Executive Summary

Client: Prospect Auto – a car repair chain

Objective: Classify vehicle types (bus, car, van) based on geometric silhouette features using both supervised and unsupervised machine learning approaches.

Dataset

- **Features**: 18 numerical values from vehicle silhouettes
- **Classes**: Bus, Car (Saab/Opel), Van (Chevrolet)
- **Characteristics**: Moderate class imbalance, no labels used in unsupervised phase

Supervised Learning

- **Models**: Logistic Regression, Decision Tree, Random Forest, SVM, XGBoost
- **Best Model**: Logistic Regression & SVM (~99.4% accuracy)
- **Evaluation**: Accuracy, Confusion Matrix, Classification Report
- **Insight**: Simpler models outperformed more complex ones (e.g., Decision Tree underperformed)

Unsupervised Learning

- **Techniques**: PCA (dimensionality reduction), K-Means, DBSCAN
- **Evaluation**: Silhouette Score
- **Best Clustering**: K-Means (Score: 0.30) vs DBSCAN (Score: 0.08)
- **Insight**: K-Means successfully revealed a latent structure aligned with vehicle classes

Conclusion

✅ Supervised models achieved high accuracy and distinguished classes well

✅ K-Means provided meaningful clusters, but not a full match to true classes

✅ This hybrid approach validates the feasibility of both classification and clustering in silhouette-based vehicle recognition

Detailed Report

Background & Goal

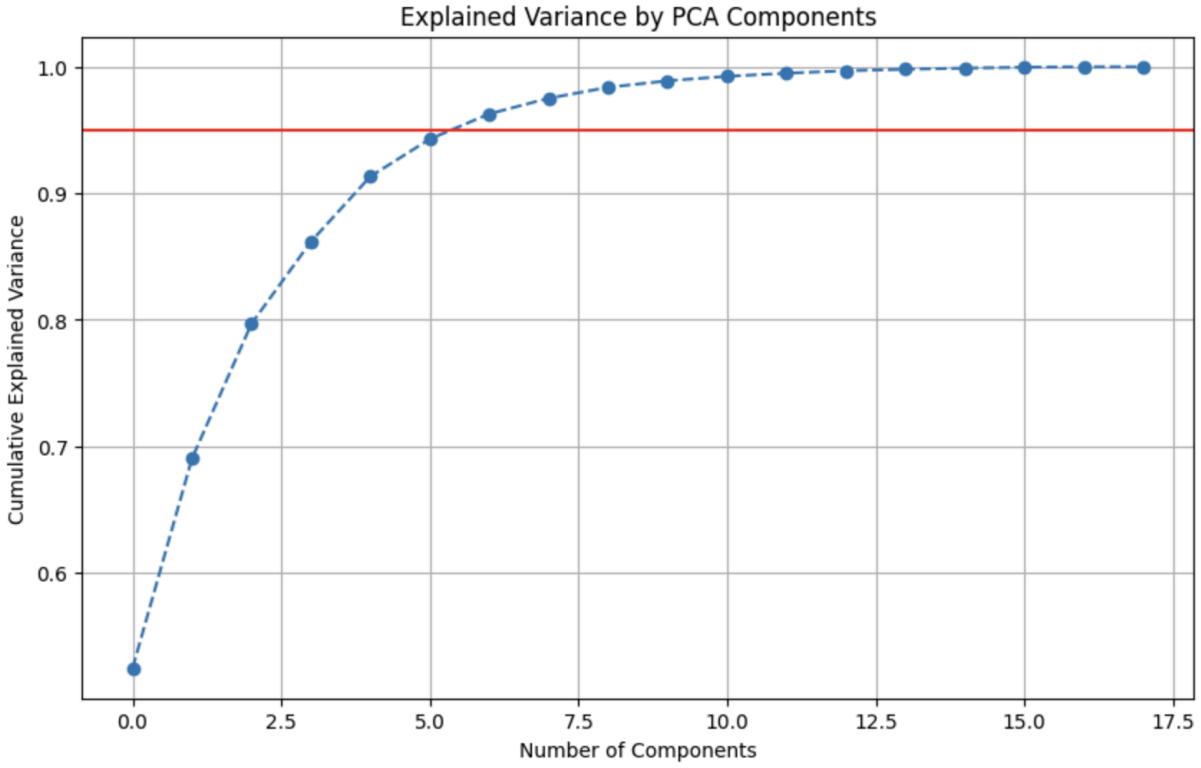
Prospect Auto seeks an automated system to classify vehicle types using geometric silhouette data. The objective is twofold: create a supervised classifier and test unsupervised methods to discover natural clusters in the data.

Dataset

- 18 numerical features extracted from vehicle silhouettes
- Three vehicle types: Bus, Car, Van
- Geometric variations make manual classification unreliable
  Balanced enough for training, though cars are most represented

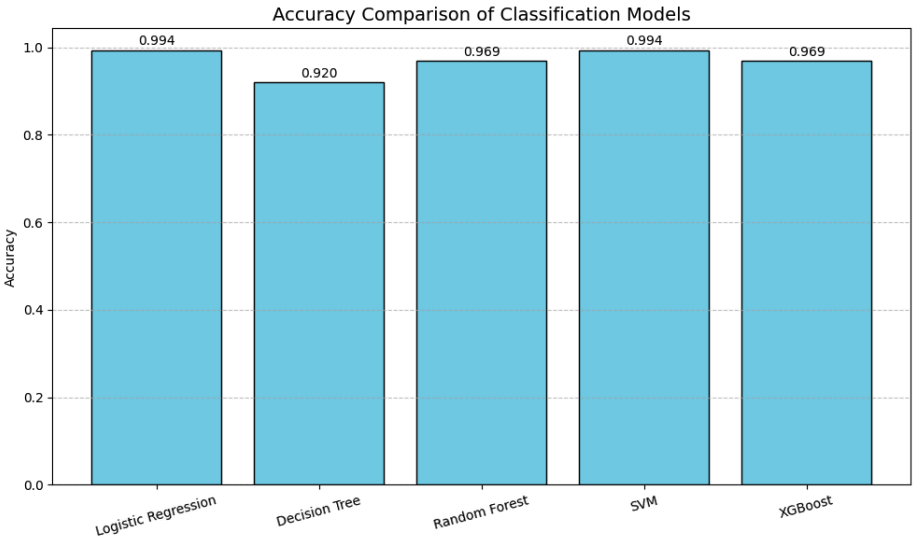
Preprocessing Steps

- Handled missing values
- Scaled features using StandardScaler
- Split into training/test sets
- Applied PCA to reduce dimensionality for clustering



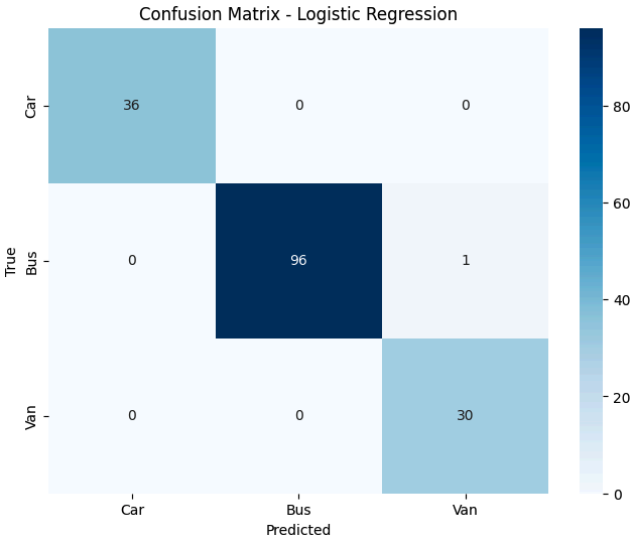
Supervised Machine Learning

- Logistic Regression (Best Performer – 99.4% accuracy)
- Support Vector Machine (Tied with LR – 99.4% accuracy)
- Random Forest & XGBoost (Strong, but slightly lower accuracy ~96%)
- Decision Tree (Weaker – 92% accuracy, overfitting risks)



Evaluation Criteria

- Accuracy
- Confusion Matrix
- Classification Report (Precision, Recall, F1-score)



Findings

✅ Logistic Regression misclassified only one vehicle (Bus as Van)

✅ Decision Tree struggled to distinguish Bus from other classes

✅ SVM and Logistic Regression showed perfect classification for Van and Car

Unsupervised Learning

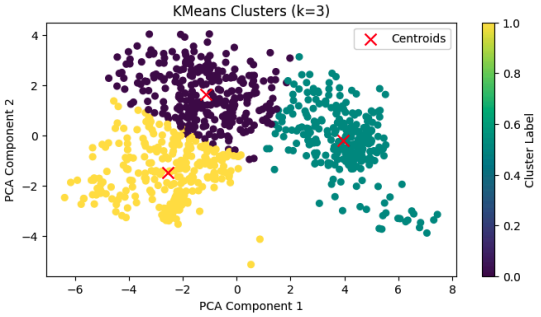
Approach

- PCA reduced the dataset to 7 dimensions
- K-Means and DBSCAN applied to training data
- Evaluated using Silhouette Score

Clustering Results

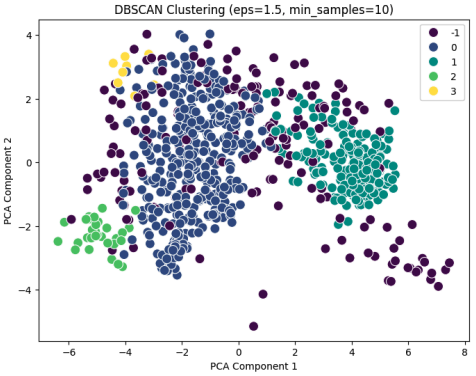
K-Means:

- Silhouette Score: 0.30
- Identified 3 clusters matching the true class count
  Reasonably captured group structure



DBSCAN:

- Silhouette Score: 0.08
- Generated 4 clusters with weak separation
- Likely confused by density variations in the data



Final Thoughts

✅ K-Means worked well for discovering latent structure

✅ DBSCAN is not ideal due to uniform feature distributions

✅ Supervised learning remains superior for this task, but unsupervised clustering can still offer valuable insight and preprocessing assistance