Vehicle Classification Report

1. Overview of Models Performance

Model	Accuracy	Precision	Recall	F1 Score
Logistic Regression	44.6%	41.4%	44.6%	39.9%
Decision Tree	82.2%	82.2%	82.2%	82.2%
Naive Bayes (*Multinomial*)	11.7%	22.6%	11.7%	7.7%
Naive Bayes (*GaussianNB*)	31.8%	19.6%	31.8%	20.2%
Naive Bayes (*ComplementNB*)	20.1%	9.5%	20.1%	12.2%
Naive Bayes (*BernoulliNB*)	34.1%	33.4%	34.1%	26.9%

2. Hyperparameter Tuning

In the hyperparameter tuning process, we identified the optimal parameters for the Logistic Regression and Decision Tree models, which achieved the best performance on this dataset.

- **Logistic Regression**: The best parameters were found to be `C=1` and `solver='liblinear'`. This configuration balances regularization and model complexity, resulting in improved stability and performance.
- **Decision Tree**: The optimal settings were `criterion='entropy'`, `max_depth=None`, `min_samples_leaf=1`, and `min_samples_split=2`. These parameters allow the model to fully grow, leveraging the entropy criterion to create pure splits and achieve strong performance.

No hyperparameter tuning was conducted for the **Naive Bayes** models, as they primarily differ in their assumptions and are less dependent on tuning. The optimal settings for **Logistic Regression** and **Decision Tree** confirmed these models' suitability for the dataset.

3. Analysis

The analysis for each model is as follows:

Logistic Regression (39.9% F1-Score)

Logistic Regression achieved moderate performance with an F1 score of 39.9%. This model assumes a linear relationship between features and class probabilities, which may limit its ability to capture the full complexity of the dataset. The lack of feature interactions and non-linear decision boundaries could be contributing to its relatively lower performance compared to more flexible models.

Decision Tree (82.2% F1-Score)

The Decision Tree model performed the best overall, achieving an F1 score of 82.2%. Decision Trees are well-suited to this dataset as they can capture non-linear relationships and interactions between features without assuming any particular data distribution. While this model demonstrated strong performance, further improvement might be possible by using ensemble techniques, such as Random Forests or Gradient Boosting, to reduce variance and improve robustness.

Naive Bayes (Best of 26.9% F1-Score)

The Naive Bayes models struggled with this dataset, showing low F1 scores across all variants. Naive Bayes classifiers generally assume that features are conditionally independent given the class, an assumption that is often unrealistic in complex datasets with correlated features. Furthermore, each variant makes specific distributional assumptions:

- *MultinomialNB* is designed for discrete count data, such as word counts in text, making it a poor fit for continuous or categorical features.
- *GaussianNB* assumes that features follow a normal distribution, which may not align with the actual distributions in this dataset.
- *GaussianNB* assumes that features follow a normal distribution, which may not align with the actual distributions in this dataset.
- *ComplementNB* is typically used for imbalanced text data and underperformed here, likely due to the data structure not matching its intended use.
- *BernoulliNB*, which assumes binary features, showed slightly better performance than the others, likely due to the presence of binary features in the dataset (e.g., one-hot encoded columns).

4. Conclusion

The analysis shows that Decision Trees are the most effective model for classifying vehicle types in this dataset, achieving an F1 score of 82.2% due to their ability to capture non-linear relationships and feature interactions. Logistic Regression performed moderately, while Naive Bayes models struggled due to their assumptions about feature independence.