Republic of the Philippines
Laguna State Polytechnic University
Province of Laguna

Machine Problem No. 2			
Topic:	**Fundamentals of Machine Learning**	**Week No.**	
Course Code:	**CSST102**	**Term:**	1st Semester
Course Title:	**Basic Machine Learning**	**Academic Year:**	2025-2026
Student Name	Capili, Judeelyn M.	**Section**	BSCS 3A
Due date		**Points**	

Fundamentals of Machine Learning

Topic: What is ML? Types of ML and Core Challenges

Objectives

By the end of the lab, students should be able to:

1. Load and explore a dataset using Scikit-Learn.
2. Perform a train-test split to prepare data.
3. Train a simple baseline ML model.
4. Evaluate model performance using metrics.
5. Relate the task to supervised vs. unsupervised learning.

Lab Outline (3 hours)

Hour 1 – Setup & Dataset Exploration

- Install/verify Python, Jupyter/Colab, and Scikit-Learn.
- Load the **Iris dataset** (classification) or **California Housing dataset** (regression).
- Explore dataset (features, targets, summary statistics).

In this laboratory exercise, I verified my Python setup and installed essential machine learning libraries such as Scikit-Learn, NumPy, and Pandas. I then worked with the California Housing dataset, which is designed for regression tasks. The dataset represents different housing districts in California, with each record containing valuable information about demographic and geographic factors that influence housing prices. Specifically, there are 20,640 samples, each with eight numerical features: median income, house age, average number of rooms, average number of bedrooms, population, average occupancy, latitude, and longitude. The output or target label is the median house value, expressed in units of $100,000.

The dataset was automatically downloaded using Scikit-Learn's built-in function and saved locally as housing.csv for easier reuse. After loading the file, I explored the data by viewing the first few rows, column names, and summary statistics. From this exploration, I observed that the median income feature had the strongest correlation with house value, while location (latitude and longitude) also significantly influenced pricing patterns. The dataset is an example of supervised learning, since it contains both input features and their corresponding output values that the model can learn from.

Mini-task: Students answer:

- What is the **input (features)**?

  The input features are median income, house age, average number of rooms, average number of bedrooms, population, average occupancy, latitude, and longitude.

- What is the **output (label)**?

  The output is the median house value of each district (in $100,000 units).

- Is this **supervised or unsupervised learning**?

  This task is an example of Supervised Learning – Regression, because the dataset provides labeled output values that the model learns to predict based on input features

Hour 2 – Train-Test Split & Baseline Model •

Perform train-test split (80% train, 20% test).

- Train a simple baseline model:
  - Logistic Regression (for Iris) ○ Linear Regression (for Housing)
- Make predictions.

Mini-task: Students compute model accuracy.

After completing dataset exploration, I prepared the data for model training by dividing it into 80% training data and 20% testing data using Scikit-Learn's train_test_split() function. This ensures that the model learns patterns from one portion of the data and is tested on unseen examples to evaluate its predictive performance.

For the model, I selected Linear Regression, a simple and interpretable algorithm that works well for establishing relationships between continuous variables. The model was trained on 16,512 training samples and tested on 4,128 testing samples. During training, the Linear Regression model calculated the best-fitting line that minimized the difference between the actual and predicted median house values. After training, the model made predictions on the test set, estimating housing prices based on the given input features.

This step demonstrates a key part of the machine learning process, building a baseline model that establishes reference performance. More advanced models like Decision Trees or Random Forests could later be compared to this baseline to determine whether they perform better.

Hour 3 – Evaluation & Reflection

- Evaluate model with different metrics:
  - Classification: Confusion matrix, precision, recall.
  - Regression: RMSE (Root Mean Squared Error).
- Discuss ML challenges: overfitting, underfitting, and bad data.
- Students reflect:
  - "What would happen if the dataset had missing or wrong values?" ○ "How does this relate to real-world ML applications?"

To evaluate the model's performance, I used the Root Mean Squared Error (RMSE), a common metric for regression tasks. RMSE measures the average difference between predicted and actual values, with lower values indicating better performance. The model achieved an RMSE of 0.722, meaning that on average, the predictions were off by approximately $72,200. While this level of error is acceptable for a simple linear model, it also shows that the relationship between features and housing prices is not purely linear. More complex models might reduce this error further.

Several important machine learning challenges were observed during this evaluation. The first is overfitting, where a model learns the training data too well and fails to generalize to new, unseen data. The second is underfitting, which occurs when a model is too simple, as Linear Regression might be in this case, failing to capture more complex patterns or interactions between features. The third issue is bad data — missing, incomplete, or incorrect values that can distort the model's understanding of relationships and lead to inaccurate predictions.

If the dataset had missing or wrong values, the model's performance would deteriorate. For example, missing entries in the median income or house age columns could lead to biased predictions or failed training. This highlights how critical data quality and preprocessing are in machine learning.

In real-world applications, regression models like this one can be used by real estate developers, financial analysts, and policy planners to estimate property values, understand housing market trends, or make data-driven investment decisions. However, success in these applications depends not only on model design but also on clean, accurate, and representative data.

Deliverables (Lab Submission)

1. Python notebook (Jupyter/Colab) with:
    - Dataset loading & exploration ○ Train-test split ○ Model training & evaluation
2. Short reflection (3–5 sentences):
    - What ML type did you use?
    - What challenge might affect the model?

    In this laboratory activity, I applied Supervised Learning – Regression using a Linear Regression model trained on the California Housing dataset. The model achieved an RMSE of 0.722, indicating that its predictions deviated by about $72,200 on average. The main challenges identified were overfitting, underfitting, and the effects of missing or inaccurate data. If the dataset contained incomplete or wrong values, the model's accuracy would drop significantly. Through this activity, I learned how to load a dataset, perform a train-test split, train a regression model, and evaluate its performance using RMSE, gaining insight into the importance of data quality in real-world ML systems.

Assessment (30 points)

- Dataset Exploration (5 pts)
- Train-Test Split (5 pts)
- Baseline Model Training (10 pts)
- Evaluation Metrics (5 pts)
- Reflection/Discussion (5 pts)

Total: 30 points