=shopping_trends.csv

Key Feature:

Customer_ID:
A unique identifier for each customer, used to track individual transactions and customer behavior.

Age:
The age of the customer at the time of the transaction, providing demographic insights.

Gender:
The gender of the customer, helping to analyze purchase patterns across different genders.

Item_Purchased:
The specific product bought by the customer during the transaction

.

Category:
The broader category to which the purchased item belongs, aiding in product-level analysis.

Purchase_Amount_USD:
The total amount spent by the customer in USD, indicating transaction value.

Location:
The geographical location where the transaction took place, useful for regional analysis.

Size:
The size of the purchased item, particularly relevant for apparel and similar products.

Color:
The color of the purchased item, which can be used to track color preferences and trends.

Season:
The season during which the purchase was made, offering insights into seasonal shopping behavior.

Review_Rating:
The customer's rating of the product or transaction, reflecting satisfaction and quality perceptions.

b) Data Preprocessing:

Data preprocessing involves leveraging its understanding of real-world entities and metadata to efficiently clean and prepare data. Here's a step-by-step approach.

1. Data Type Validation and Correction:

Ensuring that each column in the dataset has the correct data type is crucial for accurate analysis. For example, a date column stored as an object type should be converted to a datetime type. This correction prevents errors in time-based operations and analyses. Implementing a Python script to automatically validate and convert data types helps maintain consistency and integrity across the dataset.

2. Handling Missing Values:

Missing values can skew analysis results and lead to incorrect conclusions. Common strategies include:

Imputation:
Filling missing values with mean, median, or mode (e.g., replacing missing salary values with the median salary).

Removal:
Deleting rows or columns with excessive missing values (e.g., dropping columns where more than 50% of the data is missing).

Interpolation:
Using surrounding data points to estimate missing values in time series data.

3.Error Checking:

Feature engineering transforms raw data into meaningful features that enhance model performance. For example,

Revenue Calculation:
Creating a new feature by multiplying quantity and price to derive revenue.

Date Features:

Extracting year, month, and day from a datetime column to provide additional temporal insights.

Automated techniques, such as using NLP for understanding column semantics, can streamline this process.

4. Outlier Detection and Treatment:

Outliers can distort statistical analyses and modeling. Techniques for detecting and managing outliers include:

Isolation Forest:

Identifies anomalies by isolating observations in feature space.

Z-Score:

Measures how many standard deviations a data point is from the mean.

Cook's Distance:

Detects influential data points that significantly affect regression results.

Treating outliers may involve removing them or applying transformations to reduce their impact.

5. Data Normalization and Scaling:

Normalization and scaling ensure that features contribute equally to model performance. Techniques include:

Min-Max Scaling:

Rescales feature values to a range between 0 and 1 (e.g., normalizing age to a 0-1 scale).

Standardization:

Transforms features to have a mean of 0 and a standard deviation of 1 (e.g., standardizing test scores).

These methods are essential for algorithms sensitive to the scale of input features.

6. Feature Selection and Reduction:

Selecting relevant features and reducing dimensionality improves model efficiency and accuracy. Techniques include:

Principal Component Analysis (PCA):

Reduces the number of features while retaining most of the variance in the data.

Feature Importance:

Uses algorithms like Random Forest to rank features by their impact on the target variable.

Selecting the most significant features helps in simplifying models and enhancing interpretability.

7. Encoding Categorical Variables:

Machine learning models require numerical input, so categorical variables need to be encoded. Common methods include:

One-Hot Encoding:

Converts categorical values into binary vectors (e.g., turning colors into binary columns for red, green, and blue).

Label Encoding:

Assigns unique integers to each category (e.g., assigning 1 to 'low', 2 to 'medium', and 3 to 'high').

Proper encoding ensures that categorical data is appropriately represented in models.

A. Algorithmic Analytics:

This algorithmic analytics employs advanced machine learning techniques to interpret data and analyze customer behavior. Isolation Forest and other segmentation algorithms are used for identifying anomalies and grouping customers based on their behavior. Cohort analysis further segments customers into groups with similar characteristics. Regression models are applied for churn analysis, predicting customer retention and identifying factors influencing customer attrition. These methods collectively enhance the accuracy of insights and support targeted business strategies.

a) Visualizations (EDA):

Task: Exploratory Data Analysis (EDA) involves visually inspecting data to understand its patterns, distributions, and relationships. Visualizations help in identifying trends, outliers, and potential areas for further analysis.

- Histograms showing the distribution of purchase amounts.
- Scatter plots depicting the relationship between customer age and purchase frequency.
- Heatmaps displaying correlations between different product categories purchased together.

b) Facts and Dimensions:

Task: Grouping data allows for aggregating information based on specific features or attributes. It helps in identifying patterns and summarizing data within different groups.

- Customer segments (e.g., new customers vs. returning customers).
- Time periods (e.g., daily, weekly, or monthly sales).
- Product categories (e.g., electronics, clothing, or groceries).

Example Analysis: Calculating the average purchase amount per customer segment to identify which segment spends more on average.

c) Finding Anomalies:

Task: Anomalies or outliers in data can indicate errors, fraud, or unusual patterns that require further investigation. Detecting anomalies is essential for maintaining data quality and ensuring accurate analysis results.

- Analyzing transaction amounts for unusually high or low values.

- Identifying transactions with significant deviations from the norm.

Example Analysis: Flagging transactions with amounts significantly higher than the average purchase amount for further investigation.

d) Hypothesis Testing:

Task: Hypothesis testing is a statistical method used to make inferences about population parameters based on sample data. It involves formulating hypotheses and using sample data to assess the validity of these hypotheses.

- Test if offering discounts during holidays leads to a significant increase in sales.
- Assess if there's a significant difference in sales between different customer segments.

Example Analysis: Comparing sales data during holiday periods with non-holiday periods to determine if there's a statistically significant difference in sales.

e) Forecasting:

Task: Forecasting involves predicting future values based on historical data. It helps in planning and decision-making by anticipating future trends and patterns.

- Using time series analysis techniques to model sales trends over time.
- Applying machine learning models to predict future sales based on various factors.

Example Analysis: Predicting the demand for specific products during the upcoming holiday season to optimize inventory management and marketing strategies.

f) Predictive Modeling:

Task: Predictive modeling involves building models to predict outcomes based on input variables. It helps in understanding relationships between variables and making predictions for future events or behaviors.

- Predict customer churn based on factors such as purchase history, demographics, and customer interactions.
- Forecast future sales based on factors like marketing spend, economic indicators, and seasonality.

Example Analysis: Using machine learning algorithms to predict the likelihood of a customer churning in the future, allowing for targeted retention strategies.

The Analytic Agent can leverage various data analysis techniques to gain insights from retail transaction-customer data, ultimately enabling retailers to make informed decisions and optimize business performance.

Thus Finally the analytic agent will provide the abstracted file provides a comprehensive overview of the data extracted from the client's database and preprocessed for detailed analysis. The dataset has undergone a series of transformations aimed at ensuring data integrity and consistency. This includes the identification of data types (categorical, numerical, date, etc.), correction of inconsistencies, and imputation of missing values using advanced techniques like mean, median, or K-nearest neighbor imputation, depending on the nature of the missing data. Each feature within the dataset has been meticulously described, detailing its role, relevance, and distribution, alongside statistical summaries such as minimum, maximum, mean, median, and quantiles. In addition to basic statistics, more in-depth exploratory data analysis (EDA) has been performed to uncover underlying patterns, trends, and anomalies. Descriptive statistics are complemented by visualizations like histograms, box plots, and scatter plots, providing a clearer understanding of the distribution and relationships between variables.

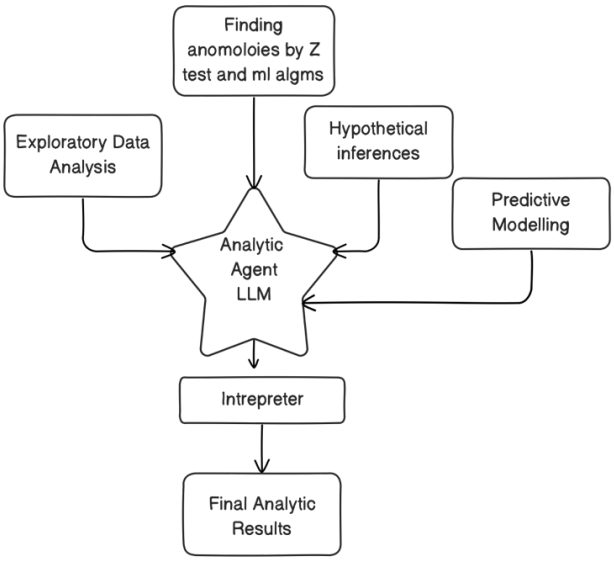

Figure 3.2: Overview of Agent Planning

B. Large Language Model for Analytics:

In the automated data analysis pipeline, the role of the Large Language Model (LLM) is pivotal in generating a well-constructed analytic report through the application of Retrieval-Augmented Generation (RAG). Here's how it contributes to the process:

Incorporating Few-Shot Prompting: Once the base text file, containing abstracted and structured information, is embedded into vectors and indexed in the vector database, few-shot prompting is applied to the LLM. This technique allows the LLM to understand the context of the analysis by providing a small number of relevant examples. It primes the model to perform specific analytical tasks based on the type of data it encounters and the examples provided.

Retrieving Relevant Contexts: Through the use of RAG, the LLM is not only tasked with generating coherent text but is also capable of retrieving the most semantically relevant information from the vector database. This ensures that the generated report is not just based on the LLM's training data but is augmented with real-time, contextually pertinent data from the indexed vectors. This retrieval step helps the LLM produce highly tailored insights that are specific to the dataset being analyzed.

Dynamic Analysis and Insights Generation: The LLM uses the retrieved information to perform detailed analysis and generate insights that are grounded in the data. This includes identifying patterns, trends, and anomalies, and making

predictions or recommendations based on the data's statistical and descriptive characteristics. The LLM ensures that the analysis is aligned with the business goals and client requirements, offering a level of flexibility and depth that static machine learning models may not provide.

Automated Reporting: The LLM's ability to produce human-like text is leveraged to generate an articulate, well-structured analytic report. The report can cover various aspects of the data, such as feature-level analysis, statistical tests, clustering results, RFM analysis, and potential customer churn. The narrative is designed to be clear and actionable, catering to both technical and non-technical stakeholders.

Interactive Clarification and Refinement: In addition to generating the initial report, the LLM is integrated into a chatbox that enables interactive question-answering. Clients can ask follow-up questions or request clarifications, and the LLM can retrieve and provide additional insights or refine the analysis based on specific queries. This feedback loop ensures that the final report is not only comprehensive but also customizable to the client's evolving needs.

Enhanced Decision-Making: By combining RAG with the LLM's analytical capabilities, the final output is a robust analytic report that aids decision-makers in interpreting data effectively. The model ensures that all insights are backed by data and the report offers strategic recommendations to guide business decisions in a competitive market.

. By combining the analytical prowess of the Analytic Agent with the communication finesse of the language model, this integrated approach ensures that data-driven insights are comprehensible and impactful for informed decision-making in the retail industry.

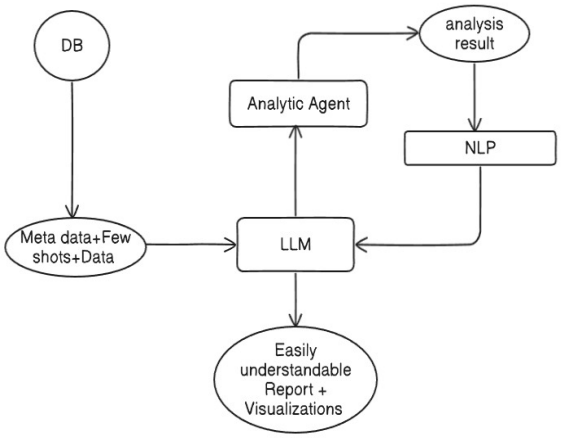

Figure 3.3: Overview of LLM Working

c) EXPERIMENTS

A. Experimental Setup

Dataset: We use the retail datasets which are available at https://www.kaggle.com/datasets/prasad22/retail-transactions-dataset/data. The dataset includes transaction details, product lists, total items and costs, payment information, location details, discounts and promotions, and customer insights categorizing customers based on background or age group.

Method: We employed the Pandas AI agent in tandem with the GPT-4 API key to analyze the provided dataset, resulting in a comprehensive set of insights. These insights shed light on various aspects of the data, uncovering patterns, trends, and correlations that offer valuable strategic implications for retail businesses.

B. Experimental Results

The trend of sales for each year

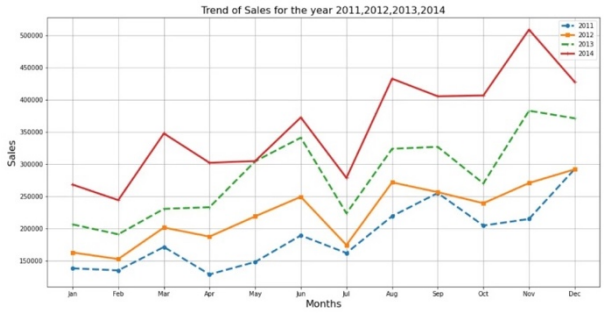

Figure 4.1: Result

Insights: In 2011, December emerged as the peak sales month for the company, while April marked the lowest sales period. Similarly, in 2012, December stood out as the month with the highest sales, contrasting with February, which experienced the lowest sales. Moving to 2013, November witnessed the highest sales, whereas February registered the lowest. This trend continued into 2014, with November again marking the highest sales and February recording the lowest. Notably, 2014 represented a peak sales year for the company overall.

Additionally, across all years, July consistently experienced a significant drop in sales. Understanding the root cause behind this recurring dip in July sales is imperative for the company's strategic planning and mitigation efforts.

c) CONCLUSION AND DISCUSSIONS

In this paper, we have introduced a robust framework for automating the data analytics pipeline, leveraging cutting-edge techniques such as data abstraction, retrieval-augmented generation (RAG), and large language models (LLM). This approach seamlessly integrates data preprocessing, statistical analysis, and customer behavior segmentation to deliver dynamic insights, facilitating faster and more accurate decision-making for businesses.

Future research could focus on enhancing the system's flexibility by incorporating more sophisticated models and expanding its capabilities to handle diverse and complex datasets. This framework showcases the transformative potential of AI-driven automation, setting a foundation for smarter, more efficient data analytics in various industries.

d) REFERENCES

[1] *Shuangyu Pang*, Retail Sales Forecast Based on Machine Learning Methods, *2022 6th Annual International Conference on Data Science and Business Analytics (ICDSBA).*

[2] Arisa Shollo, Using Business Intelligence in IT Governance Decision Marking, *IFIP WG 8.6 International Working Conference, Hamburg, Germany, September 22-24, 2011, Proceedings.*

[3] *Mortadha M, Hamad, Banaz A. Qader,* Data Pre-processing for knowledge discovery, *Tikrit Jourual of Pure Science l9 (5) 2011.*

[4] *Xuanhe Zhou, Zhaoyan Sun, Guoliang Li,* DB-GPT: Large Language Model Meets Database, *Published online: 19 January 2024*

[5] *Miss Sahrish Saifi Tandel*, Unveiling Patterns and Insights in a Retail Dataset:A Data Analytics Approach, *SJ Impact Factor: 7.538,Volume 11 Issue VII Jul 2023.*

[6] *Humza Naveeda, Asad Ullah Khana, Shi Qiub, Muhammad Saqibc, Saeed Anware, Muhammad Usmane, Naveed Akhtarg, Nick Barnesh, Ajmal Miani,* A Comprehensive Overview of Large Language Models, *arXiv:2307.06435v9 [cs.CL] 9 Apr 2024.*

[7] *Cheonsu Jeong*, Study on the Implementation of Generative AI Services Using an Enterprise Data-Based LLM Application Architecture*, Advances in Artificial Intelligence and Machine Learning; Research 3 (4) 1588-1618, Published 29-10-2023.*

[8] *Liying Cheng, Xingxuan Li, Lidong Bing,* Is GPT-4 a Good Data Analyst?, *arXiv:2305.15038v2 [cs.CL] 23 Oct 2023.*