Diamond Price Prediction

---

1. Executive Summary :

This project aims to develop a machine learning model to predict diamond prices based on their key attributes, such as carat, cut, color, clarity, and dimensions. By accurately predicting diamond prices, this project can assist stakeholders in the jewelry industry with inventory valuation, pricing strategies, and fraud detection.

The project will involve data cleaning, exploratory analysis, feature engineering, and the development of multiple regression models to achieve reliable price predictions.

2. Problem Statement :

- **Background**: In the diamond industry, accurate pricing is a critical challenge due to the variability of diamond features such as carat, cut, colour, clarity, and dimensions. The pricing process often involves manual evaluation, which can be subjective and inconsistent. This creates the need for a robust and data-driven solution to estimate diamond prices reliably and transparently.
- **Objective**: Build a predictive model to estimate diamond prices using machine learning techniques.
- **Scope**: Use a publicly available dataset with key diamond attributes (e.g., carat, cut, clarity) to design and compare predictive models.

3. Data Source :

- **Dataset**: Kaggle (Diamond Price Dataset with 50,000 records).
- **Features**:
    - **Numerical**: Carat, Dimensions (x, y, z).
    - **Categorical**: Cut (Fair, Good, Very Good, Premium, Ideal), Color (J to D), Clarity (I1 to IF).
    - **Target Variable**: Price in USD.

4. Methodology

1. **Data Preparation**:
    - Perform exploratory data analysis (EDA) to identify patterns and trends.
    - Handle missing data (if any) and ensure consistent formatting.
    - Retain outliers as they reflect valid variability in diamond characteristics.

2. **Model Development**:
    - Build and evaluate both **tree-based models** (e.g., Decision Tree, Random Forest, XGBoost) and **scaling-based models** (e.g., Linear Regression, KNN, SVR).
    - Perform hyperparameter tuning using GridSearchCV to optimize model performance.
3. **Tools and Libraries**:
    - Python (pandas, scikit-learn, matplotlib, seaborn, XGBoost).
    - Evaluate using metrics like R^2 score and Mean Squared Error (MSE).

5. Expected Outcomes

- **Key Deliverables**:
    - A trained machine learning model capable of accurately predicting diamond prices.
    - Insights into the most influential features (e.g., carat, cut, clarity) affecting pricing.
- **Practical Applications**:
    - **E-Commerce Platforms**: Automating pricing recommendations for diamonds listed on online marketplaces.
    - **Inventory Management**: Helping jewelers value their inventory based on consistent pricing.
    - **Fraud Detection**: Identifying discrepancies in diamond pricing for authentication and fraud prevention.
    - **Customer Decision Support**: Assisting customers in understanding and justifying diamond prices during purchases.

6. Challenges

- **Outliers**: Managing outliers to ensure they do not overly bias the model while retaining their valid variability.
- **Categorical Features**: Encoding qualitative attributes like cut, color, and clarity effectively for both tree-based and scaling-based models.
- **Model Tuning**: Balancing computation time during hyperparameter tuning for a large dataset.

7. Conclusion :

This project aims to showcase the application of machine learning in the jewelry industry, addressing real-world challenges in pricing diamonds. By leveraging a robust dataset and advanced modeling techniques, the expected outcomes include a reliable predictive model and actionable insights for stakeholders.