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. 3: K-Nearest Neighbors (KNN) Classifier on Breast Cancer Diagnosis Dataset

Objective:
The goal of this task is to assess your ability to apply the K-Nearest Neighbors (KNN) algorithm to predict breast cancer diagnosis based on tumor characteristics. You will preprocess the dataset, implement the KNN algorithm, and evaluate its performance. You are also required to compare the KNN model with a Logistic Regression model.

Dataset:
The provided **Breast Cancer Diagnosis Dataset** includes tumor characteristics such as radius, texture, perimeter, and area, along with the diagnosis label (malignant or benign).

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.
   - Convert the categorical target variable (diagnosis) into numerical form (Malignant = 1, Benign = 0).
   - Normalize or scale the data as required.
   - Split the data into training (80%) and testing (20%) sets.

2. **Model Development**:
   - Implement the **K-Nearest Neighbors (KNN)** algorithm to classify the tumor diagnosis.
   - Train the model with appropriate hyperparameters (start with n_neighbors=3).
   - Compare the performance of the KNN model with a **Logistic Regression** model.

3. **Model Evaluation**:
    - Evaluate both models using accuracy, precision, recall, and F1-score.
    - Present the confusion matrix for both models.
    - Discuss the performance of each model and explain which model performs better for this task.

4. **Report and Visualizations**:
    - Provide a detailed report that includes:
        - The steps taken for preprocessing, modeling, and evaluation.
        - The comparison between KNN and Logistic Regression.
        - Confusion matrices and classification reports for both models.
    - Visualize the decision boundaries (if possible) and display any relevant charts for data distribution and model performance.

Key Grading Areas:

- **Data Preprocessing**: Ensures that the dataset is correctly cleaned and prepared for model training.
- **Model Implementation**: Correctly implementing and optimizing the KNN and Logistic Regression models.
- **Model Evaluation**: Measuring and comparing model performance based on metrics such as accuracy and F1-score.
- **Critical Thinking**: Demonstrating a strong understanding of model performance and its practical implications.
- **Report Quality**: Presenting findings clearly and supporting them with appropriate 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.

Republic of the Philippines
Laguna State Polytechnic University
Province of Laguna

Rubric for K-Nearest Neighbors (KNN) Classifier Assessment Task

Criteria	Excellent (90-100%)	Good (75-89%)	Satisfactory (60-74%)	Needs Improvement (0-59%)
Data Preprocessing (20%)	Data is thoroughly cleaned, normalized, and processed; all ssig values handled appropriately; dataset split correctly for training and testing.	Most preprocessing steps correctly implemented; minor issues in handling missing values or feature scaling; appropriate train-test split.	Basic preprocessing; some steps missed or handled poorly (e.g., missing values or scaling issues); incorrect or imbalanced train-test split.	Poor or missing preprocessing; significant issues in handling missing values, data scaling, or improper train-test split.
Model Implementation (40%)	KNN and Logistic Regression models implemented accurately; hyperparameters optimized appropriately; code is efficient, well-organized, and well-documented.	Models implemented correctly but with minor errors or lack of optimization; code generally organized and readable.	Basic model implementation; noticeable errors or inefficient code; little or no hyperparameter optimization attempted.	Poor or incorrect model implementation; disorganized or ineffective code; models do not run or produce meaningful results.
Model Evaluation (20%)	Comprehensive evaluation of both models using accuracy, precision, recall, F1-score, and confusion matrices; insightful interpretation of results.	Good evaluation; all required metrics calculated correctly; some minor issues in interpretation or missing depth in the analysis.	Basic evaluation; some metrics missing or calculated incorrectly; limited interpretation and analysis of model performance.	Minimal or missing evaluation; metrics not calculated or interpreted; no meaningful analysis of model performance.
Critical Thinking and Insights (10%)	Deep analysis comparing KNN and Logistic Regression; insightful discussion on model strengths, weaknesses, and real-world applicability.	Good comparison between models; some analysis of strengths and weaknesses, though lacking depth.	Basic comparison between models; limited analysis or incorrect conclusions drawn about model performance.	Minimal or no critical thinking; no meaningful comparison between models or incorrect conclusions drawn.
Report Quality and Visualizations (10%)	Report is well-organized, clear, and professional; all steps documented; visualizations (confusion matrices, charts) effectively support the analysis.	Report is organized and clear with minor issues; most steps documented; visualizations present but may not fully support the analysis.	Basic report; lacks organization or depth; limited or unclear visualizations; minimal documentation of steps taken.	Report is unclear, disorganized, or incomplete; visualizations are missing or do not support the analysis; little to no documentation.