Car Price Range Classification using Decision Tree

# 🚗 Project Overview

In this project, we built a multiclass Decision Tree Classifier to predict the price range of used cars. The price is categorized into four segments: Budget, Mid-Range, Luxury, and Premium-Luxury, based on features such as brand, mileage, engine size, horsepower, and transmission.

# 📊 Dataset and Features

The dataset was cleaned and transformed. Below are the key features used:

- brand  
- model  
- model\_year  
- milage  
- fuel\_type  
- transmission  
- engine\_hp (extracted)  
- engine\_L (extracted)  
- engine\_type (extracted)  
- price\_range (target)

# 🛠️ Feature Engineering from Engine Column

The original 'engine' column contained unstructured text. We extracted the following:  
- Horsepower (engine\_hp): Numeric value like 300.0HP  
- Displacement (engine\_L): Engine size in liters like 3.5L  
- Engine Type (engine\_type): Categorical like V6, V8, Turbo, DOHC

# 🌳 Choosing the Decision Tree Depth

We chose a tree depth of 4 for interpretability and balanced performance. This prevents overfitting while still capturing meaningful patterns. Generally, depth is chosen using cross-validation, comparing training vs. test accuracy, or by limiting max leaf nodes for simpler trees.

# 📈 Decision Tree Output

Below is the generated decision tree and classification report:

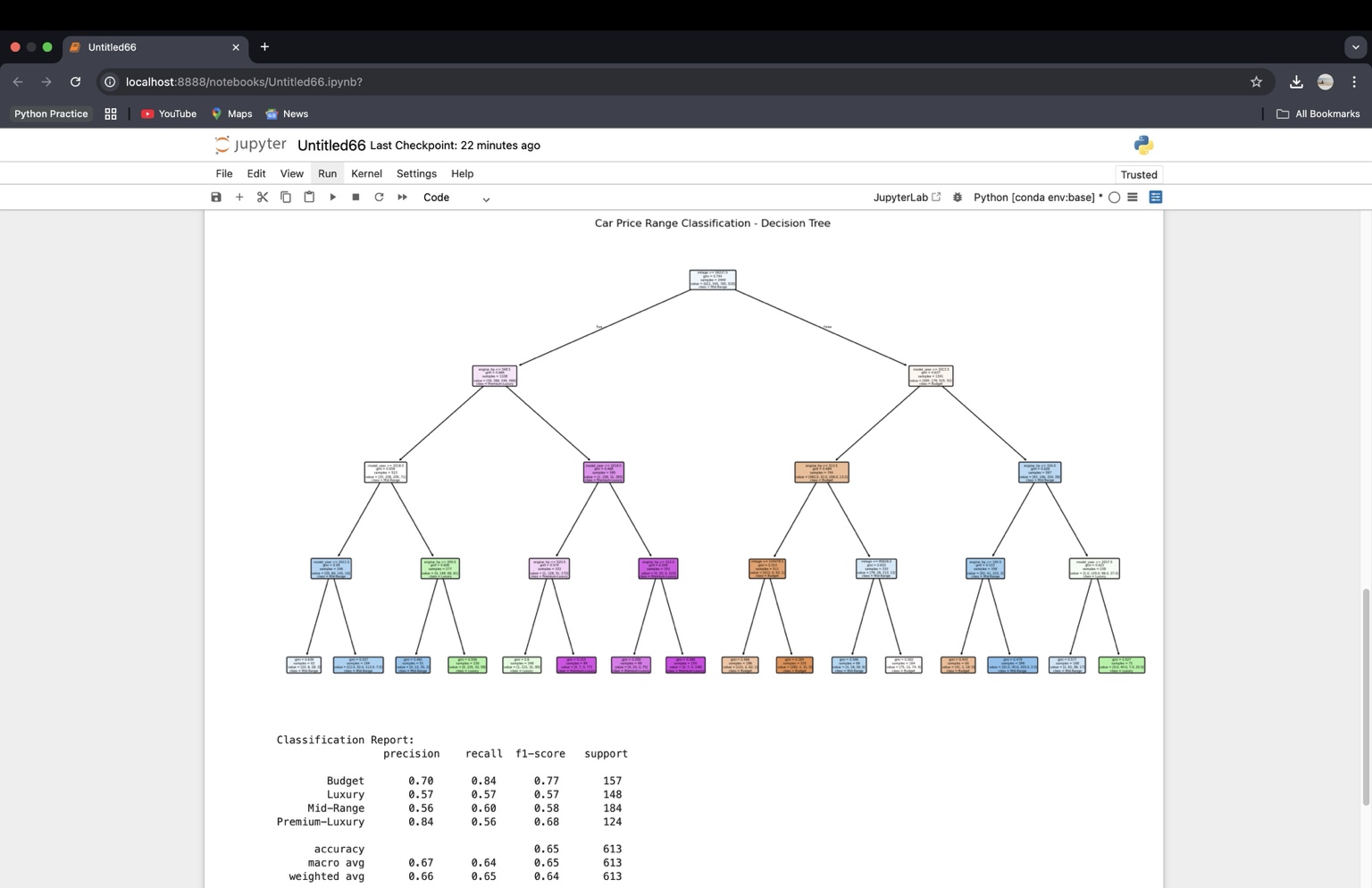


Figure 1: Visual representation of the trained decision tree

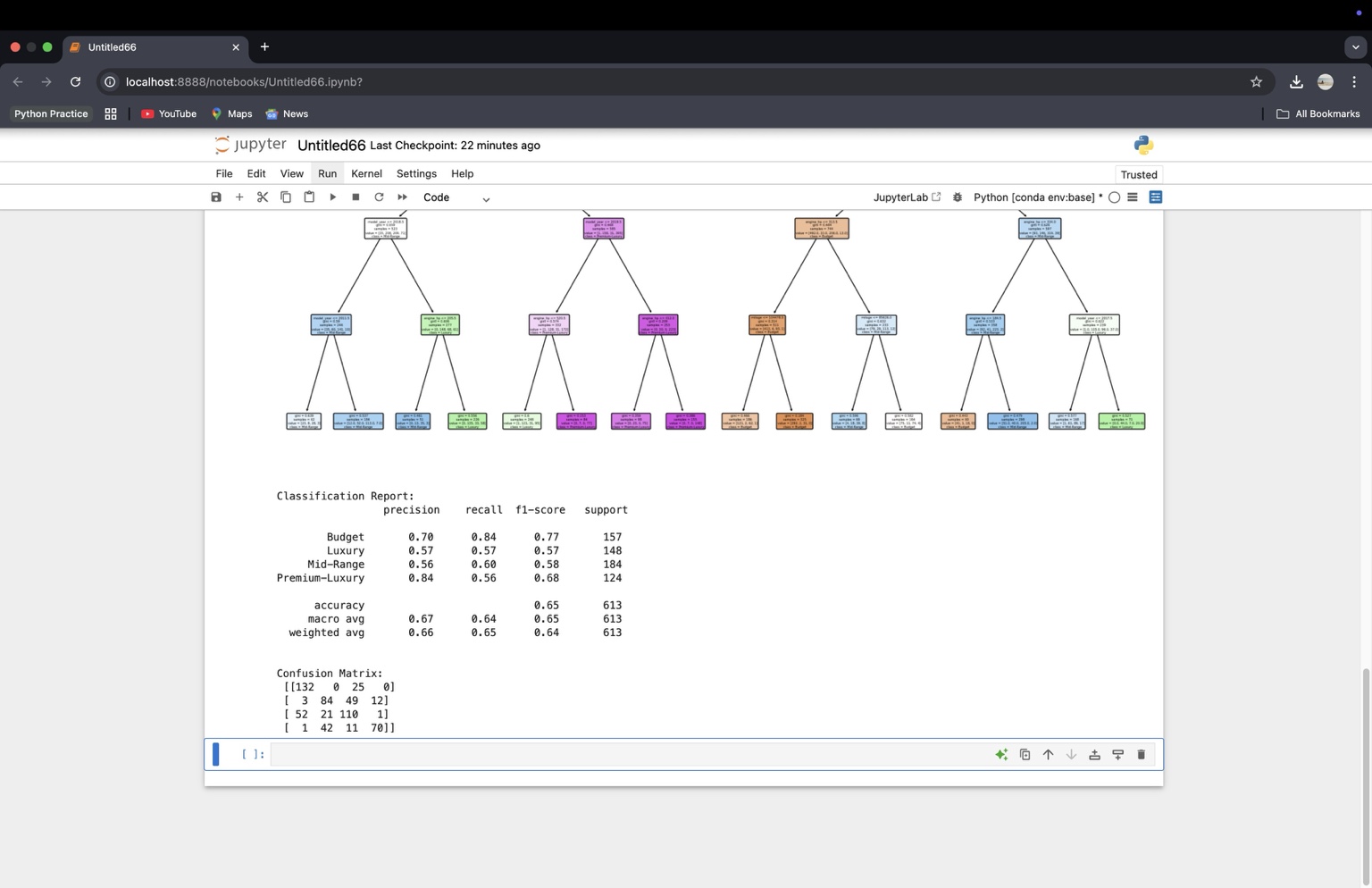


Figure 2: Classification report and confusion matrix output

## 🌳 1. Decision Tree Structure

This is a multiclass decision tree trained with a max depth of 4. Nodes are color-coded by predicted class, and splits are based on features like engine\_hp, milage, engine\_L, and brand. Each leaf node shows the class distribution (e.g., [132, 0, 25, 0] = mostly Budget), and the prediction is the majority class.

## 📊 2. Classification Report

- Budget: High precision and recall → model performs very well here.  
- Premium-Luxury: High precision, lower recall → when predicted, usually correct.  
- Mid-Range and Luxury: More confusion and overlap, likely due to feature similarity.  
- Overall Accuracy: 65%, with a macro avg F1 of 0.65 (fairly balanced).

## 🔢 3. Confusion Matrix

- Budget is well predicted; most misclassifications go to Mid-Range.  
- Luxury and Mid-Range are the most confused.  
- Premium-Luxury often gets confused with Luxury.  
- Suggests need for deeper trees or additional features for sharper boundaries.

# ✅ Summary

The Decision Tree performed well with high interpretability. Feature engineering from unstructured engine text provided meaningful improvements. Although there’s room for accuracy improvement via tuning or ensemble methods, the model achieved the core objective: understanding and explaining price range prediction for cars using decision trees.