

# Raport WUM 1

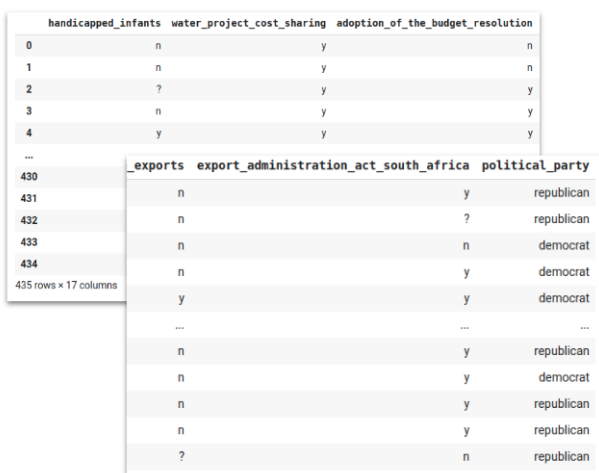
Jan Gąska, Kacper Grzymkowski

## 1.1 Opisywany problem

Naszym zadaniem było utworzenie modelu, który dokonywałby zadania klasyfikacji przynależności do partii politycznej demokratów bądź republikanów na podstawie oddanych głosów na ustawy w Kongresie Stanów Zjednoczonych Ameryki podczas jego obrad.

Naszą ramkę danych uzyskaliśmy z następującego źródła:

<https://www.apispreadsheets.com/datasets/121>



	handicapped_infants	water_project_cost_sharing	adoption_of_the_budget_resolution	
0	n		y	n
1	n		y	n
2	?		y	y
3	n		y	y
4	y		y	y
...				
430		exports	export_administration_act_south_africa	political_party
431	n		y	republican
432	n		?	republican
433	n		n	democrat
434	n		y	democrat
	y		y	democrat
...			...	...
	n		y	republican
	n		y	democrat
	n		y	republican
	n		y	republican
	?		n	republican

## 1.2. Zbiór danych

Zbiór danych jest dosyć nietypowy - wszystkie zmienne są ciągami znaków, ale przedstawiają pewne binarne zachowania - jeżeli polityk zagłosował za ustawą, to nie zagłosował przeciwko.

Dane składały się z 16 kolumn, które opisują jak dany polityk głosował nad pewnymi kluczowymi ustawami w Kongresie Stanów Zjednoczonych:

- “y” - polityk głosował za ustawą
- “n” - polityk głosował przeciw ustawie
- “?” - polityk wstrzymał się od głosu lub nie głosował

Politycy w Stanach Zjednoczonych należą do jednej z dwóch partii (nasza zmienna celu):

- “republican” - Partia republikańska
- “democrat” - Partia demokratyczna

Dane pochodzą z 1984 r. z Izby Reprezentantów USA z 98 posiedzenia Kongresu.

Sprawdziliśmy kilka rzeczy na temat tego okresu:

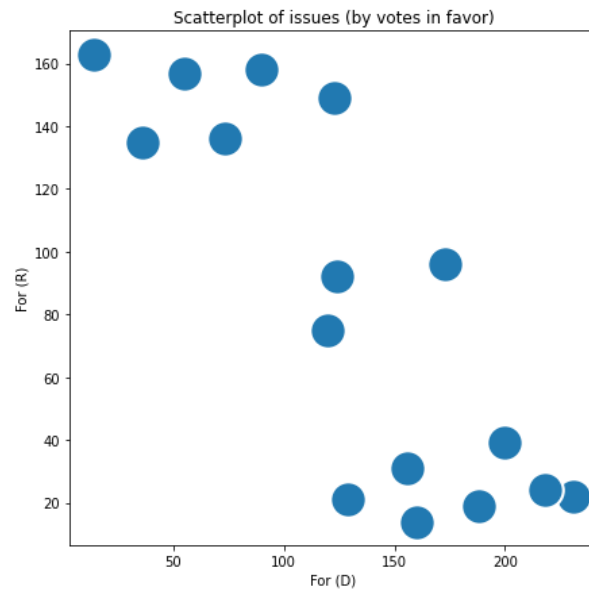
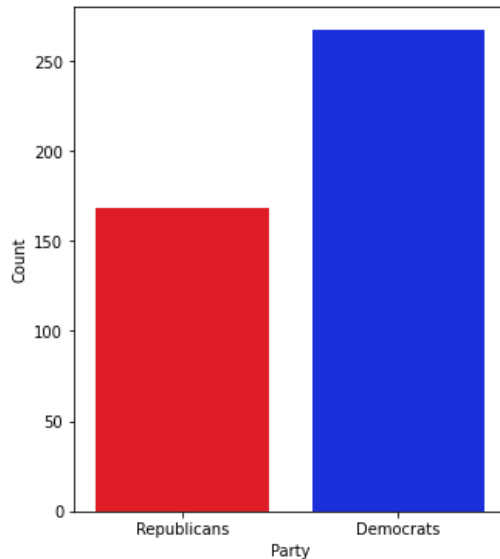
- Republikanie mieli przewagę w Senacie.
- W Afryce Południowej miała niedawno miejsce kontrowersyjna zmiana konstytucji
- W Ameryce Środkowej rebelianckie grupy wspierane przez USA walczą przeciwko domniemanym pro-komunistycznym rządóm

## 2. EDA

Zbiór okazał się niezbalansowany, ale nie jest bardzo przechylony. Wykorzystaliśmy parametr stratify przy dzieleniu zbiorów, ale korzystaliśmy z metryk accuracy i ROC AUC.

Republican politicians: 168

Democrat politicians: 267



Umieszczając ustawy na wykres punktowy, gdzie każdy punkt jest ustawą, a na osiach x i y jest liczba głosów za tą ustawą z każdej z partii politycznych, dostajemy trzy klastry - “republikański”, “centrystyczny” i “demokratyczny”. Takie klastrowanie daje duże szanse na to że nasze modele będą dobrze działać, i dokładnie to zaobserwowaliśmy.

### 3. Preprocessing danych

W przypadku obróbki danych dokonaliśmy podejścia zastosowania OrdinalEncoder, jednakże w kwestii alternatywnego testowania, zastosowaliśmy także OneHotEncoding dla sprawdzenia wyników oraz obciążenia czasowego dla modeli przy wykorzystaniu różnych form kodowania. Naszym podstawowym podejściem było użycie OrdinalEncoding dla naszych modeli. Zostało ono wykonane w sposób następujący dla wartości labeli:

"N" --> 0

"?" --> 1

"Y" --> 2

Przy enkodowaniu klas zastosowaliśmy następujące przypisanie:

"Republican" --> 1

"Democrat" --> 0

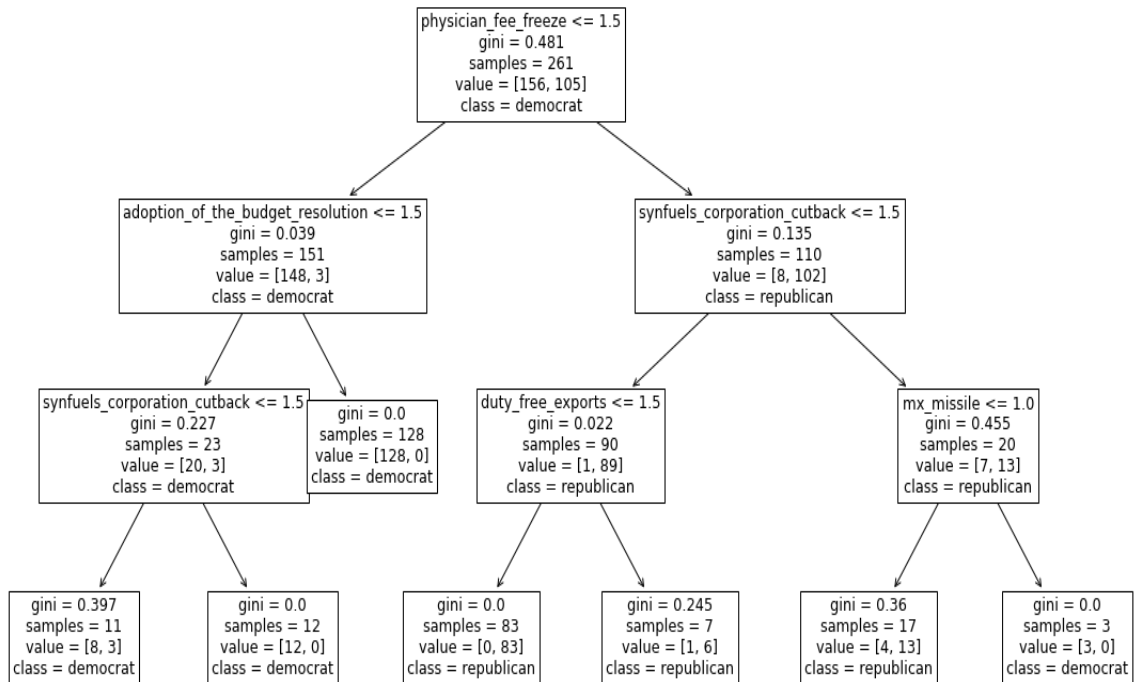
A co z przypadkiem użycia OneHotEncoding? Jak widać w zamieszczonej poniżej grafice, obydwa sposoby przypisywania wartości zmiennym kategorycznym zwracają identyczne wyniki dla modeli, jednakże OneHotEncoding wykazał się mniejszym obciążeniem czasowym.

```
OrdinalEncoding
Accuracy score (train): 0.977
Accuracy score (test): 0.937
F1 score (train): 0.977
F1 score (test): 0.937
ROC AUC score (train): 0.994
ROC AUC score (test): 0.973
Tree depth: 3
Leaf count: 7

OneHotEncoding
Accuracy score (train): 0.977
Accuracy score (test): 0.937
F1 score (train): 0.977
F1 score (test): 0.937
ROC AUC score (train): 0.994
ROC AUC score (test): 0.958
Tree depth: 3
Leaf count: 7
```

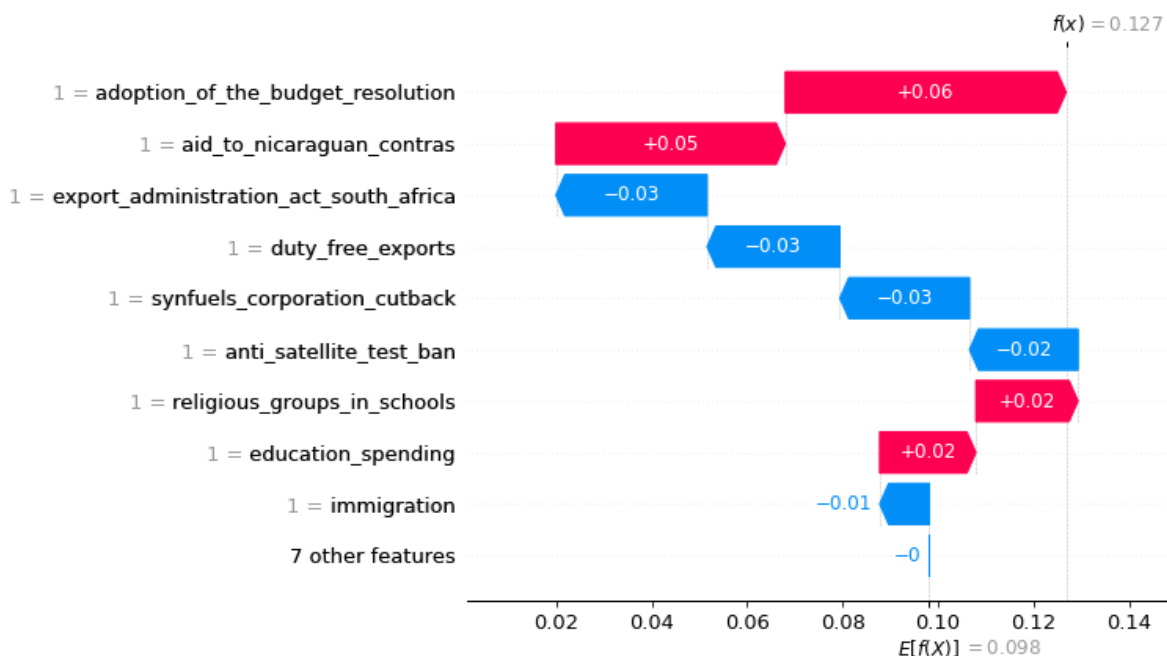
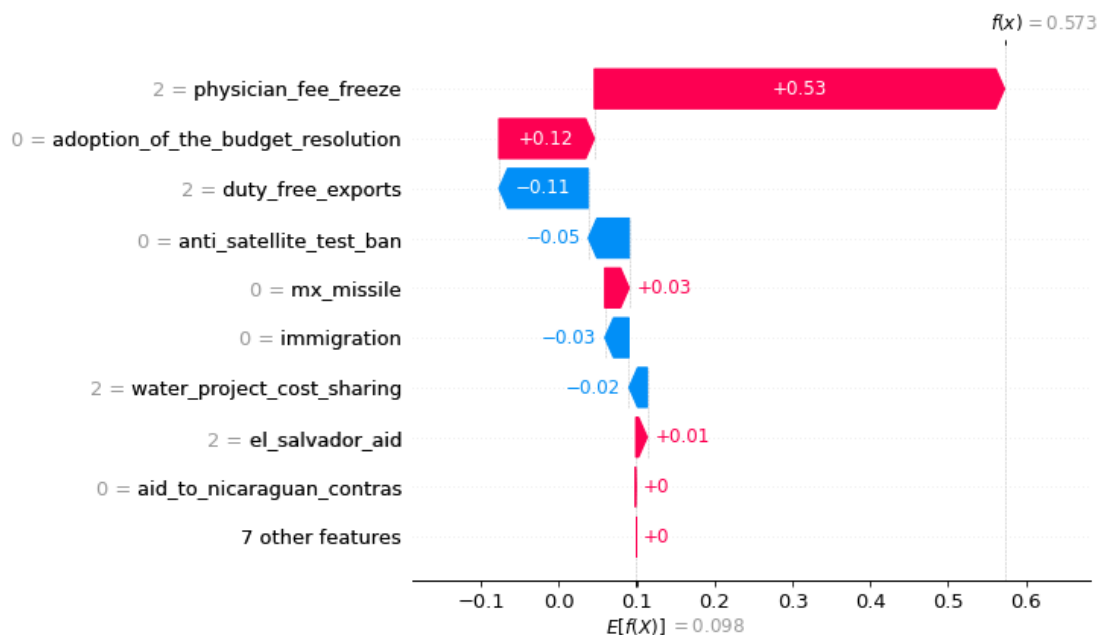
## 4. Modele drzewiaste

Przygotowaliśmy kilka modeli drzewiastych i za najlepszy uznaliśmy model drzewiasty o głębokości 3 z siedmioma liśćmi. Przeprowadziliśmy również tuning hiper parametrów i zwrócił model bardzo podobny. Plusem tego modelu jest łatwość użycia i interpretacji wyników - można przedstawić ten model jako serię zapytań. Ten model jednak jest słaby pod względem pewności, przez co jego wyniki AUC ROC są gorsze niż



## 5. Las losowy i SHAP

Naszym najlepszym modelem okazał się model lasu losowego i wykorzystaliśmy pakiet SHAP do wizualizacji tego modelu. Ponieważ ten model jest typu “czarne pudełko”, to ta biblioteka jest najlepszym sposobem na przyjrzenie się, dlaczego ten model działa. Dzięki temu mogliśmy też obserwować, dlaczego model nie działał dla niektórych polityków. Zwrócił uwagę na kilka przypadków, w których politycy mieli dużo wstrzymań od głosów oraz polityków, którzy głosowali inaczej niż partia. SHAP również pokazał, że model zakłada, że polityk jest demokratą, chyba że głosowaniami pokazał inaczej, co miało by sens biorąc pod uwagę imbalans etykiet.



## 6. Podsumowanie

Modele, które ostatecznie przygotowaliśmy miały wysoką celność i jakość AUC ROC. Modele drzewiaste były łatwo interpretowalne, lasy losowe były bardzo efektywne. Zbiór danych okazał się prosty, ale miał swoje trudności. Analiza wstępna jest utrudniona oraz niektóre modele nie działają dobrze na danych głównie binarnych - takich jak w naszym zbiorze. Pomimo tego, okazało się, że politycy są dosyć przewidywalni; nawet po jednym głosowaniu byliśmy w stanie przewidzieć z wysoką dokładnością, do której partii należeli politycy.