ML Task: Classification & Regression Using Decision Trees

You are required to build two machine learning models using decision trees:

1. Classification on the Iris dataset
2. Regression on the Diamond dataset (Price prediction)

Part 1: Iris Classification

Goal: Predict the species of Iris flower using the following features:
- SepalLengthCm
- SepalWidthCm
- PetalLengthCm
- PetalWidthCm

Steps to follow:

1. Exploratory Data Analysis (EDA) & Visualization:
   - Check data types, nulls, value counts
   - Visualize feature distributions using histograms or KDE plots
   - Use pair plots or scatter plots to explore relationships between features
   - Visualize class distribution (species)

2. Model Building:
   - Use DecisionTreeClassifier
   - Perform GridSearchCV to tune hyperparameters such as:
     - max_depth
     - min_samples_split
     - min_samples_leaf
     - criterion (gini or entropy)

3. Model Evaluation:
   - Perform cross-validation using the best model
   - Evaluate using:
     - Accuracy
     - Precision, Recall, F1-score
     - Confusion Matrix

4. Model Interpretation:
   - Print the decision tree using plot_tree or export_text
   - Show feature importances
   - Check for overfitting by comparing training vs validation scores

Part 2: Diamond Price Regression

Goal: Predict the price of a diamond using features like:
- carat, cut, color, clarity, x, y, z, depth, table

Steps to follow:

1. EDA & Visualization:
   - Explore the distribution of the price and numeric features
   - Plot price vs carat, price vs depth, etc.
   - Boxplots for price by cut, color, and clarity
   - Detect and handle any anomalies or outliers

2. Data Cleaning and Preprocessing:
   - Encode categorical features: cut, color, clarity
   - Try scaling numerical features (especially for linear regression comparison)

3. Model Building:
   - Train a DecisionTreeRegressor and tune using GridSearchCV with parameters like:
     - max_depth, min_samples_split, min_samples_leaf, criterion (mse or mae)
   - Train a Linear Regression model for comparison

4. Model Evaluation:
   - Use cross-validation to compare both models (e.g., using cross_val_score)
   - Report metrics like:
     - Mean Squared Error (MSE)
     - R^2 score

5. Model Interpretation:
   - Print the decision tree
   - Show feature importances
   - Analyze if model is overfitting