TP: Predicting Used Car Prices with Linear Regression

Problem:

Predict the price of used cars based on their features like mileage, year, model, brand, and fuel efficiency.

Data:

Utilize a publicly available dataset such as Kaggle's "Used Cars" or UCI's "Car Evaluation" dataset.

Tasks:

1. Data Exploration and Pre-processing:
   - Load the dataset and explore its structure.
   - Handle missing values, outliers, and perform necessary data cleaning.
   - Visualize the distribution of the target variable (car prices) and key features.
2. Feature Engineering:
   - Extract relevant information from features (e.g., create a new feature for car age).
   - Encode categorical variables using label encoding.
   - Normalize or scale numerical features.
3. Train a Linear Regression Model:
   - Split the dataset into training and testing sets.
   - Train a linear regression model using a suitable library (e.g., scikit-learn) and using the one we wrote on class.
4. Model Evaluation:
   - Evaluate the model's performance on the testing set using metrics such as Mean Squared Error (MSE), R-squared, and any other relevant metrics.
   - Visualize the predicted prices against the actual prices using scatter plots.
5. Feature Importance Analysis:
   - Analyze the coefficients of the linear regression model to understand which features have the most impact on the car prices.
   - Discuss the implications of these findings for car buyers and sellers.
6. Bonus: Make Predictions for New Unseen Cars:

- Take a set of new, unseen car data and use the trained model to predict their prices.
- Discuss potential use cases and limitations of the model for real-world applications.

Deliverables:

- Collab Notebook containing the code for data exploration, pre-processing, model training, and evaluation.
- Visualizations supporting the analysis, such as distribution plots, scatter plots, and feature importance charts.
- A report summarizing the key findings, insights from the feature importance analysis, and potential improvements to the model.

This TP provides a comprehensive hands-on experience, covering data cleaning, feature engineering, model training, evaluation, and interpretation. It also includes a bonus task to encourage students to apply the learned concepts to new data.