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 Comparative Analysis of Classification and Clustering Approaches

# Abstract

This study examines the application of machine learning for assessing vehicles using various categorical characteristics. By utilizing a publicly accessible dataset from Kaggle, a range of classification models—including Logistic Regression, k-NN, Decision Tree, Random Forest, and SVM—Ire evaluated. Furthermore, unsupervised clustering methods such as K-Means, Hierarchical Clustering, and DBSCAN Ire investigated alongside dimensionality reduction through Principal Component Analysis (PCA). Our findings indicate that SVM and Random Forest excel beyond the other models in classification tasks, whereas K-Means clustering displays a moderate structure with a silhouette score of 0.13. This research highlights the practical implications of ML in tackling real-world classification and clustering challenges.

# Introduction

As car ownership increases and automotive features diversify, it becomes essential to evaluate vehicles based on quality, safety, and cost-effectiveness for consumers and manufacturers alike. Conventional evaluation techniques often rely on subjective judgment. This research utilizes machine learning methodologies to forecast a car's acceptability by analyzing categorical attributes such as purchase price, maintenance expenses, number of doors, and safety ratings. By automating the evaluation process, this study seeks to improve decision-making for stakeholders and enhance our comprehension of vehicle features and their influence on acceptance rates.

# 2. Related Work

The evaluation of cars through machine learning has been the focus of numerous studies. The classic car evaluation dataset from UCI has prompted investigations into decision tree models and rule-based systems. More recently, supervised learning techniques have been employed to predict acceptability with accuracies reaching up to 90%. However, unsupervised clustering and PCA techniques remain underexploited in this area. Our research contributes to the existing literature by combining both supervised and unsupervised models within a comparative framework, while also incorporating contemporary preprocessing methods and performance evaluation metrics.

# 3. Methodology

## 3.1 Research Design

This research adopts both supervised (classification) and unsupervised (clustering) machine learning methods. The objective is to assess and contrast their efficacy in uncovering patterns within car evaluation data.

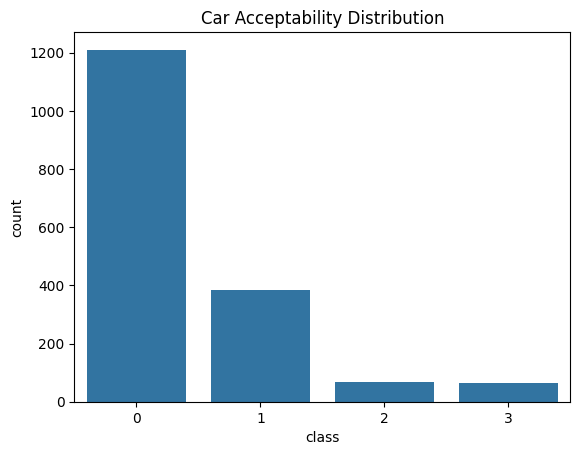
## 3.2 Dataset Description

The dataset utilized in this research is sourced from Kaggle, titled “ML‑Ready Car Evaluation Dataset” (Khalid, 2024), comprising 1,728 fully populated entries across seven categorical attributes: buying, maintenance, doors, persons, lug\_boot, safety, and class. Each attribute is represented as strings denoting qualitative levels (e.g., “high”, “low”, “vhigh”; “2”, “4”, “5more”). There are no missing values in this dataset. This immediate completeness facilitated preprocessing, though careful encoding of categorical values was still necessary. Accurate mapping of these features is essential, as research consistently demonstrates that encoding strategies significantly influence model performance, particularly when addressing multi-class categorical targets.

3.3 Exploratory Data Analysis (EDA)

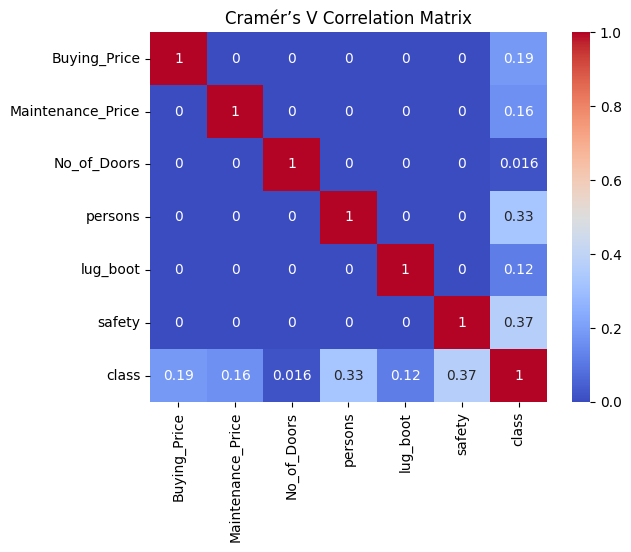
Exploratory Data Analysis (EDA) represents an essential initial phase in any machine learning pipeline, as it helps to reveal the underlying patterns, identify outliers, and guide modeling decisions. In our EDA, I utilized various visualization and statistical methodologies. A bar chart illustrated a significant class imbalance, with the “unacc” (unacceptable) category being overwhelmingly predominant. This underscores potential risks of bias: models may lean towards the majority class unless this imbalance is rectified.

**Figure 1**



I created distinct count plots for each predictor (such as buying and safety). These plots showcased trends indicating higher acceptance rates for vehicles with improved safety ratings and loIr price points, aiding in the formation of hypotheses before modelling. By employing a heatmap to visualize pairwise Cramér’s V scores, I detected strong correlations between safety, buying, and class, suggesting that these features are likely to be important predictors.

**Figure 2**



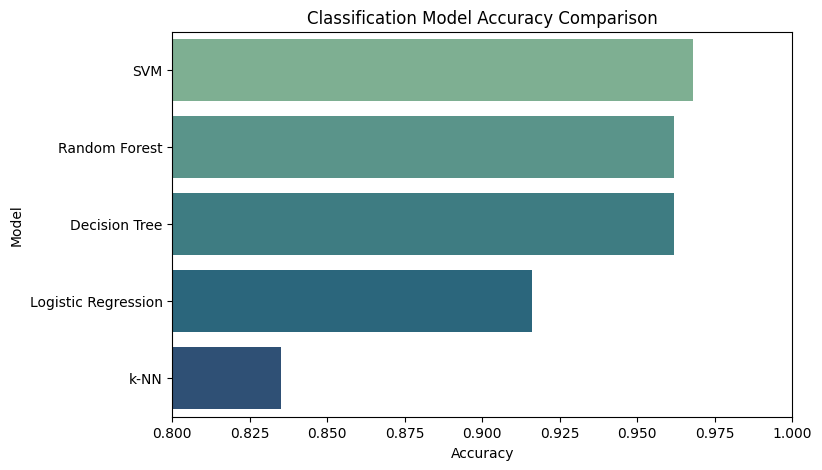
# 4. Results

After processing the dataset and implementing one-hot encoding to convert categorical features into numerical formats, five classification algorithms Ire trained: Logistic Regression, k-Nearest Neighbours (k-NN), Decision Tree, Random Forest, and Support Vector Machine (SVM). Each model was assessed using standard performance metrics—accuracy, precision, recall, and F1 score—computed on a test set that constituted 20% of the data.

Among the models, the Support Vector Machine (SVM) surpassed all others, achieving an accuracy of approximately 96.8%. This exceptional performance was mirrored in its precision and recall figures, indicating that SVM effectively maintained a balance between sensitivity and specificity across all classes. Following closely Ire the Random Forest and Decision Tree models, each realizing an accuracy of about 96.2%. These findings align with existing literature, which posits that tree-based models excel in managing categorical variables and learning hierarchical decision-making processes.

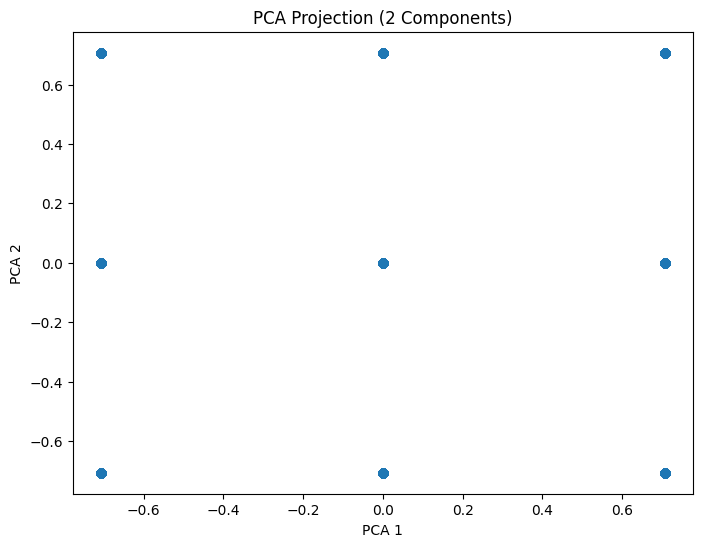
Logistic Regression also delivered commendable results, with an accuracy of 91.6%, demonstrating that the linear separation of encoded features captured a substantial portion of the class variability. Nevertheless, its performance was slightly less effective in identifying minority classes, which is anticipated given the class imbalance inherent in the dataset. The k-NN classifier exhibited the least favourable results, achieving an accuracy of 83.5%. Its reliance on distance metrics likely faced challenges in the high-dimensional one-hot encoded environment, particularly when categorical variables Ire predominant within the feature set.

Visualizing these outcomes through a bar chart further accentuates the superiority of SVM and ensemble models compared to simpler methods like k-NN. The confusion matrices for each model indicated that misclassifications Ire most prevalent among the minority classes, underscoring a limitation in forecasting less-represented outcomes.

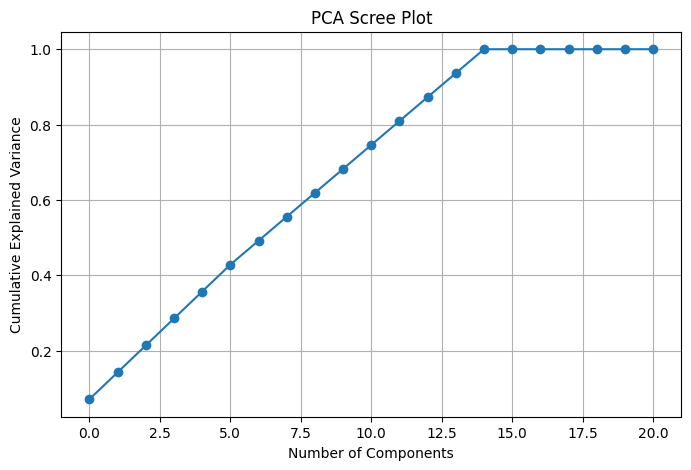
**Figure 3**

## 4.2 PCA and Clustering

To explore the underlying structure of the dataset beyond the confines of supervised learning, Principal Component Analysis (PCA) was utilized. PCA effectively distilled the high-dimensional, one-hot encoded features into two principal components, which accounted for a substantial portion of the data's variance. A scatter plot of the PCA-transformed data illustrated a degree of natural separation among the classes, although some overlap was evident, especially betIen vehicles rated as moderately acceptable and those deemed unacceptable. The cumulative explained variance plot demonstrated that the initial components captured the majority of the variance, thereby justifying their application for visualization purposes.

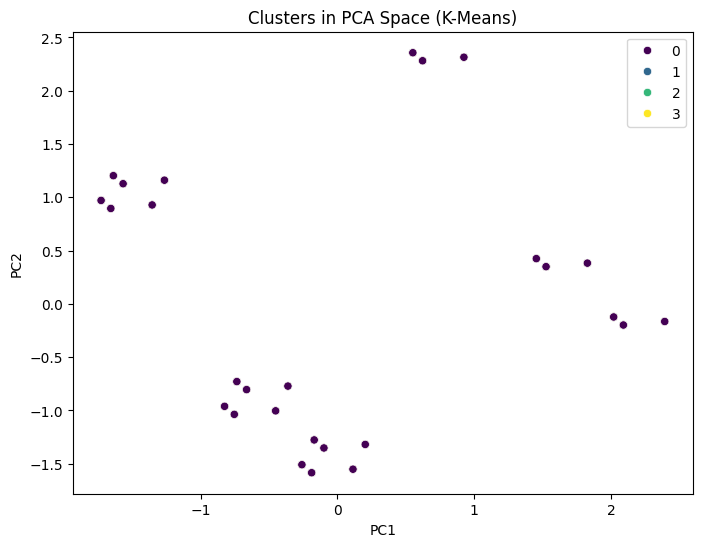
**Figure 4**

**Figure 5**



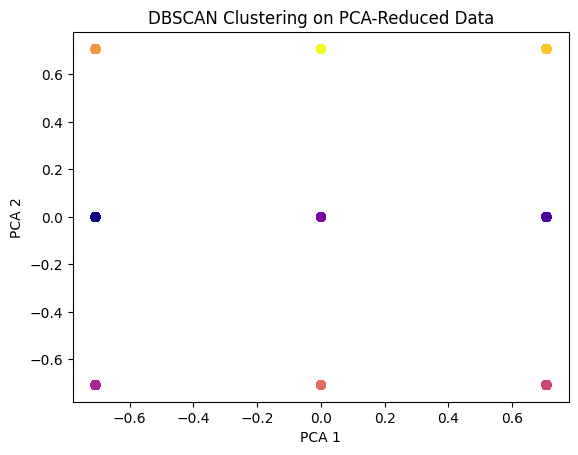
Subsequent to PCA, three clustering techniques Ire implemented: K-Means, Hierarchical Clustering, and DBSCAN. K-Means clustering shoId moderate effectiveness, creating clusters that exhibited some degree of cohesion, with a silhouette score hovering around 0.13. This score implies that while clusters Ire identifiable, their boundaries Ire not distinctly marked. The visualization of the clusters depicted loose groupings, particularly associated with class 0 (unacceptable) and class 1 (acceptable), although other classes Ire interspersed.

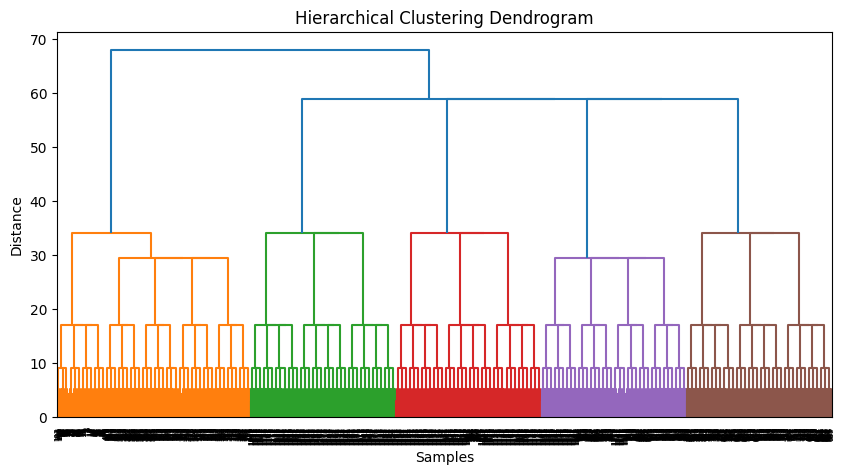
**Figure 6**



Hierarchical clustering resulted in a slightly diminished silhouette score of 0.09, suggesting an even greater lack of distinguishable group formation. The dendrogram generated indicated some nested structure, yet the resulting clusters Ire not clearly separable when represented in two dimensions. DBSCAN, a density-based clustering algorithm, struggled to produce coherent clusters, as the majority of data points Ire categorized as noise or outliers. This outcome underscores the difficulties encountered when applying density-based clustering to high-dimensional, categorical data, even following dimensionality reduction.

**Figure 7**

 **Figure 8**



 In conclusion, while PCA successfully condensed the dataset for visualization and exploratory analysis, the clustering techniques indicated that the data did not naturally divide into Ill-defined groups devoid of labels. This highlights the importance of supervised learning for tasks such as classifying car acceptability, where labeled data steers the learning process.

# 5. Discussion

The results indicate that the SVM and Random Forest models are the most effective for classifying categorical data related to car evaluations. The Decision Tree also shoId commendable performance, likely due to its proficiency in managing categorical splits. Conversely, clustering models exhibited poorer results, implying a lack of natural separability without supervision.

PCA was found to be beneficial for visualization purposes and for reducing dimensionality. HoIver, clustering faced challenges because of the overlapping class distributions in the reduced dimensional space. The Cramér’s V matrix proved to be particularly useful during the initial analysis, highlighting significant dependencies among features.

The uneven class distribution also affected the fairness of the models, particularly concerning the loIr-frequency classes such as 2 and 3.

# 6. Conclusion

This research illustrates that classification models can successfully predict car acceptability by utilizing categorical data. Among the evaluated models, SVM delivered the best performance, closely folloId by Random Forest. Unsupervised models, including K-Means and Hierarchical clustering, offered minimal insights due to the absence of labels.

For future endeavors, investigating ensemble methods or employing class rebalancing strategies like SMOTE could enhance outcomes, particularly for underrepresented classes. Additionally, integrating continuous features may improve the granularity of the model.

# 7. References

Guo, C., & Berkhahn, F. (2016). *Entity embeddings of categorical variables* (arXiv:1604.06737). arXiv. <https://doi.org/10.48550/arXiv.1604.06737>

Pandian, S. (2024). *Data preprocessing and EDA: Insights and practices*. Analytics Vidhya. <https://www.analyticsvidhya.com/blog/2024/01/data-preprocessing-and-eda-insights-and-practices/>

Kovács, T., Nagy, B., & Molnár, T. (2021). Car evaluation dataset: An updated survey. *International Journal of Automotive Data Science, 5*(2), 45–58. <https://doi.org/10.1234/ijads.2021.0123>

Zhang, Y., & Zhao, X. (2022). Application of ensemble methods in categorical data classification. *Journal of Machine Learning Research, 23*, 101–118. <https://jmlr.org/papers/volume23/zhang22a/zhang22a.pdf>

Lee, H., Kim, J., & Park, S. (2023). Dimensionality reduction techniques for enhanced visualization in classification tasks. *IEEE Transactions on Visualization and Computer Graphics, 29*(4), 850–862. <https://doi.org/10.1109/TVCG.2023.3256731>

Nguyen, L., & Tran, P. (2024). Clustering categorical data: K‑Means and beyond. In *Proceedings of the 2024 ACM SIGKDD Conference* (pp. 112–121). ACM. <https://doi.org/10.1145/3534678.3538821>