Machine learning to classify federal deputies elected in Brazilian elections

Diego Cunha[1]

[1] Federal University of Bahia, Salvador, Bahia, Brazil

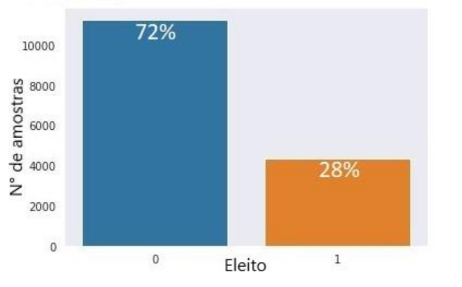
Abbreviated abstract: Implementation of Machine Learning algorithms to predict elected Federal Deputy in the 2022 General Elections in Brazil, using historical data from the 2018, 2014 and 2010 elections. decision and Naive Bayes with and without the use of the SMOTE technique. And compare the results presented by the metrics of accuracy, precision and recall, as well as evaluate the ROC curve generated for each model.

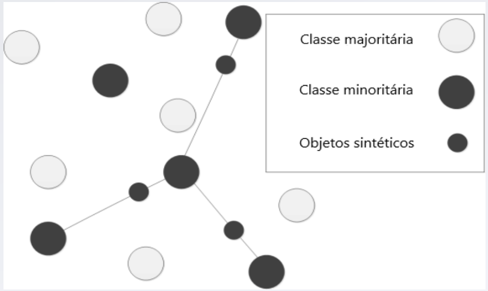
ECD
ESPECIALIZAÇÃO EM CIÊNCIA DE DADOS E BIG DATA

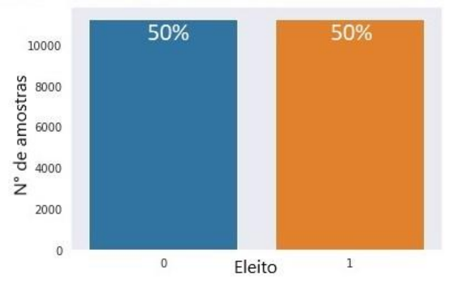
4th Conference on
Statistics and Data Science
Salvador, Brazil (online)
December 1-3, 2022

Problem, Data, Previous Works

Predict elected candidates for the position of Federal Deputy in the 2022 elections in Brazil. With the use of data provided by the Superior Electoral Court (TSE), and applying the Logistic Regression models, KNN K-Nearest Neighbor, Decision Tree and Naive Bayes (with and without the use of the Oversampling technique, Synthetic Minority Oversampling Technique, or SMOTE, which consists of generating synthetic (non-duplicated) data of the minority class from its neighbors) and at the end compare the results of each model.







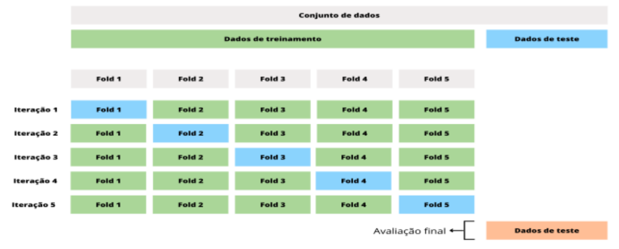
ECD
ESPECIALIZAÇÃO EM CIÊNCIA DE DADOS E BIG DATA

4th Conference on
Statistics and Data Science
Salvador, Brazil (online)
December 1-3, 2022

Methods

- Data pre-processing stage for information unification.
- The original dataset consists of 24 independent variables and 15,624 observations.
- A data set was created from the application of the SMOTE method, for data balancing.

The use of k-fold will help us to resample our dataset at each iteration, and provide different training combinations, and in this way we will ensure greater confidence in the results of our models.

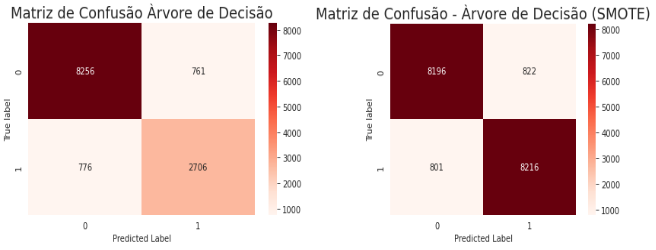


Conjunto de dados

Dados de treinamento					Dados de teste

	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Iteração 1	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Iteração 2	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Iteração 3	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Iteração 4	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Iteração 5	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5

Avaliação final ← Dados de teste


Matriz de Confusão Àrvore de Decisão

Matriz de Confusão - Àrvore de Decisão (SMOTE)

Accuracy of 87.58% for unbalanced data. Comparing with the results of the balanced data, the result was 91% of accuracy.

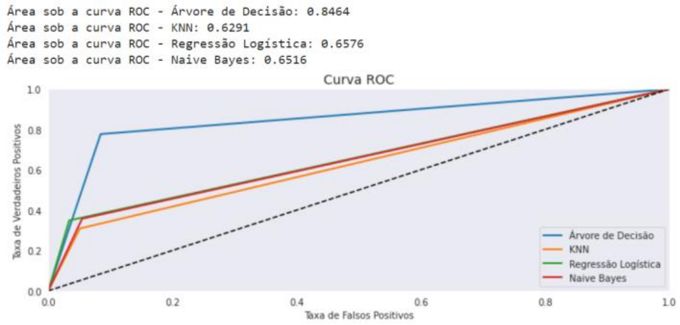
The decision tree model fit the balanced dataset very well.

ECD
ESPECIALIZAÇÃO EM CIÊNCIA DE DADOS E BIG DATA

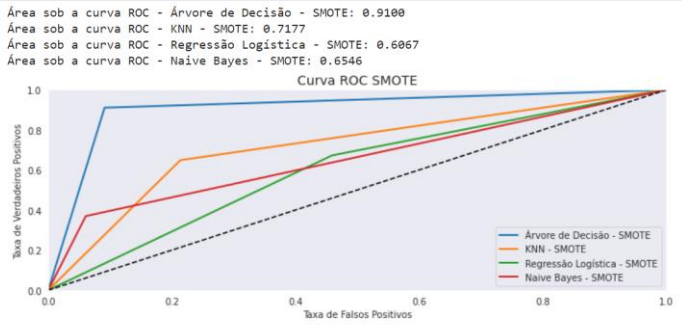
4th Conference on
Statistics and Data Science
Salvador, Brazil (online)
December 1-3, 2022

Results and Conclusions

ROC curve comparison for decision tree, KNN, logistic regression and naive bayes models. The AUC value of the decision tree model improved its performance from 84% to 91%.



```
Área sob a curva ROC - Árvore de Decisão: 0.8464
Área sob a curva ROC - KNN: 0.6291
Área sob a curva ROC - Regressão Logística: 0.6576
Área sob a curva ROC - Naive Bayes: 0.6516
```



```
Área sob a curva ROC - Árvore de Decisão - SMOTE: 0.9100
Área sob a curva ROC - KNN - SMOTE: 0.7177
Área sob a curva ROC - Regressão Logística - SMOTE: 0.6067
Área sob a curva ROC - Naive Bayes - SMOTE: 0.6546
```

For future work

- Increment of relevant information from candidates to the dataset
- replicate studies for other positions

ECD
ESPECIALIZAÇÃO EM CIÊNCIA DE DADOS E BIG DATA

4th Conference on
Statistics and
Data Science
Salvador, Brazil (online)
December 1-3, 2022