Name- MD ASHRAF **3 MONTH INTERN** **PROJECT REPORT**

1 Project Report: **Customer Segmentation**

1. Introduction

Customer segmentation helps businesses understand and categorize their customers based on behaviour and demographics, enabling personalized marketing strategies. This project aims to segment mall customers into distinct groups based on their spending and income characteristics using clustering techniques.

2. Dataset and Preprocessing

- **Dataset Used:** A synthetic dataset of 200 mall customers containing demographic and purchasing behaviour data such as age, gender, annual income, and spending score.
- **Data Characteristics:**
    - Features: Age, Gender, Annual Income, Spending Score.
    - Label: No explicit labels, as this is an unsupervised learning task.
- **Preprocessing Techniques:**
    - **Normalization:** Features were scaled to standardize data, aiding the performance of clustering algorithms.
    - **Handling Categorical Data:** Gender was encoded numerically.

3. Methodology

- **Feature Selection:** Annual Income and Spending Score were identified as the primary features impacting customer segmentation.
- **Clustering Algorithm:** K-Means Clustering was employed to segment customers.
    - **Elbow Method:** Used to determine the optimal number of clusters, found to be 5.
    - **Algorithm Details:**
        - Initialized with 5 centroids.
        - Iteratively assigned customers to the nearest cluster center.
        - Adjusted centroids based on the mean of assigned points.

4. Results and Analysis

- **Cluster Interpretation:**
    - Cluster 1: High spenders with high income – targeted for premium marketing strategies.
    - Cluster 2: Low spenders with high income – potential for upselling opportunities.
    - Cluster 3: Moderate spenders with moderate income – stable customer base.
    - Cluster 4: Low income, low spenders – require budget-friendly offerings.
    - Cluster 5: Young, average spenders – potential long-term loyal customers.

- **Visual Representation:** Cluster distribution was visualized using scatter plots, showing distinct groupings based on income and spending.

5. Conclusion

The segmentation revealed key customer groups within the mall's demographic, providing valuable insights for targeted marketing strategies. Future work includes enhancing segmentation with additional features such as shopping frequency, time spent in the mall, and integrating advanced clustering methods like Hierarchical or DBSCAN for improved segmentation.

2 .Project Report: Digit Recognition System

1. Introduction

The Digit Recognition project aims to develop a system that can accurately identify handwritten digits. This is essential in applications such as automated form processing, postal mail sorting, and digit recognition in various other fields. The objective is to train a machine learning model that can classify handwritten digits from 0 to 9 using image data.

2. Dataset and Preprocessing

- **Dataset Used:** The MNIST dataset, which consists of 70,000 images of handwritten digits (60,000 training images and 10,000 testing images).
- **Data Characteristics:** Each image is a grayscale 28x28 pixel array, and each pixel value ranges from 0 to 255, representing the intensity of the pixel.
- **Preprocessing Techniques:**
  - **Normalization:** The pixel values are scaled to a range of 0 to 1 to improve model performance.
  - **Reshaping:** Each image is reshaped to add a channel dimension, suitable for the CNN input.
  - **Label Encoding:** The target labels are one-hot encoded to represent the classes.

3. Feature Extraction and Model Development

- **Feature Extraction:** The raw pixel intensities were used as features for CNN training, capturing spatial hierarchies in images.
- **Model Used:** Convolutional Neural Network (CNN).
  - **Architecture:**
    - Input Layer: 28x28x1 input images.
    - Convolutional Layer: 32 filters of 3x3 size, activated by ReLU.
    - Max Pooling: 2x2 pooling layer to down-sample the feature maps.
    - Fully Connected Layer: A dense layer with 128 neurons activated by ReLU.
    - Output Layer: A softmax layer with 10 neurons representing each digit.
- **Training and Evaluation:** The model was trained using the Adam optimizer with a learning rate of 0.001, categorical cross-entropy loss, and a batch size of 128 over 10 epochs.

4. Results

- **Performance Metrics:**
  - Accuracy on test data: **98.5%**
  - Confusion Matrix: High accuracy across all classes, minor misclassifications between similar digits.
- **Visualizations:** Training and validation accuracy plots indicate strong performance and good convergence.

5. Conclusion

The CNN-based digit recognition system demonstrated high accuracy and efficiency in recognizing handwritten digits, making it suitable for deployment in real-world applications. Future improvements could include integrating advanced data augmentation techniques and experimenting with deeper architectures.