Republic of the Philippines
Laguna State Polytechnic University
Province of Laguna

Machine Problem No. 3			
Topic:	**Topic 3: Unsupervised Learning Techniques**	**Week No.**	4
Course Code:	**CSST102**	**Term:**	1st Semester
Course Title:	**Basic Machine Learning**	**Academic Year:**	2024-2025
Student Name		**Section**	
Due date		**Points**	

Machine Problem No. 4: K-Means Clustering on a Customer Segmentation Dataset

Objective:
The goal of this task is to assess your ability to apply the K-Means clustering algorithm to perform customer segmentation. You will preprocess the dataset, apply the K-Means algorithm, and evaluate the results. Additionally, you are required to visualize the clusters formed and discuss the characteristics of each group.

Dataset:
The provided **Customer Segmentation Dataset** includes customer demographic and purchasing behavior characteristics (e.g., age, income, spending score). The dataset contains no target labels.

Task Instructions:

1. **Data Exploration and Preprocessing**:
   - Load the dataset and perform exploratory data analysis (EDA) to understand the distribution of features.
   - Handle any missing values.
   - Normalize or scale the data if necessary.
   - Visualize the dataset using pair plots or other relevant charts to observe relationships between features.

2. **Model Development**:
   - Implement the **K-Means Clustering** algorithm to segment the customers into different groups. Start with **k=3** clusters.
   - Try different values of **k** (e.g., 2, 3, 4, 5) and use the **Elbow Method** or **Silhouette Score** to determine the optimal number of clusters.

3. **Model Evaluation**:
    - Evaluate the model using metrics such as **inertia** (sum of squared distances to centroids) and **silhouette score** to assess the quality of clusters.
    - Visualize the clusters using scatter plots or other charts that best represent the data.
    - Identify the characteristics of each cluster based on the input features (e.g., which customers belong to each group and why).

4. **Report and Visualizations**:
    - Provide a detailed report that includes:
        - The steps taken for preprocessing, model implementation, and evaluation.
        - Discussion on the chosen value of **k** and why it was selected.
        - Interpretation of the clustering results (i.e., the distinct customer segments and their characteristics).
        - Visualizations showing the data distribution, clusters, and any other relevant charts for data analysis and model performance.

Key Grading Areas:

- **Data Preprocessing** (20%): Proper cleaning, handling of missing values, normalization, and visualization of data.
- **Model Implementation** (40%): Correct implementation of K-Means algorithm, determination of optimal **k**, and clustering analysis.
- **Model Evaluation** (20%): Use of appropriate metrics like inertia and silhouette score, clear visualization of clusters.
- **Critical Thinking** (10%): Interpretation of results and understanding of customer segments.
- **Report Quality and Visualizations** (10%): Well-organized report with clear documentation and supporting visualizations.

Inability to follow this instruction will be deducted 5 points each for filename format and late submission per day. Also, cheating and plagiarism will be penalized.

Rubric for K-Means Clustering Assessment Task:

Criteria	Excellent (90-100%)	Good (75-89%)	Satisfactory (60-74%)	Needs Improvement (0-59%)
Data Preprocessing	Data is thoroughly cleaned, normalized, and visualized; no missing values; all EDA performed.	Most preprocessing steps correctly implemented; minor issues in EDA or missing value handling.	Basic preprocessing with some steps missed; minor errors in EDA.	Poor or missing preprocessing; significant issues in data cleaning or visualization.
Model Implementation	K-Means clustering implemented accurately with optimized **k**; code is efficient, well-organized.	Model implemented with minor errors or lack of optimization; code generally organized.	Basic model implementation; errors in code; limited optimization.	Poor or incorrect model implementation; models do not run or produce meaningful results.
Model Evaluation	Comprehensive evaluation using metrics and visualizations; insightful interpretation of results.	Good evaluation: metrics calculated correctly but minor issues in interpretation.	Basic evaluation with limited analysis and missing metrics.	Minimal or missing evaluation; metrics not calculated or interpreted.
Critical Thinking	Deep analysis of clusters; insightful discussion of characteristics and real-world applicability.	Good analysis with some insights but lacking depth.	Basic analysis with limited insights; superficial discussion of clusters.	Minimal or no critical thinking; poor or no interpretation of clusters.
Report and Visualizations	Well-organized, clear report with effective visualizations supporting the analysis.	Report is organized with minor issues; visualizations present but may not fully support the analysis.	Basic report with limited visualizations; minimal documentation.	Unclear, disorganized, or incomplete report; missing visualizations.