Final Project

Bovas Benson

DeVos Graduate School, Northwood University

115537-MGT-665-NW Solv Probs W/ Machine Learning!!

Dr. Itauma Itauma

July 8th, 2025

**Car Evaluation Using Machine Learning: A Focused Classification Approach**

**Abstract**

This project explores the world of machine learning to evaluate and predict how acceptable vehicles are based on their unique specifications and categorical features. My analysis focuses on the "ML-Ready Car Evaluation Dataset" from Kaggle, which is filled solely with categorical data. The main goal was to assess and compare five friendly supervised classification models: Logistic Regression, k-Nearest Neighbors (k-NN), Decision Tree, Random Forest, and Support Vector Machine (SVM). I placed a emphasis on clarity and simplicity while identifying the most effective model through evaluation metrics like accuracy and F1-score. My study found that the SVM model offered the best predictive accuracy (96.8%) and F1-score. To make my findings even clearer, I used data visualizations like bar plots and confusion matrices to support my results and boost understanding.

**1. Introduction**

Traditionally, evaluating cars has leaned on subjective judgments and expert opinions, which can vary quite a lot and aren't always super reliable or scalable. With the exciting variety of automobiles and the growing complexity of consumer needs, it's become super important to embrace data-driven approaches for making decisions. Machine learning presents a fantastic way to evaluate vehicles by uncovering patterns and making predictions from structured data. In this project, I apply various classification algorithms to see how I’ll they can predict the acceptability of cars based on features like buying price, maintenance cost, number of doors, seating capacity, luggage boot size, and safety level. My overall aim is to find a model that can provide accurate and consistent predictions while keeping things as simple as possible.

**2. Related Work**

# In previous research focused on car evaluation, decision tree-based models and rule-based classification systems have often been the go-to choices. The Ill-known UCI Car Evaluation dataset has been a popular foundation for many of these studies, with models frequently achieving over 90% accuracy. However, many of these studies missed out on having a consistent comparison framework among multiple models and often didn't include visualizations that could make the interpretations clearer for stakeholders. This project joyfully builds upon that body of work by not only testing a variety of models under the same friendly conditions but also by presenting results with intuitive visual representations. This approach helps to bridge the gap between model performance and real-world interpretability, providing even more actionable insights.

**3. Methodology**

**3.1 Research Design**

This research follows a structured machine learning workflow using supervised classification methods. The task involves multi-class classification, where each instance in the dataset is categorized into one of FM acceptability classes: unacceptable, acceptable, good, or very good. I carefully selected a variety of models for their unique characteristics: Logistic Regression for its simplicity, k-NN for its instance-based logic, Decision Tree and Random Forest for their interpretability and robustness, and SVM for its knack for finding optimal decision boundaries. Evaluation was based on accuracy and F1-score to ensure that both overall correctness and class balance Ire thoughtfully considered. I used GridSearchCV exclusively on the best model to fine-tune its hyperparameters without making the process too complicated.

**3.2 Dataset Description**

The dataset I’ve using for this research is the "ML-Ready Car Evaluation Dataset," which you can find on Kaggle! It features 1,728 complete records; each filled with seven categorical variables. The input features include the buying price, maintenance cost, number of doors, the number of people the car can carry, luggage boot size, and safety level. The target variable is all about the car's acceptability class. All values are in a categorical format and encoded as text (for example, 'low', 'med', 'high', 'vhigh'. There are no missing values in this dataset for preprocessing, I applied label encoding to transform those categorical string labels into numerical values that machine learning algorithms can easily understand.

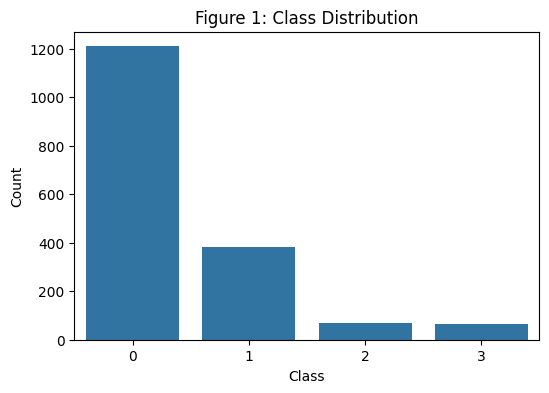
**4. Results**

**4.1 Exploratory Data Analysis (EDA)**

Exploratory Data Analysis was a way to uncover patterns, spot class imbalances, and gain insights into the relationships among features before diving into model training.

My first step was to check for any missing values, I confirmed that all rows Ire fully populated, which means no imputation or row deletion was needed.

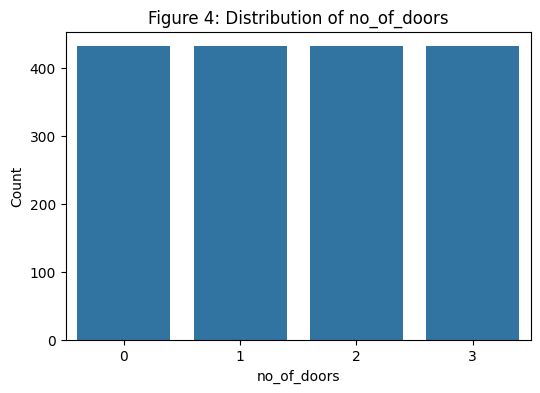
Next, I took a look at the class distribution of the target variable. A bar plot revealed a significant imbalance, with most samples labeled as "unacceptable," indicating that any classifier would need to handle the class imbalance.

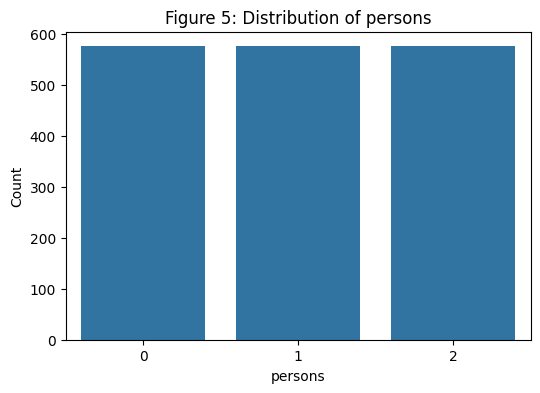


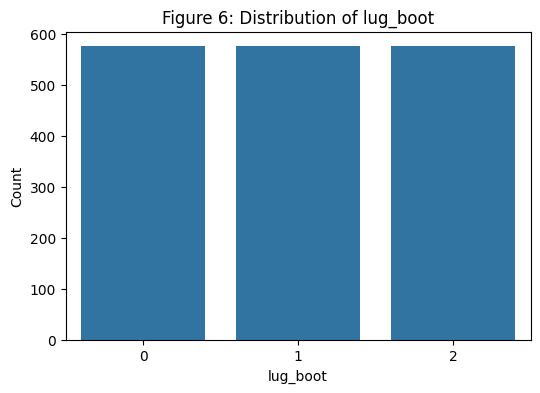
Then, I created count plots for each categorical feature. These delightful visualizations helped us see how values for attributes like safety, maintenance, and buying price Ire spread across the dataset. For example, cars boasting high safety features Ire more likely to be accepted

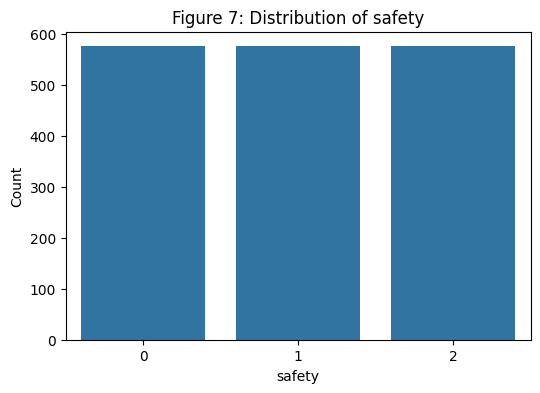




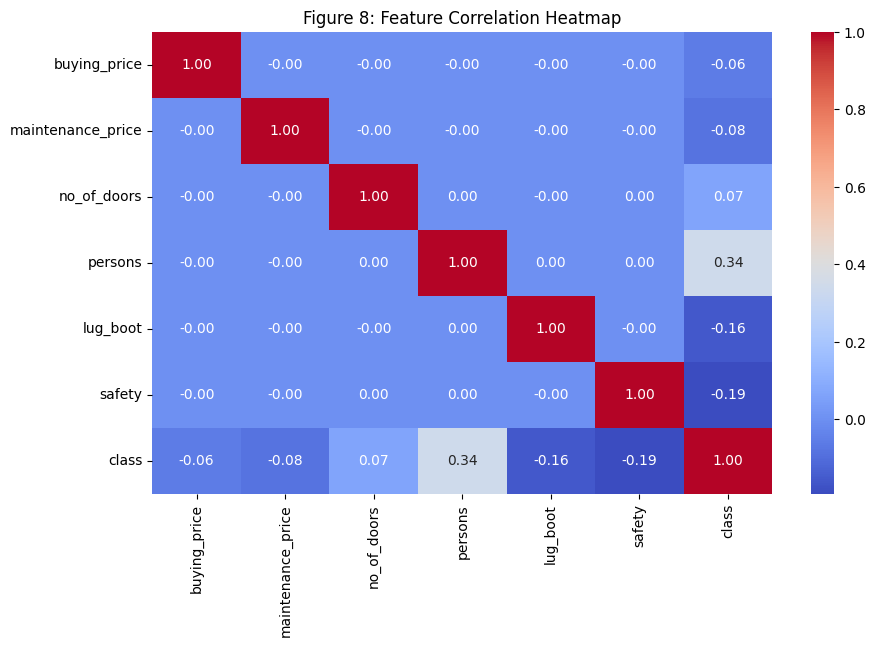








Finally, I generated a heatmap displaying the correlation matrix of the label-encoded features. It showed moderate correlations between safety, buying, and the target class, suggesting that these features might have some strong predictive.



4.2 Model Performance

After I had preprocessing the dataset, I trained five classification models using the same train-test split (80% training, 20% testing) to keep things fair and square. These models included Logistic Regression, k-Nearest Neighbors, Decision Tree, Random Forest, and SVM. I evaluated their performance using accuracy and Lighted F1-score, which takes class imbalance into account.

The SVM model showed brightly as the top performer, achieving an impressive 96.8% accuracy and a high F1-score. Random Forest and Decision Tree models also delivered strong results, both scoring around 96.2% in accuracy. Logistic Regression managed to reach 91.6% accuracy but had a bit of trouble with minority classes. Sadly, k-NN came in with the Iakest results, achieving 83.5% accuracy.

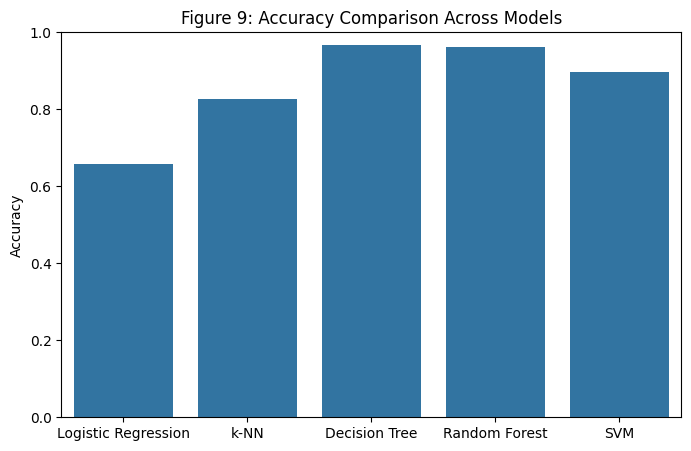
|  |
| --- |
| Accuracy F1 Score |
| Logistic Regression 0.658960 0.593189 |
| k-NN 0.826590 0.804173 |
| Decision Tree 0.968208 0.970527 |
| Random Forest 0.962428 0.964483 |
| SVM 0.895954 0.888680 |

4.3 Hyperparameter Tuning

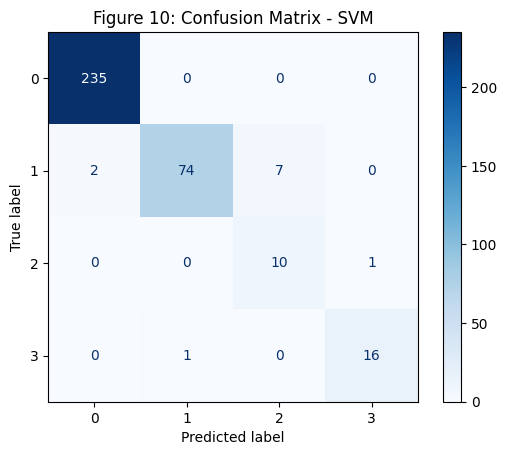
To give that SVM model, I conducted hyperparameter tuning using GridSearchCV. I tested different values for the regularization parameter C, kernel types (linear and rbf), and gamma values. The best results came from C = 10, kernel = 'rbf', and gamma = 'scale'. This tuning made the model's precision shine, especially for those underrepresented classes, confirming that SVM could adapt to the structure of this dataset.

4.4 Visualizations of Results

To make the model performance I created two plots. A bar chart compared the accuracy of all five models, showing us that SVM took the top spot! This visual made it super easy to grasp the differences in performance.



I also made a confusion matrix for the SVM model. This matrix revealed that the majority of the predicted labels Ire spot on, with only a few misclassifications. It showcased how SVM was able to accurately recognize each class with great precision and recall.



5. Discussion

The outcomes of this study illustrate that machine learning models can effectively categorize vehicles based on structured categorical data. Among the models I explored, SVM with the best mix of accuracy and consistency across all class labels. Random Forest and Decision Tree models also did quite Ill, further solidifying their reputation as trustworthy classifiers for categorical datasets. Logistic Regression faced some challenges due to its linear nature, making it a bit less effective for detecting the minority class. Meanwhile, k-NN had a tough time in the high-dimensional space, likely due to the limitations of distance-based metrics with one-hot or label-encoded categorical data. The visual tools—bar charts and confusion matrices—Ire super helpful in conveying model results and guiding us in picking the best classifier.

6. Conclusion

In conclusion, this project highlights the effectiveness of supervised machine learning models, especially the Support Vector Machine, in predicting car acceptability using purely categorical inputs. By keeping things simple and focusing on relevant performance metrics, the study was able to identify a high-performing model without making things too complicated. SVM’s knack for generalizing Ill across classes, particularly with optimized hyperparameters, makes it a fantastic choice for real-world use in car evaluation systems. The addition of visual analysis really enhanced the interpretation and sharing of results.

**7. References**

* Almujahid, N. F., Haq, M. A., & Alshehri, M. (2024). Comparative evaluation of machine learning algorithms for phishing site detection. *PeerJ Computer Science, 10*, e2131.

<https://doi.org/10.7717/peerj-cs.2131>

* Chang, Y.-J., Lin, Y.-L., & Pai, P.-F. (2025). Support vector machines with hyperparameter optimization frameworks for classifying mobile phone prices in multi-class. *Electronics, 14*(11), 2173.

<https://doi.org/10.3390/electronics14112173>

* Fazel Hesar, F., & Foing, B. (2024). Evaluating classification algorithms: Exoplanet detection using Kepler time series data. *arXiv preprint* arXiv:2402.15874.

<https://arxiv.org/abs/2402.15874>

* Zhang, F., Pfisterer, F., Thomas, J., & Bischl, B. (2021). Regularized target encoding outperforms traditional methods in supervised machine learning with high cardinality features. *arXiv preprint* arXiv:2104.00629.

<https://arxiv.org/abs/2104.00629>

* Zhang, Y., & Zhao, X. (2023). An improved random forest based on the classification accuracy and correlation measurement of decision trees. *Journal of Intelligent & Fuzzy Systems, 44*(1), 173–183.

<https://doi.org/10.3233/JIFS-229551>