Klasyfikacja zażywania narkotyków w zależności od cech demograficznych, cech osobowości i zdrowia psychicznego

Malwina Juchiewicz
Informatyka stosowana, Politechnika Wrocławska
272660@student.pwr.edu.pl

May 2024

Contents

1 Wprowadzenie

Badania dotyczące zażywania narkotyków i czynników wpływających na to zjawisko stanowią istotny obszar zainteresowania ze względu na swoje znaczenie społeczne, zdrowotne i prawne. W dzisiejszym społeczeństwie problem używania substancji psychoaktywnych jest coraz bardziej rozpowszechniony i złożony, co sprawia, że analiza różnorodnych czynników determinujących to zachowanie staje się niezwykle istotna.

Podobnie jak w przypadku analizy danych w sporcie czy innych dziedzinach, gromadzenie informacji na temat demograficznych, osobowościowych i zdrowotnych czynników może posłużyć do opracowania zaawansowanych modeli predykcyjnych. Takie podejście umożliwia nie tylko lepsze zrozumienie samego zjawiska zażywania narkotyków, ale również pozwala na identyfikację czynników ryzyka oraz opracowanie skuteczniejszych strategii prewencji i interwencji.

Niniejszy raport ma na celu przybliżenie metody analizy danych w kontekście zażywania narkotyków, wykorzystując zaawansowane modele predykcyjne oparte na danych demograficznych, osobowościowych i zdrowotnych. Poprzez analizę tych czynników oraz ich wzajemnych powiązań, będziemy dążyć do odpowiedzi na pytania dotyczące czynników determinujących ryzyko zażywania narkotyków oraz możliwości przewidywania tego zachowania na podstawie dostępnych danych. Ostatecznie, celem jest wypracowanie narzędzi, które mogą wspierać działania prewencyjne oraz interwencyjne w obszarze problematyki narkotykowej, przyczyniając się do poprawy zdrowia publicznego, a także bezpieczeństwa społecznego.

2 Dane

2.1 Zbiór danych

Dostępne są dwa zbiory danych, z których jeden dotyczy konsumpcji narkotyków, a drugi koncentruje się na zdrowiu psychicznym, ze szczególnym uwzględnieniem zaburzeń lękowych. Finalny zbiór danych został stworzony poprzez połączenie tych dwóch zbiorów.

2.1.1 Zbiór I - konsumpcja narkotyków

Zbiór dotyczący konsumpcji narkotyków został uzyskany ze strony *kaggle.com*. Zbiór danych zawiera informacje dotyczące 1885 respondentów, z których każdy jest opisany za pomocą 12 cech. Cechy obejmują pomiary osobowości, ocenione głównie za pomocą skali NEO-FFI-R, poziom wykształcenia, wiek, płeć, kraj zamieszkania i pochodzenie etniczne stanowią kolejne cechy opisujące respondentów. Dodatkowo, uczestnicy zostali zapytani o swoje doświadczenie związane z 18 różnymi substancjami, w tym legalnymi i nielegalnymi narkotykami, a także fikcyjnym lekiem (Semeron), który został wprowadzony w celu identyfikacji osób zgłaszających nadmierną ilość spożycia. Dla każdej substancji respondent miał wybór między kilkoma odpowiedziami, określającymi, czy nigdy nie używał

danej substancji, czy też używał jej w przeszłości (przed dekadą, w ciągu ostatniej dekady, roku, miesiąca, tygodnia lub dnia). Początkowo wszystkie dane w zbiorze były reprezentowane jako wartości liczbowe. Jednak w pierwszej części analizy zostały one przekształcone z powrotem na ich kategoryczne odpowiedniki.

Każdy wiersz reprezentuje jednego respondera i zawiera następujące informacje:

- **ID:** unikalny identyfikator respondenta
- **Age:** Grupa wiekowa uczestnika.
- **Gender:** Płeć uczestnika).
- **Education:** Poziom wykształcenia uczestnika.
- **Country:** Kraj pochodzenia uczestnika.
- **Ethnicity:** Etniczne pochodzenie uczestnika.
- **Nscore:** Pomiar neurotyczności wg skali NEO-FFI-R.
- **Escore:** Pomiar ekstrawersji wg skali NEO-FFI-R.
- **Oscore:** Pomiar otwartości na doświadczenia wg skali NEO-FFI-R.
- **Ascore:** Pomiar ugodowości wg skali NEO-FFI-R.
- **Cscore:** Pomiar sumienności wg skali NEO-FFI-R.
- **Impulsive:** Poziom impulsywności mierzony przez BIS-11.
- **SS:** Poziom skłonności do poszukiwania wrażeń mierzony przez ImpSS.
- **Alcohol:** Konsumpcja alkoholu.
- **Amphet:** Konsumpcja amfetamin.
- **Amyl:** Konsumpcja nitrytu amylu.
- **Benzos:** Konsumpcja benzodiazepin.
- **Caff:** Konsumpcja kofeiny.
- **Cannabis:** Konsumpcja marihuany.
- **Choc:** Konsumpcja czekolady.
- **Coke:** Konsumpcja kokainy.
- **Crack:** Konsumpcja crack.
- **Ecstasy:** Konsumpcja ekstazy.

- **Heroin:** Konsumpcja heroiny.
- **Ketamine:** Konsumpcja ketaminy.
- **Legalh:** Konsumpcja legalnych narkotyków.
- **LSD:** Konsumpcja LSD.
- **Meth:** Konsumpcja metadonu.
- **Mushroom:** Konsumpcja grzybów magicznych.
- **Nicotine:** Konsumpcja nikotyny.
- **Semer:** Klasa fikcyjnego leku Semeron.
- **VSA:** Klasa używania substancji lotnych.

2.1.2 Zbiór II - zdrowie psychiczne

Zbiór danych dotyczących zdrowia psychicznego został pozyskany ze strony internetowej *ourworldindata.org*. Zawiera on informacje na temat udziału populacji cierpiącej na zaburzenia lękowe. Dane te pozwalają zgłębić temat zdrowia psychicznego na skalę globalną, przybliżając zasięg i powszechność tego problemu.

Podobnie jak w przypadku danych o narkotykach, każdy rekord w tym zbiorze zawiera kluczowe informacje, takie jak kraj, przedział wiekowy oraz udział populacji z zaburzeniami lękowymi.

2.2 Wstępne przetwarzanie danych

Przygotowanie danych do analizy wymagało wykonania kilku kroków przetwarzania. W przypadku danych dotyczących konsumpcji narkotyków, pierwszym krokiem było usunięcie kolumny ID, która zawierała numer rekordu w oryginalnej bazie danych i nie wnosiła istotnych informacji do analizy. Następnie przeprowadzono filtrację, usuwając wszystkich respondentów, którzy zaznaczyli spożywanie fikcyjnego narkotyku Semer.

Po tej operacji również usunięto kolumnę zawierającą informacje o spożyciu Semer, aby zachować spójność struktury danych.

Kolejnym krokiem było dokonanie klasyfikacji narkotyków do trzech głównych grup: opioidy, ekstazy i benzodiazepiny.Podział został dokonany ze względu na podobną strukturę, działania substancji, ich specyficzne efekty, a także ryzyko związane z ich zażywaniem.

1. **Opioidy** są głównie używane jako środki przeciwbólowe i mają wysoki potencjał uzależniający.
2. **Ekstazy** są znane ze swoich efektów euforycznych i halucynogennych.
3. **Benzodiazepiny** są często stosowane jako leki przeciwlękowe i nasenne.

Do uwzględnienia wartości w nowo wygenerowanych kolumnach wzięto pod uwagę odpowiedź o najwyższej wartości (im dawniejsza odpowiedź, tym niższa waga). Każdy respondent został przypisany do odpowiednich grup, zgodnie z udzielonymi odpowiedziami, a następnie usunięto wszystike zbędne kolumny z konkretnymi nazwami narkotyków.

W późniejszych etapach dane kategoryczne należało zamienić na numeryczne oraz odpowiednio je przeskalować.

W przypadku danych dotyczących zdrowia psychicznego, przetwarzanie polegało na zachowaniu kolumn z nazwą państwa oraz przedziałami wiekowymi. Z kolumny zawierającej informacje o państwie usunięto wszystkie kraje niepasujące do tych, które znajdowały się w pierwszym zbiorze. Przedziały wiekowe zostały przeliczone w taki sposób, aby współgrać z danymi dotyczącymi konsumpcji narkotyków. Użyto do tego średniej ważonej. Następnie przydzielono dane do następujących kategorii:

- 0% - 5% - 1
- 5% - 6% - 2
- 6% - 7% - 3
- $>$ 7% - 4

Zbiory danych zostały połączone na podstawie wspólnych kryteriów wieku i kraju.

2.3 Analiza eksploracyjna

2.3.1 Dystrybucje cech demograficznych

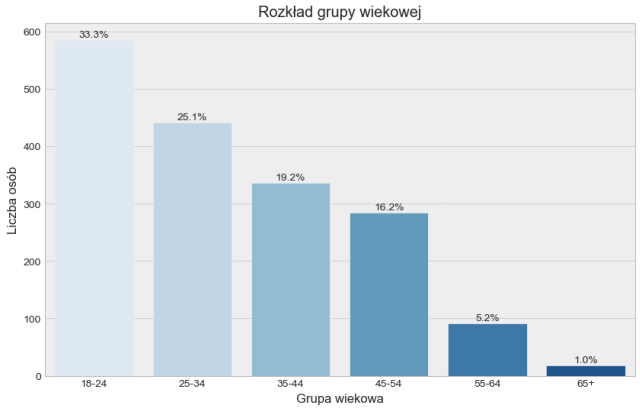

Figure 1: Histogram grup wiekowych

Histogram przedstawia rozkład wieku respondentów. Najliczniejszą grupę stanowią osoby w wieku 18–24 lat, a liczba respondentów maleje wraz z wiekiem. Oznacza to, że starsze grupy wiekowe są mniej reprezentowane w badaniu, co należy uwzględnić przy interpretacji wyników.

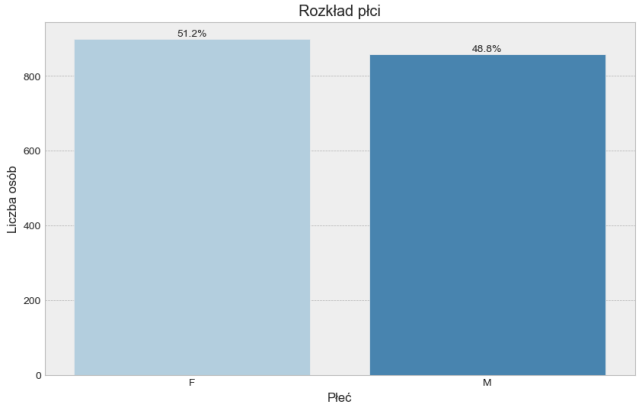

Figure 2: Histogram płci

Struktura płci respondentów jest zbliżona, z niewielką przewagą kobiet (51,2%) nad mężczyznami (48,8%).

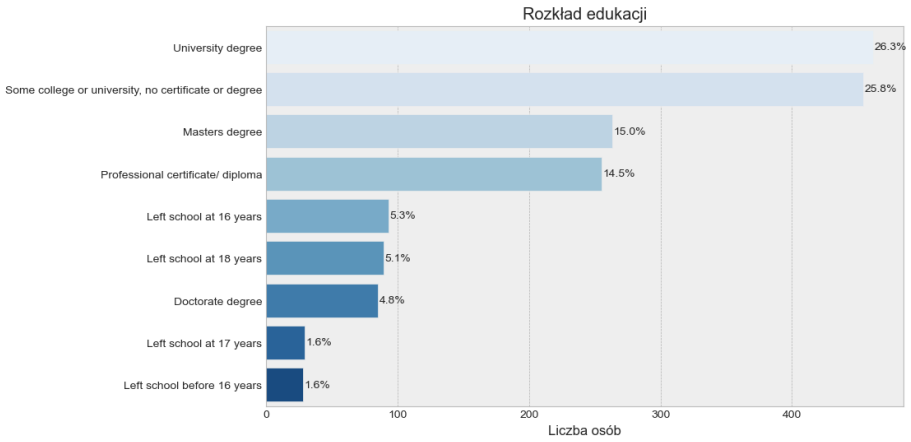

Figure 3: Histogram edukacji

Dane dotyczące pochodzenia etnicznego i kraju pochodzenia
Pochodzenie etniczne oraz kraj pochodzenia mają bardzo niezróżnicowane dane. W przypadku pochodzenia etnicznego ponad 90% to ludzie biali. Zdecydowana większość respondentów pochodzi z UK (około 60%) oraz USA(około 30%).

2.4 Dystrybucja cech osobowościowych

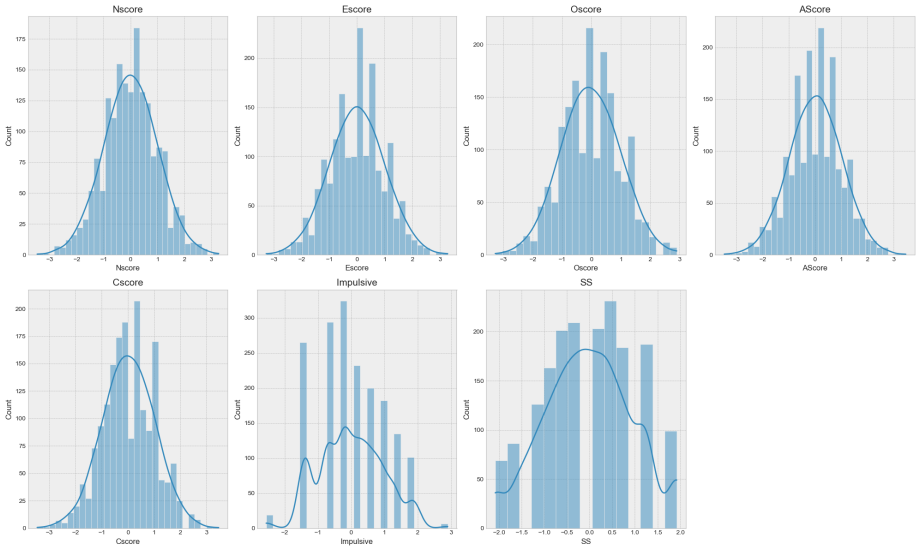

Figure 4: Dystrybucja cech osobowościowych

Wszystkie cechy osobowościowe, czyli AScore, Nscore, Escore, Oscore, Cscore , Impulsive oraz SS w przybliżeniu mają rozkład normalny.

2.4.1 Zażywanie narkotyków

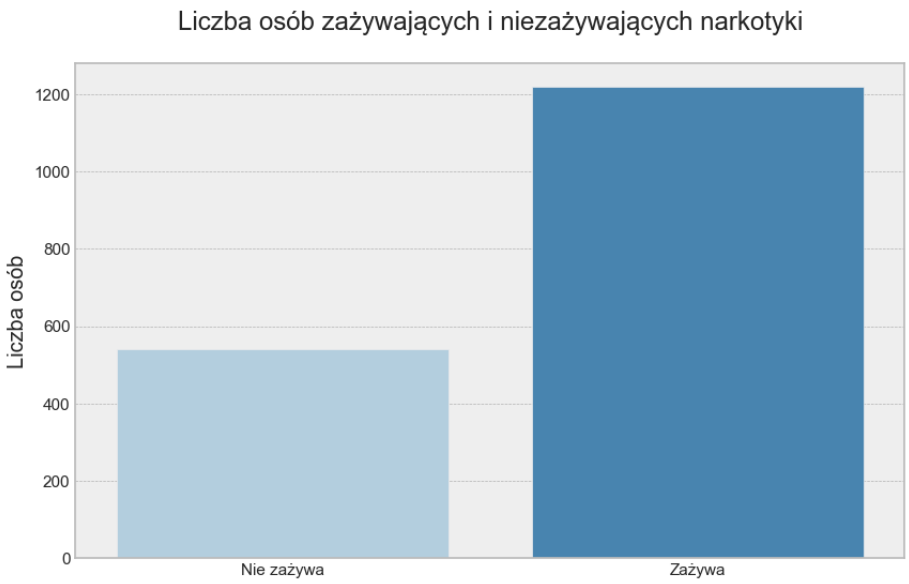

Figure 5: Podział osób na zażywających i niezażywających narkotyki

Wykres przedstawia liczbę osób zażywającą oraz niezażywającą narkotyki. Dane zostały przeliczone na podstawie deklaracji respondentów dotyczących czasu ostatniego zażycia narkotyków.

W kategorii "**Niezażywający**" (non-user) uwzględniono następujące odpowiedzi:

- Nigdy
- Dekadę temu

W kategorii "**Zażywający**" (user) uwzględniono następujące odpowiedzi:

- W ostatniej dekadzie
- W ostatnim roku
- W ostatnim miesiący
- W ostatnim tygodniu
- Wczoraj

2.4.2 Zażywanie narkotyków w poszczególnych grupach narkotykowych

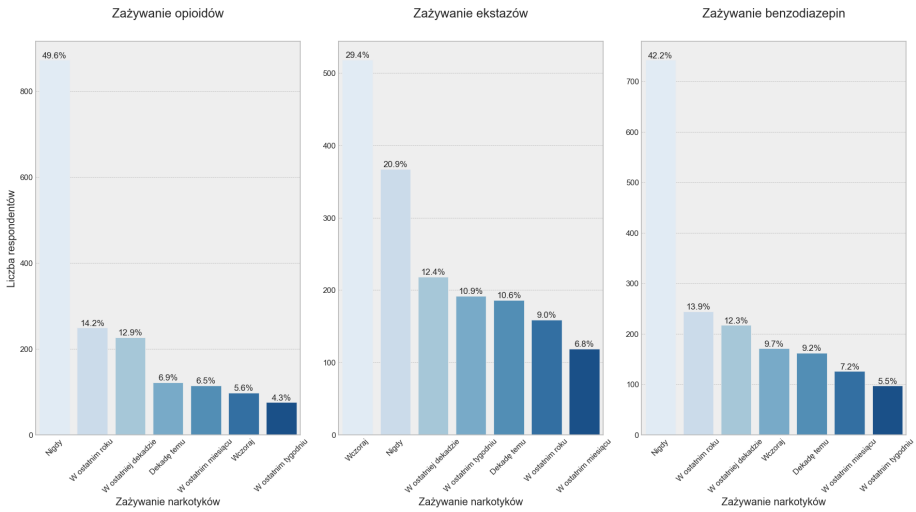

Figure 6: Częstotliwość zażywania narkotyków

Wykres przedstawia ostatni czas zażycia narkotyku dla każdej z trzech grup narkotykowych: opioidów, ekstazów i benzodiazepin. Zarówno w przypadku opioidów, jak i benzodiazepin zdecydowanie najczęstrzą odpowiedzią było 'Nigdy', natomiast w przypadku ekstazów około 1/3 respondentów zaznaczyła odpowiedź 'Wczoraj'.

Podział ten można uprościć, dzieląc respondentów na tych, którzy używają narkotyków, i tych, którzy ich nie używają, w sposób wcześniej opisany.

Na podstawie tych definicji, wartości procentowe respondentów w każdej grupie narkotykowej wyglądają następująco:

Opioidy:

- Respondenci, którzy nie zażywają narkotyków: 56%
- Respondenci zażywający narkotyki: 44%

Ekstazy:

- Respondenci, którzy nie zażywają narkotyków: 30%
- Respondenci zażywający narkotyki: 70%

Benzodiazepiny:

- Respondenci, którzy nie zażywają narkotyków: 51%
- Respondenci zażywający narkotyki: 49%

2.4.3 Zależności poszczególnych cech od grup narkotykowych

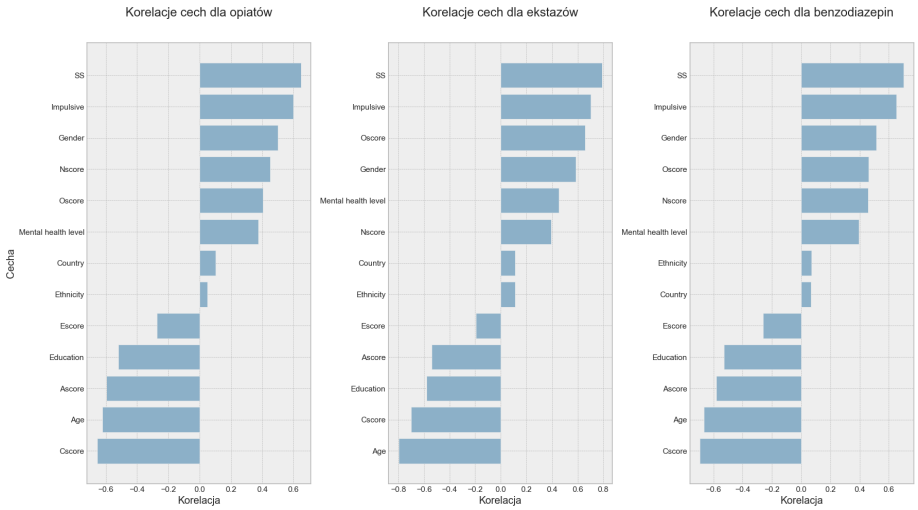

Figure 7: Korelacja cech

Wykres przedstawia korelację cech demograficznych i osobowościowych dla każdej z grup narkotykowych: opioidów, ekstazów i benzodiazepin. Najbardziej skorelowaną cechą we wszystkich trzech przypadkach jest sensacja poszukiwań (SS). Najmniej skorelowanymi cechami są kraj pochodzenia oraz pochodzenie etniczne, prawdopodobnie ze względu na ich małą różnorodność.

2.4.4 Zależność wieku od zażywania narkotyków

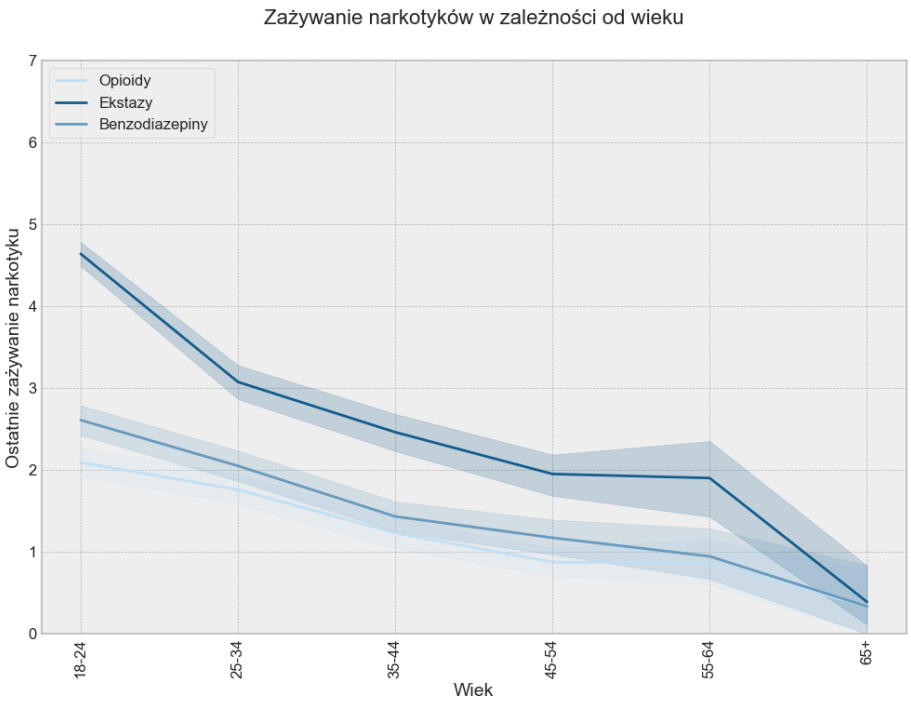

Figure 8: Średnie zażycie narkotyków w zależności od wieku

Wykres przedstawia zależność zażywania narkotyków od wieku dla trzech grup narkotykowych: opioidów, benzodiazepin i ekstazów. Najmłodsza grupa respondentów wykazuje najwyższą skłonność do zażywania narkotyków, szczególnie ekstazów. Wraz z wiekiem odsetek osób deklarujących zażywanie narkotyków maleje, jednak liczba respondentów w starszych grupach wiekowych jest mniejsza, co może wpływać na interpretację tego trendu.

2.4.5 Zależność sensacji poszukiwań od zażywania narkotyków

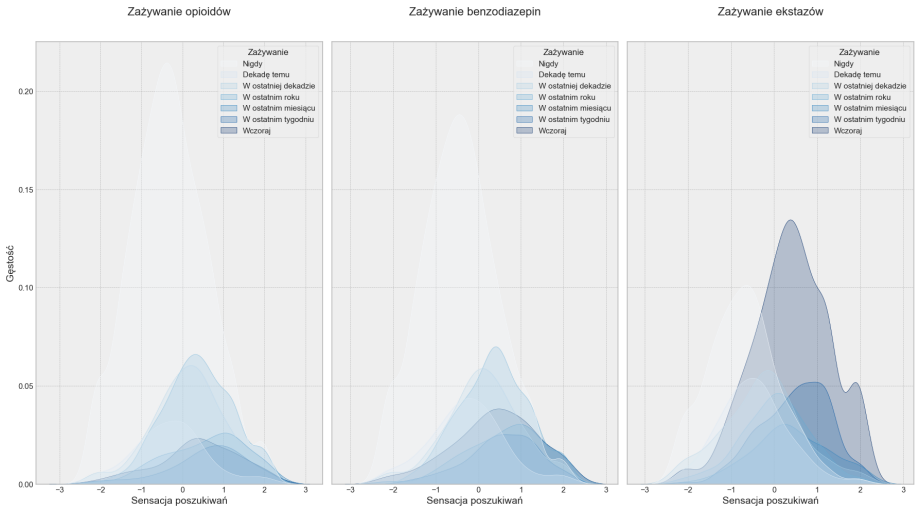

Figure 9: Wpływ sensacji poszukiwań (SS) na zażywanie narkotyków

Wykres przedstawia zależność sensacji poszukiwań (SS) od zażywania narkotyków. Z wykresu wynika (w szczególności dla ekstazów), że im wyższy poziom SS, tym ludzie są bardziej skłonni do zażywania niebezpiecznych substancji.

Poszukiwanie wrażeń to cecha osobowościowa, która charakteryzuje się poszukiwaniem nowych i ekscytujących doznań oraz gotowością do podejmowania ryzyka w celu doświadczenia silnych emocji i stymulacji. Osoby o wysokim poziomie sensation seekingu często szukają nowych wrażeń i doświadczeń, które mogą prowadzić do podjęcia zachowań eksperymentalnych, takich jak zażywanie narkotyków.

2.4.6 Zależność sensacji poszukiwań od wieku

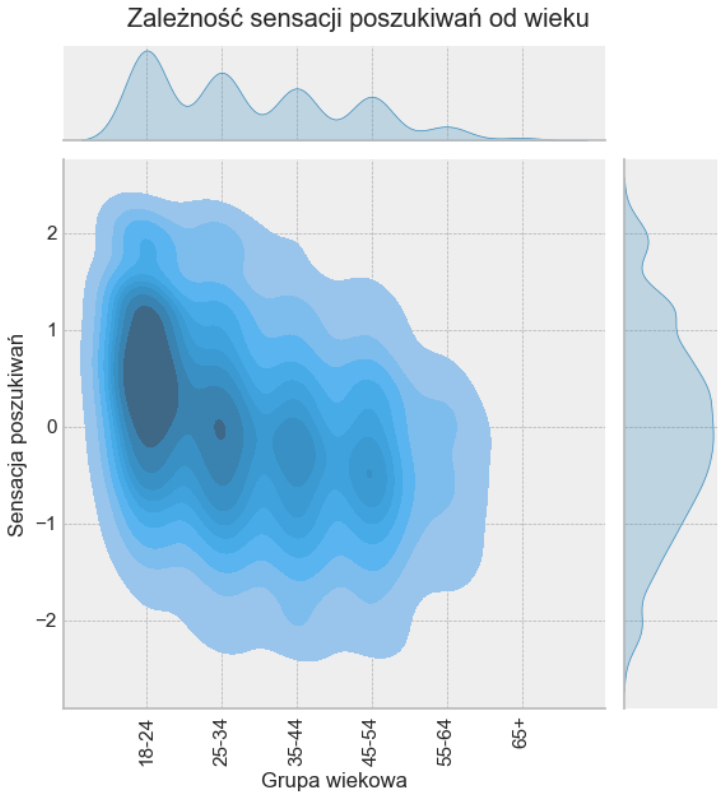

Figure 10: SS w zależności od wieku

Wykres przedstawia zależność poziomu SS (poszukiwania doznań) od grupy wiekowej. Istnieje tendencja do zmiany poziomu poszukiwania doznań wraz z wiekiem, przy czym młodsze osoby często wykazują wyższe poziomy SS niż osoby starsze. Młodsze osoby mogą być bardziej skłonne do poszukiwania nowych doznań i doświadczeń, co może prowadzić do większej skłonności do eksperymentowania z narkotykami. W miarę jak osoby starzeją się i doświadczają różnych aspektów życia, ich poziom SS może zmniejszać się, co może skutkować zmniejszeniem częstotliwości zażywania narkotyków. Należy jednak zauważyć, że liczba respondentów w starszych grupach wiekowych jest mniejsza, co może wpływać na wyniki, a ta zależność może być złożona. Dodatkowo, indywidualne różnice w poziomie SS mogą występować w każdej grupie wiekowej.

2.4.7 Zależność zdrowia psychicznego od zażywania narkotyków

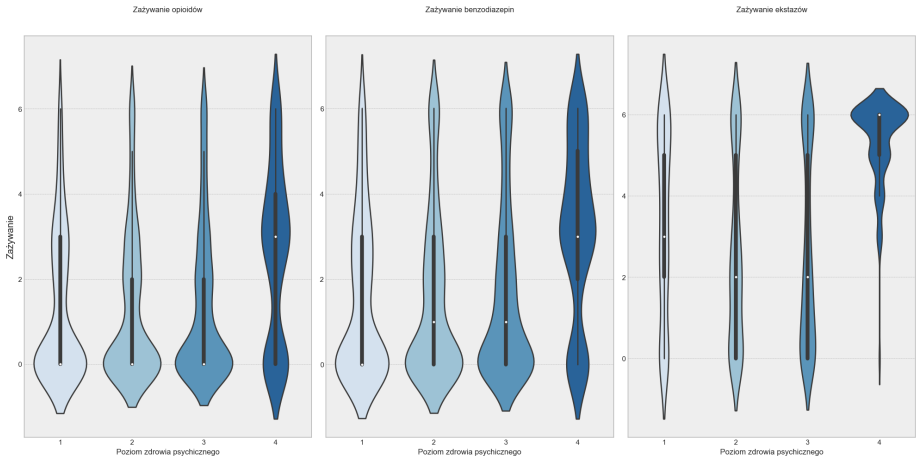

Figure 11: Zażywanie narkotyków w zależności od zdrowia psychicznego

Wykres przedstawia zależność zażywania narkotyków od zdrowia psychicznego (w szczególności zaburzeń lękowych) dla trzech grup narkotykowych: opioidów, benzodiazepin i ekstazów.

Zdrowie psychiczne, a w szczególności zaburzenia lękowe, odgrywa znaczącą rolę w skłonności do zażywania narkotyków. Dane wskazują, że osoby z wyższym poziomem zaburzeń lękowych są bardziej skłonne do używania narkotyków.

3 Macierz korelacji

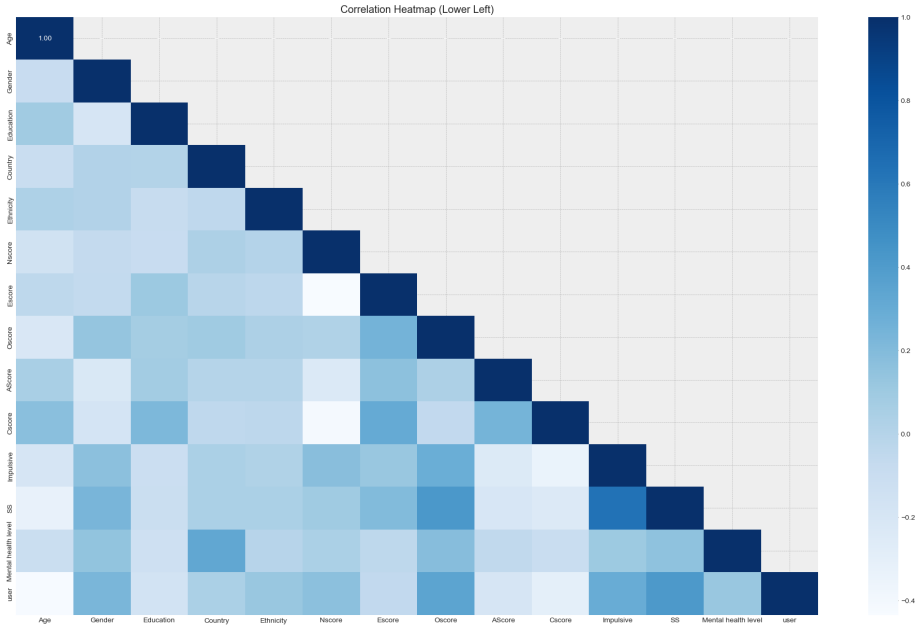

Figure 12: Macierz korelacji cech

Jedną z najmocniej skorelowanych cech z zażywaniem narkotyków jest wiek użytkownika. Nie ma w tym nic dziwnego, ponieważ w dzisiejszych czasach młodzi ludzie częściej popadają w nałogi.Wiele cech osobowościowych (takich jak OScore, CScore, Impulsive, SS) są także stosunkowo mocno skorelowane. Ostatnią cechą mającą bardziej znaczący wpływ jest edukacja respondenta.

4 Modele

Przed przystąpieniem do trenowania modeli, należy odpowiednio przygotować dane. Zbiór danych został podzielony na zmienne zależne (wynik, który chcemy przewidzieć) i zmienne niezależne (cechy, na podstawie których dokonujemy przewidywań). Następnie dane zostały podzielone na dwie części: zbiór przeznaczony do trenowania modelu oraz zbiór przeznaczony do oceny jego efektywności w stosunku 80:20, z wykorzystaniem ziarna losowości o wartości 12 oraz wykorzystując przetasowanie danych. Dane zostały przeskalowane przy pomocy klasy *StandardScaler*.

4.0.1 Model Regresji Logistycznej

W celu poprawnego doboru modelu skorzystano z GridSearch, który sprawdza wszsystkie możliwe kombinacje parametrów wejściowych. Wejściowa siatka parametrów wyglądała nastepująco: C = [0.001, 0.01, 0.1, 1, 10, 100, 1000], penalty = [l1, l2, None], solver = [lbfgs, liblinear] max iter = [5000, 10000]

Uzyskano następujące najlepsze parametry:

- C: 0.05
- max iter: 5000
- penalty: l2
- solver: liblinear

Z wybranymi parametrami uzyskano następujące wyniki:

	Precision	**Recall**	**F1-Score**
non-user	0.68	0.66	0.67
user	0.86	0.88	0.87
Accuracy		0.81	
Śr. arytm.	0.77	0.77	0.77
Śr. ważona	0.81	0.81	0.81

Table 1: Wyniki modelu regresji logistycznej

Model regresji logistycznej osiągnął skuteczność na poziomie 81%, korzystając z optymalnych parametrów wybranych za pomocą metody *GridSearch*.

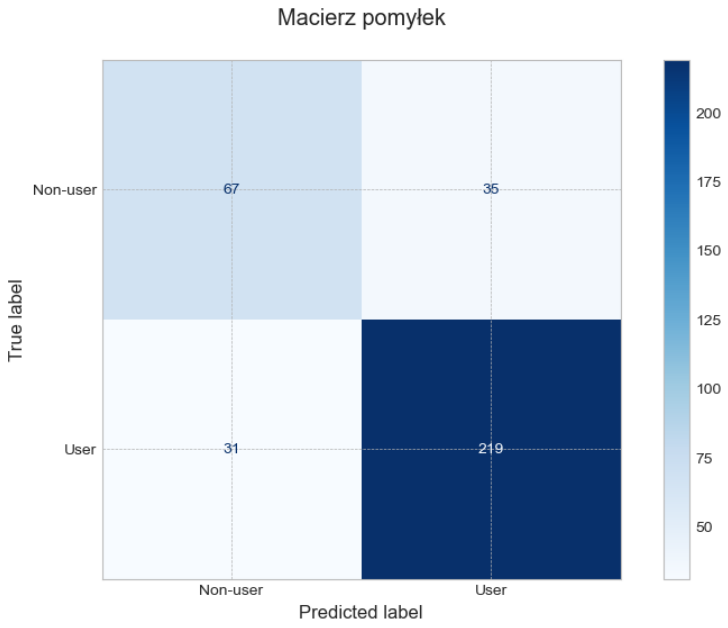

Figure 13: Macierz pomyłek dla modelu regresji logistycznej

4.1 Klasyfikator maszyny wektorów nośnych

Podobnie jak w przypadku poprzedniego modelu, uzyskano najlepsze paramtry przy pomocy GridSearch:

- C: 0.1
- gamma: auto
- penalty: l2
- kernel: linear

Z wybranymi parametrami uzyskano następujące wyniki:

	Precision	**Recall**	**F1-Score**
non-user	0.73	0.69	0.70
user	0.86	0.88	0.87
Accuracy		0.82	
Śr. arytm.	0.79	0.78	0.78
Śr. ważona	0.81	0.82	0.82

Table 2: Wyniki modelu SVM

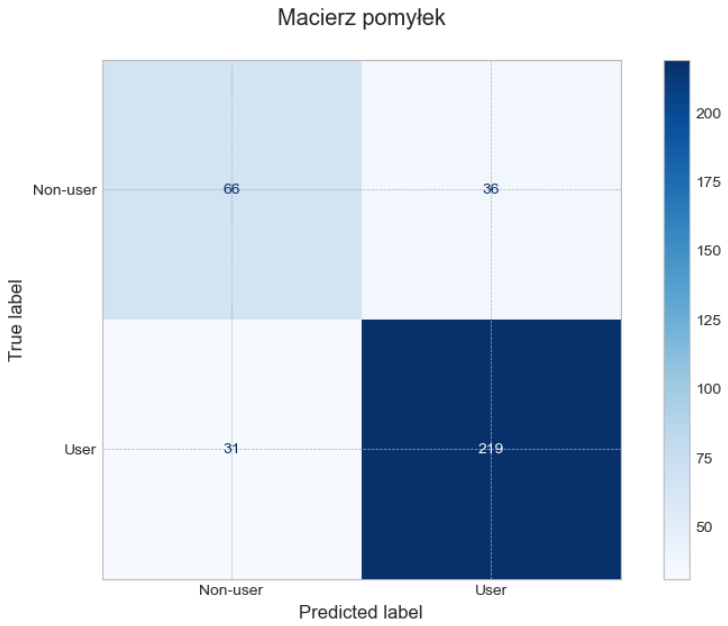

Figure 14: Macierz pomyłek klasyfikatora wektorów nośnych

Model klasyfikatora wektorów nośnych osiągnął skuteczność na poziomie 82%, korzystając z optymalnych parametrów wybranych za pomocą metody *GridSearch*.

5 Wnioski

Analizując wyniki modeli klasyfikacyjnych dla danych dotyczących narkotyków, możemy wyciągnąć kilka istotnych wniosków.

Oba modele poradziły sobie stosunkowo dobrze. Pierwszy model lepiej przewidział, który użytkownik nie zażywa narkotyków (non-user), natomiast drugi model lepiej przewidział, który z użytkownik zażywa narkotyki (user).

Model regresji logistycznej, z optymalnymi parametrami dobranymi za pomocą metody GridSearch, uzyskał 81% skuteczności w klasyfikacji. Z drugiej strony, model SVM, również z optymalnie dobranymi parametrami, osiągnął skuteczność na poziomie 78%.