# Predictive Modeling Analysis: Car Evaluation Dataset

**Overview**

This study explores the effectiveness of various classification techniques in predicting car evaluations based on six vehicle characteristics using the **Car Evaluation Dataset**. The dataset consists of **1,728 records**, with features including **buying price, maintenance cost, number of doors, passenger capacity, luggage boot size, and safety level**. The target variable categorizes cars as **unacceptable, acceptable, good, or very good**.

**Two Modeling Approaches**

To handle categorical data, we evaluated two different encoding methods:

1. **One-Hot Encoding:**
   * Converts categorical variables into binary indicator variables (dummy variables).
   * Preserves the independence of categories but increases feature dimensionality.
   * Useful when categories do not have a clear order.
2. **Ordinal Encoding:**
   * Assigns numerical values to categories while maintaining their natural order.
   * Helps distance-based models like k-NN but assumes that category differences are linear.
   * May introduce misleading relationships if the order does not imply true magnitude differences.

**Modeling Process**

We tested **Naïve Bayes (NB), Decision Tree (DT), Logistic Regression, k-NN, and SVM**, optimizing hyperparameters using **nested cross-validation**. Since the dataset is imbalanced, **balanced accuracy** was used as the key metric for model selection.

**Results and Findings**

**One-Hot Encoding Results**

| **Model** | **Accuracy** | **Precision** | **Recall** | **F1 Score** |
| --- | --- | --- | --- | --- |
| **SVM** | **99%** | **99%** | **99%** | **99%** |
| **Decision Tree** | 98% | 98% | 98% | 98% |
| **Logistic Regression** | 94% | 94% | 94% | 94% |
| **k-NN** | 90% | 89% | 90% | 89% |
| **Naïve Bayes** | 87% | 87% | 87% | 87% |

**Ordinal Encoding Results**

| **Model** | **Accuracy** | **Precision** | **Recall** | **F1 Score** |
| --- | --- | --- | --- | --- |
| **SVM** | **99%** | **99%** | **99%** | **99%** |
| **Decision Tree** | 98% | 98% | 98% | 98% |
| **k-NN** | 96% | 96% | 96% | 95% |
| **Naïve Bayes** | 81% | 79% | 81% | 78% |
| **Logistic Regression** | 79% | 78% | 79% | 79% |

**Model Selection**

Among all models, **SVM with one-hot encoding was chosen as the best model**, as it:

* Achieved the **highest balanced accuracy score**, which is crucial for handling class imbalance.
* Provided **consistent high performance** across **all classes**, minimizing misclassification.
* **Performed well even with ordinal encoding**, but one-hot encoding preserved the categorical nature of the features better.

**Why One-Hot Encoding is Better than Ordinal Encoding?**

1. **Avoids Misleading Order Assumptions:** Ordinal encoding assumes a **linear relationship between categories** (e.g., "low" < "medium" < "high"), which may not hold for features like "buying price" or "safety."
2. **Enhances Performance in Non-Distance-Based Models:** Decision Trees and SVM performed **better with one-hot encoding** as they treat categories independently rather than assuming a ranking.
3. **Prevents Unintended Bias:** Assigning numbers to categories (ordinal encoding) may cause models to **misinterpret feature importance** in algorithms like Logistic Regression.

**Conclusion**

Since the dataset is **imbalanced**, **balanced accuracy was used as the key metric** to ensure fair evaluation. **SVM with one-hot encoding was selected as the best model**, achieving nearly perfect classification. While ordinal encoding worked well for some models (e.g., k-NN), one-hot encoding provided **better feature representation**, especially for SVM and Decision Trees.