Industry Standard Documentation

Technical Requirements Document (TRD)

Data Sources : Mall Customers dataset

Technologies :

- Programming Language : Python
- Development Environment : Jupyter notebook
- Libraries and Tools :

Pandas : for data manipulation.

Numpy : for numerical operations.

Matplotlib , Seaborn, Matplotlib, Seaborn, Power BI : for visualization.

Scikit-learn : For building the model using machine learning algorithm, training and Evaluation.

Architecture :

- Data Collection :

Import the Mall Customers dataset

- Exploratory Data Analysis (EDA)**:**

Understanding the dataset through statistical summaries and visualizations.

- Data Preprocessing :

Handle missing values, normalize data formats, and remove any outliers to ensure data quality.

- Feature Engineering : Modify and select features to improve effectiveness of clustering

- Exploratory Data Analysis :

Visualize the statistical data to understand data distribution.
Identify patterns and gain insights from data.

- Clustering :

Apply K-Means clustering algorithm to segment customers.
Train the model on cleaned data.
Evaluate clusters using inertia and silhouette scores to improve their quality.

- Visualization :

Create plot to identify different customer segments.

Data Flow :

1. Import Mall Customers data :
   Load the Mall Customers dataset into the environment

2. Clean Data:

- Handle missing values, if any.
- Correct data types.
- Normalize or scale features if necessary.

3. Perform EDA for analysis :
    - Perform exploratory data analysis to understand the data distribution and relationships.
    - Use statistical methods and visualizations to gain insights.

4. Segment Customers using K-Means Clustering:
    - Apply clustering algorithms (e.g., K-Means) to segment customers based on selected features.

- Determine the optimal number of clusters using techniques like the elbow method or silhouette score.

5. Visualize Results:
    - Visualize the clusters using 2D and 3D plots.
    - Create detailed visualizations to present the findings.
    - Use Power BI for interactive dashboards and reports.

Performance considerations :

Optimize data processing and clustering to handle large datasets efficiently.
Intuitive visualizations to easily interpret the results.

Security and Compliance :

Ensure that the data is handles in accordance with privacy regulations.
Maintain accuracy and consistency throughout the process.