Electric Vehicle Data Analysis

Buğra Mutluer 201911040
Ahmet Berkay Aslan 201911009

Electric Vehicle Data

The dataset, derived from an electric vehicle survey, includes 6108 participants and 74 features.

This dataset offers insights into electric vehicle acceptance, preferred features, societal perceptions, and biases, aiding in understanding user attitudes and guiding strategies for increased adoption.

Data Preprocessing

Handling Missing Values:

- Replaced "?" values with NaN.
- Replaced "D" (Don't Know) responses in Q11, Q12, and Q13 with NaN for imputation.

Understanding the Target (Q16):

- Removed "D" values from the target.
- Combined two positive answers as 0.
- Replaced "C" with 1, representing the negative.

Imputation:

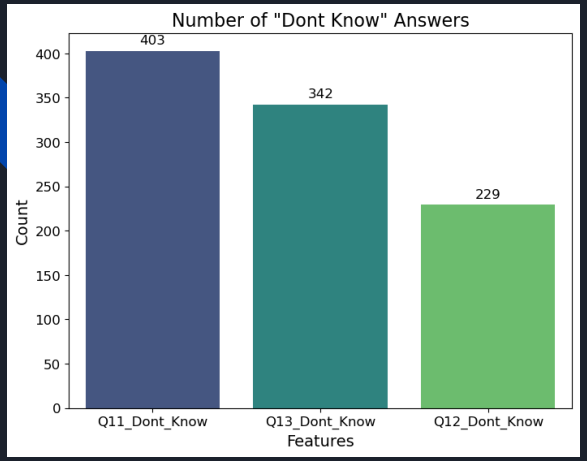
- Utilized KNN imputation to estimate and replace NaN values in the dataset.

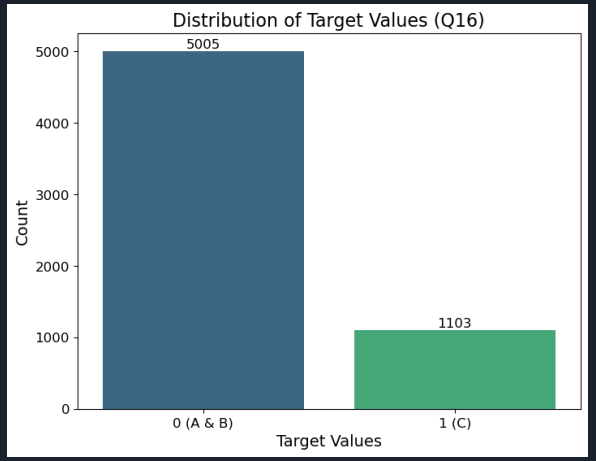
Feature Selection:

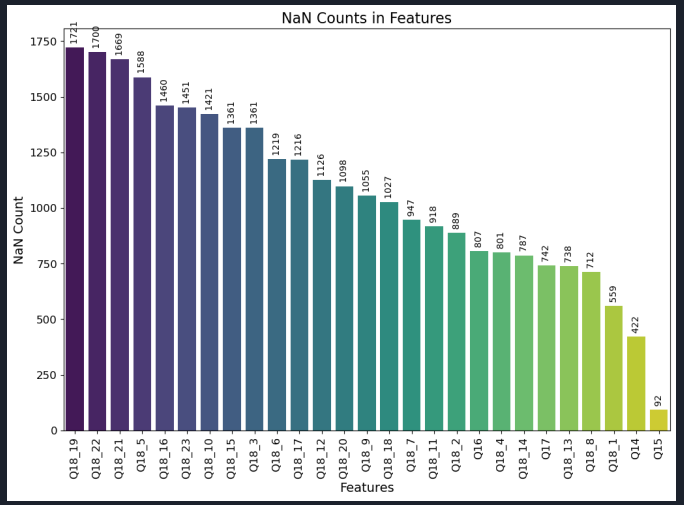
- Used the chi-square (chi2) statistical test for feature selection.
- Selected the best 60 features based on the chi-square test.

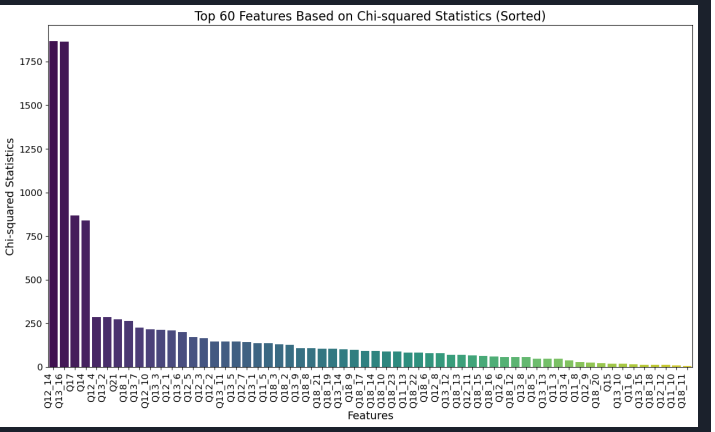
Handling Imbalanced Data:

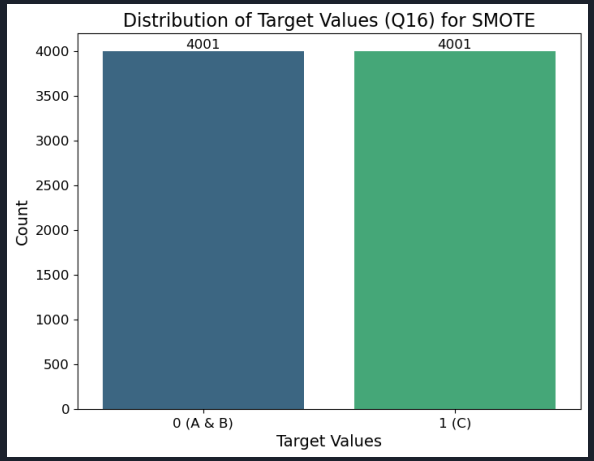
- Applied Random Oversampling and SMOTE oversampling techniques to address the imbalance issue in the target variable.

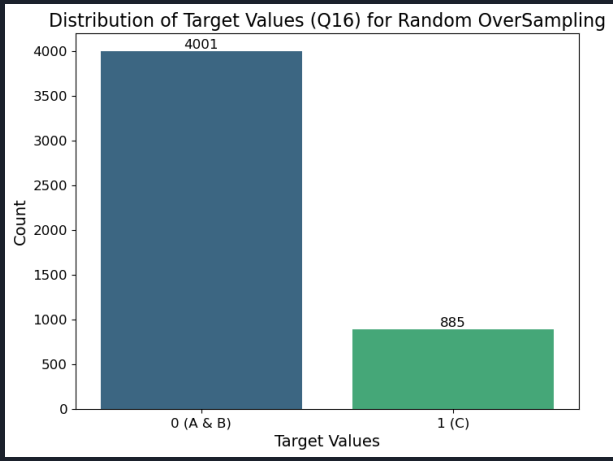

Number of "Dont Know" Answers
Count
Features
403
342
229
Q11_Dont_Know
Q13_Dont_Know
Q12_Dont_Know


Distribution of Target Values (Q16)
Count
Target Values
5005
1103
0 (A & B)
1 (C)

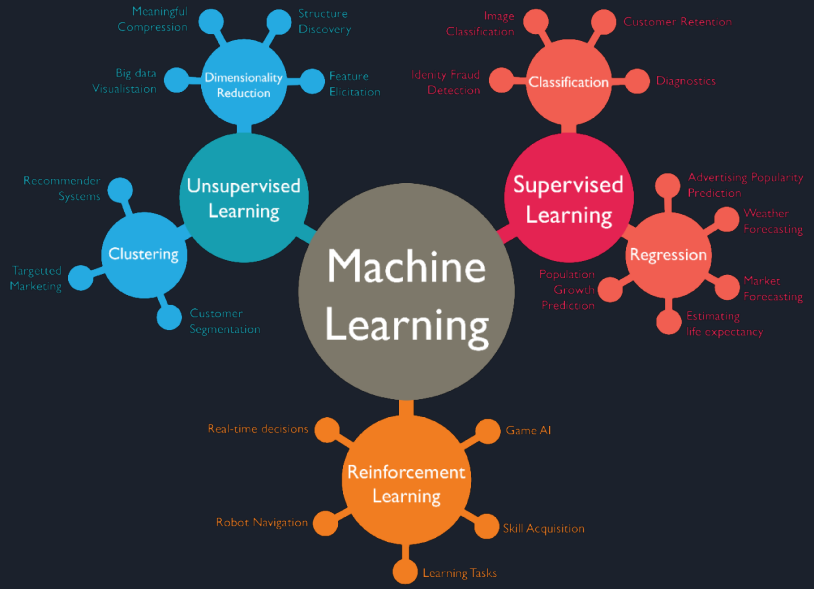

NaN Counts in Features
NaN Count
Features


Top 60 Features Based on Chi-squared Statistics (Sorted)
Chi-squared Statistics
Features


Distribution of Target Values (Q16) for SMOTE
Count
Target Values
4001
4001
0 (A & B)
1 (C)

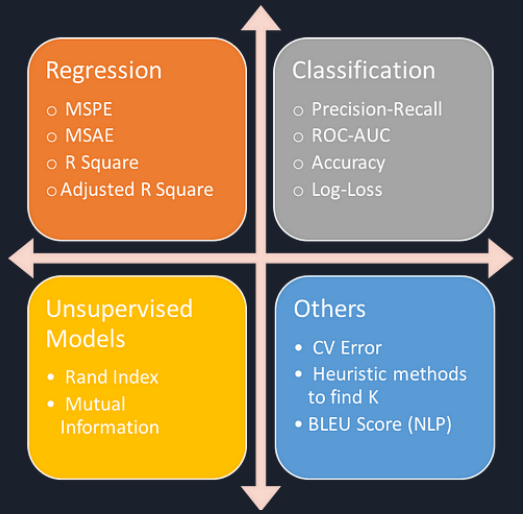

Distribution of Target Values (Q16) for Random OverSampling
Count
Target Values
4001
885
0 (A & B)
1 (C)

Classification Models



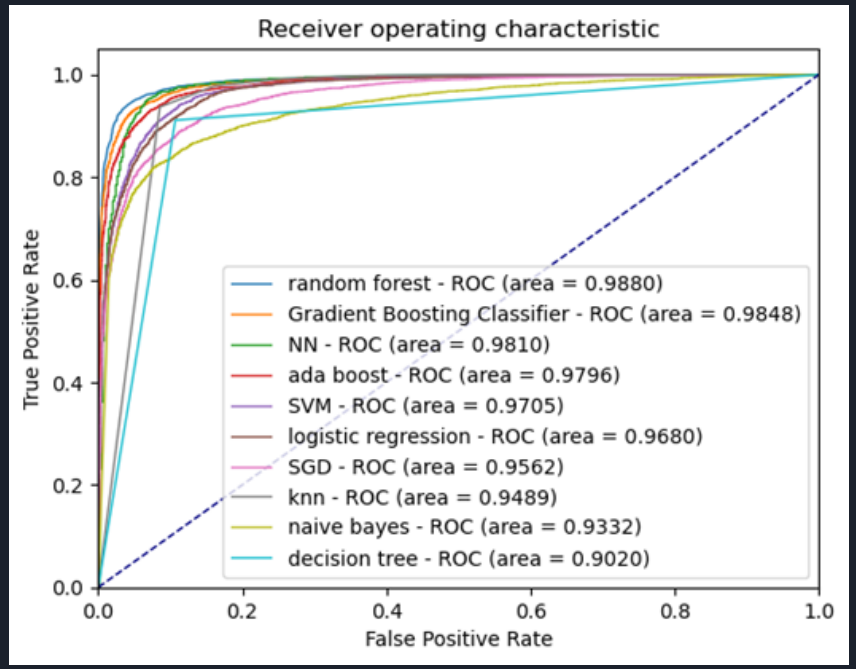
Ten different classifiers were selected to evaluate their performance on the original data, random oversampled data and the SMOTE-resampled data. These classifiers encompass a variety of algorithms, ranging from ensemble methods like Random Forest and AdaBoost to neural network-based approaches such as MLP. SVM, Logistic Regression, SGD, KNN, Naive Bayes, and Decision Tree were also included to provide a diverse set of comparisons.

Classification Models

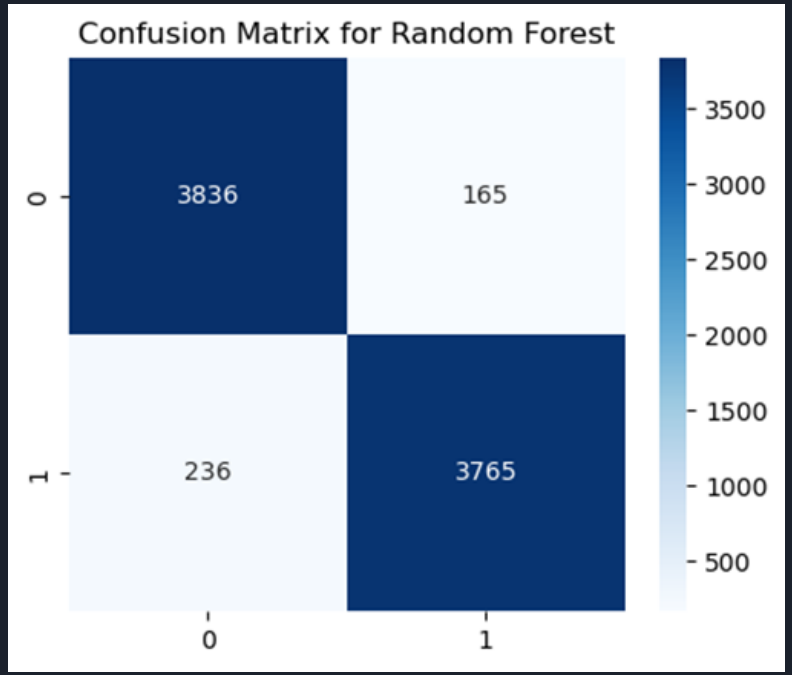


ROC curves were generated to visually assess the discriminative capabilities of each classifier. Key classification metrics, such as accuracy, area under the ROC curve (AUC), and F1 score, were computed to provide a comprehensive evaluation of model performance. Notably, a Confusion Matrix was exclusively crafted for the Random Forest algorithm, identified as the top performer based on classification accuracy.

ROC Curve



Confusion Matrix



Model Performance Metric

Classification accuracy, AUC, and F1, Precision, Recall and MCC score were calculated for each classifier to provide a comprehensive understanding of their overall performance on the original data, Random oversampling, SMOTE-oversampling data.

Random Forest

Metrics	Original Data	SMOTE Data	Random Sampling Data
CA	0.9160	0.9499	0.9147
AUC	0.8159	0.9499	0.8159
F1	0.7392	0.9494	0.7372
Precision	0.8414	0.9580	0.8333
Recall	0.6591	0.9410	0.6610
MCC	0.6973	0.8999	0.6937

Neural Network

Metrics	Original Data	SMOTE Data	Random Sampling Data
CA	0.9009	0.9410	0.8995
AUC	0.8210	0.9410	0.8110
F1	0.7104	0.9420	0.7079
Precision	0.7525	0.9263	0.7475
Recall	0.6727	0.9583	0.6723
MCC	0.6524	0.8826	0.6488

Gradient Boost

Metrics	Original Data	SMOTE Data	Random Sampling Data
CA	0.9163	0.9336	0.9190
AUC	0.8316	0.9396	0.8383
F1	0.7511	0.9394	0.7609
Precision	0.8116	0.9433	0.8171
Recall	0.6990	0.9355	0.7119
MCC	0.7040	0.8793	0.7148

Ada Boost

Metrics	Original Data	SMOTE Data	Random Sampling Data
CA	0.9119	0.9278	0.9073
AUC	0.8204	0.9278	0.8131
F1	0.7352	0.9273	0.7223
Precision	0.8041	0.9330	0.7895
Recall	0.6772	0.9218	0.6655
MCC	0.6865	0.8556	0.6706

Support Vector Machines (SVM)

Metrics	Original Data	SMOTE Data	Random Sampling Data
CA	0.9173	0.9124	0.9122
AUC	0.8195	0.9124	0.8113
F1	0.7443	0.9125	0.7293
Precision	0.8429	0.9119	0.8257
Recall	0.6664	0.9130	0.6531
MCC	0.7026	0.8248	0.6844

Decision Tree

Metrics	Original Data	SMOTE Data	Random Sampling Data
CA	0.8731	0.9020	0.8676
AUC	0.7904	0.9020	0.7858
F1	0.6529	0.9029	0.6427
Precision	0.6451	0.8949	0.6258
Recall	0.6609	0.9110	0.6576
MCC	0.5754	0.8042	0.5617

Logistic Regression

Metrics	Original Data	SMOTE Data	Random Sampling Data
CA	0.9157	0.9019	0.9140
AUC	0.8290	0.9031	0.8306
F1	0.7553	0.9031	0.7565
Precision	0.8385	0.9131	0.8349
Recall	0.6872	0.8933	0.6915
MCC	0.7128	0.8085	0.7132

Stochastic Gradient Descent (SGD)

Metrics	Original Data	SMOTE Data	Random Sampling Data
CA	0.9075	0.8874	0.8878
AUC	0.8135	0.8874	0.8074
F1	0.7224	0.8871	0.6876
Precision	0.7886	0.8896	0.6939
Recall	0.6664	0.8845	0.6814
MCC	0.6707	0.7748	0.6193

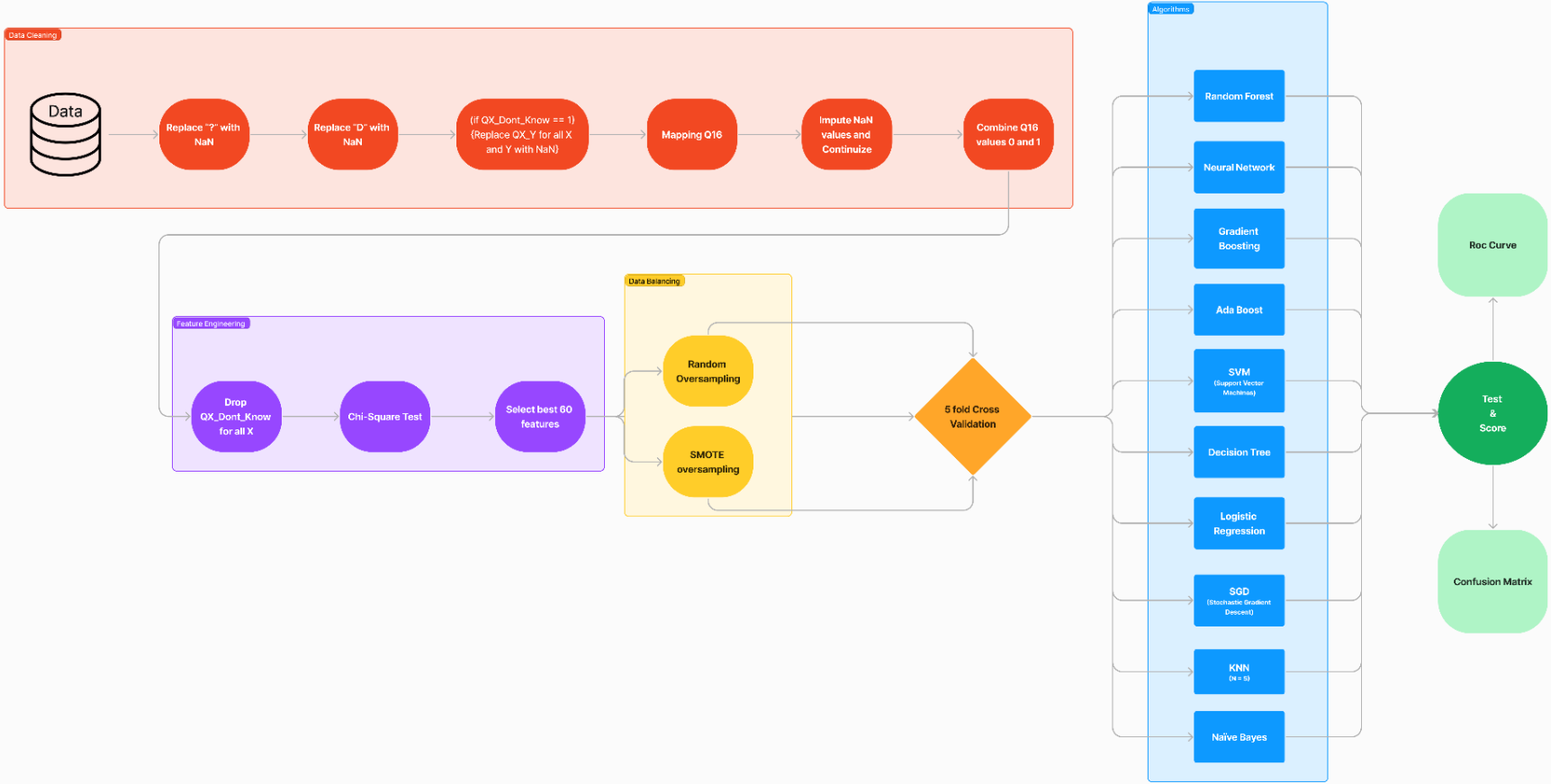
K-Nearest Neighbours (n= 5)

Metrics	Original Data	SMOTE Data	Random Sampling Data
CA	0.8988	0.8677	0.8946
AUC	0.8245	0.8677	0.8239
F1	0.7165	0.8823	0.7102
Precision	0.7252	0.7946	0.7074
Recall	0.7081	0.9918	0.7130
MCC	0.6550	0.7591	0.6458

Naive Bayes

Metrics	Original Data	SMOTE Data	Random Sampling Data
CA	0.8422	0.8672	0.8432
AUC	0.8245	0.8672	0.8268
F1	0.6458	0.8671	0.6493
Precision	0.5429	0.8676	0.5458
Recall	0.7969	0.8665	0.8011
MCC	0.5657	0.7343	0.5698

Data Mining Flow Chart



Thank You