**🧩 Customer Segmentation Project – Detailed Summary**

**🔍 1. Project Objective**

The primary goal of this project was to **segment customers** based on their **demographics, purchasing behavior, and engagement**, enabling the business to:

* Understand different customer personas
* Design **targeted marketing strategies**
* Improve personalization and customer retention

**🗂️ 2. Data Exploration and Preprocessing**

**✅ Data Overview**

The dataset included:

* Demographics: Income, Education, Age\_Group, Family\_Size
* Spending: MntWines, MntFruits, MntGoldProds, etc.
* Behavior: Recency, NumWebPurchases, Total\_Purchase, etc.

**✅ Preprocessing Steps**

* Converted Dt\_Customer to datetime and calculated customer tenure
* Engineered features:
  + Spent = total amount spent
  + Family\_Size, Children, Is\_Parent
* Encoded categorical features using:
  + OrdinalEncoder for Education and Age\_Group
* Scaled numerical features using StandardScaler
* Removed outliers using the IQR method for Income and Total\_Purchase

**🧠 3. Feature Selection**

Final features used:

* Scaled numeric: Income, Spent, Gold, Total\_Purchase
* Encoded categorical: Education, Age\_Group

All transformations were encapsulated using a **ColumnTransformer pipeline** for clean workflow.

**🔬 4. Clustering Techniques Applied**

**📌 KMeans Clustering**

* Optimal clusters: **2 (determined via Elbow Method)**
* Silhouette Score: **0.500**
* Clear distinction:
  + **Cluster 0**: High-income, high-spending customers
  + **Cluster 1**: Budget-conscious segment
* Visualized using scatter plots, bar plots, and swarm plots

**🌳 Agglomerative Hierarchical Clustering**

* Linkage method: complete and average
* Dendrogram used to confirm **2 clusters**
* Silhouette Score: **0.537** ✅ (Best-performing)

**📈 DBSCAN (Density-Based Clustering)**

* Parameters: eps = 4, min\_samples = 24
* Detected **dense clusters** and **noise points**
* Silhouette Score: **0.527**
* Good separation + outlier detection

**🔁 Mean Shift Clustering**

* Bandwidth estimated via estimate\_bandwidth
* Clusters auto-detected
* Silhouette Score: **0.158** ⚠️
* Weak separation, possibly due to overlapping clusters or too many small segments

**🔍 Gaussian Mixture Model (GMM)**

* Flexible clustering with ellipsoidal shapes
* Soft assignment (probabilistic cluster membership)
* Silhouette Score: **0.488**
* Close performance to KMeans; useful when clusters slightly overlap

**🤖 5. Supervised Classification Model**

**✅ Purpose:**

To **predict customer segment (0 or 1)** using a classification model trained on features + KMeans cluster labels.

**✅ Steps:**

* Split data: train\_test\_split()
* Built pipeline: ColumnTransformer + GradientBoostingClassifier
* Trained on labeled cluster data
* Achieved strong accuracy (evaluated with classification\_report and .score())

**📦 6. Model Export & Deployment Prep**

* Saved final pipeline using pickle → classifier.pkl
* Model is deployment-ready (can be loaded for Streamlit/Flask apps)

**📈 7. Business Insights from Clusters**

| **Segment** | **Traits** | **Strategy** |
| --- | --- | --- |
| **Cluster 0** | High income, high spending, gold purchases, larger families | Target with loyalty programs, premium offers, and early access |
| **Cluster 1** | Lower income, more educated, budget-conscious | Offer bundle discounts, referral bonuses, and value campaigns |

**🧠 8. Final Cluster Comparison Summary**

| **Method** | **Silhouette Score** | **Notes** |
| --- | --- | --- |
| KMeans | 0.500 | Balanced, clear structure |
| Agglomerative | 0.537 | ✅ Best performer |
| DBSCAN | 0.527 | Flexible clusters + outlier detection |
| Mean Shift | 0.158 | Weak structure, many small blobs |
| GMM | 0.488 | Soft boundaries, flexible shape |

**🧠 9. Key Takeaways**

* Clustering revealed **two clear customer personas**: premium vs budget-conscious
* **Agglomerative Clustering** gave the best silhouette score
* DBSCAN added value by detecting noise
* Built a **reliable classifier** to predict customer segments
* Developed a **complete and deployable machine learning pipeline**

Top of Form

Bottom of Form