Republic of the Philippines
Laguna State Polytechnic University
Province of Laguna

Machine Problem No. 2			
Topic:	**Topic 2: Supervised Learning Fundamentals**	**Week No.**	3
Course Code:	**CSST102**	**Term:**	1st Semester
Course Title:	**Basic Machine Learning**	**Academic Year:**	2024-2025
Student Name		**Section**	
Due date		**Points**	

Machine Problem No. 1: Predicting House Prices with Multiple Regression

Objective: To apply multiple regression techniques learned in the course to predict house prices based on various factors such as location, size, number of bedrooms, and age of the property. This problem simulates a real-world scenario where you need to create a predictive model that can assist real estate agents in estimating house prices more accurately.

Problem Statement: You are a data scientist working for a real estate company. Your task is to develop a predictive model that can estimate the prices of houses in a particular city. The dataset provided includes multiple features, such as the size of the house (in square feet), the number of bedrooms, the age of the house, the proximity to downtown (in miles), and the house price. Your goal is to build a multiple regression model that can accurately predict house prices based on these features.

Dataset Overview: The dataset consists of the following columns:

- **Size (sq. ft.):** Size of the house in square feet.
- **Bedrooms:** Number of bedrooms in the house.
- **Age:** Age of the house in years.
- **Proximity to Downtown (miles):** Distance of the house from the downtown area.
- **Price:** Actual price of the house (in thousands of dollars).

Task Instructions:

1. Data Exploration and Visualization:

- Perform an exploratory data analysis (EDA) to understand the relationships between the features and the house prices.
- Visualize the data using scatter plots, histograms, and correlation matrices to identify potential patterns and relationships.

2. **Data Preprocessing:**
   - Handle any missing data appropriately.
   - Normalize or standardize the data to ensure that all features are on a similar scale.
   - Encode any categorical variables if necessary.

3. Model Development:
   - Implement a multiple regression model using Python (you may use libraries such as Scikit-learn for this task).
   - Train the model on a training set (use a 70-30 split between training and test data).
   - Perform feature selection if necessary to identify the most significant predictors.

4. Model Evaluation:
   - Evaluate the model's performance using metrics such as Mean Squared Error (MSE), R-squared, and Adjusted R-squared.
   - Provide a detailed interpretation of the model coefficients and their significance.
   - Plot the predicted prices against the actual prices to visualize the model's accuracy.

Report:
   - Document your process, including the steps you took for data preprocessing, model development, evaluation, and any attempts at model improvement.
   - Discuss the challenges faced and how you overcame them.
   - Include visualizations and plots that support your findings.
   - Conclude with a discussion on the model's applicability in real-world scenarios and any potential limitations.

Submission Requirements:
   - Python script (.py file) or Google Colab/ Jupyter Notebook (.ipynb file) with your code implementation.
   - Report document (PDF format) detailing your analysis, findings, and conclusions.
   - Upload the files to the GitHub repository as instructed.
   - Filename Format: **2A-BERNARDINO-EXER1**

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 Laboratory Machine Problem: Predicting House Prices with Multiple Regression

Criteria	Excellent (90-100%)	Good (75-89%)	Satisfactory (60-74%)	Needs Improvement (0-59%)
Data Exploration and Visualization	Comprehensive exploration with insightful analysis; visualizations are clear, well-labeled, and effectively convey patterns and relationships.	Good exploration: visualizations are mostly clear and relevant, with minor issues in labeling or interpretation.	Basic exploration: visualizations are present but may lack clarity or depth in analysis.	Minimal or no exploration; visualizations are missing, unclear, or irrelevant.
Data Preprocessing	Data is thoroughly cleaned and preprocessed; all necessary steps, including handling missing values and scaling, are correctly implemented.	Good preprocessing with minor issues; most necessary steps are implemented correctly.	Basic preprocessing: some steps are missing or incorrectly implemented.	Poor or no preprocessing; significant issues in handling data or missing steps.
Model Development	Model is accurately implemented with appropriate feature selection; code is efficient and well-organized; experimentation with advanced techniques (optional) is well-executed.	Model is implemented with minor issues; code is generally efficient; some feature selection is attempted.	Basic model implementation; some errors in code or feature selection; lacks advanced experimentation.	Poor or incorrect model implementation; code is disorganized or ineffective; no feature selection.
Model Evaluation	Thorough evaluation with insightful interpretation of metrics (MSE, R-squared, etc.); visualizations effectively illustrate model performance.	Good evaluation: metrics are correctly calculated but with minor issues in interpretation; visualizations are mostly clear.	Basic evaluation; some errors in metrics calculation or interpretation; visualizations are present but not fully clear.	Minimal or incorrect evaluation; metrics are missing or incorrectly calculated; visualizations are unclear or missing.
Critical Thinking and	Demonstrates strong problem-solving skills;	Good problem-solving skills;	Basic problem-solving; applies concepts but	Weak problem-solving; poor

Criteria	Excellent (90-100%)	Good (75-89%)	Satisfactory (60-74%)	Needs Improvement (0-59%)
Problem-Solving	effectively applies multiple concepts to address the problem; creative solutions and critical analysis are evident.	applies learned concepts effectively with minor issues; some evidence of critical analysis.	with errors or omissions; limited critical analysis.	application of concepts; little to no critical analysis.
Report Quality	Report is well-organized, clear, and professional; thoroughly documents the process, findings, and conclusions; visualizations are well-integrated and support the narrative.	Report is organized and clear with minor issues; documentation is mostly complete; visualizations are included but not fully integrated.	Basic report; somewhat clear but may lack organization or completeness; visualizations are present but may not support the narrative well.	Report is unclear, disorganized, or incomplete; visualizations are missing or do not support the narrative.