# Organizacja pracy

Na samym początku spotkaliśmy się i zapoznaliśmy się z problemem, następnie zadania na mniejsze subproblemy. Alicja, mająca największe doświadczenie, zaproponowała wykorzystanie platformy Kaggle do pracy nad modelem i danymi, co pozwoliło nam na łatwą współpracę oraz dostęp do narzędzi do analizy danych.

kaggle

link

GitHub

link

# Wykorzystanie DataFrame

Jako strukturę danych wybraliśmy DataFrame (z biblioteki pandas), ponieważ oferuje on dużą elastyczność i wydajność w pracy z dużymi zbiorami danych. DataFrame umożliwia łatwe manipulowanie danymi, filtrowanie, grupowanie oraz wykonywanie operacji agregacyjnych.


# Analiza danych

**1.** Rozpoczęliśmy od wczytania danych do DataFrame i wykonania wstępnej analizy za pomocą narzędzi takich jak Pandas i Matplotlib. Wykonaliśmy losowy sampling danych oraz przejrzeliśmy opisy kolumn, aby zdobyć ogólne i szczegółowe informacje na temat zbioru danych.

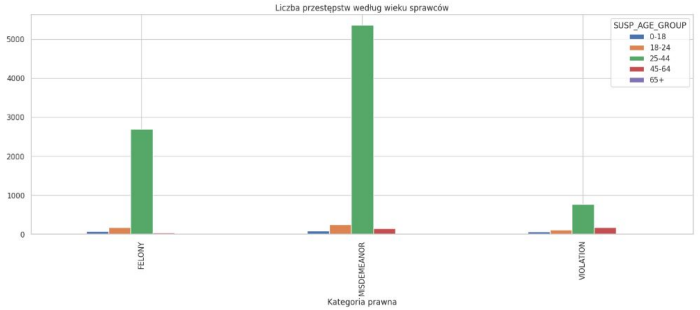
**2.** Analiza danych pozwoliła nam zobrazować nasz cel oraz ocenić przydatność danych.
Czyszczenie danych zawierało:
- Usuwanie wartości typu 'null'.
- Naprawa wartości logicznie niemożliwych, np. wiek równy 900 lat.
- Poprawa formatowania wartości w kolumnach.
- Usuwanie zbędnych wartości.
- Zastępowanie brakujących wartości średnimi lub medianą, ponieważ te metody minimalizują zniekształcenie rozkładu danych.
- Standaryzacja wartości narzędziami takimi jak MinMaxScaler, aby zapewnić, że wszystkie cechy mają porównywalne skale

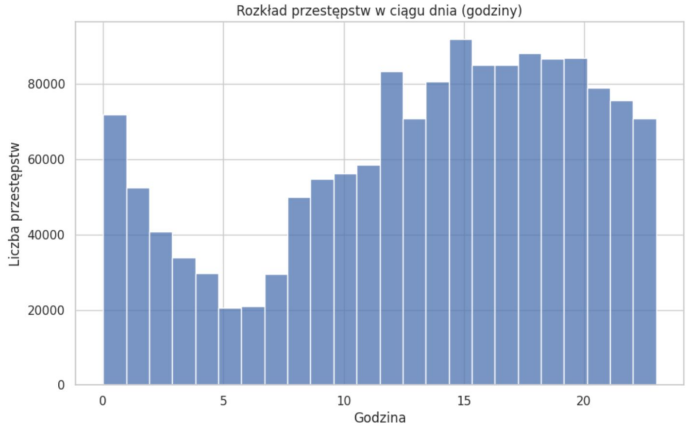
# Wizualizacje

Po obróbce i czyszczeniu danych wykonaliśmy parę interesujących wizualizacji:

Na wykresie pierwszym widać ilość konkretnych typów przestępstw według wieku sprawców. Okazuje się, że bez znaczenia na kategorię, w zdecydowanej większości przestępstwa popełniały osoby w wieku od 25 do 44 lat.

Wykres drugi przedstawia liczbę przestępstw w konkretnych godzinach w ciągu dnia. Widać, że najmniej przestępstw było dokonanych w godzinach porannych - od 5 do 7 rano. Najwięcej dokonywano w godzinach popołudniowych.



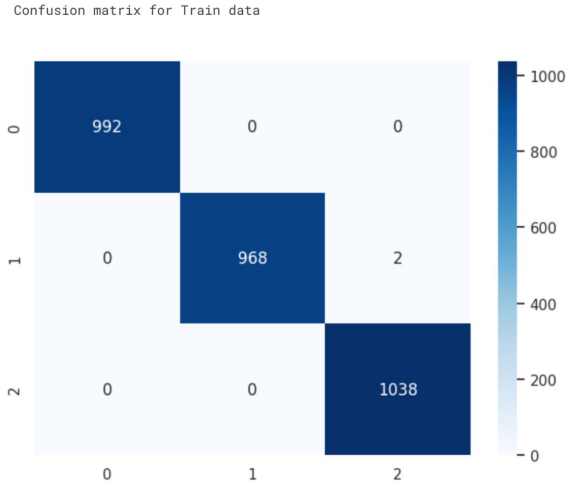
# Model

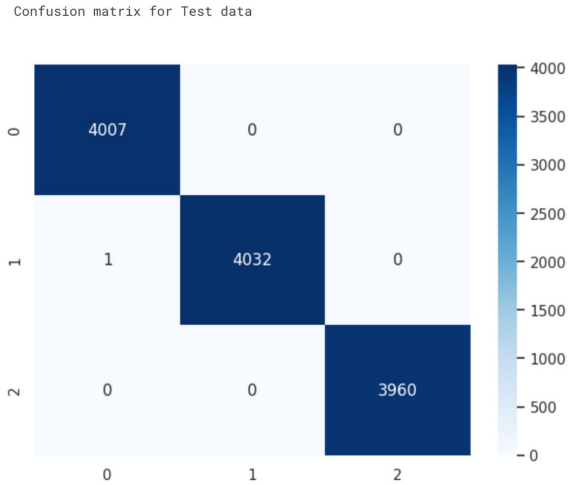
**Wybór modelu**
Po analizie danych, zdecydowaliśmy się na wykorzystanie modelu k-NN do klasyfikacji. Wybraliśmy ten model ze względu na jego prostotę i skuteczność w klasyfikacji wieloklasowej.

**Podzielenie zbioru na dane treningowe i testowe**
Podzieliliśmy dane na zbiór treningowy i testowy, stosując współczynnik 80/20, aby zapewnić równowagę między trenowaniem modelu a jego walidacją.

**Trening modelu**
Przetestowaliśmy różne wartości parametru k dla k-NN, aby znaleźć optymalną liczbę sąsiadów. Najlepszy wynik uzyskaliśmy dla k = 5.

**Wyniki modelu**
Model osiągnął bardzo wysoką dokładność na zbiorze testowym. (można tutaj wkleić macierz z kaggle ostatnią)



# Podsumowanie

Nasza analiza dostarczyła cennych wglądów w naturę i rozkład zgłoszeń do NYPD. Model klasyfikujący, choć wstępny, wykazał obiecujące wyniki i może być dalej doskonalony. Kolejne kroki mogłyby obejmować głębszą analizę konkretnych typów przestępstw oraz rozbudowę modelu o inne zadania klasyfikujące. Projekt ten podkreślił znaczenie dokładnego czyszczenia danych i przemyślanego wyboru modelu.