# **Project Report**

**Name:** Aravindh V  
 **Course:** B.Sc CS (AI & ML) 2nd year  
 **Registration Number:** TU6243202111006

### **Project Title:**

Customer Segmentation using K-Means Clustering

## **1. Project Description**

This project focuses on performing customer segmentation using the K-Means clustering algorithm. The aim is to group customers based on their annual income and spending score to identify different types of customers, such as high-value, moderate, and low-value segments. This type of clustering is widely used in marketing and business analytics to enable personalized strategies and better customer relationship management.

## **2. Learning Objectives**

* Objective 1: Understand how clustering works and determine the optimal number of clusters.
* Objective 2: Implement the K-Means algorithm and interpret the resulting clusters.
* Objective 3: Visualize and evaluate the clustering performance using metrics and plots.

## **3. Timeline**

* **Start Date:** October 16, 2025
* **Submission date:** October 16, 2025

## **4. Algorithm Used**

* Algorithm Name: K-Means Clustering
* K-Means is an unsupervised learning algorithm that partitions data into K distinct clusters. Each cluster is represented by a centroid, and each data point belongs to the cluster with the nearest centroid. It iteratively refines cluster centers to minimize the sum of squared distances between data points and their respective cluster centers. It is simple, efficient, and highly interpretable, making it ideal for customer segmentation tasks.

## **5. Tools & Libraries**

* **Programming Language:** Python
* Libraries Used:  
  • Pandas  
  • NumPy  
  • Scikit-learn  
  • Matplotlib  
  • Seaborn
  + Pandas
  + NumPy
  + Scikit-learn
  + Matplotlib / Seaborn
  + TensorFlow / PyTorch (if used)
  + [Any other relevant tool]

## **6. Dataset Description**

* Source: Synthetic dataset generated using Scikit-learn’s make\_blobs function.  
  Size: 300 rows × 2 features (Annual\_Income, Spending\_Score)  
  Target Variable: None (unsupervised learning)  
  Description: Each record represents a customer characterized by annual income and spending score.
* **Size:** [Number of rows and features]
* **Target Variable:** [If supervised learning]
* **Description of Features:** Provide a short description of key features.

## **7. Methodology**

* Data Preprocessing: Missing values were handled using column mean imputation, and features were scaled using StandardScaler.  
  Model Training: K-Means algorithm was applied to the scaled dataset with K determined by the Elbow and Silhouette methods.  
  Evaluation: Inertia and Silhouette scores were used to evaluate clustering performance.  
  Visualization: Scatter plots were used to display customer segments and cluster centers.
* **Model Training:** Split dataset (train/test), chosen model, training procedure.
* **Evaluation:** Metrics used (Accuracy, RMSE, AUC, F1-score, etc.).
* **Hyperparameter Tuning:** If performed, mention GridSearchCV/RandomSearch/Optuna.

## **8. Results**

* Optimal Number of Clusters: 4  
  Performance Metrics: The Silhouette score indicated well-separated clusters.  
  Insights: Customers were divided into four key groups representing different income and spending behaviors. The model effectively identified premium, moderate, and budget segments, aiding marketing strategies.
* **Visualizations:** Include confusion matrix, ROC curve, error distribution plots, feature importance, etc.
* **Insights:** Interpret results—what do they mean in the project’s context?

Attach screenshot of outputs and short description

## **9. Questions**

**Q1:** How many clusters are optimal?  
**Q2:** What is the elbow method?  
**Q3:** What defines each cluster?  
**Q4:** How to visualize clusters?  
**Q5:** What is inertia?  
**Q6:** How to scale features?  
**Q7:** What is silhouette score?  
**Q8:** How to interpret results?  
**Q9:** What is the business impact?  
**Q10:** How to use clustering in marketing?  
**Q11:** How to load and prepare the dataset?  
**Q12:** How to handle missing values?  
**Q13:** How to assign cluster labels to customers?  
**Q14:** How to describe cluster centers (centroids)?  
**Q15:** How to predict the segment of a new customer?

## **10. Challenges & Improvements**

* Challenges: Selecting the correct number of clusters, dealing with outliers, and interpreting overlapping groups.  
  Future Improvements: Use advanced clustering algorithms such as DBSCAN or Hierarchical Clustering for better accuracy. Incorporate more features like age, gender, or region for deeper insights.
* **Future Improvements:** Suggest improvements such as using larger datasets, advanced algorithms, ensemble models, feature engineering, or deep learning.

## **11. References**

* Dataset Link: Synthetic dataset generated internally.  
  Research Papers / Documentation:  
  - Scikit-learn documentation: https://scikit-learn.org/  
  - K-Means Clustering - Machine Learning Mastery
* **Research Papers / Documentation:** [Insert APA/IEEE references]

## **12. GitHub Link**

https://github.com/aravindhvinayagam2007-crypto/Customer-Segmentation