ASSIGNMENT

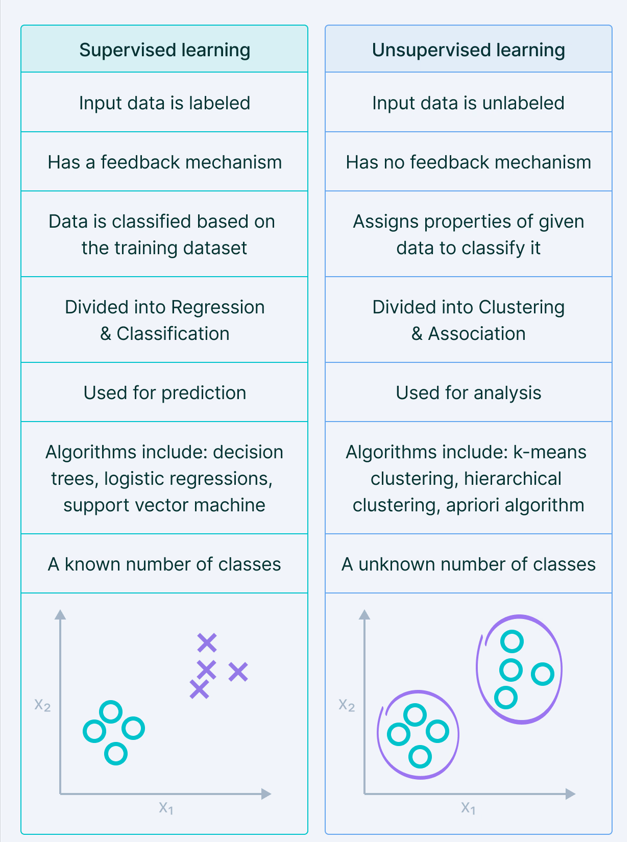
QUIZ 1

Submitted By:

Shubhranshu Sahoo

1. **What is the difference between supervised and unsupervised learning?**

**Answer:**



1. **What is a feature?**

**Answer:**

In the context of Machine Learning (ML), a feature is an individual measurable property or characteristic of the phenomenon being observed. In simple terms, features are the input variables or independent variables used by a machine learning model to make predictions or decisions. They represent the attributes or properties of the data that help the model understand the patterns in the dataset.

For example, if we are building a machine learning model to predict the price of a house, features could include:

* Size of the house (in square feet)
* Number of bedrooms
* Location
* Age of the house
* Number of bathrooms

Each of these attributes contributes to the final prediction and is considered a feature. The better and more relevant the features, the more accurate the machine learning model tends to be.

In practical ML workflows, features are often stored in tabular format, where each row represents a data sample, and each column represents a feature. Feature engineering the process of selecting, transforming, and creating new features plays a critical role in improving model performance.

1. **What is a label?**

**Answer:**

A label in Machine Learning is the output variable or dependent variable that the model is trying to predict. It is the answer or ground truth associated with each input data point during training. The model learns from the features to correctly predict the labels.

For example:

* In a house price prediction model, the label would be the actual price of the house.
* In an email spam detection system, the label would be either "spam" or "not spam".
* In an image classification task, the label could be the category the image belongs to, such as "cat", "dog", or "car".

During supervised learning, the model is trained on a dataset containing both features and labels. The goal is to allow the model to learn a mapping from features to labels so that it can predict labels for new, unseen data accurately.

1. **Which library do we use to implement ML models in Python?**

**Answer:**

In Python, there are several libraries used to implement machine learning models, but one of the most widely used and powerful libraries is Scikit-learn (also written as sklearn).

Scikit-learn is an open-source Python library that provides simple and efficient tools for data mining and data analysis. It is built on top of popular scientific Python libraries like NumPy, SciPy, and matplotlib.

Besides Scikit-learn, other commonly used Python libraries in machine learning include:

* TensorFlow – developed by Google, mainly used for deep learning.
* Keras – a high-level API built on top of TensorFlow.
* PyTorch – developed by Facebook, popular for deep learning and research.
* Pandas – used for data manipulation and analysis.
* NumPy – used for numerical computations.

Among these, Scikit-learn is ideal for beginners and traditional ML tasks because of its simplicity, ease of use, and comprehensive documentation.

**House Price Prediction:**

**Summary:-**

**1. Imports and Dataset Loading**

**from sklearn.datasets import fetch\_california\_housing**

* Loads a real-world dataset containing California housing data.
* Converts it to a pandas DataFrame for analysis.

**2. Exploratory Data Analysis (EDA)**

**df.describe()**

**sns.pairplot(...)**

**sns.heatmap(...)**

* Displays data types, missing values, and summary statistics.
* Uses pairplot and heatmap to visualize relationships and correlations.
* Helps identify strong predictors like MedInc.

**3. Feature Selection**

**X = df[['MedInc', 'AveRooms', 'HouseAge']]**

**y = df['MedHouseValue']**

* Uses 3 features that are likely to affect housing price.
* We can experiment with adding or removing features based on heatmap insights.

**4. Model Training**

**model = LinearRegression()**

**model.fit(X\_train, y\_train)**

* Trains a Linear Regression model on training data.

**5. Model Evaluation**

**mse = mean\_squared\_error(y\_test, y\_pred)**

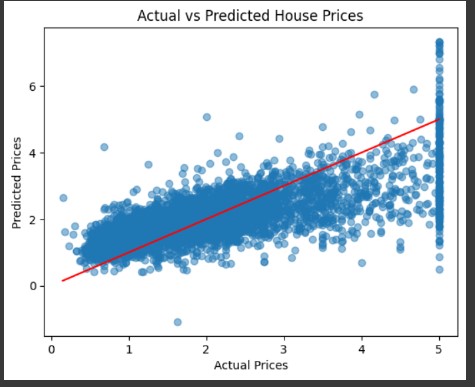
**r2 = r2\_score(y\_test, y\_pred)**

* Evaluates how well the model performs:
  + Mean Squared Error (MSE): Measures average squared difference between actual and predicted.
  + R² Score: Measures proportion of variance explained (closer to 1 is better).

**6. Actual vs Predicted Plot**

**plt.scatter(...)**

* Visualizes how close the predictions are to actual values.
* A red diagonal line helps interpret perfect prediction alignment.

****

* "MedInc (Median Income)" gives the best R² score in theory because it has the strongest linear relationship with house prices.  
  Other features like AveRooms and HouseAge help improve the model when combined, but on their own, they are weaker predictors.