Republic of the Philippines
Laguna State Polytechnic University
Province of Laguna

Machine Problem No. 1

Topic:	**Fundamentals of Machine Learning**	**Week No.**	2
Course Code:	**CSST102**	**Term:**	1st Semester
Course Title:	**Basic Machine Learning**	**Academic Year:**	2025-2026
Name:	**Shaila Patrice D. Avellaneda**		
Section	**BSCS 3A**		

Fundamentals of Machine Learning

Lab Outline (3 hours)

Hour 1 – Setup & Dataset Exploration

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

```
from sklearn.datasets import load_iris
import pandas as pd


iris = load_iris(as_frame=True)
df = iris.frame
print(df.head())


print(df.describe())
print("Target classes:", iris.target_names)
```

Mini Task: Student answer:

- What is the **input (features)?**
  - The input or features of the dataset are the measurable characteristics of each iris flower. These include *sepal length (cm)*, *sepal width (cm)*, *petal length (cm)*, and *petal width (cm)*. These numerical features describe the physical dimensions of the flowers and are used as the independent variables that help the model identify patterns in the data. In the DataFrame, these columns serve as the predictors that the model will analyze to classify the iris species.
- What is the **output (label)?**
  - The output or label of the dataset is the species of the iris flower, represented by the target column. This target variable indicates which species each flower belongs to, with numerical values assigned as follows: 0 for *setosa*, 1 for *versicolor*, and 2 for *virginica*. The label is the dependent variable that the model aims to predict based on the given feature values. It provides the correct

classifications that the model uses during training to learn how to make accurate predictions.

- Is this **supervised or unsupervised learning?**
    - This task is an example of **supervised learning** because the dataset includes both input features and known output labels. In supervised learning, the model learns from these labeled examples by finding relationships between the features and the corresponding target values. The main goal is for the model to generalize this knowledge so it can accurately predict the correct class (species) when given new, unseen data.

Hour 3 – Evaluation & Reflection

```
from sklearn.metrics import confusion_matrix, classification_report
print(confusion_matrix(y_test, y_pred))
print(classification_report(y_test, y_pred))
```

- Evaluate with different metrics:
    - Classification: Confusion matrix, precision, recall.
    - Regression: RMSE (Root Mean Squared Error).
        - **Answer:** The model is evaluated using metrics such as the confusion matrix, precision, and recall. The confusion matrix shows correct and incorrect predictions, while precision and recall measure how accurate and complete the model's classifications are. These metrics help assess the model's overall performance beyond just accuracy.
- Discuss ML challenges: overfitting, underfitting, and bad data.
        - **Answer:** Common machine learning challenges include overfitting, underfitting, and bad data. Overfitting means the model learns the training data too well and fails on new data, while underfitting means it's too simple to capture patterns. Bad or missing data can also lead to poor results.
- Students reflect:
    - "What would happen if the dataset had missing or wrong values?"
    - "How does this relate to real-world ML applications?"
        - **Answer:** If the dataset had missing or incorrect values, the model's accuracy would decrease because it would learn from flawed information. In real-world applications, this shows why data quality, proper evaluation, and model tuning are essential for reliable machine learning outcomes.

Reflection: Short Reflection (3-5 sentences):

For this activity, I used **supervised machine learning** because the model was trained using labeled data from the Iris dataset, where each set of flower measurements has a known species. One challenge that might affect the model is **overfitting**, which happens when the model performs very well on training data but poorly on new, unseen data. Another challenge could be **bad or missing data**, which can lead to inaccurate predictions. Ensuring clean data and proper model evaluation helps maintain accuracy and reliability in real-world applications.