DATA MINING PROJECT

Master in Data Science and Advanced Analytics

NOVA Information Management School

Universidade Nova de Lisboa

Data Mining (Bonus)

Group 25

Afonso Ferreira, 20240689

João Figueiredo, 20240853

Jorge Cordeiro, 20240594

Tomás Silva, 20230982

1st Semester 2024-2025

TABLE OF CONTENTS

1. INTRODUCTION

The development of this application represents an advancement in customer segmentation analysis. As part of a project focused on understanding customer behavior and preferences, this tool was designed to complement our work with an intuitive interface for exploring complex customer data. The application transforms static analysis into an interactive experience, allowing users to dynamically explore and understand customer segments.

Link to app: https://dm-group-25.streamlit.app/

Link to github: https://github.com/JoaoDargent/DataMiningInterface

1.1. Objective

The primary goal was to create a user-friendly platform that makes complex clustering analyses accessible and interpretable. Key objectives included:

- Providing an intuitive interface for exploring customer segmentation results

- Enabling visualization of multidimensional data

- Facilitating comparison between different clustering approaches

- Offering dynamic filtering capabilities for detailed segment analysis

1.2. Application Overview

The application is structured into three main sections:

1. EDA Raw Data: Provides access to exploratory data analysis visualizations
2. Segmentation & Clustering: Presents detailed clustering analyses with multiple methodologies
3. Final Clusterization: Offers an interactive 3D visualization tool for exploring final clusters

2. TECHNICAL IMPLEMENTATION

2.1. Technology Stack

The development of the interactive application was carried out using a robust and user-friendly technology stack, chosen for its compatibility with Python and its ability to handle data analysis and visualization tasks efficiently. The main technologies utilized include: Python, Streamlit (For building the interactive interface), Pandas (Data manipulation and preprocessing), Plotly(Interactive visualizations (3d scatter plot), Pickle (used to import the scaler and the preprocessed data).

2.2. Application Architecture

3. FEATURE IMPLEMENTATION

3.1. Exploratory Data Analysis (EDA)

The EDA module allows users to explore the dataset structure and contents. Key functionality is checking the distribution of every feature.

This module helps users prepare and understand the dataset before segmentation.

3.2. Segmentation & Clustering Analysis

This module shows pre-trained clustering models to each segmentation based on selected attributes. The clustering process includes:

- Hierarchical + Kmeans
- SOM + Kmeans
- SOM + Hierarchical
- DBSCAN
- Combined Results

The results provide insights into customer groups and their shared behaviors or demographics.

3.3. Interactive 3D Cluster Visualization

The final clustering visualization provides users with a plotly's 3D scatter plot to explore and analyze customer segments. This tool visually represents clusters by making the user select which features he wants on the x, y, and z axes. Each cluster is differentiated by color to make identification clear.

Users can interact with the plot by rotating it, zooming in, hovering and filtering the customer age.

For the customer age and all the data to be unscaled we dumped the scaler used on the preprocessing and used *inverse_transform* from the *Standard Scaler* scaler which undoes the scaling.

4. CONCLUSION

The project successfully integrates clustering analysis with an intuitive, interactive interface, meeting all project requirements. Challenges such as handling large datasets and module integration were resolved using Streamlit caching and efficient data handling techniques. Future improvements could be replacing the image graphs and switch them by dynamic plotly graphs, incorporate predictive models as in infering which cluster does an inserted via app customer belongs to, additional widgets for cluster exploration, and enhanced scalability for larger datasets. These advancements would further expand the application's utility and impact.