Borce, Ron Ken S. | BSCS 3B | Basic Machine Learning

Machine Problem 1

Mini Task – Setup & Dataset Exploration

- Input (features): The input features are the measurements of the iris flowers: 'sepal length (cm)', 'sepal width (cm)', 'petal length (cm)', and 'petal width (cm)'.
- Output (label): The output label is the 'species' of the iris flower, which is what you are trying to predict or classify.
- Supervised or unsupervised learning: This is a supervised learning task because the dataset includes the 'species' label for each data point. Supervised learning involves training a model on labeled data to predict the output for new, unseen data.

Student Reflect – Evaluation & Reflection

What would happen if the dataset had missing or wrong values?

Incomplete or incorrect data—often referred to as "dirty data"—can seriously degrade the effectiveness of a machine learning model. Since most algorithms aren't built to manage missing values, introducing such data can cause:

• Training Errors: Models may crash or refuse to run when they encounter missing entries, as seen with errors like (ValueError: Input X contains NaN).

• Skewed Predictions: If gaps in the data aren't properly addressed, the model may learn false patterns, resulting in biased outputs.

• Lower Accuracy: Even if the model runs, dirty data typically leads to weaker performance and less reliable results.

• Distorted Insights: Incorrect entries—such as typos or faulty inputs—can mislead the model, producing inaccurate conclusions.

How does this relate to real-world ML applications?

In practical machine learning scenarios, datasets are often messy due to their diverse origins. They may include missing data, inconsistent formatting, or incorrect entries. That’s why data cleaning and preprocessing are vital steps in the ML workflow.

Preparing data thoroughly—by addressing missing values, correcting errors, and standardizing formats—is essential before training a dependable model. The quality of your input data directly influences how well your model performs in real-world applications. Clean data leads to more accurate, stable, and trustworthy predictions.

In fact, data preparation often demands more time and effort than the modeling itself. Tackling flawed or incomplete data is a core responsibility in applied machine learning and key to building models that truly deliver.

Short reflection (3–5 sentences):

o What ML type did you use?

I applied supervised learning through a classification approach, since the dataset includes labeled entries indicating the iris flower species.

I'm using a classification model—specifically, a Decision Tree Classifier—to predict the 'ocean_proximity' feature based on other variables in the housing dataset.

o What challenge might affect the model?

One challenge that could impact the model's performance is the feature overlap between certain species—especially 'versicolor' and 'virginica'—as observed in the scatter plot. This similarity in feature distribution may hinder the model's ability to clearly differentiate between the two classes.

A potential issue for the model is managing the categorical 'ocean_proximity' feature if it's used as an input rather than the target variable. Moreover, missing data in the numerical attributes could create problems during training if not properly addressed beforehand.