Supermarket Purchase Prediction using Data Mining Techniques

Name: Harleen Nanwani
Roll No: 53
College: Vivekanand Education Society's Institute of Technology
Department: Information Technology

Abstract

This project focuses on predicting supermarket purchase behavior using various data mining techniques. It aims to uncover purchasing patterns, classify customer types, and generate insights through visualization, clustering, and association rule mining. The analysis was implemented using Python and RapidMiner.

Introduction

Supermarkets generate massive amounts of transaction data daily. Analyzing this data helps identify consumer patterns, popular products, and factors influencing sales. This project applies classification, clustering, and association rule mining techniques to understand and predict customer purchasing behavior.

System Design

Star Schema: Central fact table linked to dimension tables like Customer, Product, and Time.
Snowflake Schema: Normalized version of Star Schema where dimension tables are further split.

Methodology

1. Data Preprocessing: Cleaned missing values, converted categorical variables, and normalized numeric fields.
2. Exploratory Data Analysis (EDA): Visualized sales trends and customer distribution.
3. Classification: Implemented Decision Tree and Naïve Bayes algorithms to classify high-value purchases.
4. Clustering: Applied K-Means, Agglomerative, and DBSCAN to group similar customers.
5. Association Rule Mining: Used Apriori algorithm to find product co-purchase patterns.

Results and Discussion

The Naïve Bayes classifier achieved good accuracy for predicting high-value customers. The confusion matrix visualization confirmed correct classification for most test cases. Clustering revealed distinct customer groups based on spending behavior, and Apriori association rules showed strong relationships between frequently bought product lines.

Conclusion

The supermarket purchase prediction project effectively demonstrated the application of data mining techniques to real-world retail data. Through classification, clustering, and association analysis, the project provided insightful findings that can support marketing strategies and inventory planning.

References

1. Han, J., Kamber, M., & Pei, J. (2012). Data Mining: Concepts and Techniques.
2. Supermarket Sales Dataset – Kaggle Open Data Source.
3. Scikit-learn, Pandas, and Matplotlib Documentation.