Republic of the Philippines

Laguna State Polytechnic University

Province of Laguna

Exercise 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**	

Exercises for K-Nearest Neighbors (KNN) and Logistic Regression on Breast Cancer Diagnosis Dataset

Exercise 1: Data Exploration and Preprocessing

1. **Load the Dataset**
   Load the **customer_segmentation.csv** dataset into your preferred programming environment.

2. **Data Exploration**
   - Display the first few rows of the dataset.
   - Check for missing values in the dataset. If there are any, handle them appropriately.
   - Explore the distribution of features such as Age, Annual Income, and Spending Score using histograms or box plots.

3. **Data Normalization**
   - Normalize or standardize the numerical columns (Age, Annual Income, Spending Score) to ensure all features have equal weight during clustering.

Exercise 2: Implementing K-Means Clustering

1. **Initial Model Implementation**
   - Implement the K-means algorithm on the dataset with **k=3** clusters. Use an appropriate library such as scikit-learn in Python.

2. **Choosing Optimal k**
   - Experiment with different values of **k** (e.g., 2, 3, 4, 5).
   - Use the **Elbow Method** to determine the optimal number of clusters. Plot the within-cluster sum of squares (inertia) for each value of **k**.

3. **Cluster Visualization**
   - Visualize the clusters in a 2D scatter plot based on any two features (e.g., Annual Income vs. Spending Score).
   - Assign colors to distinguish different clusters.

Exercise 3: Model Evaluation

1. **Silhouette Score**
   - Calculate the **silhouette score** for each value of **k** (e.g., 2, 3, 4, 5) and determine which value of **k** yields the best clustering result.

2. **Cluster Analysis**
   - Identify the characteristics of each cluster. For example:
     - Which group tends to have the highest Annual Income?
     - Which group has customers with the lowest Spending Score?

Exercise 4: Interpretation and Reporting

1. **Cluster Interpretation**
   - Provide a brief interpretation of what each cluster represents. For example, a cluster may represent high-income, low-spending customers or young, high-spending customers.

2. **Report**
   - Write a report summarizing:
     - The data exploration process.
     - The results of the K-means clustering and the optimal value of **k**.
     - The characteristics of each cluster.
     - Any insights or observations from the clustering analysis.

3. **Visualizations**
   - Include relevant visualizations such as the Elbow Method plot, silhouette scores, and cluster scatter plots in your report.

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 Machine Problem

Criteria	Excellent (90-100%)	Good (75-89%)	Satisfactory (60-74%)	Needs Improvement (0-59%)
Data Preprocessing	Thorough data cleaning, normalization, and handling of missing values. Dataset is well prepared for clustering.	Data is cleaned and normalized, but minor errors in handling missing values or scaling.	Some preprocessing steps are missed or done incorrectly, such as improper scaling or missing value handling.	Minimal or no preprocessing. Significant issues with missing values, scaling, or data integrity.
Exploratory Data Analysis (EDA)	Extensive use of plots (histograms, box plots) and insightful interpretation of the data.	Appropriate plots and analysis, though lacking some depth in interpretation.	Basic plots provided, but limited exploration of the data.	No or very minimal visual exploration and analysis of the dataset.
K-Means Implementation	K-means clustering implemented correctly with efficient code. Used multiple values of **k** and justified the final choice using the Elbow Method or Silhouette Score.	K-means clustering implemented correctly, but lacks thorough optimization or justification for final **k** value.	Basic implementation of K-means. Some issues in the process of selecting **k** or clustering.	Incorrect or poor implementation of K-means, without optimization or justification of **k**.
Cluster Visualization	Clear, meaningful visualizations of clusters (e.g., 2D scatter plot) that clearly distinguish different groups.	Visualizations provided but could be clearer or lack depth in explanation.	Visualizations are present but do not effectively convey cluster separation.	Poor or missing visualizations. Clusters are not represented clearly.
Optimal k Selection	Elbow Method and Silhouette Score used accurately to determine the best value of **k**. Clear explanation of the process.	Elbow Method or Silhouette Score used but lacks depth in analysis.	Basic attempt to find optimal **k**, but may not be well-justified or fully explained.	No or incorrect method used for determining the optimal number of clusters.
Model Evaluation	Thorough evaluation using silhouette scores, inertia, and other relevant metrics. Clear and insightful explanation of cluster quality.	Evaluation provided with all key metrics, but with less in-depth analysis of the results.	Basic evaluation with limited metrics or explanation of cluster quality.	Minimal or missing evaluation. Poor or incorrect use of evaluation metrics.
Cluster Interpretation	Detailed and meaningful interpretation of the customer segments. Clear understanding of the characteristics of each cluster.	Good interpretation of customer segments but lacks depth in describing each cluster's characteristics.	Basic interpretation of clusters with limited insights on customer segments.	Poor or incorrect interpretation of clusters. No meaningful insights gained from the clustering.
Report Quality	Clear, well-organized, and professional report. All steps in the process are well-documented with supporting visualizations and analysis.	Report is well-organized, but may lack depth in certain sections or explanations. Visualizations are present but not fully integrated.	Basic report, but some sections lack clarity or organization. Visualizations may not be fully explained.	Disorganized or incomplete report. Lacks essential steps, analysis, or visualizations.
Critical Thinking & Insights	Demonstrates a deep understanding of K-means clustering and its practical implications. Offers insightful conclusions about customer segments and their real-world applicability.	Shows a good understanding of K-means with some insights into the clusters, though not fully developed.	Basic understanding of K-means with limited real-world implications discussed.	Minimal or no critical thinking displayed. Clusters not analyzed in practical or meaningful terms.