Politechnika Wrocławska
Wydział Informatyki i Telekomunikacji

Metody i Systemy Decyzyjne

Sprawozdanie z laboratorium

Autor

Piotr Bonar

nr albumu: **272720**

kierunek: **Informatyka Stosowana**

4 czerwca 2024

Streszczenie

Celem pracy jest analiza danych dotyczących używek wśród ludzi mieszkających w Stanach Zjednoczonych oraz próba stworzenia modelu kategoryzacjącego jakie substancje ktoś kiedykolwiek zażywał, na podstawie podanych danych dotyczących jego osoby. Dane do badań zostały pobrane ze strony SAMHSA i pochodzą z 2022 roku. Wyniki zostały przedstawione w formie wykresów i tabel, które umożliwiają zrozumienie wzorców zachowań związanych z używaniem substancji w różnych grupach społecznych.

1 Wstęp – sformułowanie problemu

Zjawisko zażywania narkotyków stanowi jedno z najpoważniejszych wyzwań zdrowotnych i społecznych współczesnego świata. Narkotyki, zarówno legalne, jak i nielegalne, mają potencjał do wywoływania poważnych konsekwencji zdrowotnych, psychologicznych i społecznych. Problemy te dotyczą nie tylko jednostek, które bezpośrednio zażywają substancje, ale także ich rodzin, społeczności i systemów opieki zdrowotnej.

Z tego właśnie powodu uważam, że warto pochylić się nad tym problemem i spróbować zrozumieć jak sytuacja wygląda w zależności od części społeczeństwa. Podobnego zdania była SAMHSA, czyli Substance Abuse and Mental Health Services Administration. Zebrała ona w 2022 roku dane 58 000 obywateli Stanów Zjednoczonych dotyczące ich doświadczeń z różnego rodzaju używkami.

Analiza tych danych będzie polegała na próbie znalezienia zależności między czynnikami środowiskowymi i demograficznymi, a zażywaniem substancji wśród badanych.

Tymi czynnikami w tym przypadku będą:

- Wiek (AGE3)
- Płeć (IRSEX)
- Identyfikacja seksualna (SEXIDENT)
- Mniejszość etniczna (NEWRACE2)
- Wielkość aglomeracji, w której znajduje się miejsce zamieszkania (COUTYP4)
- Całkowity zarobek rodziny rocznie (INCOME)
- Czy kiedykolwiek palił papierosy (CIGFLAG)
- Czy kiedykolwiek spożywał alkohol (ALCFLAG)

Substancje, których zażywanie będzie badane oraz dodatkowe informacje to:

- Problemy zdrowotne związane z alkoholem (HLTINALC)
- Problemy zdrowotne związane z narkotykami (HLTINDRG)
- Czy kiedykolwiek zażywał marihuanę (MJEVER)
- Czy kiedykolwiek zażywał kokainę (COCEVER)
- Czy kiedykolwiek zażywał heroinę (HEREVER)

- Czy kiedykolwiek zażywał LSD (LSD)
- Czy kiedykolwiek zażywał PCP (PCP)
- Czy kiedykolwiek zażywał pejotl (PEYOTE)
- Czy kiedykolwiek zażywał meskalinę (MESC)
- Czy kiedykolwiek zażywał psylocybinę (PSILCY)
- Czy kiedykolwiek zażywał ecstasy (ECSTMOLLY)
- Czy kiedykolwiek zażywał ketaminę (KETMINESK)
- Czy kiedykolwiek zażywał DMT, AMT lub foxy (DMTAMTFXY)
- Czy kiedykolwiek zażywał szałwię divinorum (SALVIADIV)
- Czy kiedykolwiek zażywał halucynogeny (HALLUCEVR)
- Czy kiedykolwiek zażywał środki wziewne (INHALEVER)
- Czy kiedykolwiek zażywał metamfetaminę (METHAMEVR)
- Czy kiedykolwiek zażywał narkotyczne środki przeciwbólowe (PNRNMLIF)
- Czy kiedykolwiek zażywał trankwilizery (TRQNMLIF)
- Czy kiedykolwiek zażywał środki stymulujące (STMANYLIF)
- Czy kiedykolwiek zażywał środki uspokajające (SEDANYLIF)

2 Opis rozwiązania

2.1 Zbiór danych

Dane zostały pobrane ze strony `https://www.datafiles.samhsa.gov/data-sources`. Baza została pobrana w postaci pliku o rozszerzeniu .sav.

2.2 Przetwarzanie wstępne

Przetwarzanie wstępne było wykonywane etapowo z powodu wyjątkowo dużych danych i ograniczeń systemowych

W celu odczytania danych należało pobrać program SPSS, gdyż był to jedyny sposób na odczytanie wszystkich danych zawartych w pliku, czyli danych o badanych oraz opisy kolumn. Została użyta biblioteka *pyreadstat* do przetłumaczenia pliku na rozszerzenie .csv, aby łatwiej się nim było posługiwać.

Następnie należało wybrać z ponad 2600 kolumn te, które nas będą interesować. Zdecydowałem się na 29 z nich i zapisałem ich nazwy i numery w pliku *data/columns_selected_names.csv*. Następnie została wykorzystana biblioteka *Pandas* w celu umieszczenia danych w ramce danych i usunięcia kolumn, które nie będą dla nas użyteczne w późniejszych etapach.

Przeprowadzono również walidację wierszy, mającą na celu wykluczenie niekompletnych wierszy oraz poprawienie tych, które posiadają różne kodowania tych samych informacji. W tym celu dane o tym zapisano w pliku *data/columns_verification_values.txt*.

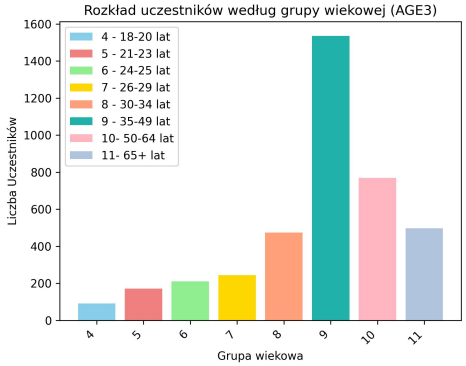
Po wykonaniu tych kroków liczba rekordów gotowych do użycia zmniejszyła się do 3991 i została zapisana w pliku *data/NSDUH_2022_selected_columns_validated.csv*

2.3 Analiza eksploracyjna

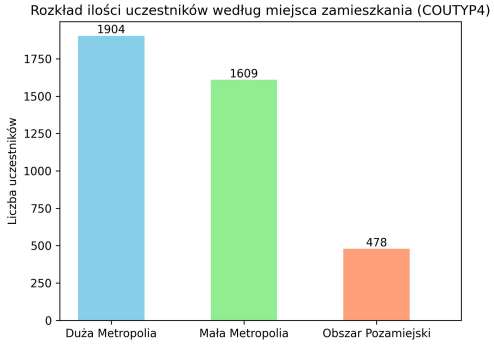
Przeprowadzono wstępną analizę eksploracyjną danych, aby zrozumieć rozkład oraz wzorce występujące w danych. Poniżej znajdują się kluczowe wykresy wygenerowane podczas analizy.

2.3.1 Analiza dystrybucji

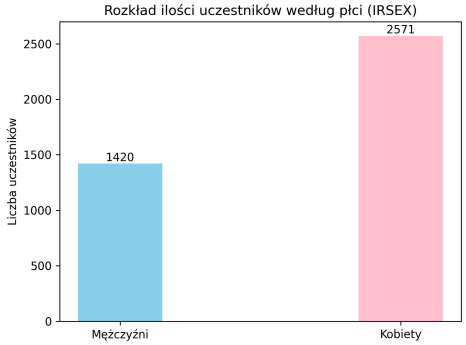
W tej sekcji przedstawiono rozkład uczestników badania według różnych kategorii demograficznych i środowiskowych. Wykresy pozwalają zrozumieć, jak różne grupy społeczne są reprezentowane w zbiorze danych, co jest istotne dla późniejszych analiz.



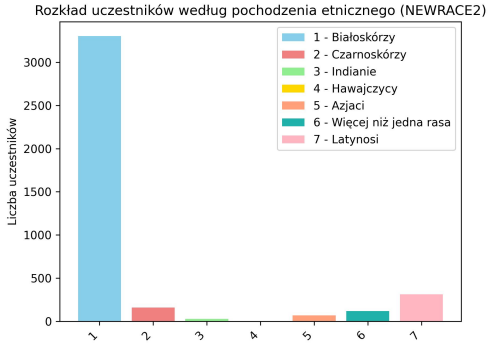
(a) Rysunek 1



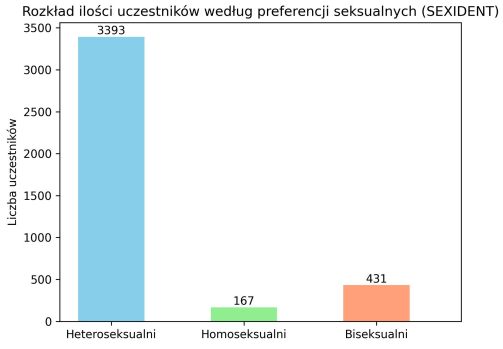
(b) Rysunek 2



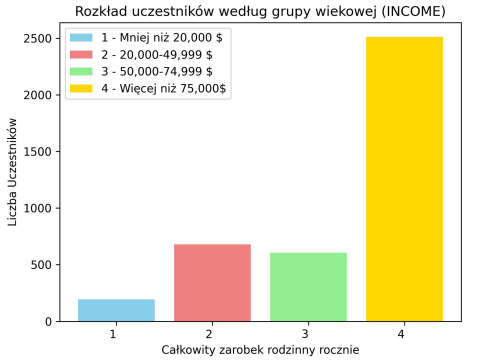
(c) Rysunek 3



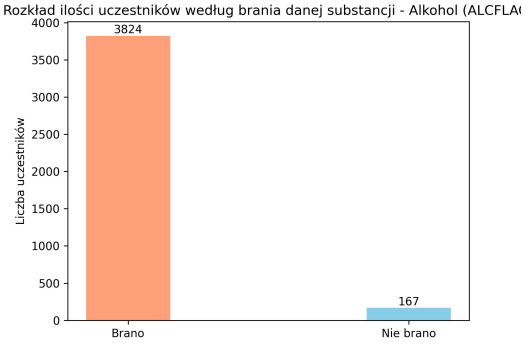
(d) Rysunek 4



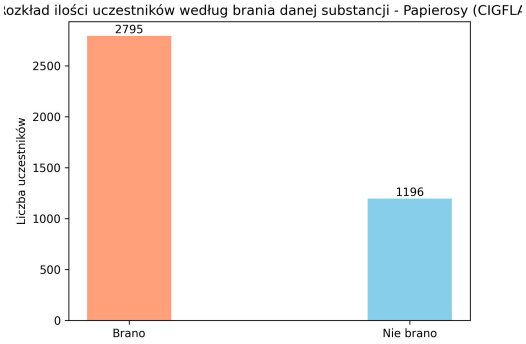
(e) Rysunek 5



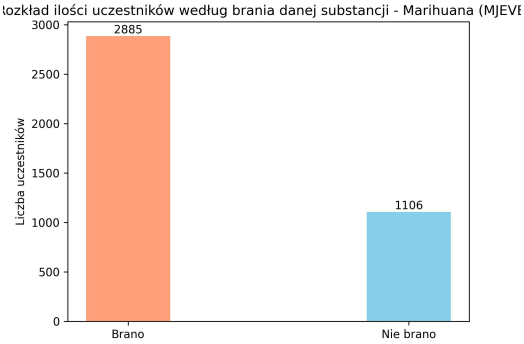
(f) Rysunek 6



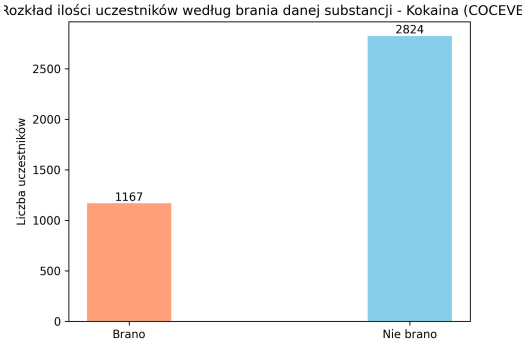
(g) Rysunek 7



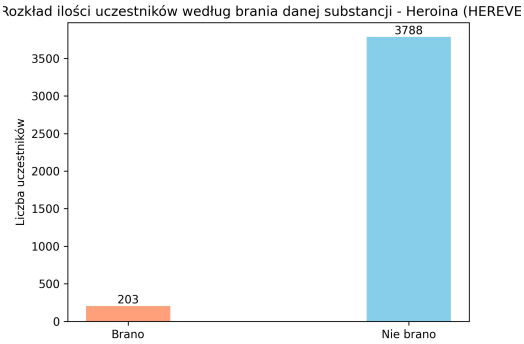
(h) Rysunek 8



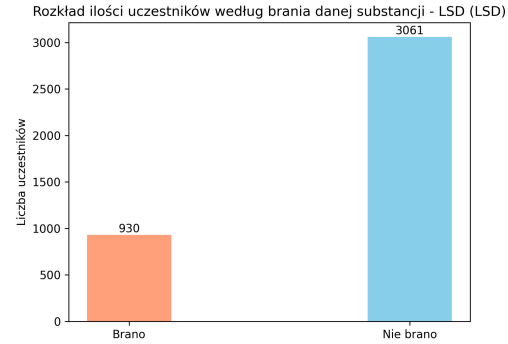
(i) Rysunek 9



(j) Rysunek 10

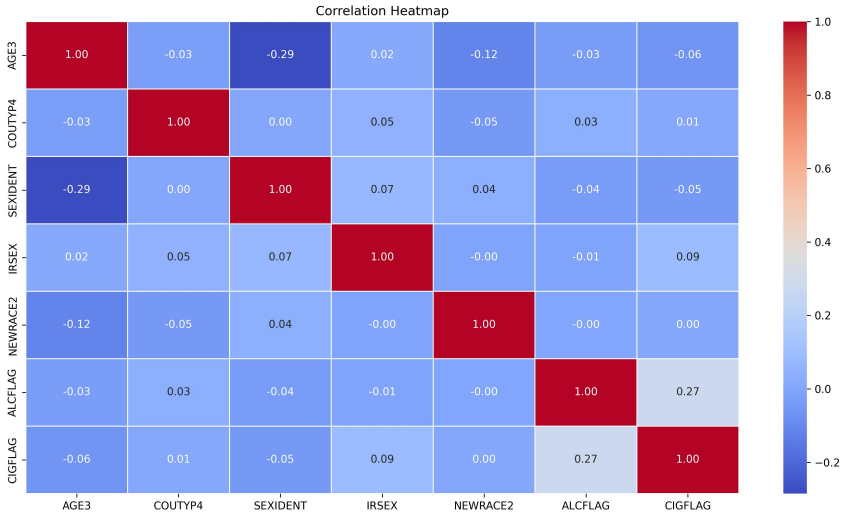


(k) Rysunek 11



(l) Rysunek 12

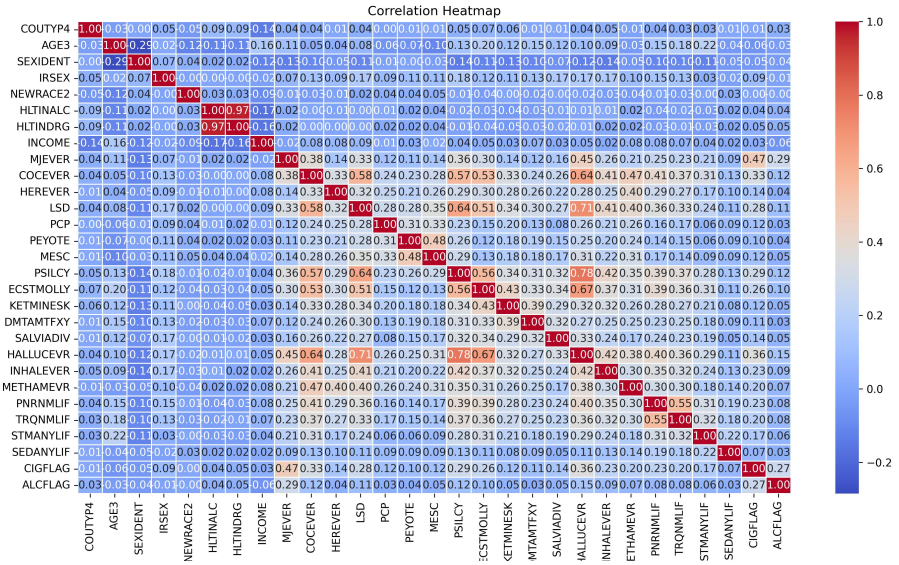
2.3.2 Analiza korelacyjna



Rysunek 1: **Mapa korelacji dla wybranych czynników środowiskowych i demograficznych.**

Na Rysunku 1. przedstawiono macierz korelacji dla czynników środowiskowych i demograficznych. Warto zauważyć kilka rzeczy, odnośnie tego, co możemy z tego wyciągnąć a propos zależności pomiedzy nimi dla ogółu społeczeństwa:

- Silne ujemne korelacje pomiędzy wiekiem (AGE3) oraz identyfikacja seksualna (SEXINDENT) sugerują, że wśród osób starszych jest mniej osób homoseksualnych i biseksualnych
- Silne dodatnie korelacje pomiędzy spożyciem alkoholu, a zapaleniem papierosa wskazują, że ludzie którzy sięgają po jedno, często sięgają również po drugie.
- Słaba ujemna korelacja pomiędzy grupą etniczną (NEWRACE2) a wiekiem (AGE3) wskazuje, że ludzie nie będący białoskórzy są statystycznie młosi w społeczeństwie amerykańskim. Prawdopodobnie wynika to z faktu napływu imigrantów z innych części świata
- Słaba dodatnia korelacja pomiędzy płcią (IRSEX) oraz paleniem pamierosów (CIIGFLAG) wskazuje na to, że kobiety częściej po nie sięgają.



Rysunek 2: **Mapa korelacji wszystkich kolumn.**

Na Rysunku 2. przedstawiono macierz korelacji między różnymi zmiennymi w zbiorze danych. Kilka kluczowych obserwacji:

- Silne dodatnie korelacje między używaniem różnych substancji psychoaktywnych (np. LSD, PCP, peyotl, meskalina, psylocybina, ecstasy) sugerują, że osoby, które próbowały jednego rodzaju substancji, są bardziej skłonne do eksperymentowania z innymi.
- Korelacja między wiekiem (AGE3) a zażywaniem substancji wskazuje, że młodsze osoby są bardziej skłonne do eksperymentowania z narkotykami. Świadczy to jednak nie o różnicy w podejściu do substancji przez osoby młodsze, a z późniejszych pokoleń
- Korelacja między dochodem (INCOME) a zażywaniem substancji wskazuje, że osoby o niższych dochodach mogą być bardziej narażone na zażywanie narkotyków.
- Zmienna identyfikacja seksualna (SEXIDENT) oraz płeć (IRSEX) również wykazują pewne korelacje z zażywaniem substancji, co może sugerować różnice w zachowaniach związanych z używkami w różnych grupach demograficznych.
- Zaskakująco słabe korelacje mniejszości etnicznej (NEWRACE2) z używaniem różnych substancji. Oznacza to, że ludzie niezależnie od pochodzenia etnicznego, zażywają w podobnej części te same substancje.

3 Rezultaty obliczeń

3.1 Plan badań

Celem badania jest stworzenie modelu kategoryzacji, który pozwoli na przewidzenie, jakie substancje ktoś kiedykolwiek zażywał na podstawie podanych danych demograficznych i środowiskowych. Model będzie oceniany na podstawie jego dokładności oraz zdolności do poprawnej klasyfikacji różnych grup użytkowników substancji. Liczyć będzie się nie tylko dokładność, ale również precyzja, czułość oraz F1-score

- **Przygotowanie danych:**
  1. Wybór istotnych zmiennych.
  2. Przetwarzanie brakujących danych.
- **Eksploracyjna analiza danych:**
  - Analiza rozkładów zmiennych.
  - Analiza korelacji między zmiennymi.
- **Modelowanie:**
  - Wybór i trenowanie modeli kategoryzacji (regresja logistyczna, las losowy, SVM).
  - Walidacja modeli za pomocą techniki cross-validation (podział danych: 80% nauka, 20% testy).
- **Ocena wyników:**
  - Porównanie wyników różnych modeli.
  - Wybór najlepszego modelu na podstawie wskaźników jakości.

3.2 Wybór i trenowanie modeli kategoryzacji

Dla problemu klasyfikacji używek wybrałem trzy modele: regresję logistyczną, las losowy (Random Forest) oraz SVM (Support Vector Machine). Zostały one zaimplementowane z użyciem biblioteki *sklearn*. Każdy z tych modeli ma unikalne cechy, które mogą uczynić go odpowiednim do tego zadania:

- **Regresja Logistyczna:** przez prostotę, efektywność na zbiorach danych każdego rozmiaru oraz odporność na przetrenowanie
- **Random Forest (Las Losowy):** przez to, że łączy elementy wielu drzew decyzyjnych oraz dobrze radzi sobie z dużą ilościa danych
- **SVM (Support Vector Machine):** przez wysoką efektywność w przestrzeniach wielowymiarowych oraz elastyczność

3.3 Wyniki obliczeń

3.3.1 Regresja Logiczna - Zalety

- **Wysoka średnia dokładność:** 0.826
- **Wysoka dokładność dla wielu zmiennych:** Model osiąga wysoką dokładność dla zmiennych takich jak HEREVER (0.947), PCP (0.967), PEYOTE (0.967), KETMINESK (0.951), DMTAMTFXY (0.960), SALVIADIV (0.940), METHAMEVR (0.884).
- **Stabilność wyników:** Model osiąga stabilne wyniki w wielu kategoriach, co świadczy o jego niezawodności.

3.3.2 Regresja Logiczna - Wady

- **Niska czułość i precyzja dla niektórych zmiennych:** Czułość i precyzja dla niektórych zmiennych jest bardzo niska (0.0) świadcząca o braku pozytywnego wykrycia, np. HEREVER, PCP, PEYOTE, MESC, SALVIADIV, METHAMEVR, SEDANYLIF.

Tabela 1: **Logistic Regression Results**

Metric	Accuracy	Precision (TRUE)	Recall (TRUE)	F1-Score (TRUE)
MJEVER	0.805	0.837	0.902	0.868
COCEVER	0.713	0.542	0.191	0.282
HEREVER	0.947	0.000	0.000	0.000
LSD	0.770	0.500	0.147	0.227
PCP	0.967	0.000	0.000	0.000
PEYOTE	0.967	0.000	0.000	0.000
MESC	0.942	0.000	0.000	0.000
PSILCY	0.750	0.571	0.244	0.342
ECSTMOLLY	0.785	0.415	0.168	0.239
KETMINESK	0.951	0.333	0.026	0.049
DMTAMTFXY	0.960	0.500	0.031	0.059
SALVIADIV	0.940	0.000	0.000	0.000
HALLUCEVR	0.667	0.551	0.495	0.522
INHALEVER	0.792	0.621	0.104	0.178
METHAMEVR	0.884	0.000	0.000	0.000
PNRNMLIF	0.741	0.597	0.195	0.294
TRQNMLIF	0.812	0.463	0.129	0.202
STMANYLIF	0.685	0.573	0.249	0.347
SEDANYLIF	0.655	0.000	0.000	0.000
HLTINALC	0.806	0.808	0.995	0.892
HLTINDRG	0.809	0.811	0.995	0.894
Average	**0.826**	-	-	-

3.3.3 Las Losowy - Zalety

- **Wysoka dokładność dla wielu zmiennych:** Model osiąga wysoką dokładność dla zmiennych takich jak HEREVER (0.944), PCP (0.961), PEYOTE (0.959), KETMINESK (0.945), DMTAMTFXY (0.947), SALVIADIV (0.932).
- **Dobra precyzja dla rzadkich klas:** Random Forest wykazuje dobrą precyzje dla klas występujących rzadko, z którymi poprzedni model miał problemy

3.3.4 Las Losowy - Wady

- **Lekko niższa średnia dokładność:** 0.802
- **Niższa dokładność dla niektórych zmiennych:** W porównaniu z regresją logistyczną, model ma niższą dokładność dla zmiennych takich jak COCEVER (0.682), LSD (0.743), INHALEVER (0.762).
- **Zmienne czułości i precyzje:** Model ma zmienną czułość i precyzję w różnych kategoriach, co wskazuje na pewną niestabilność w przewidywaniach.

Tabela 2: **Random Forest Results**

Metric	Accuracy	Precision (TRUE)	Recall (TRUE)	F1-Score (TRUE)
MJEVER	0.766	0.819	0.861	0.840
COCEVER	0.682	0.437	0.263	0.328
HEREVER	0.944	0.333	0.071	0.118
LSD	0.743	0.389	0.201	0.265
PCP	0.956	0.000	0.000	0.000
PEYOTE	0.960	0.000	0.000	0.000
MESC	0.934	0.111	0.022	0.036
PSILCY	0.728	0.485	0.296	0.367
ECSTMOLLY	0.766	0.370	0.230	0.284
KETMINESK	0.937	0.125	0.053	0.074
DMTAMTFXY	0.942	0.111	0.062	0.080
SALVIADIV	0.930	0.250	0.083	0.125
HALLUCEVR	0.652	0.528	0.481	0.504
INHALEVER	0.762	0.417	0.249	0.312
METHAMEVR	0.857	0.111	0.032	0.050
PNRNMLIF	0.715	0.473	0.276	0.349
TRQNMLIF	0.782	0.378	0.286	0.326
STMANYLIF	0.661	0.495	0.349	0.410
SEDANYLIF	0.588	0.345	0.214	0.264
HLTINALC	0.768	0.812	0.927	0.866
HLTINDRG	0.773	0.815	0.930	0.869
Average	**0.802**	-	-	-

3.3.5 SVM - Zalety

- **Wysoka średnia dokładność:** 0.823
- **Wysoka dokładność dla większości zmiennych:** Model osiąga wysoką dokładność dla zmiennych takich jak HEREVER (0.947), PCP (0.967), PEYOTE (0.967), KETMINESK (0.952), DMTAMTFXY (0.960), SALVIADIV (0.940), METHAMEVR (0.884).
- **Dobra czułość:** Wysoka czułość oznacza, że model dobrze identyfikuje rzeczywiste pozytywne przypadki.
- **Dobra identyfikacja negatywnych przypadków**

3.3.6 SVM - Wady

- **Niska precyzja:** Model ma bardzo niską precyzję, co oznacza problem z identyfikacją prawdziwych pozytywnych przypadków.

Tabela 3: **SVM Results**

Metric	Accuracy	Precision (TRUE)	Recall (TRUE)	F1-Score (TRUE)
MJEVER	0.792	0.853	0.856	0.854
COCEVER	0.705	0.000	0.000	0.000
HEREVER	0.947	0.000	0.000	0.000
LSD	0.770	0.000	0.000	0.000
PCP	0.967	0.000	0.000	0.000
PEYOTE	0.967	0.000	0.000	0.000
MESC	0.942	0.000	0.000	0.000
PSILCY	0.733	0.000	0.000	0.000
ECSTMOLLY	0.798	0.000	0.000	0.000
KETMINESK	0.952	0.000	0.000	0.000
DMTAMTFXY	0.960	0.000	0.000	0.000
SALVIADIV	0.940	0.000	0.000	0.000
HALLUCEVR	0.666	0.547	0.512	0.529
INHALEVER	0.783	0.000	0.000	0.000
METHAMEVR	0.884	0.000	0.000	0.000
PNRNMLIF	0.723	0.000	0.000	0.000
TRQNMLIF	0.816	0.000	0.000	0.000
STMANYLIF	0.663	0.000	0.000	0.000
SEDANYLIF	0.655	0.000	0.000	0.000
HLTINALC	0.806	0.806	1.000	0.893
HLTINDRG	0.809	0.809	1.000	0.894
Average	**0.823**	-	-	-

4 Wnioski

4.1 Analiza danych

Przeprowadzone analizy potwierdziły, że demograficzne i środowiskowe zmienne mają istotny wpływ na zażywanie różnych substancji. Największy wpływ miały wiek, płeć oraz wcześniejsze doświadczenia z alkoholem i papierosami.

4.2 Modelowanie

Każdy z modeli ma swoje zalety i wady:

- **Regresja Logistyczna** dobrze radzi sobie z przewidywaniem negatywów, ale ma problemy z kategoriami, gdzie jest mało przypadków pozytywnych.
- **Random Forest** dobrze radzącym sobie z rzadkimi klasami, ale ma zmienne wyniki w różnych kategoriach i niższą precyzję
- **SVM** osiąga wysoką dokładność i czułośćdla negatywów, ale ma duże problemy z precyzją dla nielicznych pozytywów.

Według mojej opinii najlepszą z dostępnych opcji do rozwiązania tego problemu będzie Random Forest, ponieważ jest on jako jedyny użyteczny przy obsłudze klas rzadkich. Z tego właśnie powodu został wybrany do zaimplementowania kodu, który na podstawie informacji podanych z klawiatury próbuje przewidzieć, jakie substancje brał użytkownik.

A Dodatek

Kody źródłowe umieszczone zostały w repozytorium github:
`https://github.com/pbonar/msid_project/`.

Link do pobrania oryginalnej wersji danych (Wersja SPSS): `https://www.datafiles.samhsa.gov/dataset/national-survey-drug-use-and-health-2022-nsduh-2022-ds0001`.