Comparing the Performance of Classification Algorithms for Predicting Electric Vehicles Adoption

Ahmet Berkay Aslan[1], Buğra Mutluer[1]

[1] Çankaya University, Ankara, Türkiye
c1911009@student.cankaya.edu.tr, c1911040@student.cankaya.edu.tr

Abstract. This project evaluates the performance of various classification models on a dataset. The models are trained on the original dataset, a dataset balanced with SMOTE, and a dataset balanced with Random Oversampling. The performance of each model is evaluated using several metrics.

Keywords: Electric Vehicle Adoption, Classification Algorithms, Data Imputation, Feature Selection, Imbalanced Data Handling, ROC Analysis, Confusion Matrix, Performance Metrics, SMOTE Resampling, Random OverSampling, Machine Learning, Random Forest, Logistic Regression, Gradient Boost, KNN, Naive Bayes, Decision Tree, SVM, Ada Boost, Neural Network, Stochastic Gradient Descent, Data Mining, Hyperparameter Optimization.

1 Introduction

In this report, we present the results of our data mining project. The goal of this project is to classify the data using various machine learning algorithms and evaluate their performance. We have used ten different algorithms, including Random Forest, Logistic Regression, Gradient Boost, KNN, Naive Bayes, Decision Tree, SVM, Ada Boost, Neural Network, and Stochastic Gradient Descent, to classify the data. The models are trained on three versions of the dataset: the original dataset, a dataset balanced with Synthetic Minority Over-sampling Technique (SMOTE), and a dataset balanced with Random Oversampling with 5-fold cross-validation.

2 Methodology

2.1 Data Preprocessing

Handling Missing Values and Identifying “Don’t Know” Answers: Addressing Missing Values and Identifying "Don't Know" Answers: Prior to initiating any algorithms for analysis, all "?" values were replaced with NaN. Additionally, instances where individuals responded with "D," signifying "Don't Know" in the target value, were identified and substituted with NaN for imputation. Subsequently, individuals who answered Q11, Q12, and Q13 as "Don't Know" had their responses replaced with NaN for imputation purposes. (Shown in the ***figure 1***)

Understanding the Target: After getting rid of the D value on the target. It was observed that there are two positive answers to the target (Q16). These two values were combined as 0, while the "C" value was replaced with 1, representing the negative. (Shown in the ***figure 2***)

Imputation: The KNN imputation method was implemented to estimate and replace NaN values in the dataset, aiming for more accurate analysis across ten algorithms. (Shown in the ***figure 3***)

Feature Selection: Utilizing the chi-square (chi2) statistical test for feature selection, the best 60 features were chosen based on the chi-square test, as it yielded optimal results for the ten algorithms (Shown in the ***figure 4***).

Handling Imbalanced Data: Recognizing the prevalence of more "0" values than "1" in the target, causing data imbalance and impacting algorithm performance evaluation, Random Oversampling and SMOTE oversampling techniques were applied to address this imbalance. (Shown in the ***figure 5*** and ***figure 6***)

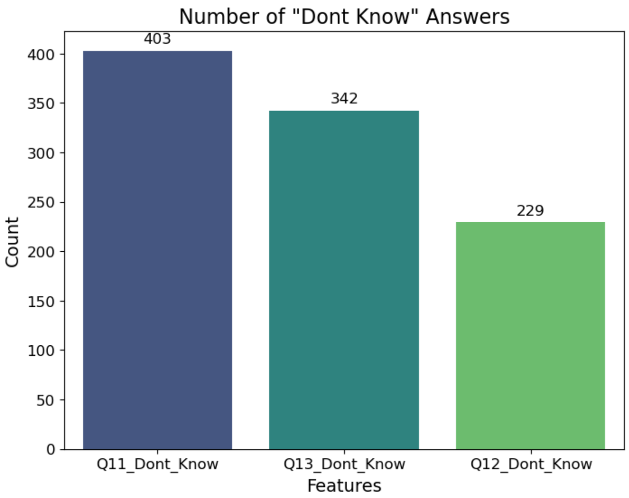


Fig. 1. Number of "Don't Know" Answers.

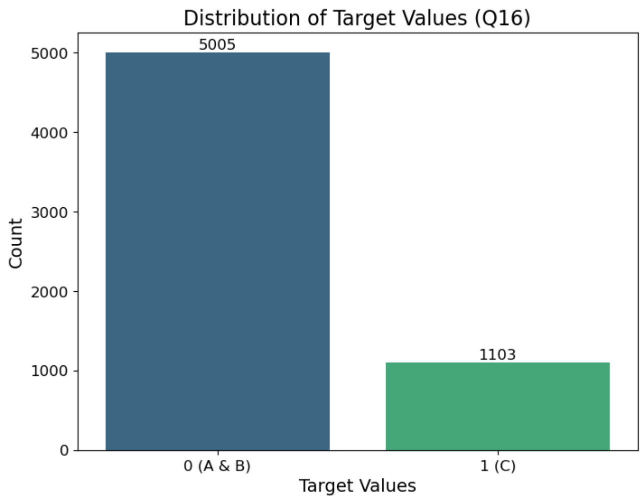


Fig. 2. Distribution of Target Values (Q16).

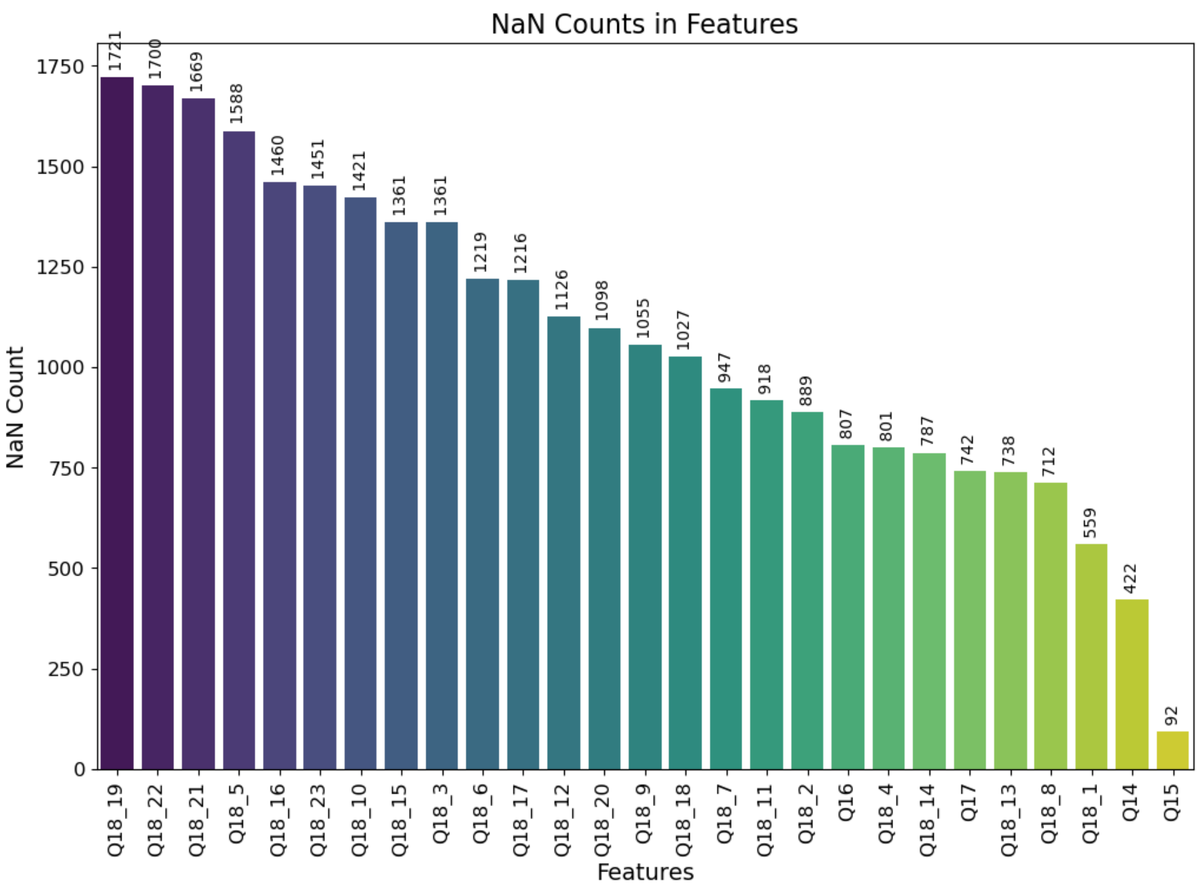


Fig. 3. Count of NaN in Features.

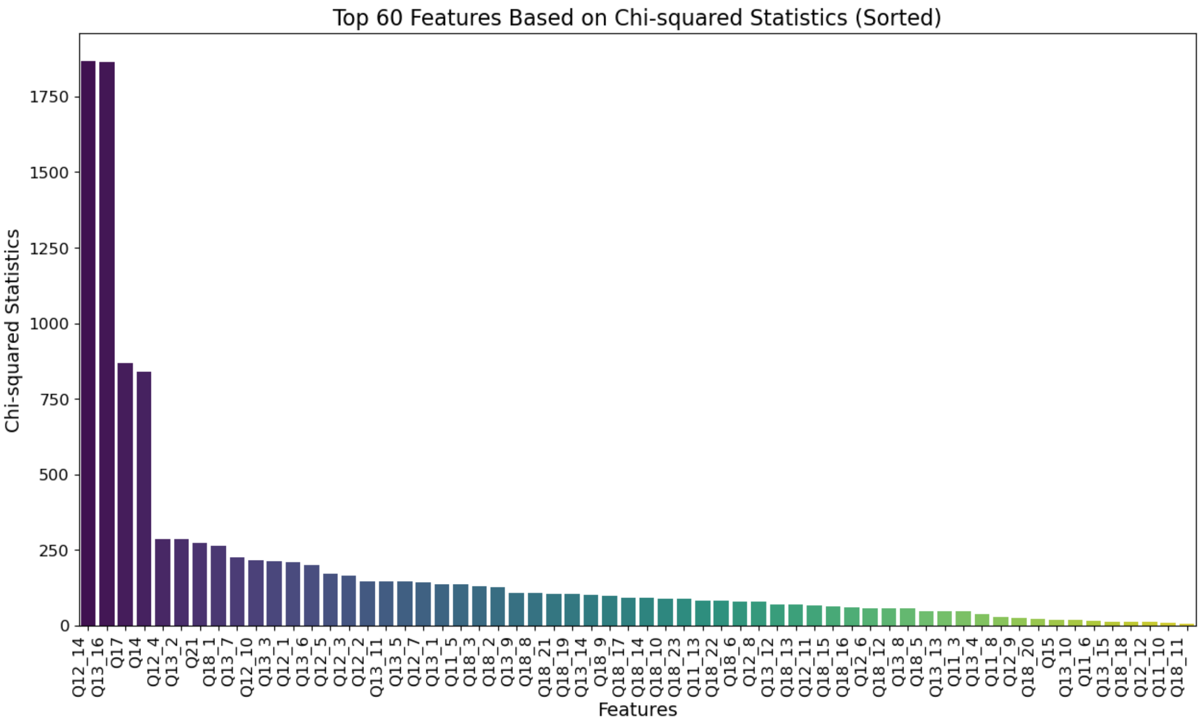
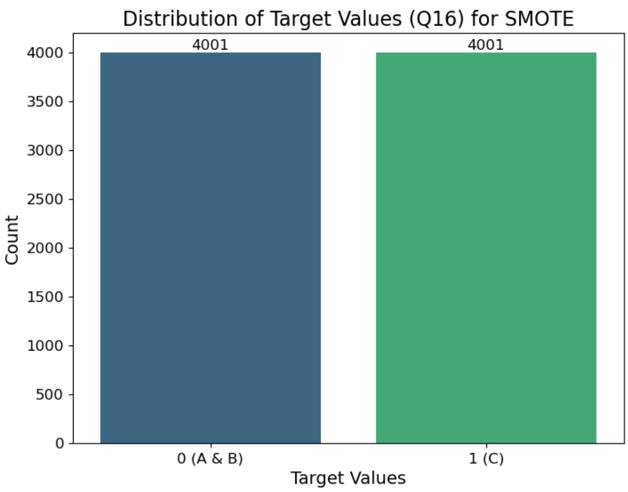


Fig. 4. Best 60 features based on Chi-Square.



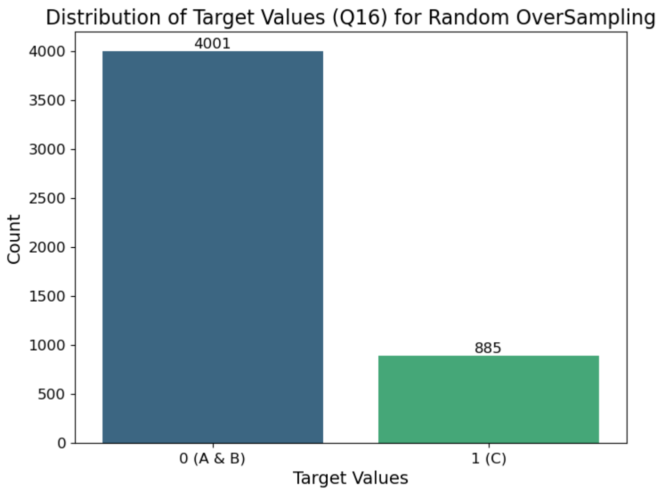


Fig. 5. New Distribution of Target Values for SMOTE. **Fig. 6.** New Distribution of Target Values for Random OverSampling.

2.2 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.

2.3 Evaluation Metrics

A thorough evaluation of classifier performance involved the calculation of multiple metrics across different datasets, including the original data, random oversampled data, and SMOTE-oversampled data. The following key metrics were computed to offer a comprehensive understanding of each classifier's effectiveness:

Classification Accuracy:

- Assessing the overall correctness of the classifier's predictions across the different datasets.

Area Under the ROC Curve (AUC):

- Providing insights into the classifier's ability to discriminate between positive and negative instances, particularly crucial in imbalanced datasets.

F1 Score:

- Balancing precision and recall, the F1 score is especially informative in scenarios where there is an uneven distribution between the classes.

Precision:

- Indicating the proportion of true positive predictions among all positive predictions, emphasizing the accuracy of positive identifications.

Recall:

- Highlighting the ratio of true positive predictions to the total actual positives, essential for assessing the classifier's ability to capture all positive instances.

MCC Score (Matthews Correlation Coefficient):

- Offering a more nuanced measure that considers both false positives and false negatives, providing a comprehensive assessment of classifier performance.

This meticulous approach allows for a nuanced understanding of each classifier's strengths and weaknesses across various datasets, contributing to a more informed decision-making process in model selection and optimization.

3 Data

This dataset originates from a survey focused on electric vehicles. It contains 6108 participants, 74 features. It encompasses various components:

1. Participant Preferences and Opinions:
   - Q11: General vehicle preferences (numerical responses).
   - Q12: Preferences regarding electric vehicles (numerical responses).
   - Q13: Factors influencing electric vehicle preferences (numerical responses).

2. Social Perspectives:
   - Q14: Social opinions on rewarding electric cars (categorical responses).

3. User Experience and Perception:
   - Q15: Participants' experience with driving electric cars (categorical responses).
   - Q17: Suitability of electric cars for daily driving needs (categorical responses).

4. Changes in Views and Opinions:
   - Q20: Changes in opinions regarding internal combustion engine cars (numerical responses).
   - Q21: Changes in opinions regarding electric cars (numerical responses).

5. Perceptions and Biases Towards Electric Vehicles:
   - Q18: Participants' general thoughts on electric vehicles (categorical responses).

6. Target Variable:
   - Q16: Target variable measuring preferences for electric vehicles (categorical responses).

This dataset provides insights into widespread acceptance of electric vehicles, preferred features, societal perceptions, and potential biases. It can be valuable for understanding user attitudes towards electric vehicles and informing strategies for increasing their adoption.

Table 1. Data fields in the dataset.

Attribute Name	Description	Data Type
Q11	If you were to buy a car tomorrow, what would be the most important criteria for you in your choice of an car? Multiple answers are possible	Numerical
Q11_1	Driving economy	Numerical
Q11_2	*Driving characteristics*	Numerical
Q11_3	*Range*	Numerical
Q11_4	*Purchase price*	Numerical
Q11_5	*Environmental and climate impact (CO2 per km)*	Numerical
Q11_6	*Expected value loss*	Numerical
Q11_7	Type and size (hatchback, sedan, stationcar etc.)	Numerical
Q11_8	Car brand	Numerical
Q11_9	Acceleration	Numerical
Q11_10	Operational costs (maintenance, insurance, etc)	Numerical
Q11_11	Design	Numerical
Q11_12	Top speed	Numerical
Q11_13	Another reason, not mentioned in this list	Numerical
Q11_14	Don't know	Numerical

Q12	If you were to buy an electric car tomorrow, what would be the most important thing for you in your choice of an electric car? Multiple answers per responder	Numerical
Q12_1	Driving economy	Numerical
Q12_2	Driving characteristics	Numerical
Q12_3	Range	Numerical
Q12_4	Purchase price	Numerical
Q12_5	Environmental and climate impact (CO2 per km)	Numerical
Q12_6	Expected value loss	Numerical
Q12_7	Type and size (hatchback, sedan, stationcar etc.)	Numerical
Q12_8	Car brand	Numerical
Q12_9	Acceleration	Numerical
Q12_10	Operational costs (maintenance, insurance, etc)	Numerical
Q12_11	Design	Numerical
Q12_12	Top speed	Numerical
Q12_13	Another reason, not mentioned in this list	Numerical
Q12_14	Would not buy an EV	Numerical
Q12_15	Don't know	Numerical
Q13	Which benefits could make you more positive about buying an electric car? Multiple answers per responder	Numerical
Q13_1	Lower registration fee than other cars	Numerical
Q13_2	Positive effect on global climate	Numerical
Q13_3	Less noise	Numerical
Q13_4	Higher acceleration	Numerical
Q13_5	Faster charging	Numerical
Q13_6	Lower operational costs than other cars	Numerical
Q13_7	Positive effect on the local environment	Numerical
Q13_8	Improved driving characteristics	Numerical
Q13_9	More charging stations	Numerical
Q13_10	Improved towing capabilities (trailer)	Numerical
Q13_11	Improved fuel economy over other cars	Numerical
Q13_12	More and improved parking possibilities for electric cars only	Numerical
Q13_13	Longer range	Numerical
Q13_14	The price	Numerical
Q13_15	Another reason, not mentioned in this list	Numerical
Q13_16	Nothing	Numerical
Q13_17	Don’t know	Numerical
Q14	Do you agree or disagree with the following statement? The society must reward electric cars instead of petrol and diesel cars	Categorical
Q15	Have you tried to drive an electric car yourself?	Categorical
Q17	If you only had an electric car, how good or bad do you think it would suit your daily driving needs?	Categorical
Q18	In what manner do you agree with the following statements	Categorical

Q18_1	Electric cars are boring	Categorical
Q18_2	The safety is not good in an electric car.	Categorical
Q18_3	There is a greater risk that electric cars will burst into flames compared to petrol and diesel cars	Categorical
Q18_4	It is difficult to buy an electric car	Categorical
Q18_5	It is difficult to get an electric car repaired	Categorical
Q18_6	It will become difficult to sell a used electric car because the technology develops so fast	Categorical
Q18_7	The selection of different models of electric cars is too small	Categorical
Q18_8	An electric car can not drive far enough to cover my daily driving	Categorical
Q18_9	An electric car can not drive fast enough to drive on the highway	Categorical
Q18_10	The weather has a too big an impact on the range to use an electric car car in my everyday life	Categorical
Q18_11	Electric cars are in an early stage of development. It's better to buy an electric car in five years	Categorical
Q18_12	Electric cars are a temporary solution. There will be a better and cleaner technology in the future	Categorical
Q18_13	Electric cars are best for short trips in cities, but are not suitable for long journeys by country roads and highways	Categorical
Q18_14	An electric car is only suitable for car number two	Categorical
Q18_15	It's too slow to charge an electric car while on the go	Categorical
Q18_16	It's too difficult to cross national borders in electric cars when you need to charge along the way	Categorical
Q18_17	It is too difficult to charge an electric car	Categorical
Q18_18	There are too few public chargers	Categorical
Q18_19	Electric cars are more expensive in daily operation than conventional petrol and diesel cars	Categorical
Q18_20	Electric cars are more expensive in purchasing compared to similar petrol and diesel cars	Categorical
Q18_21	Electric car mostly run on coal-generated power. Therefore they are not as climate-friendly as they are being portrayed.	Categorical
Q18_22	The battery in an electric car has a short life and is expensive to replace	Categorical
Q18_23		Categorical
Q20	How did your opinion about cars with a diesel or gasoline motor change during the past year?	Numerical
Q21	How did your opinion about electric cars change during the past year?	Numerical
Q16	Which of the following statements about electric car suits you the best?	Categorical(TARGET)

4 Analysis and Results

4.1 Roc Curve Analysis

ROC curves were plotted for each classifier with the SMOTE-oversampling data to visualize their performance in terms of true positive rate (sensitivity) against the false positive rate. The area under the ROC curve (AUC) was computed to quantify the classifiers' ability to discriminate between the positive and negative classes.

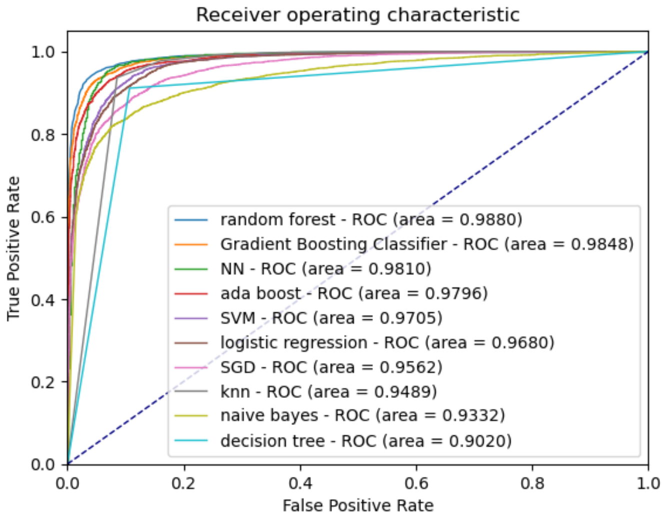


Fig. 7. ROC curve for the tested classification algorithms.

4.2 Confusion Matrix Analysis

A confusion matrix was constructed for the best algorithm to provide a detailed breakdown of the classifier's performance:

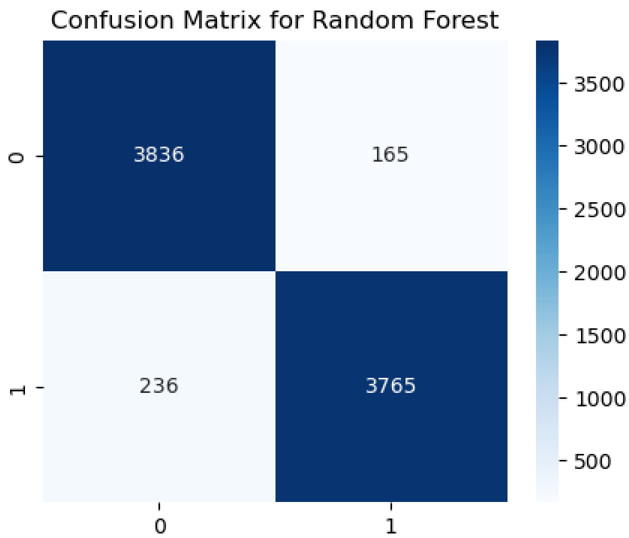


Fig. 8. Confusion matrix for the selected **Random Forest** algorithm, which has the highest CA (Classification Accuracy) value.

4.3 Model Performance Metrics

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: is a decision tree-based algorithm that uses an ensemble of decision trees to classify the data. We used the scikit-learn implementation of Random Forest to classify the data. We used 5-fold cross-validation to evaluate the performance of the algorithm. The results of our experiments are summarized below:

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: is a machine learning algorithm that is inspired by the structure and function of the human brain. We used the scikit-learn implementation of Neural Network to classify the data. We used 5-fold cross-validation to evaluate the performance of the algorithm. The results of our experiments are summarized below:

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: is a decision tree-based algorithm that uses an ensemble of decision trees to classify the data. We used the scikit-learn implementation of Gradient Boost to classify the data. We used 5-fold cross-validation to evaluate the performance of the algorithm. The results of our experiments are summarized below:

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: is a boosting algorithm that creates a powerful classifier by combining several low-performing classifiers, We used the scikit-learn implementation of Ada Boost to classify the data. We used 5-fold cross-validation to evaluate the performance of the algorithm. The results of our experiments are summarized below:

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): is a linear and non-linear algorithm that separates the data into classes using a hyperplane. We used the scikit-learn implementation of SVM to classify the data. We used 5-fold cross-validation to evaluate the performance of the algorithm. The results of our experiments are summarized below:

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: is a decision tree-based algorithm that uses a tree-like model of decisions and their possible consequences to classify the data. We used the scikit-learn implementation of Decision Tree to classify the data. We used 5-fold cross-validation to evaluate the performance of the algorithm. The results of our experiments are summarized below:

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: is a statistical algorithm that uses a logistic function to model a binary dependent variable. We used the scikit-learn implementation of Logistic Regression to classify the data. We used 5-fold cross-validation to evaluate the performance of the algorithm. The results of our experiments are summarized below:

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): is a variant of the Gradient Descent algorithm that is used for optimizing machine learning models. We used the scikit-learn implementation of SGD to classify the data. We used 5-fold cross-validation to evaluate the performance of the algorithm. The results of our experiments are summarized below:

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 Neighbors (KNN): is a non-parametric algorithm that classifies the data based on the k-nearest neighbors in the feature space. We used the scikit-learn implementation of KNN to classify the data. We used 5-fold cross-validation to evaluate the performance of the algorithm. The results of our experiments are summarized below:

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: is a probabilistic algorithm that uses Bayes' theorem to classify the data. We used the scikit-learn implementation of Naive Bayes to classify the data. We used 5-fold cross-validation to evaluate the performance of the algorithm. The results of our experiments are summarized below:

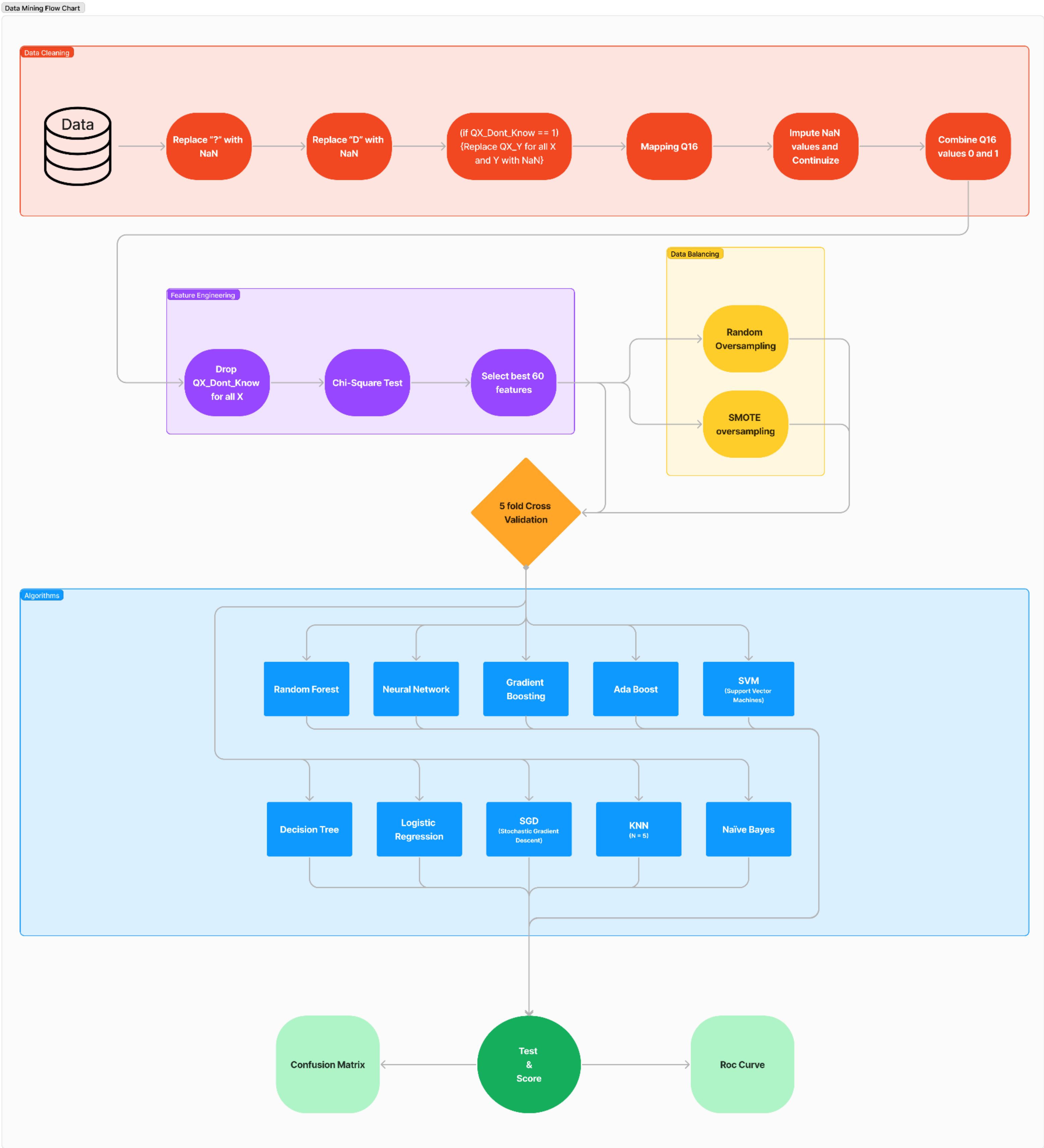
Metrics	**Original Data**	**SMOTE Data**	**Random Sampling Data**
CA	0.8422	0.8672	0.8432
AUC	0.8245	0.8672	0.8268
F1	0.6458	0.8671	0.6493
Precision	0.5429	0.8676	0.5458
Recall	0.7969	0.8665	0.8011
MCC	0.5657	0.7343	0.5698

Algorithm	CA (Classification Accuracy)	AUC (Area Under Curve)
Random Forest	*0.9499*	0.9880
Neural Network	*0.9410*	0.9810
Gradient Boosting	*0.9396*	0.9848
Ada Boost	*0.9278*	0.9796
SVM	*0.9173*	0.9705
Decision Tree	*0.9020*	0.9020
Logistic Regression	*0.9019*	0.9680
Stochastic Gradient Descent	*0.8874*	0.9562
Knn (n=5)	*0.8677*	0.9489
Naïve Bayes	*0.8672*	0.9332

5 Conclusions

The analysis reveals varying degrees of performance across different classifiers on the SMOTE-resampled data. The Random Forest classifier demonstrated the highest accuracy and F1 score, suggesting its effectiveness in handling the resampled dataset. However, the choice of the best classifier depends on the specific goals and requirements of the application. Further fine-tuning and optimization of hyperparameters could potentially enhance the performance of the classifiers. Overall, this study contributes insights into the impact of SMOTE resampling on classifier performance, aiding in the selection of suitable models for imbalanced datasets.

6 Flowchart



Data Mining Flow Chart
Data Cleaning
Data
Replace "?" with NaN
Replace "D" with NaN
(if QX_Dont_Know == 1) {Replace QX_Y for all X and Y with NaN}
Mapping Q16
Impute NaN values and Continuize
Combine Q16 values 0 and 1
Feature Engineering
Drop QX_Dont_Know for all X
Chi-Square Test
Select best 60 features
Data Balancing
Random Oversampling
SMOTE oversampling
5 fold Cross Validation
Algorithms
Random Forest
Neural Network
Gradient Boosting
Ada Boost
SVM (Support Vector Machines)
Decision Tree
Logistic Regression
SGD (Stochastic Gradient Descent)
KNN (N = 5)
Naïve Bayes
Confusion Matrix
Test & Score
Roc Curve