**INTERACTIVE DATA**

**FINAL PROJECT**

**-Divya Singireddy**

**Classifying and Grouping Products to Gain Marketing Knowledge**

**Overview**:

This research aims to explore the potential of clustering and classification approaches to enhance product strategy in corporate settings. Using data-driven insights, this study will enable business analysts, product managers, and marketing professionals to better understand customer preference, product segmentation, and overall product plans. It involves various machine learning techniques that group the products based on their features, using the "Product Classification and Clustering Dataset" that comes from the UCI Machine Learning Repository. The outcome can be used to help directly take decisions on product development, marketing campaigns, and product ordering/stock management. The final output of the project will be an interactive data visualization dashboard, which will allow the user to explore product clusters and make decisions based on that product cluster.

**Driving Question:**

How can product attributes effectively classify and cluster items to support marketing and inventory decisions, hence assisting firms in optimizing their product strategies?

This query illustrates the project's main goal, which is to demonstrate how companies may obtain insightful information by grouping products according to attributes like cost, features, and performance indicators. Businesses can find trends by putting comparable products together, which can help them with pricing, inventory management, marketing, and overall product strategy.

**Data Sources, Quality, and Structure:**

* **Data Source**: The dataset applied in the current project is UCI Machine Learning Repository's "Product Classification and Clustering Dataset". This dataset contains information about different products, including categories, prices, details of specifications, and performance indicators.

Link to Dataset: [UCI Machine Learning Repository](https://archive.ics.uci.edu/dataset/837/product+classification+and+clustering)

* **Data Quality:** The data set quality varies with missing values, potential outliers, and inconsistencies in the data. These problems could appear in the quality of the clustering results. The following measures were taken to ensure the reliability of the analysis:
  + Missing Data: The missing values were filled using the mean of the corresponding columns of numeric attributes. For categorical variables, the imputation methods such as most frequent could be considered in further

iterations.

* + Outliers: Outliers in the numeric columns were checked for through the Z-scores and removed to prevent distorted results in the clustering. In the first instance, a threshold of 3 Z scores was used but had to be adjusted to 4 after checking that the first removal of outliers did not leave any data behind.
  + Inconsistencies: Columns that contained other data types of entries than numeric were changed to numeric, if possible, otherwise left out of analysis.
* **Data Structure:** The dataset is in the form of a CSV file, where the attributes refer to product characteristics including: product categories, specifications, and performance metrics. Some of the key fields in this dataset are:
* Product Category: This refers to the kind or category of a product.
* Price: This refers to the price of the product.
  + Product Features: These include long descriptions of its specifications like dimensions, weight, and brand.
  + Performance Metrics: These refer to quantitative assessments of performance concerning user ratings and durability, etc.

This architecture affords the possibility of clustering by features that are numerical (e.g., price, performance metrics) and classification by features that are categorical (e.g., category of product).

**Data Gathering, Wrangling and Analysis:**

**Data Gathering**: The dataset was obtained from the UCI Machine Learning Repository and loaded into a panda DataFrame to see the overall view. Basic statistics were checked on the dataset to get an idea of how it looked and what went wrong.

**Data Wrangling**: There were some basic preprocessing steps done on the data before getting it ready for clustering. Here these steps are described:

* Missing Values Handling:
  + Missing numeric values were imputed based on column mean to ensure that the dataset remained complete and had no valuable information loss.
  + Non-numeric columns, which cannot be converted into numerical values, were excluded or individually handled by implementing custom imputation methods. • Outlier Detection and Removal:
  + Outliers were identified based on Z-scores, a measure of how far a data point is from the mean in terms of standard deviations. Any row containing a Z-score over 3 was removed. In instances where no data existed following the removal of outliers, the threshold for the removal was reduced to 4.
  + Z-scores are the traditional method of detecting outliers. However, IQR could also be used in future revisions.
* Feature Normalization:

The data was subjected to StandardScaler for normalization. This standardizes every feature (price, performance metric(s), etc.) to a similar scale, and hence clustering algorithms work more effectively by giving equal weightage to each feature in consideration.

**Clustering and Classification**: After feature normalization, the following technique of clustering was applied on the data to make a group of similar products under consideration:

**Principal Component Analysis (PCA):**

The dataset was reduced to two components using PCA, which simplified the visual

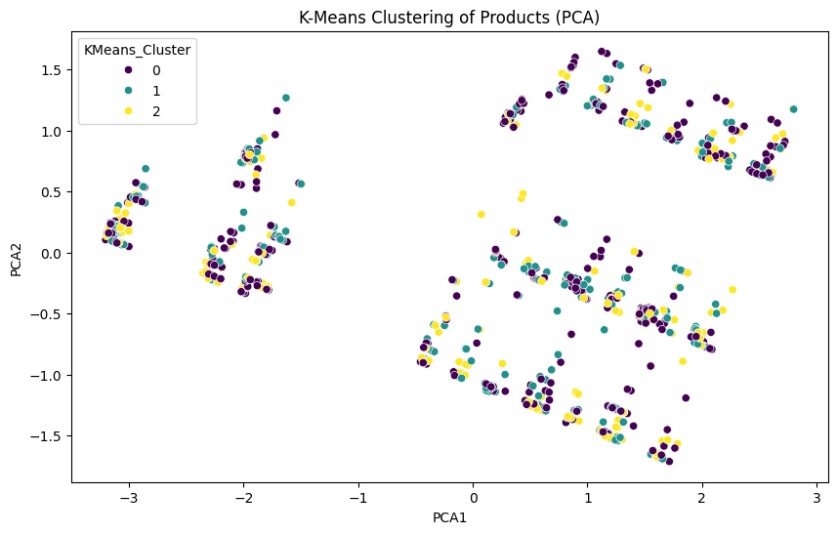
representation of the data and made it easier to see the structure of the product groups. In these high-dimensional datasets, this technique is particularly useful for reducing the number of variables in order to discover the underlying patterns.

The PCA-transformed data was added to the DataFrame for visualization purposes, with new columns representing the first and second principal components (PCA1 and PCA2).

**K-Means Clustering:**

* MiniBatchKMeans, an efficient variant of the traditional K-Means algorithm, was applied to the normalized features. MiniBatchKMeans uses smaller batches of data for faster computation, making it ideal for large datasets.
* The algorithm was set to find 3 clusters; this was an arbitrary choice. In future work, methods like the Elbow Method or Silhouette Score could be used for determining the optimal number of clusters.
* Cluster assignments were added to the DataFrame, so that each product could be labeled by its corresponding cluster.

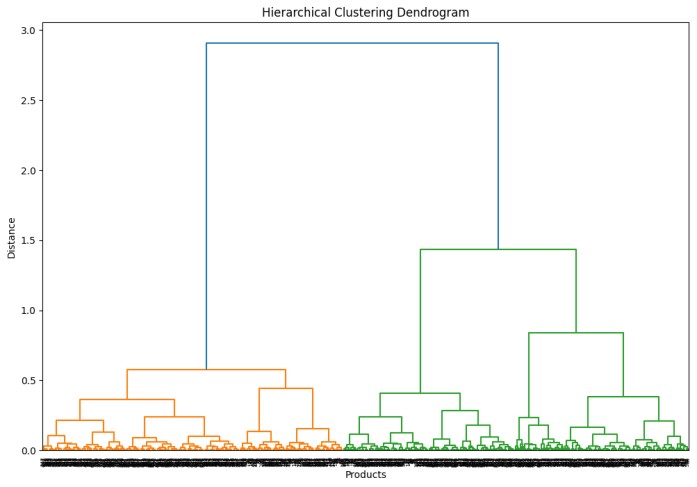
The results of K-Means Clustering are shown below:



**Hierarchical Clustering:**

* The hierarchical clustering applies Ward's method, which aims at minimizing the variance within clusters. The resulting dendrogram will be a graphical illustration of the nested grouping of products according to their similarities.
* Hierarchical clustering provides excellent insights into the relationships between the products as a way of understanding how products are hierarchically associated across different levels of similarity.

The results are shown below:



**Interactive Data Visualization:**

Dashboard Development: It has been developed as an interactive dashboard through Dash and Plotly. The dashboard will enable to explore the product clusters dynamically, providing a fun platform where business professionals can make data-driven decisions.

Key functionalities of the dashboard include:

* Dropdown Selector: Users can select the clustering method, such as K-Means or Hierarchical, from the dropdown menu, so that they can view varied results of clustering.
* Cluster Visualization: For this end, the first two PCA components are used to plot a scatter plot for displaying the identified product clusters of the selected clustering algorithm.
* Dendrogram for Hierarchical Clustering: For hierarchical clustering, users can get an idea about the hierarchical relationship among the product clusters through a dendrogram.

Interactive features of the dashboard have been powered through Plotly and Dash as the web application framework. Filtering, interaction, and exploration of data in real-time have become possible from the dashboard of the product to let users make informed decisions on segmentation and strategy.



**Application of Course Concepts:**

* Data Preprocessing: The preprocessing techniques involved in this project (processing missing data, removal of outliers, and normalization) have been a reflection of the basics in data analysis in order to make the dataset clean and prepared for analysis. With the standardization of numeric data, the analysis eliminates the differences in the scales of the features presented to prevent any feature dominating the outcome of the clustering.
* Clustering Techniques: Among the fundamental clustering algorithms implemented in this project was K-Means and Hierarchical Clustering. These clustering algorithms are important in shaping the profiles of similar products that might not be separately visualized through raw data alone.
* Dimensionality Reduction: Application of PCA helped to reduce the number of dimensions in the dataset as an approach toward easier visualization and correct interpretation. In most real-world applications, datasets contain a large number of variables.
* Data Visualization: The project applied key data visualization concepts in developing interactive dashboards to ensure that the visual representation of clusters was clear, dynamic, and user-centric-an essence of communicating insights to decision-makers in an intuitive way.

**Audiences' Needs and the Priorities of Design:**

The design of the dashboard as well as the analysis focuses on the following user needs:

* User-Centric Interface: The dashboard interface was designed simple and interactive so that users can, in their choice of clustering method, easily probe product clusters.
* Clear and Actionable Insights: The visualizations were designed to present clear, actionable insights into how products group together. This way, business professionals can enhance product strategies and campaign marketing.

The target audience for this project is mainly business analysts, product managers, and marketing specialists. Therefore, the delivery of the visualizations was not only informative but also relevant and actionable in improving the business processes.

**Limitations and Future Scope:**

* Data Size and Sampling: As the analysis is done on a sample of the data, which is 1000 products, some of the patterns existing within the complete dataset were probably lost. The full dataset could be analyzed for improved results in future.
* Outlier Detection: The outlier detection method through Z-score may not be appropriate for all types of data. Another type of outlier removal could be the IQR-based one by ensuring all valuable data is not discarded.
* Identifying Optimal Clusters: For K-Means, the number of clusters was chosen arbitrarily: 3 clusters. Future runs could incorporate additional methods, such as Elbow Method or Silhouette Score to determine an optimal number of clusters.
* Feature Engineering: Other feature engineering may allow better clustering performance. Domain-specific features, for example, could be added, as well as better encoding schemes for categorical features.

**Conclusion:**

This project really does show that clustering and classification techniques can really empower a business to optimize strategies. By putting the grouping of similar products on feature grounds into an interactive visual presentation, it allows for hidden insights to open up for businesses which drive marketing strategies, inventory management, and product development, using that interaction as input to make informed decisions that ultimately contribute to better product strategy and customer satisfaction.

**References**

Abdi, H., & Williams, L. J. (2010). Principal component analysis. *Wiley Interdisciplinary*

*Reviews: Computational Statistics*, *2*(4), 433–459. <https://doi.org/10.1002/wics.101>

*Dash Documentation & User Guide | Plotly*. (n.d.). Dash.plotly.com.

Retrieved November 26, 2024, from <https://dash.plotly.com/>

*UCI Machine Learning Repository*. (2023). Uci.edu.

<https://archive.ics.uci.edu/ml/datasets/Product+Classification+and+Clustering>

Xu, R., & WunschII, D. (2005). Survey of Clustering Algorithms. *IEEE Transactions on*

*Neural Networks*, *16*(3), 645–678. <https://doi.org/10.1109/tnn.2005.845141>