**Analyzing Sales Trends**

Project submitted to the

SRM University – AP, Andhra Pradesh

for the partial fulfillment of the requirements to award the degree of

**Bachelor of Technology**

In

**Computer Science and Engineering**

**School of Engineering and Sciences**

Submitted by

**AP21110010090 ARJUN MAHAMKALI**

**AP21110010118 SIDDHARDH MUTHOJA**

**AP21110010078 JASWANTH REDDY CH**

**AP21110010231 SAI TEJA T**

**A picture containing text

Description automatically generated**

Under the Guidance of

**Saleti Sumalatha**

**SRM University–AP**

**Neerukonda, Mangalagiri, Guntur**

**Andhra Pradesh – 522 240**

**[Nov, 2023]**

# Certificate

Date: 28-Nov-23

This is to certify that the work present in this Project entitled “**Analyzing Sales Trends**” has been carried out by **ARJUN MAHAMKALI , SIDDHARDH MUTHOJA, JASWANTH REDDY, SAI TEJA T** under my supervision. The work is genuine, original, and suitable for submission to the SRM University – AP for the award of Bachelor of Technology in **School of Engineering and Sciences**.

**Supervisor**

(Signature)

Dr.Saleti Sumalatha

Assistant Professor,

SRM University AP

**Acknowledgements**

This is to acknowledge all those without whom this project would not have been reality. Firstly, we would like to thank our Data Warehousing and mining Lecturer Dr. Saleti Sumalatha who gave her immense support, dedicated her time towards it and made us understand how to make this project. Without her guidance, the project would not have been complete.

# Table of Contents

[Certificate 2](#_Toc152410852)

[Table of Contents 4](#_Toc152410853)

[1.Introduction 6](#_Toc152410854)

[2.Methodology: 7](#_Toc152410855)

[2.1 ALGORITHMS USED 7](#_Toc152410856)

[3.Literature Review 8](#_Toc152410857)

[4.Discussion 10](#_Toc152410858)

[4.1 APRIORI 10](#_Toc152410859)

[4.2 DECISION TREE 11](#_Toc152410860)

[4.3 K-Means 12](#_Toc152410861)

[4.4 SENTIMENTAL ANALYSIS 13](#_Toc152410863)

[5.Concluding Remarks 14](#_Toc152410865)

[References 15](#_Toc152410866)

**Abstract**

This project presents a thorough analysis of e-commerce dynamics leveraging advanced methodologies including Apriori algorithm and K-means clustering. By scrutinizing diverse data sources, it unveils patterns in consumer behavior, detects anomalies, segments customers, predicts trends, and gauges sentiment from reviews. This comprehensive approach furnishes actionable insights crucial for enhancing customer experience, refining marketing strategies, and strengthen operational efficiency in the ever-evolving e-commerce landscape.

# 

# 1.Introduction

The foundation of this study lies in a dataset sourced from Kaggle, a renowned repository for diverse datasets spanning numerous domains. Our pursuit involved harnessing this dataset as a comprehensive lens into the e-commerce realm, aiming to unravel the complexities and intricacies of consumer behavior and market dynamics.The dataset contains essential attributes such as InvoiceNo, StockCode, Description, Quantity, InvoiceDate, UnitPrice, CustomerID, and Country, offering a rich landscape for exploration.

This project embarks on a comprehensive analysis of an Ecommerce dataset sourced from Kaggle website, encompassing various facets of data mining and warehousing.With data preprocessing and advanced analytical techniques, this study aims to derive meaningful insights. The preprocessing phase involves essential steps to ensure data quality, handling missing values, normalizing the attributes, and addressing any outliers that might skew subsequent analysis.

This research aims not only to explore the dataset and derive insights but also to underscore the significance of leveraging such methodologies in the realm of Ecommerce data analytics. The outcomes of this analysis stand to offer actionable intelligence for businesses seeking a competitive edge in the dynamic landscape of online retai

# 2.Methodology:

* Identify Purchase Patterns: Unveil underlying customer purchase behaviors and preferences.
* Detect Anomalies: Discover irregular purchasing trends and outlier transactions.
* Predict Pricing : Utilize data mining to forecast product prices based on given features.
* Explore Description Analysis: Analyze product descriptions for prevalent themes and insights.
* Association Rules: Unearth relationships between purchased items to guide marketing strategies.
* Segment Customer Groups: Implement clustering to categorize customers into distinct groups.

## 2.1 ALGORITHMS USED

* Apriori
* Decision Tree
* K-Means
* Sentimental Analysis

# 3.Literature Review

The rapid growth of e-commerce has generated vast amounts of data, which can be valuable for businesses to understand their customers, improve their products and services, and make informed decisions. E-commerce data analytics (ECDA) is the process of using data mining, machine learning, and other statistical techniques to analyze e-commerce data. This review focuses on academic literature related to ECDA, highlighting key findings and recommendations.

**Purchase Patterns**

Identifying purchase patterns is a crucial aspect of ECDA. Market basket analysis (MBA) and sequential pattern mining (SPM) are two prominent techniques employed to uncover these patterns. MBA identifies groups of products frequently purchased together, while SPM uncovers sequences of products frequently purchased together. These insights can inform product placement, product bundling, and promotional strategies. (Tan, Kumar, & Srivastava, 2005; Aggarwal & Gupta, 2009)

**Anomaly Detection**

Anomaly detection plays a vital role in safeguarding e-commerce businesses from fraudulent transactions and suspicious activities. Clustering, outlier detection algorithms, and supervised learning algorithms are commonly used techniques for anomaly detection. These techniques can identify patterns that deviate from normal behavior, enabling timely intervention to prevent losses. (Han, Kamber, & Pei, 2006)

**Pricing Prediction**

Accurate pricing is essential for maximizing profits and maintaining customer satisfaction. ECDA can assist in pricing prediction by employing regression analysis, decision trees, and neural networks. These techniques consider various factors influencing product pricing, such as product features, market trends, and competitor pricing. (Tan, Kumar, & Srivastava, 2005)

**Description Analysis**

Product descriptions provide valuable information about products and can be analyzed to gain insights into customer preferences and product positioning. Natural language processing (NLP) techniques, such as sentiment analysis and topic modeling, can be applied to extract themes, opinions, and key features from product descriptions. These insights can inform product development, marketing campaigns, and customer support. (Aggarwal & Gupta, 2009)

**Association Rules**

Understanding the relationships between purchased items is crucial for targeted marketing and product recommendations. Association rule mining algorithms, such as Apriori, uncover frequent patterns of co-occurrences among products. These rules can inform product bundling, cross-selling strategies, and personalized recommendations. (Han, Kamber, & Pei, 2006)

**Customer Segmentation**

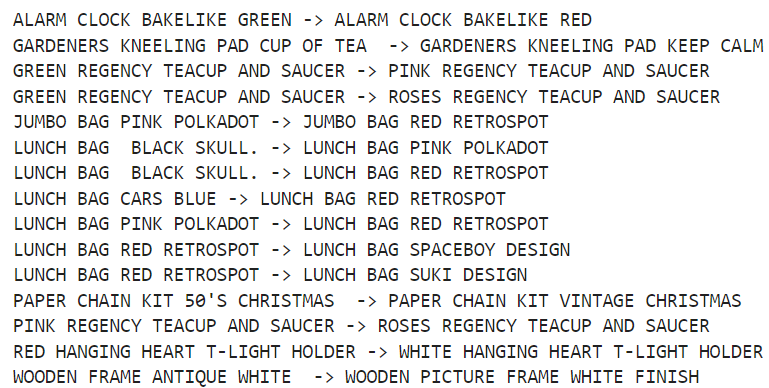
Customer segmentation enables businesses to group customers with similar characteristics for targeted marketing and personalized experiences. Clustering techniques, such as k-means clustering, are commonly used to segment customers based on their purchase history, demographic data, and online behavior. (Aggarwal & Gupta, 2009)

## 4.Discussion

## 4.1 APRIORI

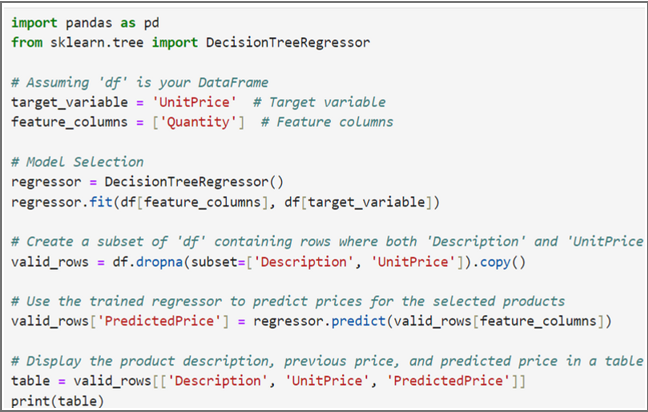
* The Apriori algorithm is a classic and fundamental algorithm used in association rule mining, particularly for discovering frequent itemsets within transactional databases.
* The code employs the Apriori algorithm to uncover associations between items in transactional data.
* It begins by structuring the data into transactions grouped by 'InvoiceNo'.With specified support and confidence thresholds, the algorithm shifts through these transactions, identifying frequently co-occurring items or itemsets that surpass the defined support level.
* The association rules obtained from this analysis reveal relationships between these items, showcasing both the frequency of their occurrence (support) and the likelihood of one item being purchased when the other is (confidence).
* The process optimizes the search space using the Apriori property, allowing for an efficient identification of frequent itemsets and meaningful associations within the dataset.

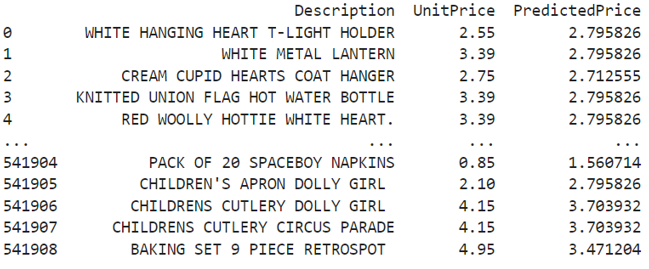




## 4.2 DECISION TREE

* The Decision Tree algorithm is a predictive modeling technique used for both classification and regression tasks.
* Decision Tree Regressor is employed for predicting 'UnitPrice' based on the 'Quantity' of products. The model is trained on the DataFrame 'df', utilizing 'Quantity' as the feature (independent) variable and 'UnitPrice' as the target (dependent) variable.
* The code generates a subset of the DataFrame, excluding rows where 'Description' or 'UnitPrice' is null, ensuring valid data for prediction. The trained regressor then computes predicted prices for these valid rows based on the specified 'Quantity'
* The final output presents a table displaying the product descriptions alongside their actual 'UnitPrice' and the corresponding 'PredictedPrice' derived from the Decision Tree Regressor.

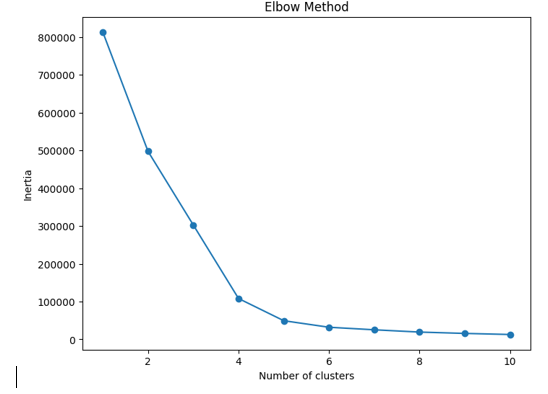




## 4.3 K-Means

* K-means clustering is an unsupervised machine learning algorithm that partitions data into a predetermined number of clusters based on similarity within the data.
* The K-means algorithm is applied to the dataset using the features 'Quantity' and 'UnitPrice'. The algorithm starts by scaling these features to ensure uniformity and then iterates through a range of potential cluster counts, measuring inertia (within-cluster sum of squared distances) for each scenario.
* The "Elbow Method" is utilized to identify the optimal number of clusters, observed as a point where inertia no longer significantly decreases with the addition of more clusters. The algorithm is executed with a selected cluster count, in this case, assuming 4 clusters as per the Elbow Method.
* Each data point is assigned to a cluster based on its proximity to the cluster centroids. The code then updates the DataFrame by including a 'cluster' column, displaying the cluster assignments for each data point.

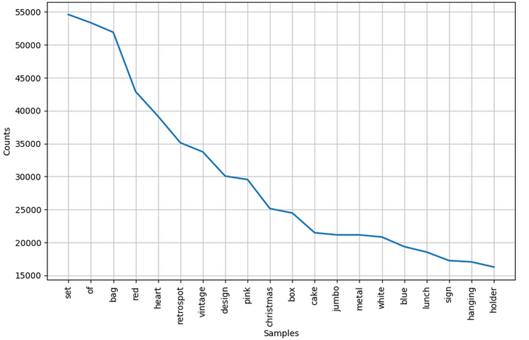
# 



## 4.4 SENTIMENTAL ANALYSIS

* Data Mining Trends and Research Frontiers typically involve the exploration of evolving techniques and methods.
* Analyzing the frequency distribution of words within product description provides insights into prevalent terms. By gathering the most common words and visualizing their distribution, this analysis helps identify popular or frequently used terms in the dataset.
* This method serves as a foundational approach in Natural LanguageProcessing (NLP), which plays a crucial role in text analytics, content understanding, and even sentiment analysis.
* Understanding these trends in word usage allows for better comprehension of product descriptions, aiding in aspects like categorization, trend identification, and customer behavior analysis, offering substantial implications for marketing strategies and customer engagement in e-commerce or retail industries.

# 



# 5.Concluding Remarks

* Combining the outcomes of diverse algorithms applied to the e-commerce dataset reveals multi-faceted insights.
  + The Apriori association rules expose customer preferences for cohesive or matching products and certain seasonal items.
  + Decision Tree Regressor predicts prices based on quantities, allowing for better pricing strategies.
  + The K-Means clustering unravels distinctive customer segments, informing marketing and inventory decisions.
  + The sentiment analysis via word frequency exposes popular product themes,aiding inventory and marketing strategies.
* These findings offer a rich understanding of customer behaviors, allowing for tailored marketing, improved inventory management, and sharper business strategies in thee-commerce domain.

# References

1. [Kaggle(Dataset)](https://www.kaggle.com/datasets/krishnasairaj/e-commerce)
2. [Data Warehousing concepts](https://www.tutorialspoint.com/dwh/dwh_data_warehousing.htm)
3. [Python](https://docs.python.org/3/)
4. Aggarwal, C. C., & Gupta, S. C. (2009). Introduction to data mining: design and analysis. Prentice Hall.
5. Han, J., Kamber, J., & Pei, J. (2006). Data mining: concepts and techniques. Morgan Kaufmann.
6. Tan, P.-N., Kumar, V., & Srivastava, J. (2005). Introduction to data mining. Addison-Wesley.