

PROJEKT Z PRZEDMIOTU EKSPLORACJA DANYCH

DRUGI ETAP: PRZYGOTOWANIE DANYCH + MODELOWANIE

SPEED DATING EXPERIMENT

13.06.2025 r.

Spis treści

1. Charakterystyka zbioru danych	2
1.1. Pochodzenie	2
1.2. Format	2
1.3. Liczba przykładów	2
1.4. Ilość zbiorów danych	2
2. Cel eksploracji i kryteria sukcesu	2
3. Założenia wstępne	2
4. Przygotowanie danych	2
4.1. Dane brakujące i dane do ujednolicenia	2
4.2. Zamiana na nominalne/numeryczne	2
4.3. Podzbiór danych	2
5. Wyniki i model	3
5.1. Krótki opis modelu	3
5.2. Parametry modelu	3
5.3. Ewaluacja wyników	3
5.4. Wyniki osiągnięte przez model	3
6. Optymalizacja modelu	6
7. Wnioski	9

1. Charakterystyka zbioru danych

1.1. Pochodzenie

<https://www.kaggle.com/datasets/annavictoria/speed-dating-experiment>

1.2. Format

.csv

1.3. Liczba przykładów

8378 rekordów

1.4. Ilość zbiorów danych

1

2. Cel eksploracji i kryteria sukcesu

Celem eksploracji danych ze zbioru „Speed Dating Experiment” jest znalezienie odpowiedzi na pytania:

- Czy ludzie potrafią dokładnie przewidzieć swoją postrzeganą wartość na rynku randkowym?
- Sprawdzenie, jaki atrybut najmocniej wpływa na dobór partnera przeciwnej płci.

Kryteria sukcesu, które zostaną przyjęte w celu oceny skuteczności eksploracji danych, obejmują:

- wysoka korelacja ($\geq 0,6$) między przewidywaną a rzeczywistą wartością uczestników na rynku randkowym
- zidentyfikowanie cech, które mają największy wpływ na postrzeganą wartość uczestników
- przeprowadzenie analizy istotności atrybutów ze wskazaniem najistotniejszego

3. Założenia wstępne

Zakładamy, że z racji na olbrzymią ilość kolumn oraz stosunkowo niewielką liczbę wierszy danych najlepiej sprawdzi się klasyfikator oparty na drzewie decyzyjnym.

Kolejnym argumentem za drzewami decyzyjnymi jest ich metodyka pracy, przewidują one wartość żądanego atrybutu w oparciu o inne atrybuty i potrafią zbudować ścieżki zależności między parametrami. Odpowie nam to na jedno z pytań - celi.

4. Przygotowanie danych

4.1. Dane brakujące i dane do ujednolicenia

Nie wystąpiła potrzeba uzupełnienia brakujących danych.

4.2. Zamiana na nominalne/numeryczne

Dla wybranych cech nie było takiej potrzeby.

4.3. Podzbiór danych

Wybrano dane z wydarzeń speed datingu o numerach: 1-5, 10-11, 15-17. Zdecydowano się na te edycje, gdyż zostały one przeprowadzone w tych samych warunkach a sposób oceniania preferencji polegał na rozdziale 100 punktów między kategorie. W innych edycjach warunki przeprowadzenia eksperymenty były inne, znacznie różniące się. Wybranie innych edycji zakłóciłoby porównywanie wyników i wyciągnięcie rzetelnych wniosków.

5. Wyniki i model

5.1. Krótki opis modelu

Wykorzystano model `DecisionTreeClassifier` z biblioteki `scikit learn`. Model ten implementuje drzewo decyzyjne. Model ten przyjmuje postać drzewa binarnego, w którym każdy węzeł odpowiada decyzji podjętej na podstawie jednej z cech opisujących dane, natomiast liść drzewa reprezentuje końcową prognozę – przypisanie do jednej z klas. Uczenie drzewa decyzyjnego polega na rekurencyjnym dzieleniu przestrzeni cech w taki sposób, aby w kolejnych krokach uzyskiwać podzbiory jak najbardziej jednorodne pod względem klas decyzyjnych. Ocenę jakości podziału realizuje się przez atrybut Gini. Wysokość drzewa może być ograniczona, tutaj nie jest.

5.2. Parametry modelu

Do odpowiedzi na pytanie użyto preferencji nt. aktywności wykonywanych przez uczestników. Dane zostały przetasowane losowo a następnie podzielone na zbiór treningowy i testowy w proporcji 80% do 20%.

5.3. Ewaluacja wyników

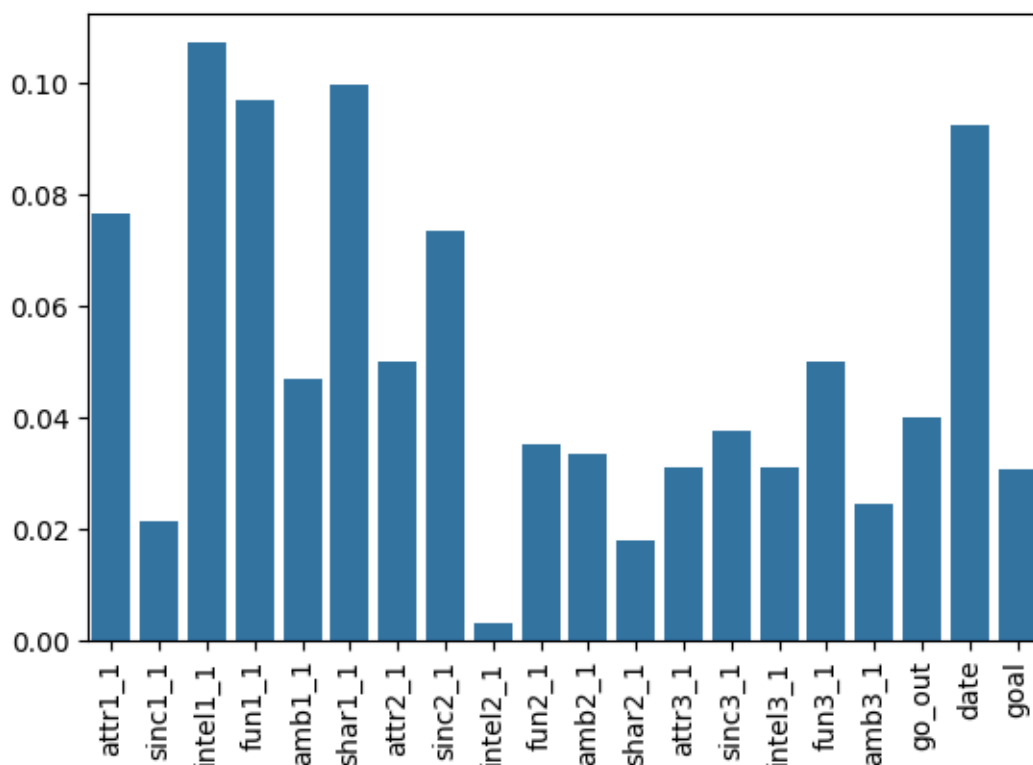
Rezultaty pracy modelu zostały ocenione na podstawie wskaźników:

- precision (precyzja) - miara dokładności klasyfikacji, określająca, ile z przewidzianych pozytywnych przypadków jest rzeczywiście pozytywnych,
- recall (czułość) - miara zdolności modelu do wykrywania pozytywnych przypadków, określająca, ile z rzeczywistych pozytywnych przypadków zostało poprawnie przewidzianych,
- F1-score - miara łącząca precyzję i czułość, która jest szczególnie przydatna w przypadku nierównomiernych klas (który tu występuje).

Wygenerowano również confusion matrix (macierz pomyłek), która pokazuje jakość przewidywań (obrazuje trafienie, poprawne odrzucenie, chybienie i fałszywe alarmy).

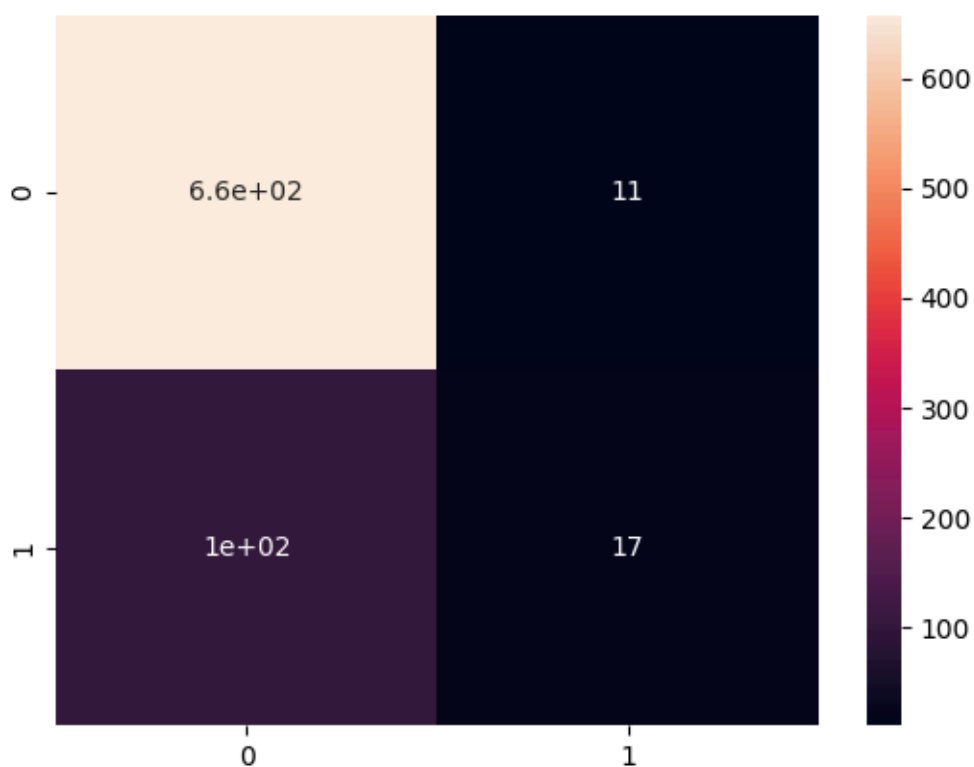
5.4. Wyniki osiągnięte przez model

Wygenerowano wykres ważności cech, w zależności od kontekstu. Konteksty obejmowały: tego szukam u partnera/ partnerki (atributy `xxxx1_1`), tego szuka płeć przeciwna (atributy `xxxx2_1`), własna ocena (atributy `xxxx3_1`). Do zakresu analizy dodatno również częstotliwość uczęszczania na randki i imprezy.

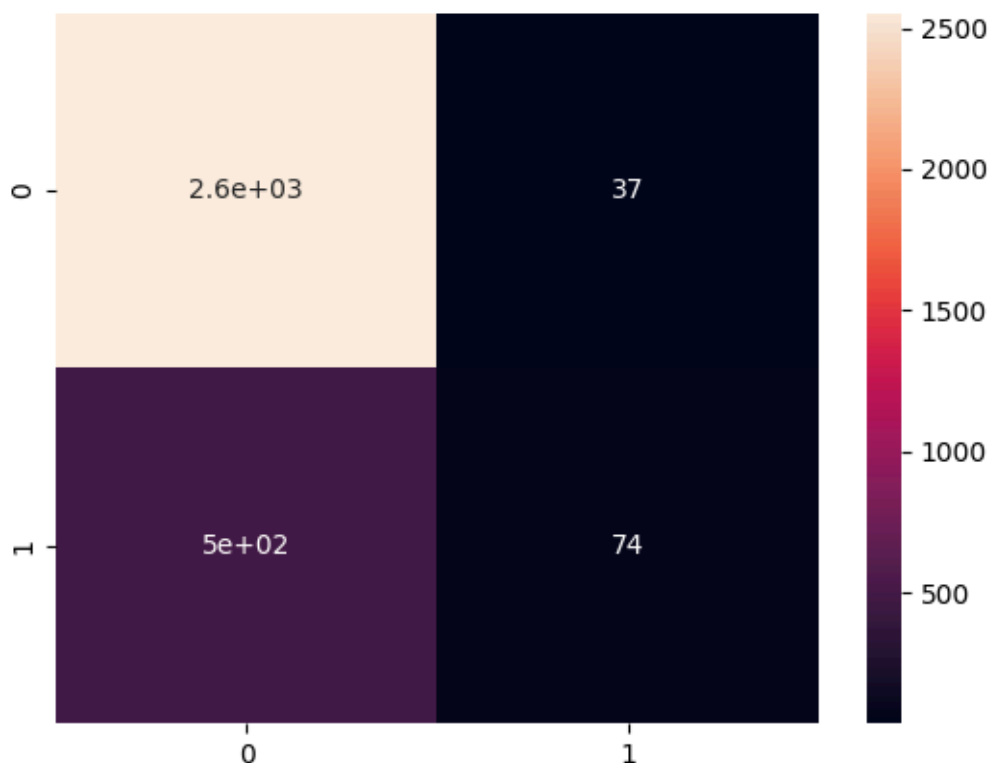


Rysunek 1: Wykres ważności cech wpływający na przewidywanie, czy uczestnicy przypadną sobie do gustu (match będzie zrealizowany).

Wygenerowano także macierz pomyłek dla zbiorów: testowego i treningowego.

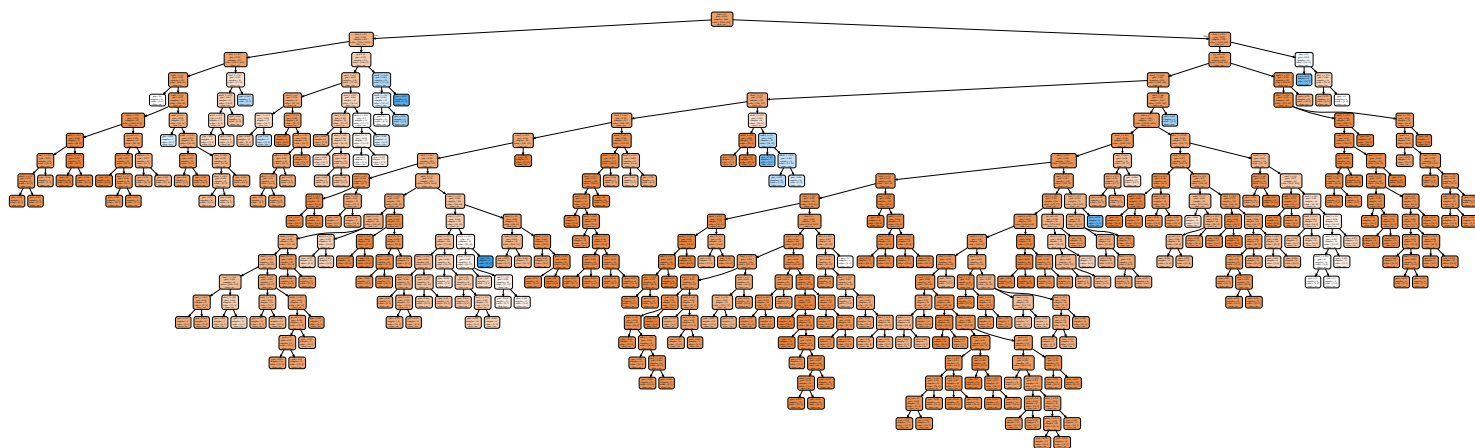


Rysunek 2: Macierz pomyłek dla zbioru testowego.



Rysunek 3: Macierz pomyłek dla zbioru treningowego.

Powstało również drzewo decyzyjne, niestety z większością liści dających rezultat „brak dopasowania” :(. Ścieżki dające pozytywny scenariusz zakończone są liśćmi w odcieniach niebieskiego.

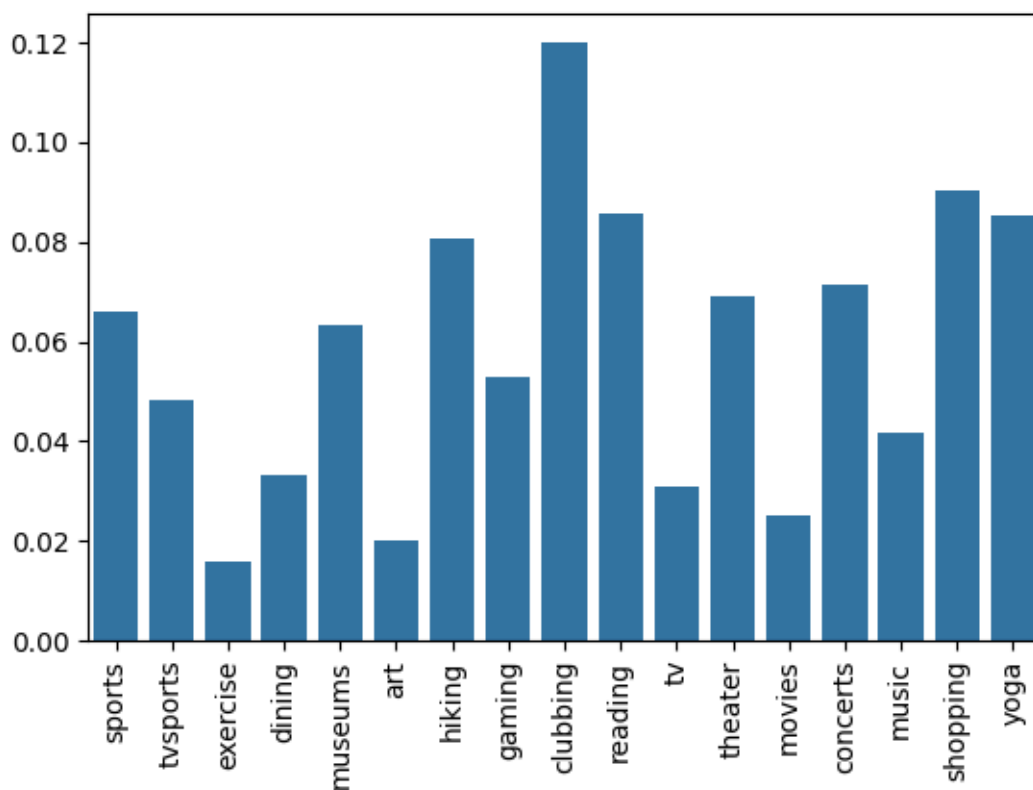


Rysunek 4: Drzewo decyzyjne.

< Opis wyników >

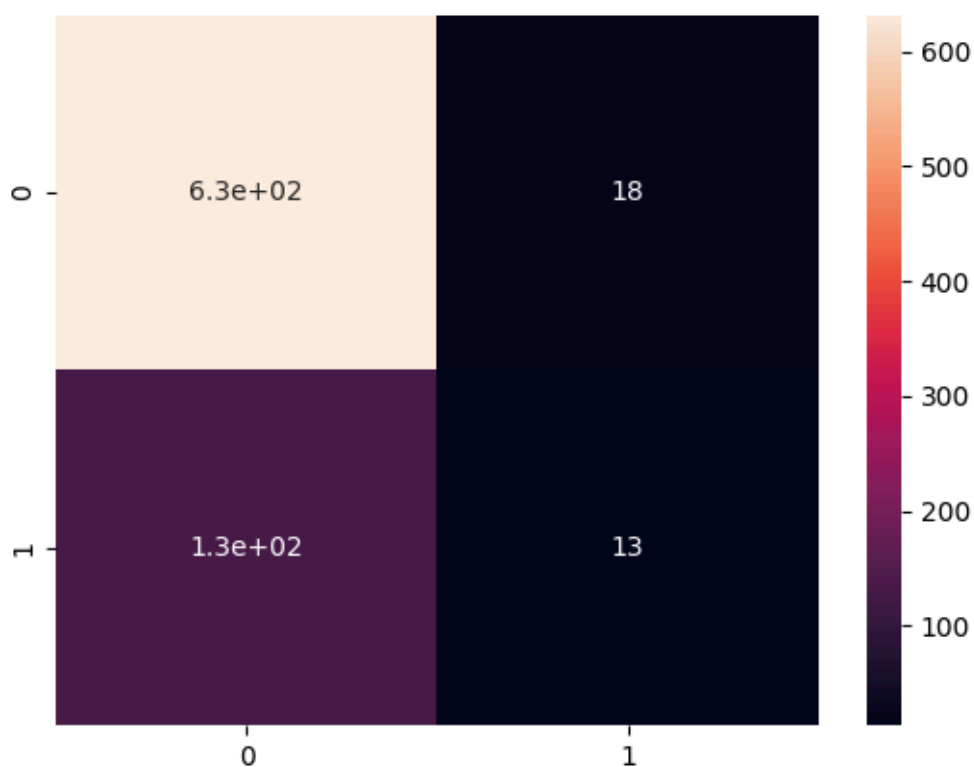
6. Optymalizacja modelu

Przeprowadzono drugie badanie z użyciem tego samego modelu, ale innych parametrów. Tym razem wybrano ocenę chęci zaangażowania się w jakąś aktywność pozanaukową.

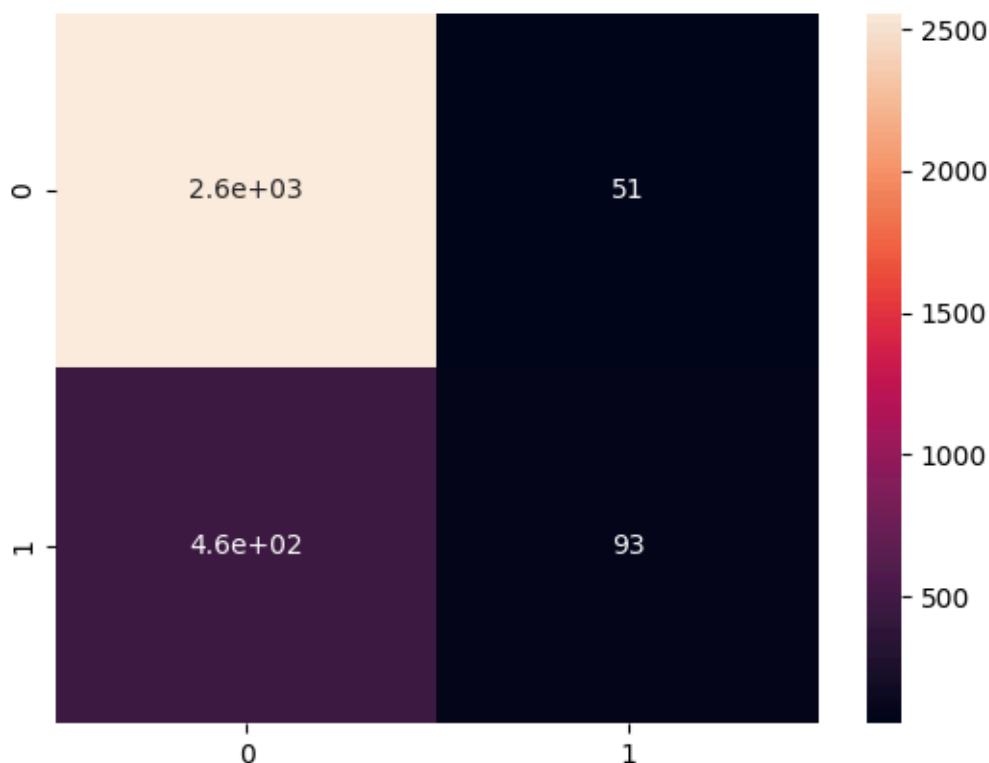


Rysunek 5: Wykres ważności cech wpływający na przewidywanie, czy uczestnicy przypadną sobie do gustu (match będzie zrealizowany).

Wygenerowano także macierz pomyłek dla zbiorów: testowego i treningowego.

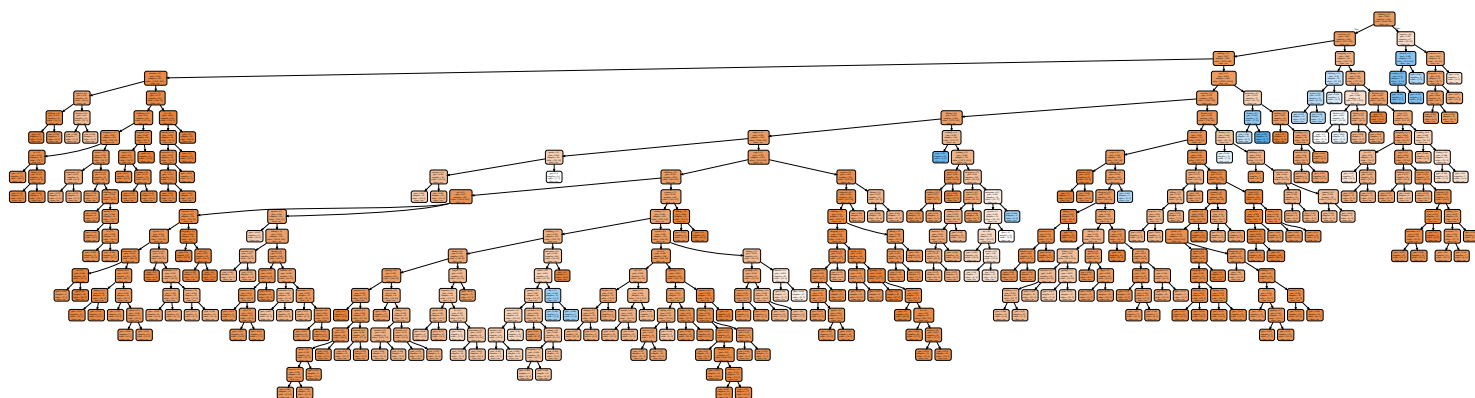


Rysunek 6: Macierz pomyłek dla zbioru testowego.



Rysunek 7: Macierz pomyłek dla zbioru treningowego.

Powstało również drzewo decyzyjne, niestety z większością liści dających rezultat „brak dopasowania” :(. Ścieżki dające pozytywny scenariusz zakończone są liśćmi w odcieniach niebieskiego.



Rysunek 8: Drzewo decyzyjne.

< Opis wyników >

7. Wnioski

< wnioski końcowe, osiągnięte wyniki, komentarz, jakie parametry były najlepsze >