POLITECHNIKA GDAŃSKA
Eksploracja Danych, informatyka 2024/2025

Jakub Bronowski, 193208
Mikołaj Klikowicz, 193264
Maciej Raciniewski, 189774

Projekt z przedmiotu Eksploracja Danych

Drugi etap: Przygotowanie danych + Modelowanie

Speed Dating Experiment

13.06.2025 r.

Spis treści

Jakub Bronowski, 193208
Mikołaj Klikowicz, 193264
Maciej Raciniewski, 189774

POLITECHNIKA GDAŃSKA

Eksploracja Danych, informatyka 2024/2025

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 (>= 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 agrumentem za drzewami decyzyjnym 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.

Jakub Bronowski, 193208
Mikołaj Klikowicz, 193264
Maciej Raciniewski, 189774

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 odpowiedzenia 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),
- Support - ilość próbek zadanej klasy.

5.3.1. Próba nr 1

Badana klasa	Precision	Recall	F1-score	Support
0 - brak dopasowania	0.83	0.97	0.90	650
1 - jest dopasowanie	0.42	0.09	0.15	140
Accuracy	-	-	0.82	790
Macro avg	0.63	0.53	0.52	790
Weighted avg	0.76	0.82	0.77	790

Tab. 1
Miary jakości modelu dla zbioru testowego. Próba numer 1

Widać, że dla zbioru testowego dokładność wskazania braku dopasowania jest na wysokim poziomie, czego nie można powiedzieć dla wskazania dobrego dopasowania.

Badana klasa	Precision	Recall	F1-score	Support
0 - brak dopasowania	0.85	0.98	0.91	2610
1 - jest dopasowanie	0.65	0.17	0.27	550
Accuracy	-	-	0.84	3160
Macro avg	0.75	0.57	0.59	3160
Weighted avg	0.81	0.84	0.80	3160

Jakub Bronowski, 193208
Mikołaj Klikowicz, 193264
Maciej Raciniewski, 189774

Tab. 2
Miary jakości modelu dla zbioru treningowego Próba numer 1.

W zbiorze treningowym model radzi sobie dużo lepiej niż w przypadku zbioru testowego.

5.3.2. Próba nr 2

Badana klasa	Precision	Recall	F1-score	Support
0 - brak dopasowania	0.86	0.98	0.92	669
1 - jest dopasowanie	0.61	0.14	0.23	121
Accuracy	-	-	0.85	790
Macro avg	0.74	0.56	0.57	790
Weighted avg	0.82	0.85	0.81	790

Tab. 3
Miary jakości modelu dla zbioru testowego. Próba numer 2

Wnioski podobne jak przy próbie 1. Widać, że dla zbioru testowego dokładność wskazania braku dopasowania jest na wysokim poziomie, wzrosła też dokładność dopasowania. Pozostałe parametry uległy poprawie wzgl. próby 1.

Badana klasa	Precision	Recall	F1-score	Support
0 - brak dopasowania	0.84	0.99	0.91	2591
1 - jest dopasowanie	0.67	0.13	0.22	569
Accuracy	-	-	0.83	3160
Macro avg	0.75	0.56	0.56	3160
Weighted avg	0.81	0.83	0.78	3160

Tab. 4
Miary jakości modelu dla zbioru treningowego Próba numer 2.

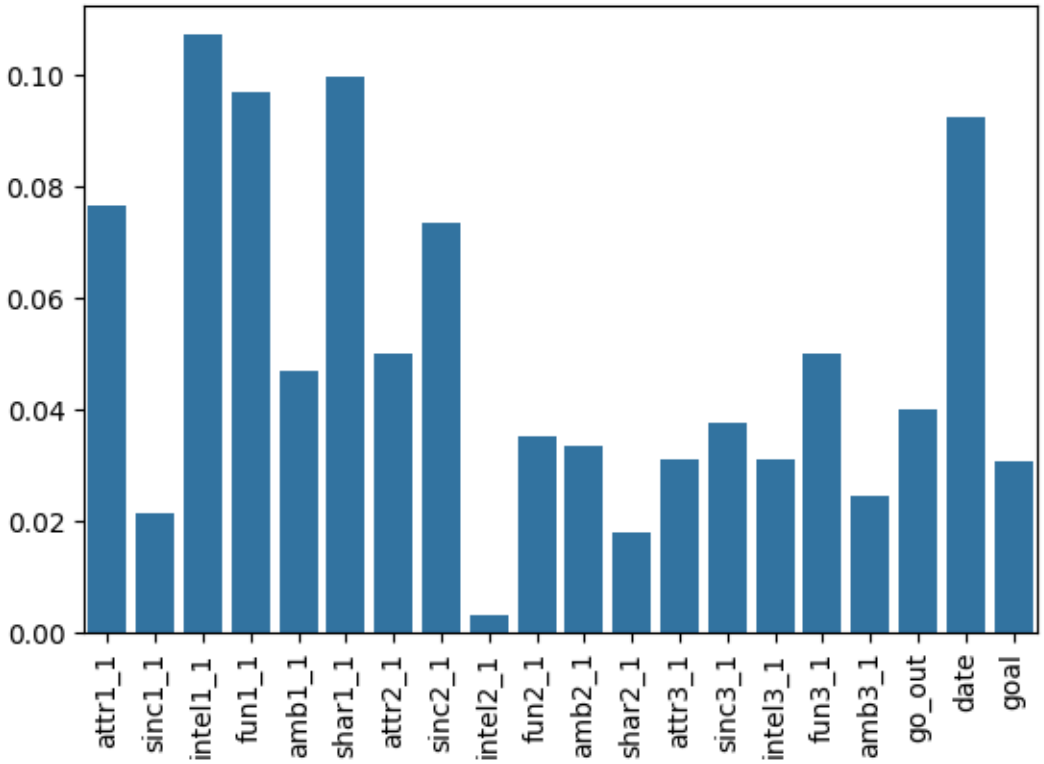
Wyniki niemalże identyczne jak w przypadku próby 1.

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

5.4. Wyniki osiągnięte przez model

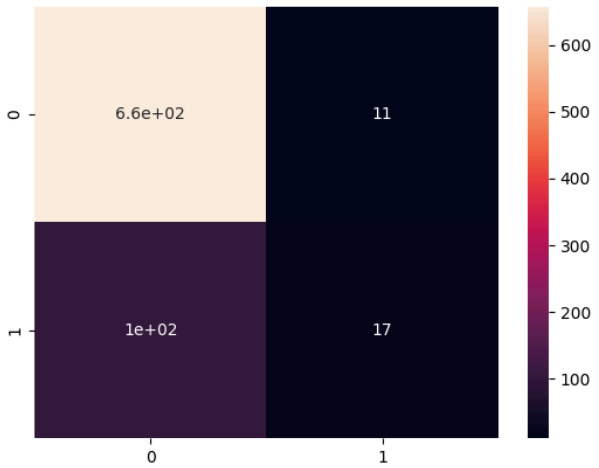
Wygenerowano wykres ważności cech, w zależności od kontekstu. Konteksty obejmowały: tego szukam u partnera/partnerki (atrybuty xxxx1_1), tego szuka płeć przeciwna (atrybuty xxxx2_1), własna ocena (atrybuty xxxx3_1). Do zakresu analizy dodatno również częstotliwość uczęszczania na randki i imprezy.

Jakub Bronowski, 193208
Mikołaj Klikowicz, 193264
Maciej Raciniewski, 189774



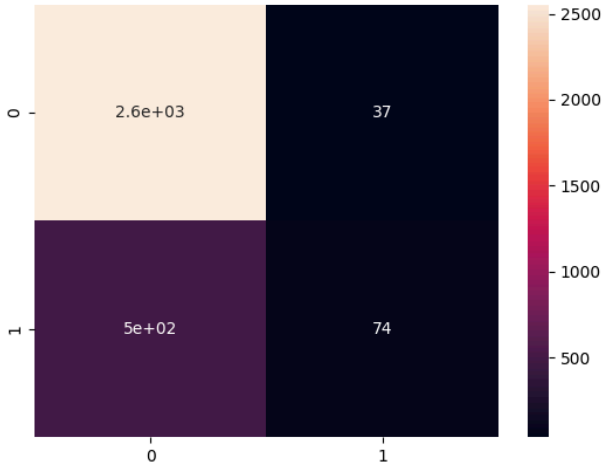
Rysunek 1: Wykres ważności cech wpływający na przewidywanie, czy uczestnicy przypadną sobie do gustu (match będzie zrealizowany).

Wygenerowano także macierz pomyłek dla zbiorów: testowego i treningowego.



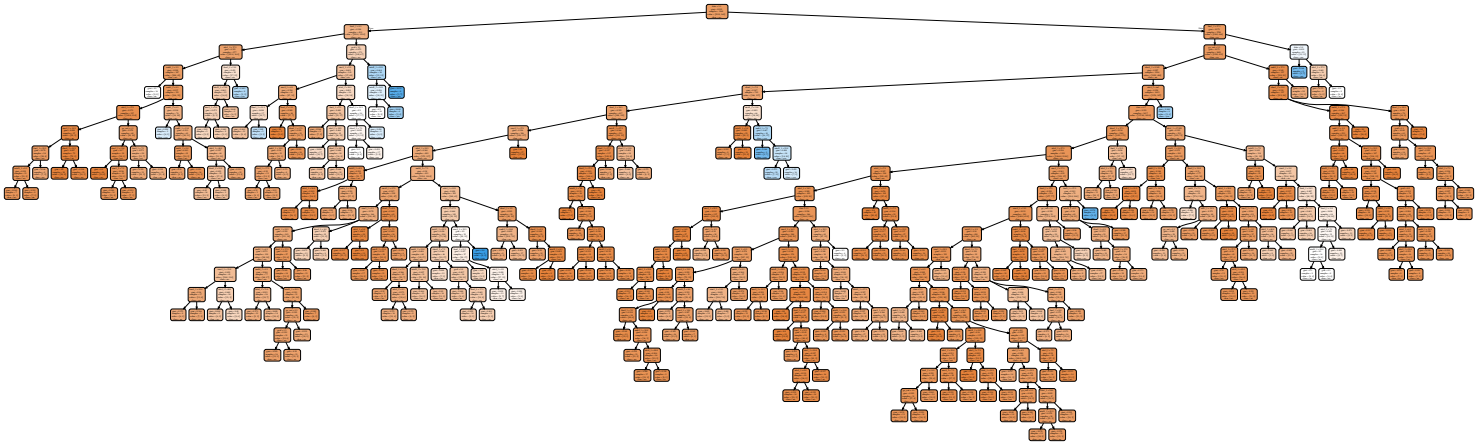
Rysunek 2: Macierz pomyłek dla zbioru testowego.

Jakub Bronowski, 193208
Mikołaj Klikowicz, 193264
Maciej Raciniewski, 189774



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.

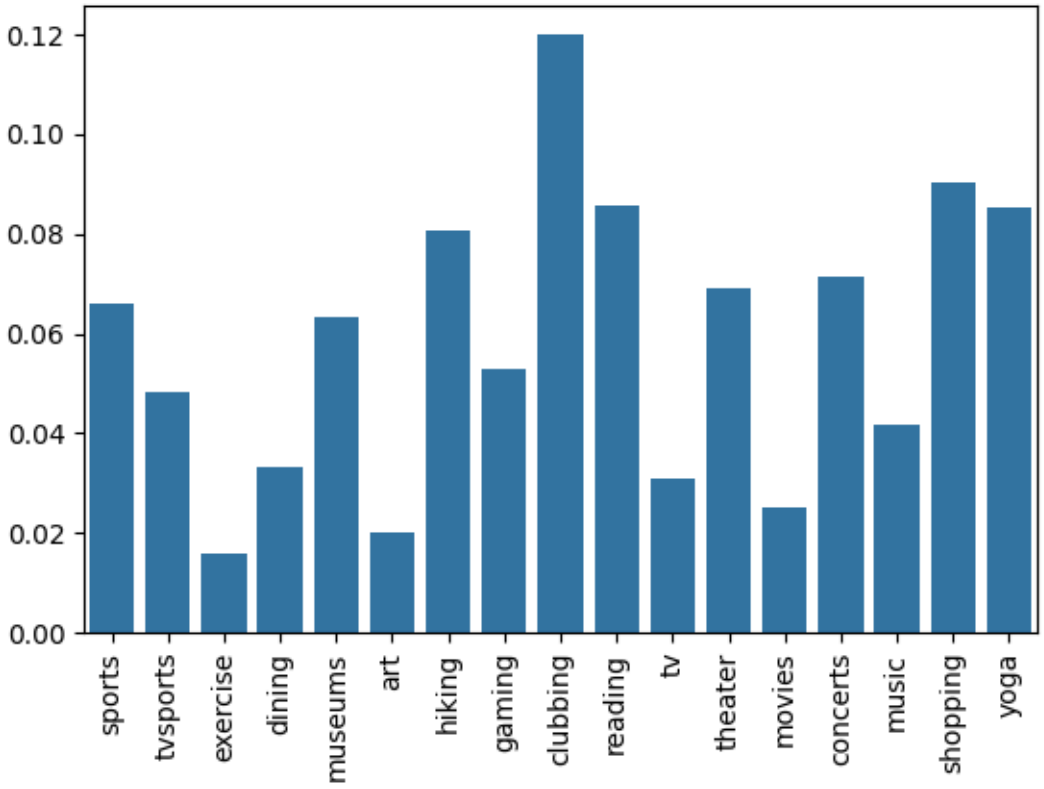


Rysunek 4: Drzewo decyzyjne.

6. Optymalizacja modelu

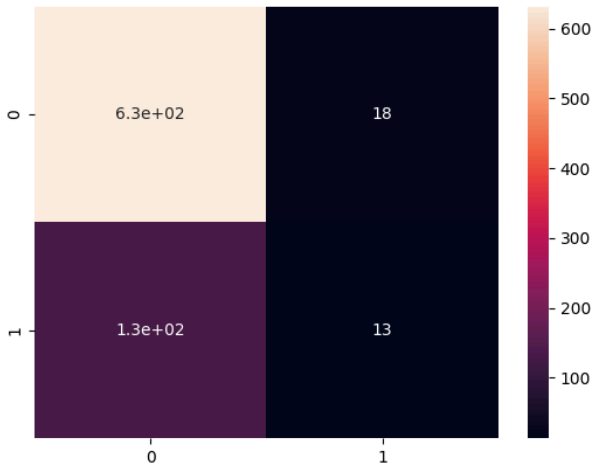
Przeprowadzono drugie badanie z użyciem tego samego modelu, ale innych parametrów. Tym razem wybrano ocenę chęci zaangażowania się w jakąś aktywność pozanaukową.

Jakub Bronowski, 193208
Mikołaj Klikowicz, 193264
Maciej Raciniewski, 189774



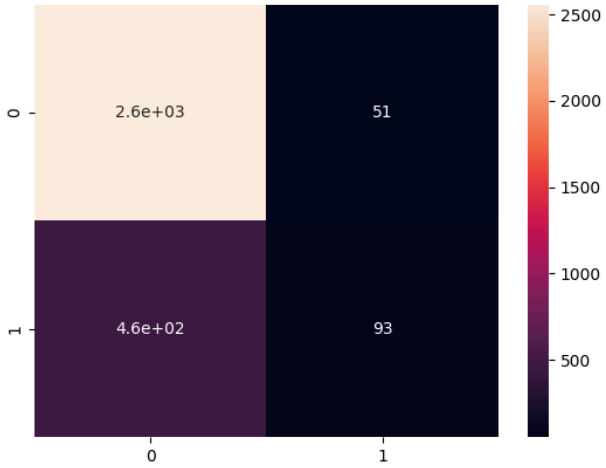
Rysunek 5: Wykres ważności aktywności wpływający na przewidywanie, czy uczestnicy przypadną sobie do gustu (match będzie zrealizowany).

Wygenerowano także macierz pomyłek dla zbiorów: testowego i treningowego.



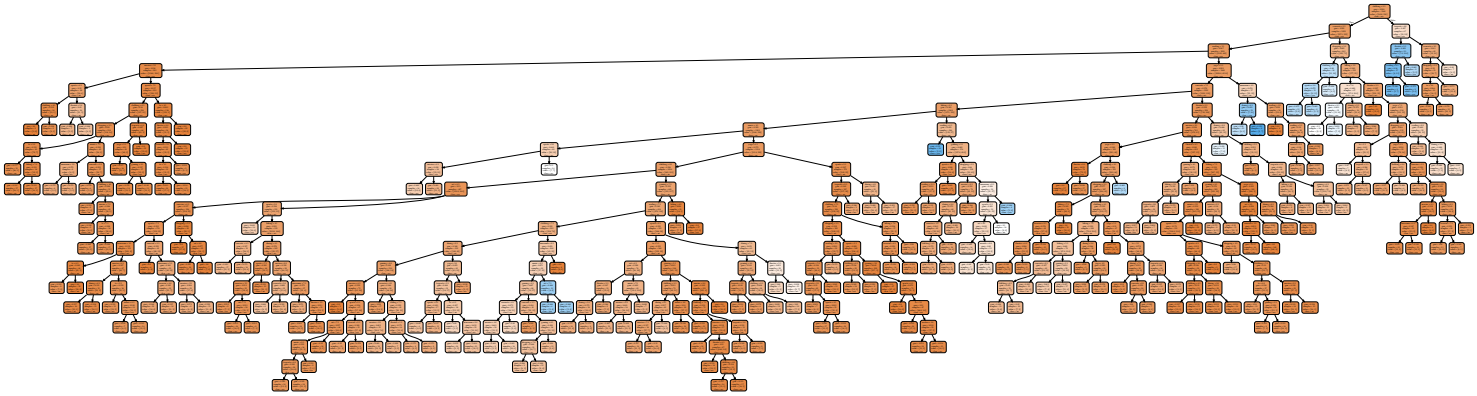
Rysunek 6: Macierz pomyłek dla zbioru testowego.

Jakub Bronowski, 193208
Mikołaj Klikowicz, 193264
Maciej Raciniewski, 189774



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.

7. Wnioski i wyniki

7.1. Analiza pierwszej próby

Na podstawie analizy *Rysunku 1* możemy zauważyć, że najważniejsze cechy wpływające na dopasowanie partnerów w eksperymencie speed dating to:

- Inteligencja płci przeciwnej – jest to najistotniejszy czynnik, z wartością wpływu przekraczającą 10%, co wskazuje, że uczestnicy wysoko cenili tę cechę w poszukiwaniu partnera.
- Wspólne zainteresowania i hobby – również odgrywały znaczącą rolę, z wartością wpływu bliską 10%, co podkreśla znaczenie wspólnego spędzania czasu i podobnych pasji w budowaniu relacji.
- Poczucie humoru – było nieco mniej istotne, ale nadal miało istotny wpływ na dopasowanie, z wartością lekko poniżej 10%, co sugeruje, że uczestnicy postrzegali tę cechę jako ważną w relacji.
- Rzadkie chodzenie na randki – wartość wpływu wynosząca około 9% wskazuje, że preferencje dotyczące częstotliwości randek odgrywały pewną rolę w ocenie potencjalnych partnerów.
- Atrakcyjność fizyczna – z wartością poniżej 8%, była ważna, ale nie kluczowa w porównaniu do innych cech.
- Szczerość – Jest to poszukiwana najbardziej cecha przez osoby będące potencjalnymi partnerami w trakcie spotkań.

Jakub Bronowski, 193208
Mikołaj Klikowicz, 193264
Maciej Raciniewski, 189774

Odpowiedzi z ankiet, które najrzadziej prowadziły do sukcesu randki, to:
- Przekonanie, że druga osoba poszukuje u nas inteligencji,
- Uważanie, że płeć przeciwna ceni wspólne zainteresowania,
- Poszukiwanie szczerości u partnera randki.

Z analizy cech wpływających na dopasowanie partnerów w eksperymencie speed dating wynika, że różne cechy mają zróżnicowany wpływ na to, czy uczestnicy uznają się za kompatybilnych. Najważniejszym czynnikiem jest inteligencja płci przeciwnej. Wartość wpływu przekraczająca 0.1 wskazuje, że uczestnicy szczególnie cenili tę cechę u potencjalnych partnerów. Może to sugerować, że w krótkich interakcjach cechy intelektualne, takie jak elokwencja czy sposób wyrażania się, były wyraźnie zauważalne i miały duże znaczenie przy podejmowaniu decyzji o dalszej znajomości. Kolejną istotną cechą okazały się wspólne zainteresowania i hobby, co podkreśla znaczenie kompatybilności stylu życia i wspólnych pasji w ocenie potencjalnych partnerów. Jest to logiczne, ponieważ podobne zainteresowania mogą być uznawane za fundament długoterminowych relacji.

Poczucie humoru, mimo że nieco mniej istotne niż inteligencja czy zainteresowania, nadal miało zauważalny wpływ. To potwierdza, że zdolność do rozładowywania napięć i budowania przyjemnej atmosfery może być postrzegana jako kluczowa cecha w relacjach międzyludzkich. Interesujące jest jednak, że cechy takie jak atrakcyjność fizyczna i szczerość, które są powszechnie uznawane za istotne w relacjach, miały niższy wpływ w tym eksperymencie. Może to wynikać z krótkotrwałości interakcji podczas speed datingu, gdzie uczestnicy mogą przywiązywać większą wagę do cech, które są natychmiast zauważalne i łatwo komunikowalne.

Z drugiej strony, analizując odpowiedzi z ankiet, zauważono, że pewne oczekiwania uczestników, takie jak przekonanie, że płeć przeciwna poszukuje u nich inteligencji, wspólnych zainteresowań czy szczerości, najrzadziej prowadziły do sukcesu randki. Może to sugerować, że istnieje rozbieżność między tym, co ludzie deklarują jako istotne, a tym, co rzeczywiście kieruje ich decyzjami w praktyce.

7.2. Wnioski dotyczące drzewa decyzyjnego drugiej próby (rysunek 8):

1. Znaczenie zainteresowań:

a. Kluczowe cechy wpływające na decyzje o wyborze partnera to zainteresowanie clubbingiem i koncertami. Osoby o wysokim zainteresowaniu clubbingiem (>9,5/10) stanowiły liczną grupę, z dużą skutecznością dopasowania wynoszącą 42%. Są to zwykle osoby energiczne i spontaniczne, dla których format krótkich spotkań był atrakcyjny.

b. Niskie zainteresowanie oglądaniem sportów w telewizji (< 4 pkt) w połączeniu z wysokim poziomem clubbingu dawało nawet 72% szans na dopasowanie, choć dotyczyło to niewielkiego grona (25 osób, czyli 0,008% wszystkich uczestników). Niemniej, stanowiło to 5% wszystkich udanych dopasowań.

2. Serie zainteresowań z najwyższym wskaźnikiem dopasowania:

a. Kombinacja: clubbing ≤ 9,5, koncerty ≤ 9,5, czytanie ≥ 5,5, joga ≥ 9,5, muzea ≤ 6,5, stołowanie się ≥ 8,5 dawała najwyższą skuteczność dopasowania na poziomie 88%. Są to prawdopodobnie osoby ciche i spokojne, co zaskakuje w kontekście ich niskiego zainteresowania muzeami.

b. Inny zestaw: clubbing ≤ 9,5, koncerty ≤ 9,5, czytanie ≤ 5,5, joga ≥ 9,5, koncerty ≤ 7,5, wędrówki ≥ 8,5 zapewniał szansę na dopasowanie na poziomie 80%.

Jakub Bronowski, 193208
Mikołaj Klikowicz, 193264
Maciej Raciniewski, 189774

3. Brak dopasowań:

a. Znacząca większość uczestników (82,6%) nie znalazła dopasowania. Spośród grup o najgorszych wynikach można wyróżnić:
- Clubbing ≤ 9,5, koncerty ≤ 9,5, czytanie ≤ 5,5, filmy ≥ 5,5 – aż 93,8% z 352 osób w tej grupie nie znalazło partnera. Możliwe, że osoby te były zbyt przeciętne i nie wyróżniały się szczególnymi cechami.
- Clubbing ≤ 9,5, koncerty ≤ 9,5, czytanie ≤ 5,5, joga ≤ 2,5, koncerty ≥ 9,5, teatr ≤ 8,0, filmy ≤ 8,5 – ta grupa miała 0% skuteczności w dopasowaniu, co sugeruje brak wspólnych zainteresowań z innymi uczestnikami.

4. Najbardziej cenione zainteresowania:

a. Clubbing i koncerty były najbardziej cenione, szczególnie wśród osób z wysokimi ocenami w tych kategoriach (≥9,5). Spośród wszystkich udanych dopasowań (550 osób), 101 (18,4%) to osoby z wyraźnym zainteresowaniem clubbingiem i koncertami. W tej grupie skuteczność dopasowania wynosiła ponad 30%, co znacznie przewyższało średnią.

7.3. Wnioski odnosnie wykresu rysunek 5

Na wykresie wyraźnie widać, że aktywność najbardziej wpływająca na dopasowanie uczestników to clubbing, który osiągnął wartość aż 12%. Inne aktywności, takie jak wspinaczka, czytanie, zakupy oraz yoga, miały zdecydowanie niższy ale podobny do siebie poziom wpływu, wynoszący 8%-9%. Z kolei aktywności, które najmniej przyczyniały się do dobrego dopasowania, to ćwiczenia fizyczne (poniżej 2%) oraz sztuka i filmy, które miały wartości w przedziale 2%-2.5%.

8. Podsumowanie

Z naszej analizy wynika, że aktywności związane z dynamiką i spontanicznością, takie jak clubbing czy wspinaczka, bardziej sprzyjały dopasowaniu partnerów. Natomiast bardziej indywidualne zainteresowania, jak sztuka czy ćwiczenia, miały niewielki wpływ na sukces w eksperymencie, co może wskazywać na ich mniejszą rolę w budowaniu relacji podczas krótkich spotkań.