**Exercise on Random Forest Regression with Hyperparameter Tunning**

**Diamond Price Prediction Regression**

**1. Business Problem**

Analyze the relationship between diamond attributes and their prices to predict the price of a diamond using Random Forest Regression. This involves identifying significant predictors and building a robust predictive model.

**2. Data Dictionary**

Dataset Link: [Diamond Price Prediction](https://drive.google.com/file/d/1_sHJKaXP6FQSSmk0hWzr7SMypEs3RoDz/view?usp=sharing)

* **price**: Price in US dollars ($326–$18,823)
* **carat**: Weight of the diamond (0.2–5.01)
* **cut**: Quality of the cut (Fair, Good, Very Good, Premium, Ideal)
* **color**: Diamond color, from J (worst) to D (best)
* **clarity**: Clarity rating (I1 (worst), SI2, SI1, VS2, VS1, VVS2, VVS1, IF (best))
* **x**: Length in mm (0–10.74)
* **y**: Width in mm (0–58.9)
* **z**: Depth in mm (0–31.8)
* **depth**: Total depth percentage = z / mean(x, y) = 2 \* z / (x + y) (43–79)
* **table**: Width of the top of the diamond relative to the widest point (43–95)

**3. Data Pre-Processing**

**3.1 Data Cleaning**

1. **Missing Values**:
   * Verify and handle any missing or inconsistent values in the dataset.

**3.2 Outlier Treatment**

1. **Outlier Detection**:
   * Identify outliers using visualizations (box plots) and statistical methods (e.g., IQR).
2. **Outlier Handling**:
   * Replace or remove extreme outliers to avoid model distortion.

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

**4.1 Summary Statistics**

* **Summary Table**:
  + Calculate mean, median, variance, and standard deviation for each numerical feature.

**4.2 Univariate Analysis**

1. **Numerical Data**:
   * Visualize distributions using histograms and box plots to identify skewness and outliers.
2. **Categorical Data**:
   * Use bar charts to observe distributions of categorical variables (e.g., "cut", "color", "clarity").
   1. **Bivariate Analysis**

**1. Numerical-Numerical Relationships**: Use scatter plots to find relationships between numerical variables (e.g., carat and price).

2. **Categorical-Numerical Relationships:** Use box or violin plots to see how categories affect numerical values.

3. **Correlation Analysis:** Create a heatmap to identify strong correlations between features.

**5. Model Building**

**5.1 Encoding Categorical Variables**

* Convert the "cut", "color", and "clarity" categorical variables into dummy variables to incorporate them into the regression model.

**5.2 Data Scaling**

* Apply standardization to ensure features are on a similar scale.

**5.3 Data Split**

* Divide the dataset into training and testing sets (e.g., 80:20 split).

**5.4 Model Training with Hyperparameter Tuning using GridSearchCV**

1. **Purpose:**
   * Optimize the Random Forest hyperparameters (e.g., number of trees, maximum depth, etc.) for better model performance.
2. **Steps:**
   * Define a parameter grid specifying the hyperparameters and their ranges.
   * Use GridSearchCV to perform an exhaustive search over the specified parameter combinations and tune the hyperparameters during model training.
   * Identify the best combination of hyperparameters based on cross-validated performance metrics.
3. **Example Parameters for GridSearch:**
   * n\_estimators: [50, 100, 200]
   * max\_depth: [None, 10, 20, 30]
   * min\_samples\_split: [2, 5, 10]
   * min\_samples\_leaf: [1, 2, 4]
4. **Evaluate the Tuned Model:**
   * Use the optimized hyperparameters to retrain the model and evaluate it on the test dataset.