Machine learning for classification of members of Colombian congress according to their number of judicial processes

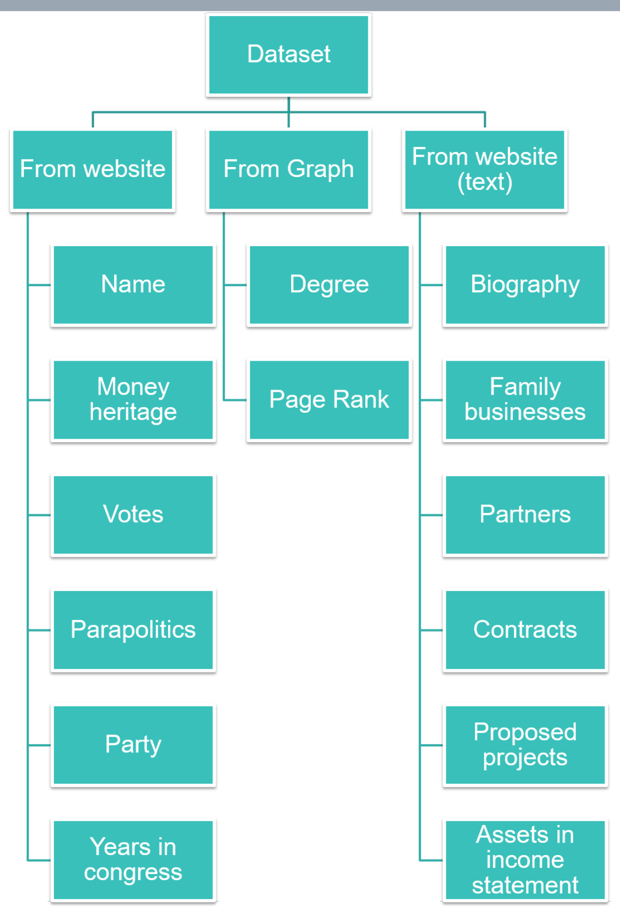
Glenn Harry Amaya and Lillian Daniela Beltrán

https://youtu.be/MiGoB2zWGEc

Introduction

Corruption in the Colombian context is one of the main inconveniences which affects the country because of, according to the Corruption Perception Index, Colombia is ranked 92 out of 180 countries worldwide, besides that 63% of Colombians consider that the members of congress are the people most involved in acts of corruption. Based on the above, the aim of this work is to train a machine learning model that provides indications of the transparency of the candidates based on the political trajectory, business, partners, biography, connections and heritage of current members of congress, this transparency will be measured by the number of judicial processes in which a member of Congress has been involved. Hence the trained model could be applied to new candidates for Congress in the upcoming elections of 2022.

Dataset



Dataset

From website

- Name
- Money heritage
- Votes
- Parapolitics
- Party
- Years in congress

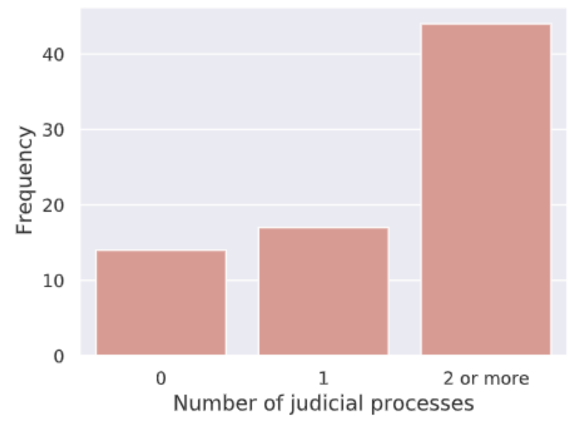
From Graph

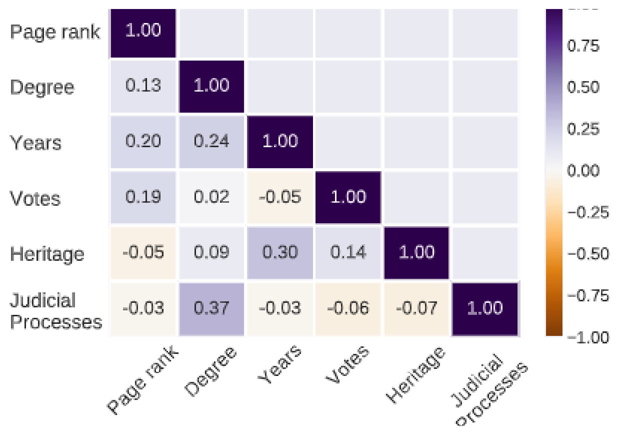
- Degree
- Page Rank

From website (text)

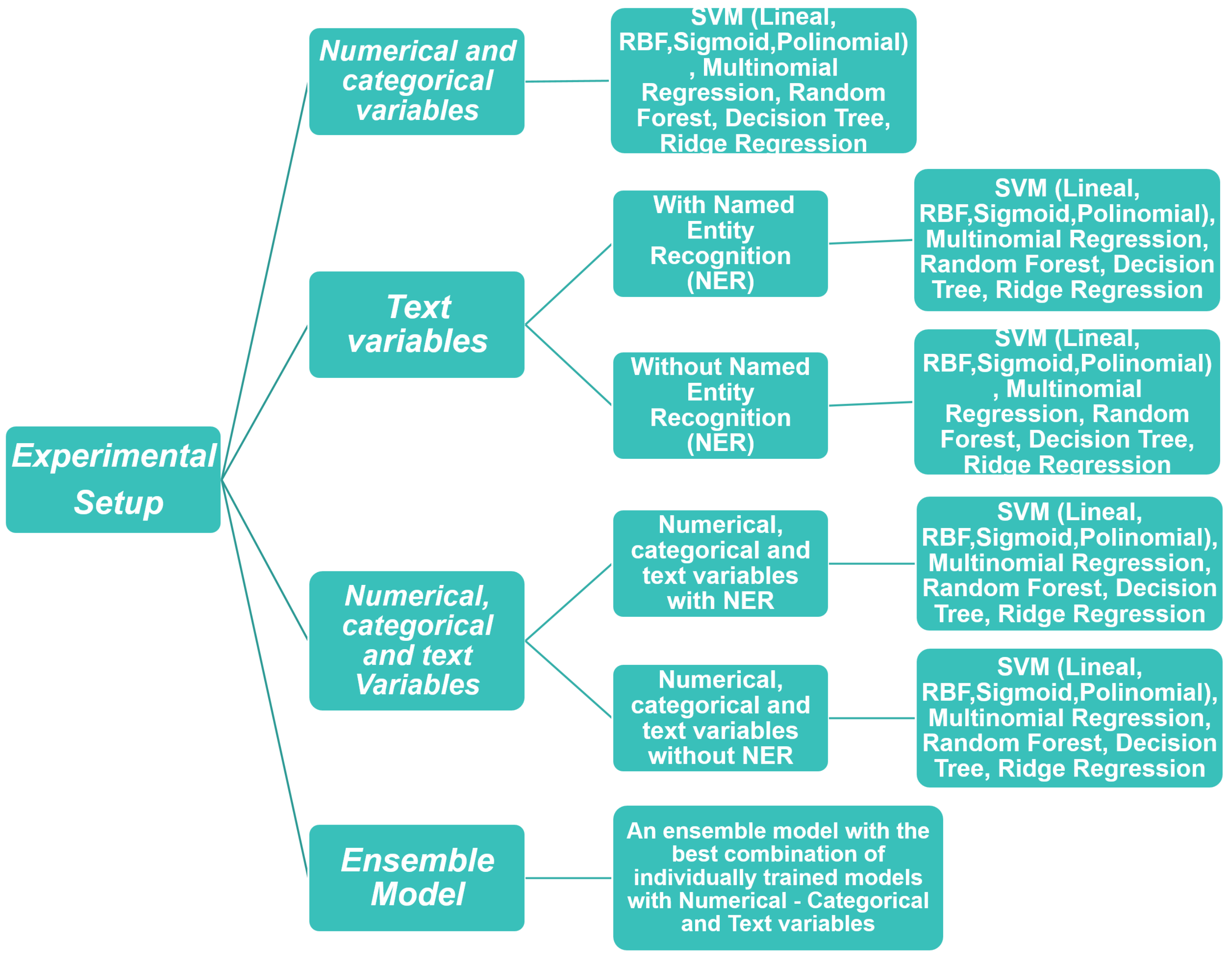
- Biography
- Family businesses
- Partners
- Contracts
- Proposed projects
- Assets in income statement

Descriptive Analysis





Experimental setup



Experimental Setup

Numerical and categorical variables
- SVM (Lineal, RBF,Sigmoid,Polinomial) , Multinomial Regression, Random Forest, Decision Tree, Ridge Regression

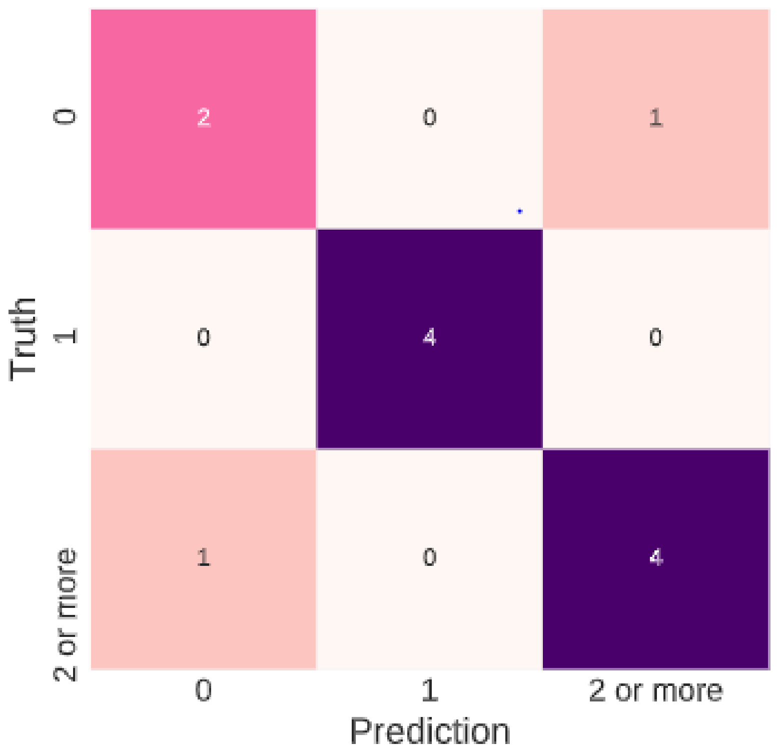
Text variables
- With Named Entity Recognition (NER)
  - SVM (Lineal, RBF,Sigmoid,Polinomial), Multinomial Regression, Random Forest, Decision Tree, Ridge Regression
- Without Named Entity Recognition (NER)
  - SVM (Lineal, RBF,Sigmoid,Polinomial) , Multinomial Regression, Random Forest, Decision Tree, Ridge Regression

Numerical, categorical and text Variables
- Numerical, categorical and text variables with NER
  - SVM (Lineal, RBF,Sigmoid,Polinomial), Multinomial Regression, Random Forest, Decision Tree, Ridge Regression
- Numerical, categorical and text variables without NER
  - SVM (Lineal, RBF,Sigmoid,Polinomial), Multinomial Regression, Random Forest, Decision Tree, Ridge Regression

Ensemble Model
- An ensemble model with the best combination of individually trained models with Numerical - Categorical and Text variables

Results

Mean F1-Score		Test Metrics	
Train	Validation	Accuracy	AUC Score
0.815	0.638	0.833	0.934



- *The data that demonstrates to have a better relationship with the number of judicial processesis the textual information due to the metrics of this type of models are the highest among the other configurations of data. Moreover, when it is combined both numerical-categorical data with text the best results are obtained.*
- *Due to the size of the data set and unbalance of the target variable, we had some issues fitting the models. However, using weights applied proportionally to the size of the classes to tackle the problem of unbalanced was a helpful strategy to solve those problems.*