Machine Learning Assignment 01

Objective:

The goal of this project is to analyze the dataset, preprocess the data, apply machine learning models, and evaluate their performance in predicting car prices.

Section 1: Data Exploration (EDA)

1.1 Loading the Data

1. Load the dataset using Pandas.
2. Display the first five rows of the dataset.
3. Check for missing values in each column.
4. What are the data types of each column? Identify numerical and categorical features.

1.2 Statistical Summary and Distributions

5. Generate a summary report using `.describe()` for numerical columns. What insights can you draw?
6. Plot histograms for numerical features. What do you observe?
7. Identify outliers using boxplots for at least three numerical columns.

1.3 Relationships Between Features

8. Compute and visualize the correlation matrix. Which features are highly correlated with price?
9. Using a pivot table, find the average price for different fuel types.
10. Use `groupby()` to analyze the average price for different car brands.
11. How does body style impact car price? Use suitable plots and analysis.
12. Which type of engine (engine-type) is most commonly used? Represent this with a bar chart.

Section 2: Data Preprocessing

2.1 Handling Missing Values

13. Identify columns with missing values and the percentage of missing data.
14. What strategy will you use to handle missing numerical values? (Mean/Median/Mode)

15. How will you handle missing values in categorical columns?
16. Replace missing values and verify that no NaN values remain in the dataset.

2.2 Encoding Categorical Features

17. List all categorical features that need to be encoded.
18. Use One-Hot Encoding or Label Encoding for categorical columns. Explain why you chose your method.
19. Verify that encoding was applied correctly by checking the transformed dataset.

Snippet for Label Encoding:

```
from sklearn.preprocessing
import LabelEncoder
encoder = LabelEncoder()
data['column_name'] = encoder.fit_transform(data['column_name'])
```

Snippet for One-Hot Encoding:

```
import pandas as pd
data = pd.get_dummies(data, columns=['column_name'], drop_first=True)
```

2.3 Feature Scaling and Selection

20. Should features like `horsepower` and `engine-size` be scaled? Why?
21. Apply standardization or normalization to the necessary columns.
22. Select relevant features for modeling by checking their correlation with price.

2.4 Splitting Data for Model Training

23. Split the dataset into **training (80%)** and **testing (20%)** sets.
24. Why is it important to keep a test set separate from training data?

Section 3: Model Building

3.1 Applying Regression Models

25. Train a **Linear Regression** model. Record the **MSE** and **R2 Score**.
26. Train a **Decision Tree Regression** model. Compare its performance with Linear Regression.
27. Train a **Random Forest Regression** model. How does it improve results?

3.2 Evaluating Model Performance

28. Create a table comparing the performance of different models based on **MSE and R2 Score**.
29. Which model performed the best? Why?

30. If a model overfits the training data, how can it be improved?

Section 4: Reporting & Insights

4.1 Summary of Findings

31. What key insights did you gain from EDA about car prices?
32. Which features had the most impact on price prediction?
33. What challenges did you face during preprocessing and modeling?
34. If given a larger dataset with more features, what additional steps would you take?

Deliverables:

- **A Jupyter Notebook or Python script** with complete code and outputs.
- **A final report** summarizing key findings and model performance.
- **Graphs and visualizations** supporting your analysis.

Note: These are recommended steps for our assignment. If you want to add more details to your work, it's up to you. You can include additional insights or improvements based on your knowledge.

If you want to do something but don't know the code, please refer to documentation and then apply it. I will attach some links to materials and articles with the assignment to make it easier for you.

Link: https://medium.com/@codekalimi/use-cases-of-linear-regression-models-b7ae3f2a3713
Link: https://www.analyticsvidhya.com/blog/2020/03/one-hot-encoding-vs-label-encoding-using-scikit-learn/

Link: https://hex.tech/templates/feature-selection/

Links: https://www.analyticsvidhya.com/blog/2024/01/best-libraries-for-machine-learning-explained/

Link: https://www.analyticsvidhya.com/blog/2021/06/understanding-random-forest/

Good Luck!