Przetwarzanie i analiza dużych zbiorów danych2021/2022

Prowadzący: mgr inż. Rafał Woźniak Czwartek, 15:45

Szymon Gruda	239661	239661@edu.p.lodz.pl
Jan Karwowski	239671	239671@edu.p.lodz.pl
Michał Kidawa	239673	239673@edu.p.lodz.pl
Kamil Kowalewski	239676	239676@edu.p.lodz.pl

Checkpoint 4

Spis treści

1. Wprowadzenie

Głównym celem projektu jest przeprowadzenie kompleksowej analizy zbioru *NYPD Complaint Data Historic* [1] poprzez wstępną analizę zbioru danych, zaobserwowanie zależności pomiędzy danymi i określenie celów projektu, których realizacja da odpowiedź na podstawione cele badawcze i na jej podstawie zostaną wyciągnie wnioski.

2. Podział obowiązków w zespole

— Szymon Gruda - Realizacja celu "Klasyfikacja rodzaju lub poziomu przestępstwa"
— Jan Karwowski - Realizacja celu "Regresja godziny wystąpienia zdarzenia"
— Michał Kidawa - Realizacja celu "Grupowanie przestępstw na podstawie podzbiorów cech"
— Kamil Kowalewski - Przygotowywanie preprocessingu, miejsca oraz infrastruktury do realizacji celów badawczych. Pomoc programistyczna oraz techniczna dla pozostałych członków zespołu w badaniach

3. Charakterystyka zbioru danych

3.1. Przegląd pierwotnie dostępnych kolumn

Zbiór danych zawiera ponad 7 mln. (7375993) wierszy i 35 kolumn różnego typu, głównie dane kategorialne. Poniżej wypisane zostały wszystkie dostępne kolumny, pogrupowane tematycznie, wraz z informacją kolejno o liczbie unikalnych wartości (włącznie z NaN) i brakujących wartości (NaN) w danej kolumnie.

1. Identyfikator
   — CMPLNT_NUM (unikalnych 7373143, brakujących 0) - Losowo generowany trwały identyfikator dla każdego zgłoszenia
2. Data i czas zdarzenia
   — CMPLNT_FR_DT (unikalnych 8607, brakujących 655) - Dokładna data wystąpienia zgłoszonego zdarzenia (lub data początkowa wystąpienia, jeżeli CMPLNT_TO_DT istnieje)
   — CMPLNT_FR_TM (unikalnych 1442, brakujących 48) - Dokładny czas wystąpienia zgłoszonego zdarzenia (lub czas rozpoczęcia wystąpienia, jeżeli CMPLNT_TO_TM istnieje)
   — CMPLNT_TO_DT (unikalnych 6826, brakujących 1704204) - Data końcowa wystąpienia zgłoszonego zdarzenia, jeżeli dokładny czas wystąpienia nie jest znany
   — CMPLNT_TO_TM (unikalnych 1442, brakujących 1699541) - Końcowy czas wystąpienia zgłoszonego zdarzenia, jeżeli dokładny czas wystąpienia nie jest znany
3. Data i czas zgłoszenia
   — RPT_DT (unikalnych 5479, brakujących 0) - Data zgłoszenia zdarzenia na policję

4. Typ i opis wykroczenia/przestępstwa
   - KY_CD (unikalnych 74, brakujących 0) - Trzycyfrowy kod klasyfikacji wykroczeń
   - OFNS_DESC (unikalnych 72, brakujących 18823) - Opis wykroczenia odpowiadający kodowi klucza
   - PD_CD (unikalnych 433, brakujących 6278) - Trzycyfrowy kod klasyfikacji wewnętrznej (bardziej szczegółowy niż Key Code)
   - PD_DESC (unikalnych 423, brakujących 6278) - Opis wewnętrznej klasyfikacji odpowiadającej kodowi PD (bardziej szczegółowy niż opis przestępstwa)
   - LAW_CAT_CD (unikalnych 3, brakujących 0) - Poziom wykroczenia: przestępstwo, wykroczenie, naruszenie
5. Czy się udało
   - CRM_ATPT_CPTD_CD (unikalnych 3, brakujących 7) - Status, czy przestępstwo zostało dokonane czy była próba jego popełnienia lub czy zostało ono przerwane
6. Otoczenie zdarzenia
   - PREM_TYP_DESC (unikalnych 75, brakujących 40745) - Rodzaj otoczenia; sklep spożywczy, domek jednorodzinny, ulica itp.
   - LOC_OF_OCCUR_DESC (unikalnych 6, brakujących 1543800) - Lokalizacja w stosunku do otoczenia; wewnątrz, naprzeciw, z przodu, z tyłu
7. Lokalizacja zdarzenia
   - ADDR_PCT_CD (unikalnych 78, brakujących 2166) - Posterunek
   - BORO_NM (unikalnych 6, brakujących 11329) - Dzielnica
   - JURIS_DESC (unikalnych 25, brakujących 0) - Opis kodu jurysdykcji
   - JURISDICTION_CODE (unikalnych 26, brakujących 6278) - Kod jurysdykcji na której miało miejsce to zdarzenie
   - PARKS_NM (unikalnych 1206, brakujących 7348330) - Nazwa parku, placu zabaw lub terenów zielonych w Nowym Jorku, jeśli dotyczy (parki stanowe nie są uwzględnione)
   - HADEVELOPT (unikalnych 280, brakujących 7029181) - Nazwa osiedla NYCHA miejsca zdarzenia, jeśli dotyczy
   - HOUSING_PSA (unikalnych 5088, brakujących 6809283) - Kod poziomu rozwoju
   - X_COORD_CD (unikalnych 71344, brakujących 17339) - Współrzędna X dla układu współrzędnych płaszczyzny stanu Nowy Jork, strefa Long Island, NAD 83, jednostki w stopach (FIPS 3104)
   - Y_COORD_CD (unikalnych 73934, brakujących 17339) - Współrzędna Y dla układu współrzędnych płaszczyzny stanu Nowy Jork, strefa Long Island, NAD 83, jednostki w stopach (FIPS 3104)
   - TRANSIT_DISTRICT (unikalnych 13, brakujących 7212494) - Okręg tranzytowy, w którym doszło do wykroczenia.
   - Latitude (unikalnych 205540, brakujących 17339) - Współrzędna szerokości geograficznej środkowego bloku dla globalnego układu współrzędnych, WGS 1984, stopnie dziesiętne (EPSG 4326)
   - Longitude (unikalnych 201312, brakujących 17339) - Współrzędna

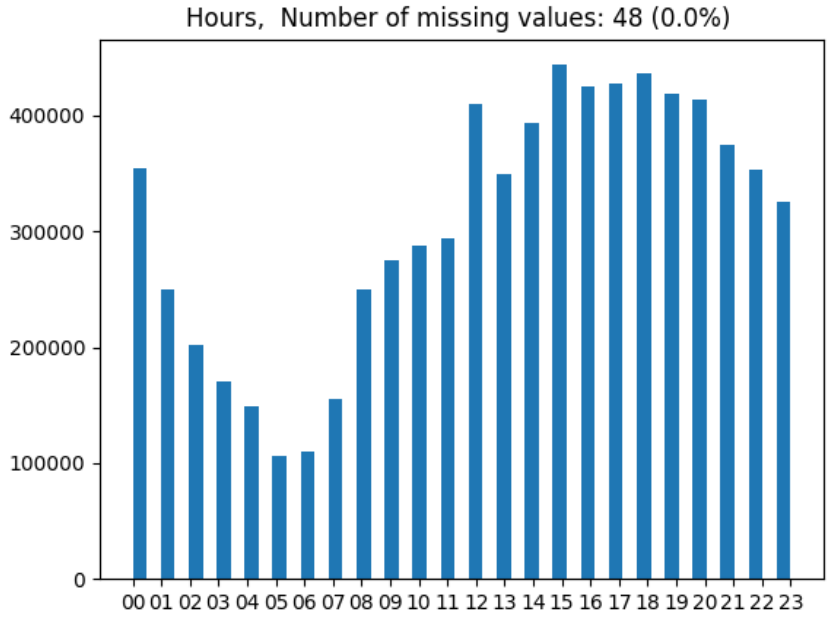
długości bloku środkowego dla globalnego układu współrzędnych, WGS 1984, stopnie dziesiętne (EPSG 4326)
   - Lat_Lon (unikalnych 198553, brakujących 17339) - Punkt lokalizacji geoprzestrzennej (łącznie szerokość i długość geograficzna)
   - PATROL_BORO (unikalnych 9, brakujących 6735) - Nazwa dzielnicy patrolowej, w której doszło do incydentu
   - STATION_NAME (unikalnych 373, brakujących 7212494) - Nazwa stacji tranzytowej
8. Cechy podejrzanego
   - SUSP_AGE_GROUP (unikalnych 112, brakujących 4795235) - Grupa wiekowa podejrzanego
   - SUSP_RACE (unikalnych 9, brakujących 3426694) - Rasa podejrzanego
   - SUSP_SEX (unikalnych 4, brakujących 3560008) - Płeć podejrzanego
9. Cechy ofiary
   - VIC_AGE_GROUP (unikalnych 203, brakujących 1638445) - Grupa wiekowa ofiary
   - VIC_RACE (unikalnych 9, brakujących 309) - Rasa ofiary
   - VIC_SEX (unikalnych 6, brakujących 308) - Płeć ofiary

3.2. Szczegółowa charakterystyka wybranych kolumn

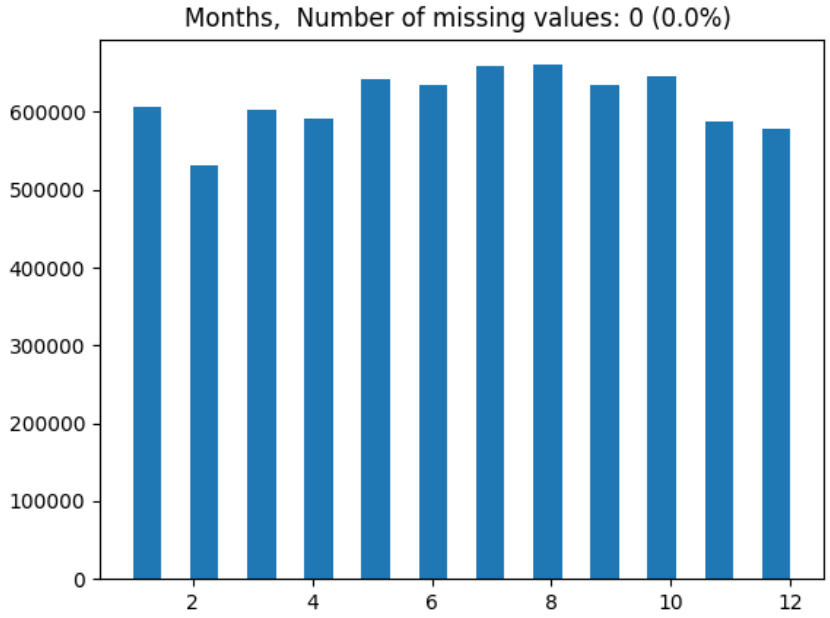
Poniżej przedstawiona została bardziej szczegółowa charakterystyka wybranych kolumn ze zbioru danych, zwłaszcza rozkłady dostępnych w nich wartości (histogramy). Omówione również zostały możliwości wstępnego oczyszczenia i przetwarzania danych, przed właściwą analizą (szczegółowy preprocessing jest omówiony osobno dla każdego celu). Wspomniane zostały także kolumny, które wydają się być nieprzydatne i nie zostały wykorzystane w dalszej analizie.

3.2.1. Data oraz czas zdarzenia i zgłoszenia

Jako, że data i czas zdarzenia co do swojej oryginalnej wartości nie niosą praktycznie żadnej przydatnej informacji, niezbędne jest dla tych kolumn przeprowadzenie prostej ekstrakcji cech. Podstawowymi cechami, które można wyekstrahować z daty i czasu wystąpienia zdarzenia jest pora dnia, dzień tygodnia czy pora roku. Dla rekordów, dla których znany jest czas zakończenia zdarzenia, można wyekstrahować również czas trwania zdarzenia.



Rysunek 1. Histogram liczby zdarzeń w zależności od godziny



Rysunek 2. Histogram liczby zdarzeń w zależności od miesiąca (pory roku)

Na podstawie rysunku 1 można zaobserwować, że największa liczba przestępstw/wykroczeń miała miejsce w godzinach popołudniowych, natomiast najmniejsza w godzinach wczesnoporonnych tzn godzina 5 i 6 rano. Na podstawie rysunku 2 można zaobserwować, że wyniki liczby przestępstw/wykroczeń są do siebie zbliżone natomiast większa liczba była w czasie miesięcy letnich.

Parametry rozkładu czasu trwania zdarzenia to:
— średnia = 248h
— odchylenie standardowe = 4769h (prawie 200 dni)
— minimum = -122h (błąd - należy dodatkowo oczyścić dane)
— maksimum = ponad 10 lat
— mediana = 1380s
— centyl 80 = 11h
— centyl 90 = 40h
Jak widać rozkład ten jest bardzo nierównomierny i zwłaszcza próba estymacji tej wartości może okazać się trudna. Data zgłoszenia zdarzenia na policję nie niesie szczególnie istotnej wartości i nie jest wykorzystywana w dalszej analizie.

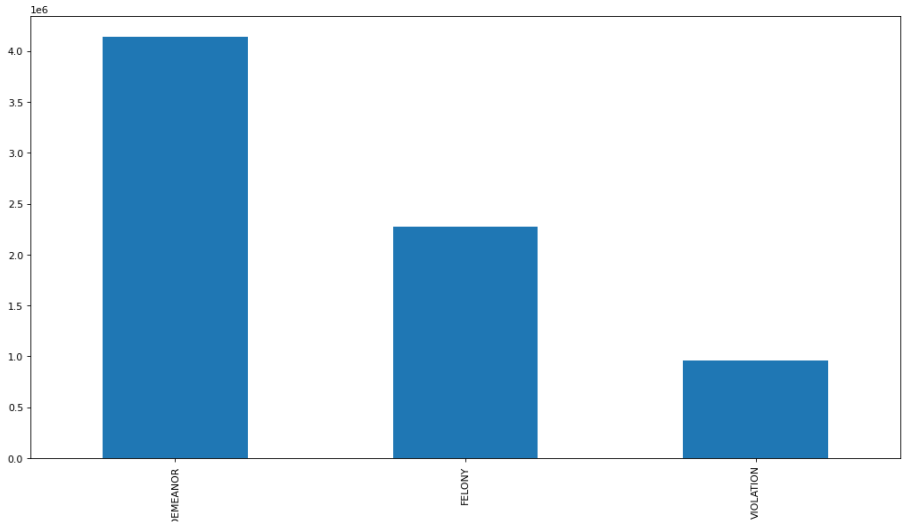
3.2.2. Typ i opis zdarzenia

Na podstawie rysunku 3 można zaobserwować, że największy odsetek stanowiły wykroczenia, na drugim miejscu uplasowały się przestępstwa natomiast najmniej było występków(jest to pośredni czyli między wykroczeniem a przestępstwem). Rysunki 4 oraz 5 pokazują, że zdecydowana większość zdarzeń mieści się w zaledwie kilku pierwszych (z kilkudziesięciu) kodów klasyfikacji zdarzenia.

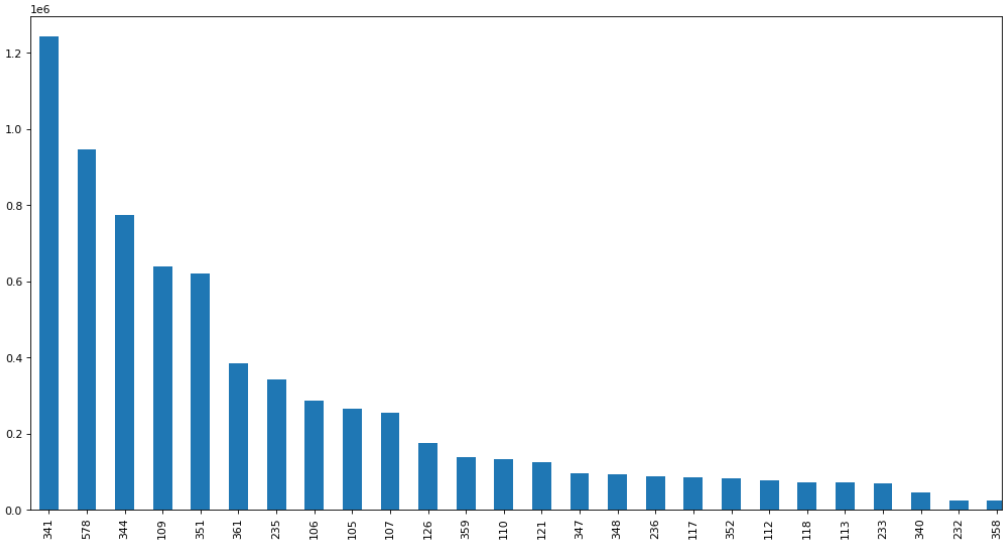
Kolumny zawierające słowne opisy przestępstwa odpowiadają praktycznie zawsze danem kodowi, tak więc na potrzeby działania algorytmu są zupełnie redundantne i nie będą wykorzystywane.

Należy wspomnieć, że występuje bardzo silne powiązanie (reguła asocjacyjna) między kolejno PD_CD -> KY_CD oraz KY_CD -> LAW_CAT_CD. Nie mają one 100% pokrycia tak więc istnieją takie przestępstwa (KY_CD), które występują we wszystkich trzech rodzajach. Jednakże pokrycie wspomnianych zależności jest bardzo duże i nie ma sensu przeprowadzać klasyfikacji LAW_CAT_CD w oparciu o KY_CD.

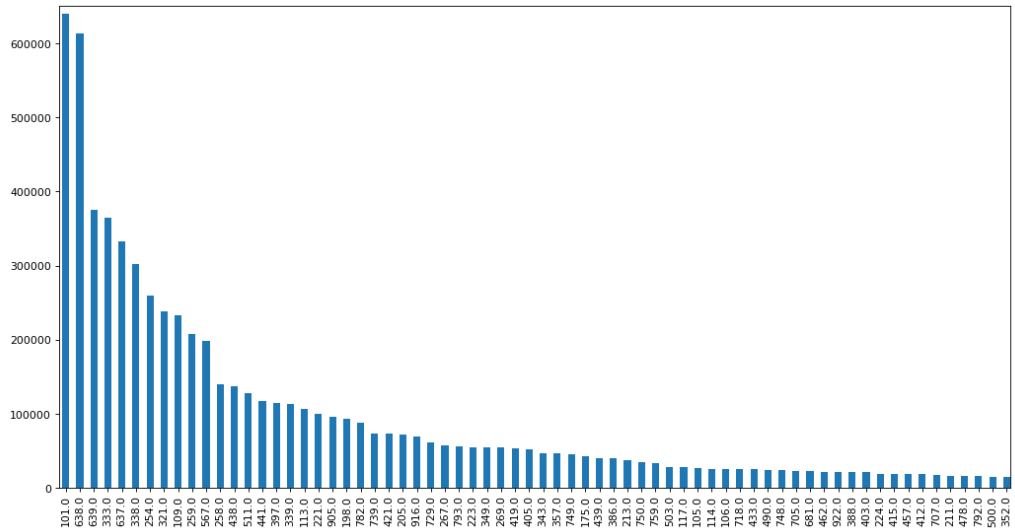
Wspomniane kolumny mają charakter kategorialny dlatego podczas ich wykorzystanie niezbędne będzie zastosowanie techniki *one hot encoding*. Prawdopodobnie nie ma również sensu wykorzystywać wszystkich trzech informacji, zwłaszcza, że jak opisano wyżej, są one zależne jedna od drugiej.



Rysunek 3. Histogram liczby zdarzeń zależnie od jego typu

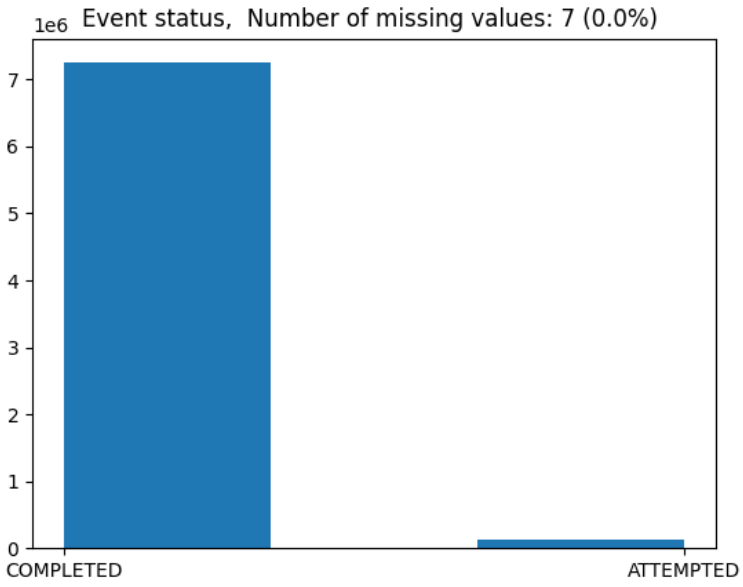


Rysunek 4. Początek histogramu liczby zdarzeń dla kodu zdarzenia (KY_CD)



Rysunek 5. Początek histogramu liczby zdarzeń dla wewnętrznego kodu zdarzenia (PD_CD)

3.2.3. Czy się udało

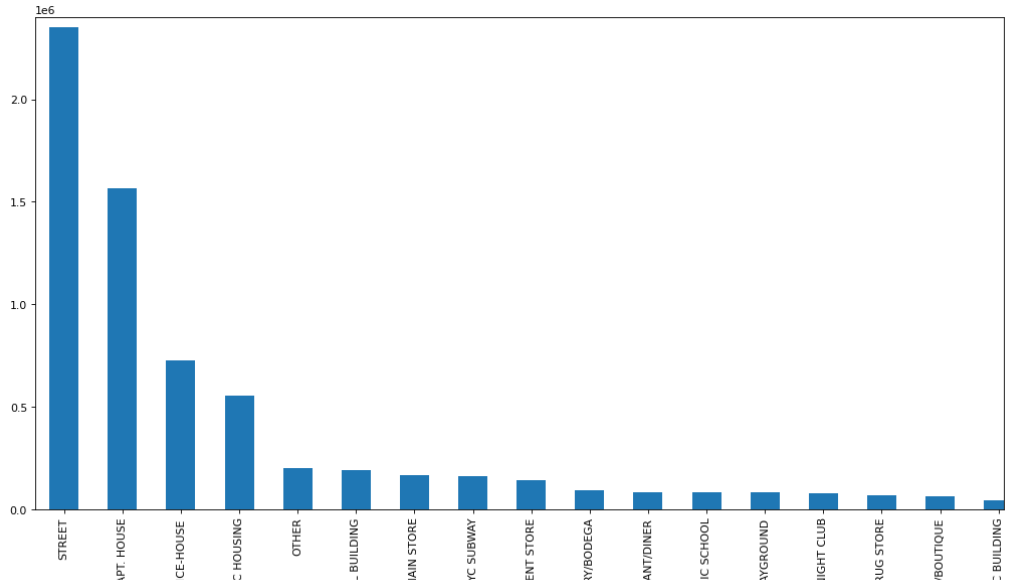


Rysunek 6. Histogram liczby zdarzeń w zależności od tego czy doszło ono do skutku czy zostało zatrzymane

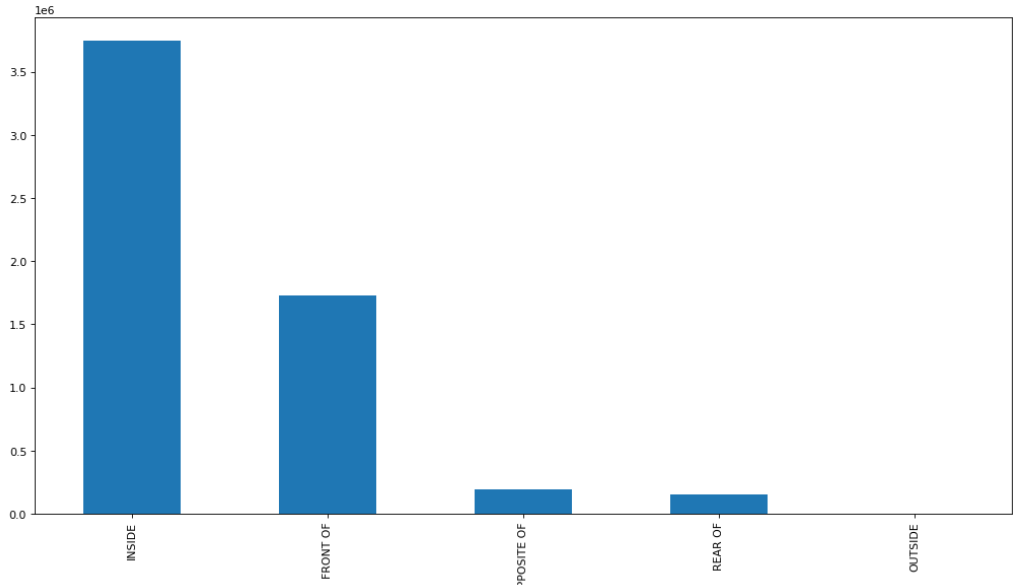
Na podstawie rysunku 6 można zaobserwować, że praktycznie wszystkie przestępstwa/wykroczenia zostały dokonane a naprawdę niewielki odsetek został udaremniony. Rozkład wartości w tej kolumnie jest więc silnie niezbalansowany.

Wartość ta może zostać zakodowana w postaci pojedynczej kolumny z wartościami binarnymi.

3.2.4. Otoczenie zdarzenia



Rysunek 7. Początek histogramu liczby zdarzeń dla otoczenia zdarzenia



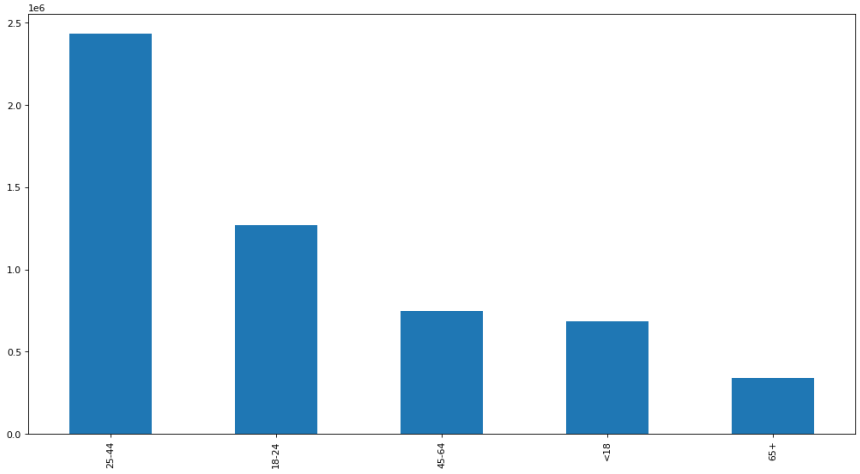
Rysunek 8. Histogram liczby zdarzeń dla lokalizacji w stosunku do otoczenia

Na rysunku 7 widać, że zdecydowana większość zdarzeń ma miejsce na ulicy i w domach. Z kolei rysunek 8 pokazuje, że najwięcej zdarzeń dzieje się wewnątrz jakiegoś miejsca. Jak widać rozkłady tych wartości znowu są silnie niezbalansowane. Typ danych pozostaje kategorialny i wymaga stosowania *one hot encodingu*.

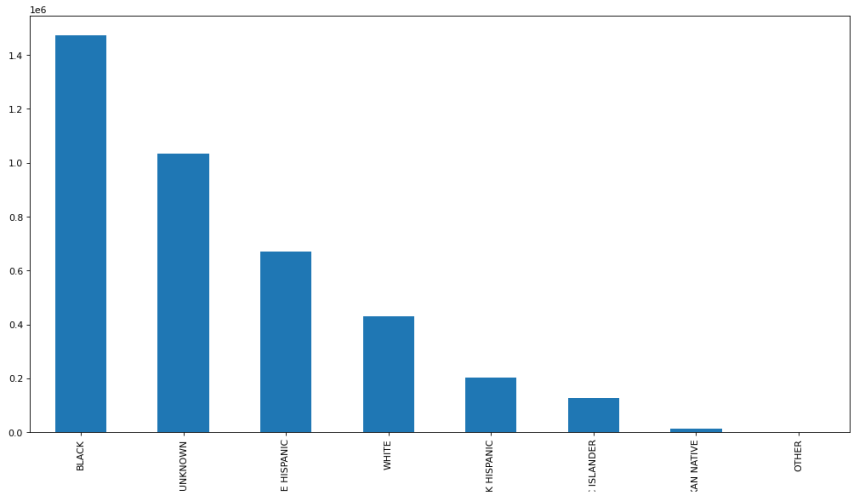
3.2.5. Lokalizacja zdarzenia

Informacje o dokładnej lokalizacji zdarzenia, w przeciwieństwie do wszystkich pozostałych informacji na temat zdarzenia dostępnych w tym zbiorze, nie mają charakteru uniwersalnego. Wykorzystanie ich oznaczałoby uzależnienie się od kontekstu miasta Noweg Jorku, z którego dane pochodzą. Dodatkowo dane te są redundantne, i w większości nie niosą żadnych przydatnych informacji dotyczących czy to rodzaju, czy też chwili popełnionego przestępstwa. Żadna z kolumn z tej grupy nie została wykorzystana w dalszej analizie.

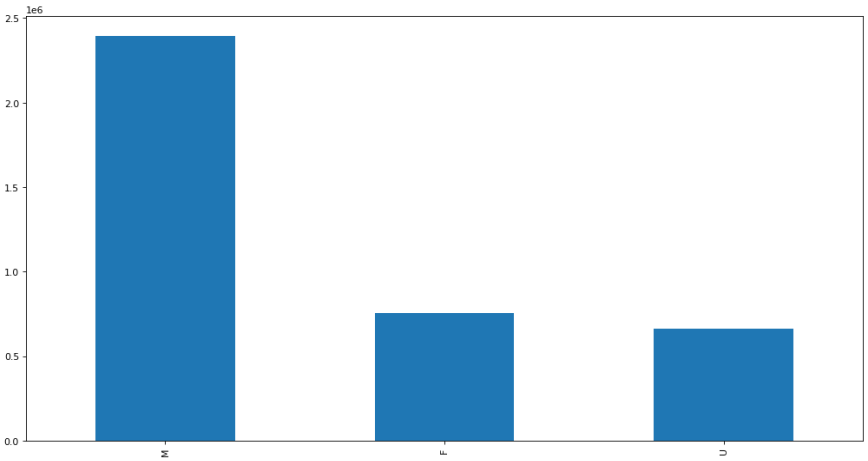
3.2.6. Cechy podejrzanego



Rysunek 9. Histogram liczby zdarzeń popełnianych przez określoną grupę wiekową (po wstępnym oczyszczeniu zbioru).



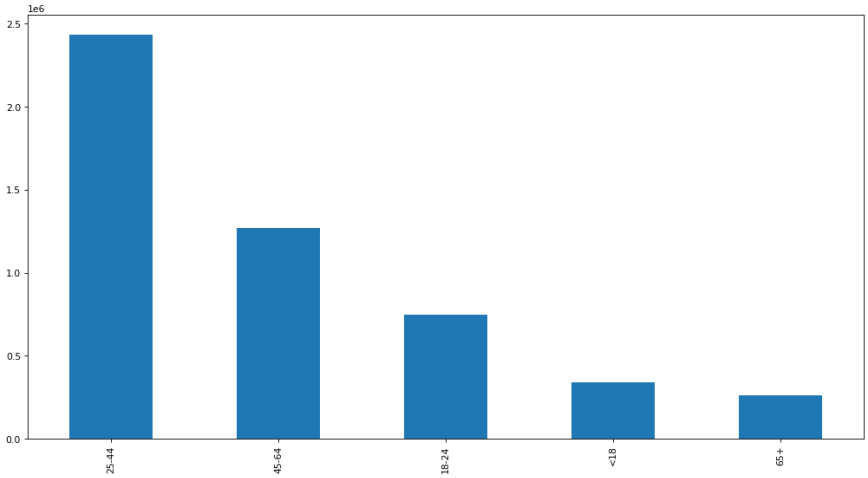
Rysunek 10. Histogram liczby zdarzeń popełnianych przez określoną rasę.



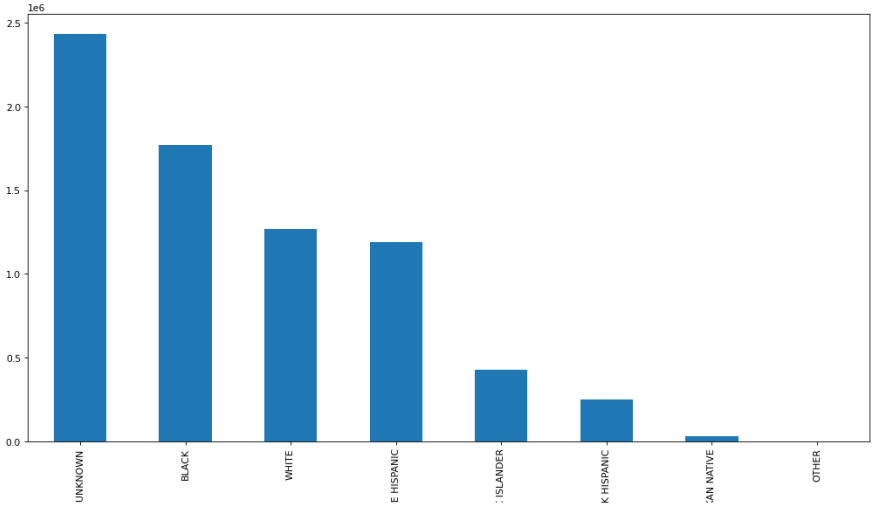
Rysunek 11. Histogram liczby zdarzeń popełnianych przez określoną płeć.

Na cechy podejrzanego składają się trzy wartości: grupa wiekowa, rasa oraz płeć. Grupa wiekowa jest silnie zanieczyszczona różnymi losowymi wartościami liczbowymi. Najprostszą metodą, stosowaną w większości przypadków, będzie ograniczenie liczby wartości do tych poprawnych (<18, 18-24, 25-44, 45-64, 65+) oraz wartości nieznanej (NaN). Rysunek 9 przedstawia rozkład wieku podejrzanych po takim oczyszczeniu tych danych. Wynika z niego, że większość ludzi uwikłanych w różnego rodzaju przestępstwa i występki mają między 20 a 40 lat. Jest to również najliczniejsza grupa w społeczeństwie, stąd też przede wszystkim wynika ten rozkład (rozkład normalny). Kolejnym spostrzeżeniem, tym razem z rysunków 10 i 11 jest fakt, że większość podejrzanych stanowią czarni mężczyźni.

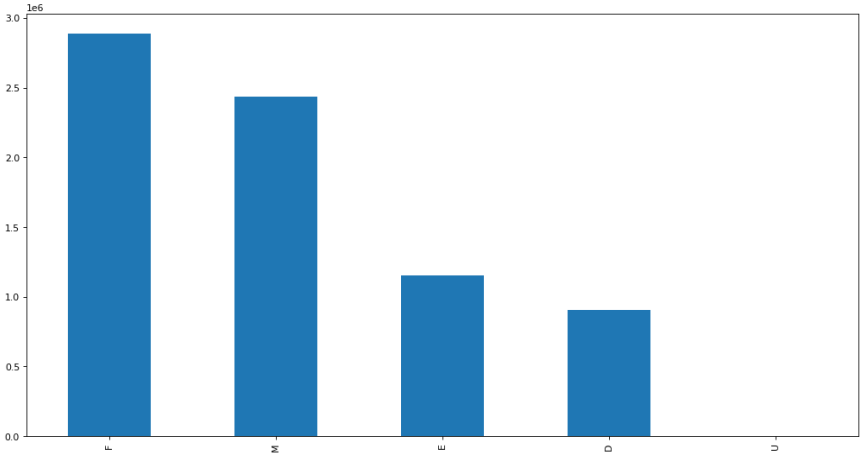
3.2.7. Cechy ofiary



Rysunek 12. Histogram liczby zdarzeń dla danych grup wiekowych poszkodowanych (po wstępnym oczyszczeniu zbioru).



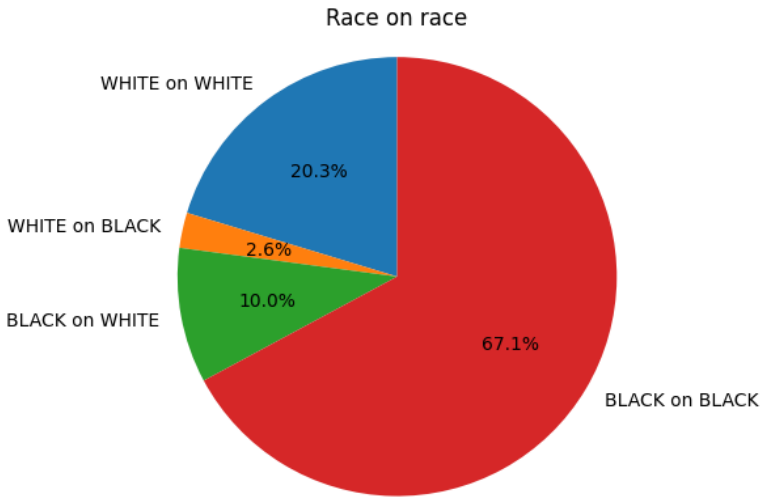
Rysunek 13. Histogram liczby zdarzeń dla danych ras poszkodowanych.



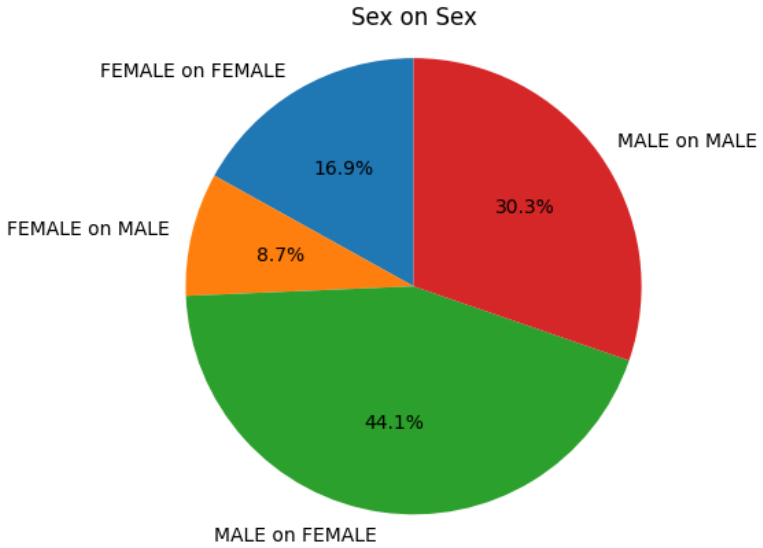
Rysunek 14. Histogram liczby zdarzeń dla danych płci poszkodowanych.

W przypadku cech ofiary/poszkodowanego sytuacja z grupami wiekowymi jest taka sama jak w przypadku cech podejrzanego. Rysunek 12 pokazuje rozkład wieku ofiar po wstępnym oczyszczeniu zbioru. Również tutaj widać wyraźnie rozkład normalny. Rysunku 13 oraz 14 pokazują, że wśród poszkodowanych najwięcej jest kobiet nieznanej rasy. Mamy tutaj do czynienia z dodatkowymi wartościami w kolumnie płci, które oznaczają grupy ludzi: instytucję lub państwo. Z powodu, że taki może być charakter poszkodowanych większość zdarzeń nie ma określonej płci poszkodowanego.

3.2.8. Zależności między podejrzanym a ofiarą



Rysunek 15. Wykres przedstawiający zależność pomiędzy podejrzanym a ofiarą, na podstawie ich rasy (uwzględnione zostały tylko dwie).



Rysunek 16. Wykres przedstawiający zależność pomiędzy podejrzanym a ofiarą, na podstawie ich płci (uwzględnione zostały tylko dwie).

Do wykresów 15 i 16 zostały wybrane dwie rasy i płcie, które największą liczbę razy uczestniczyły w przestępstwie jako ofiara lub podejrzany. W ten sposób się zaobserwować, że największy odsetek przestępstw dotyczył tej samej rasy, można zatem stwierdzić, że odsetek przestępstw między rasowych jest niższy niż można było się spodziewać. Natomiast sprawdzenie rozkładu płci, nie zaskutkowało niespodziewanymi wnioskami. Mężczyźni ogólnie częściej są podejrzanymi o przestępstwa i częściej popełniają takie, w których ofiarami są kobiety niż mężczyźni.

4. Ogólny schemat podjętego przetwarzania i analizy danych

Ogólny schemat przetwarzania i analizy danych podjętych w tym zadaniu jest następujący:

1. Pozyskanie danych - pobranie na dysk ze źródła internetowego
2. Przegląd danych i wstępna analiza statystyczna w celu rozeznania się w zawartości zbioru
3. Rozważenie możliwych metod preprocessingu (czyszczenia danych, imputacji, kodowania kolumn, etc.)
4. Dla każdego z wyznaczonych celów:
   a) Wstępne przetwarzanie danych (zależy od zastosowanych algorytmów i od realizowanego zadania)
   b) Pierwszy eksperyment: uczenie modelu z wybranymi hiperparametrami
   c) Dalsze eksperymenty: modyfikacja hiperparametrów i preprocessingu w zależności od uzyskanych wyników...
   d) Wskazanie najlepszych wyników i wnioski z przeprowadzonych badań

5. Cele projektu

W ramach projektu sformułowane zostały trzy następujące cele.

5.1. Klasyfikacja rodzaju lub poziomu przestępstwa

5.1.1. Opis

W ramach tego etapu przeprowadzony został szereg eksperymentów mających na celu stworzenie klasyfikatora typu przestępstwa (KY_CD). W ramach porównania dokonano klasyfikacji poziomu wykroczenia, stanowiącego bardziej ogólną informacje. Podczas realizacji tego celu, pod uwagę zostały wzięte kolumny, przechowujące następujące informacje:
— godzina zdarzenia
— dzień tygodnia zdarzenia
— odstęp między zgłoszeniem a zdarzeniem
— czy doszło do skutku (CRM_ATPT_CPTD_CD)
— otoczenie zdarzenia
— cechy podejrzanego
— cechy ofiary
— poziom lub typ przestępstwa w zależności od tego co klasyfikujemy

Wybrane cechy uległy drobnym modyfikacjom w trakcie trwania eksperymentów. Z powodu małej liczby danych z dostępnymi informacjami o czasie trwania przestępstwa, pomysł wykorzystania tej danej został zarzucony.

Informacja o dokładnej lokalizacji zdarzenia nie jest wykorzystana ze narzędzia względu na chęć stworzenia uniwersalnego narzędzia.

Aby zrealizować zaproponowany cel wykorzystane zostały następujące metody: metody imputacji brakujących danych, prosta ekstrakcja cech (zwłaszcza z daty), naiwny klasyfikator Bayesa (ze względu na uzyskiwane wyniki w czasie trwania projektu, został porzucony) i lasy losowe (ze względu na uzyskiwane wyniki w czasie trwania projektu, metoda ta została potraktowana priorytetowo).

5.1.2. Przygotowanie danych

Bazując na założeniach celu oraz wiedzy, wyciągniętej z poprzednich checkpoint'ów, pod uwagę w czasie eksperymentów zostały brane pod uwagę następujące kolumny:
— KY_CD - kod popełnionego przestępstwa;
— LAW_CAT_CD - poziom popełnionego przestępstwa;
— CMPLNT_FR_DT - data zgłoszenia przestępstwa;
— LOC_OF_OCCUR_DESC - opis otoczenia zdarzenia;
— SUSP_AGE_GROUP - Grupa wiekowa podejrzanego
— SUSP_RACE - Rasa podejrzanego
— SUSP_SEX - Płeć podejrzanego
— VIC_AGE_GROUP - Grupa wiekowa ofiary
— VIC_RACE - Rasa ofiary
— VIC_SEX - Płeć ofiary

Powodem odrzucenia kolumny CRM_ATPT_CPTD_CD było ogromne niezbalansowanie zawartych w niej danych. Dane poddano obróbce (preproces-

sing'owi), na którą składały się operacje różne w zależności od specyfiki przeprowadzanego eksperymentu. W celu redukcji wymiarów oraz ograniczenia liczby potencjalnych wartości, a co za tym idzie zmniejszenie liczby danych etykiet do kodowania (przetwarzanie dużego zbioru zakodowanych danych jest bardzo wymagającym wydajnościowo procesem, przy małym wpływie na jakość potencjalnych wyników) przeprowadzono następujące operacje grupowania:

— grup wiekowych, tak aby można było nadać im odpowiednią etykietę ("<18", "18-24", "25-44", "45-64", "65+", "UNKNOWN") oraz jej ewentualna imputacja najczęściej występującą wartością;
— rasy (podejrzanego i ofiary), tak aby można było nadać im odpowiednią etykietę (zależną od eksperymentu) - ["WHITE", "BLACK", "HISPANIC", "UNKNOWN", "OTHER"], lub ["WHITE", "BLACK", "HISPANIC", "UNKNOWN", "OTHER"]
— (podejrzanego i ofiary), tak aby można było nadać im odpowiednią etykietę - ["MALE", "FEMALE", "OTHER"];
— pozycji otoczenia zdarzenia, tak aby można było nadać im odpowiednią etykietę - ["UNKNOWN", "FRONT OF", "INSIDE", "OTHER"];

Po dokonaniu grupowania, ze zbioru zostały usunięte te wiersze, dla których wykorzystywane w eksperymentach wartości były puste. Następnie (w przypadku eksperymentów, w których było to badane) obróbce poddana została informacja o dacie zgłoszenia przestępstwa. W pierwszym kroku wyekstrahowano z tej kolumny informacje o dniu tygodnia i dniu roku. Następnie, aby nie stracić potencjalnie przydatnej informacji, jaką niesie cykliczność dni tygodnia, czy też roku, obie te wartości zostały zapisane przy pomocy odpowiadających sobie wartości funkcji trygonometrycznych sinus i cosinus. Ostatnim krokiem przygotowania danych było odpowiednie zakodowanie danych kolumn, w zależności od tego jaki typ danych prezentowały.

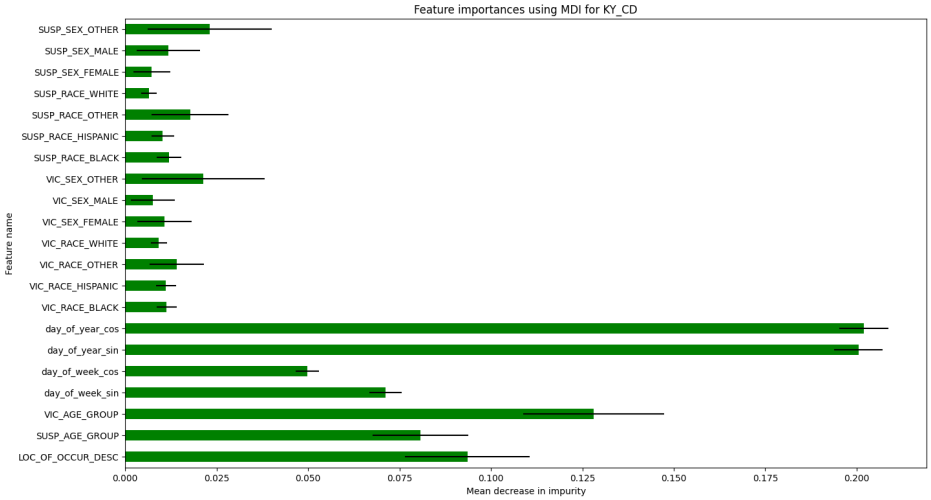
5.1.3. Przetwarzanie i analiza danych

Na wstępie przeprowadzono serię eksperymentów, w trakcie których określano najoptymalniejszy dobór cech do klasyfikacji w oparciu o dokładność klasyfikacji oraz wskazania ważności cech (ang. *feature importance*). W tym celu wykorzystano dwie przeciwstawne metody, dzięki czemu można było uzyskać pełniejszy obraz (zmniejszone zostało prawdopodobieństwo podatności któreś metody, na np. dane kategorialne), wykorzystano:

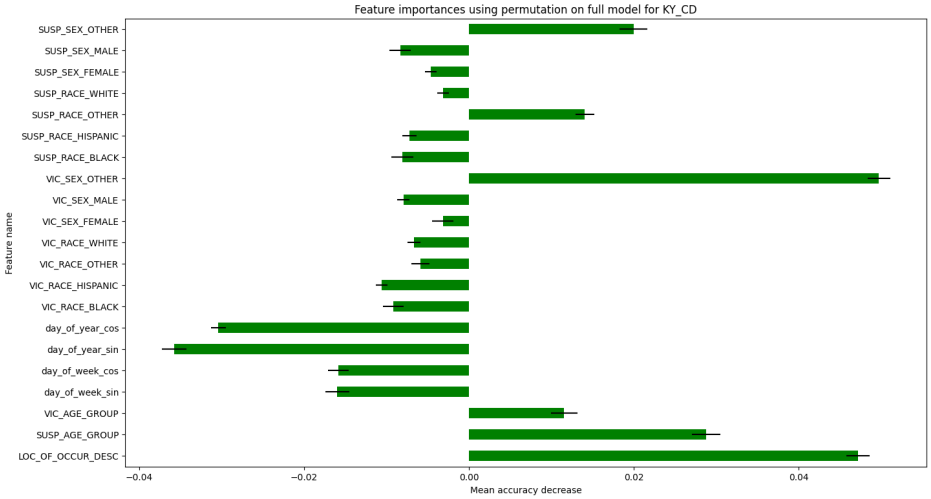
— metodę opartą o *mean decrease in impurity*;
— metodę opartą o *feature permutation*

Eksperymenty, mające na celu wyodrębnienie najważniejszych cel, zostały przeprowadzone na wycinku zbioru danych - dla 100000 wierszy danych. W tych eksperymentach były generowane pary wykresów (rezultaty działania dwóch wspomnianych metod badania ważności cech) dla klasyfikacji kodu przestępstwa (*key-code*) oraz poziomu wykroczenia (*law-breaking-level*), natomiast zamieszczano je w komplecie, tylko w sytuacji gdy różniły się od siebie w sposób znaczący i warty analizy, w przeciwnym razie zostały one w sprawozdaniu pominięte. Dodatkowo zaprezentowane zostaną tabelę z informacją o dokładności klasyfikacji, wraz z porównaniem do ostatnio przeprowadzonego eksperymentu.

Eksperyment nr 1 W tym eksperymencie pogrupowano rasy pogrupowano zgodnie z kategoriami ["WHITE", "BLACK", "HISPANIC", "UNKNOWN", "OTHER"] oraz wykorzystano dzień tygodnia i dzień roku (w postaci par wartości funkcji sinus i cosinus).



Rysunek 17. Feature importances (MDI), w klasyfikacji key-code



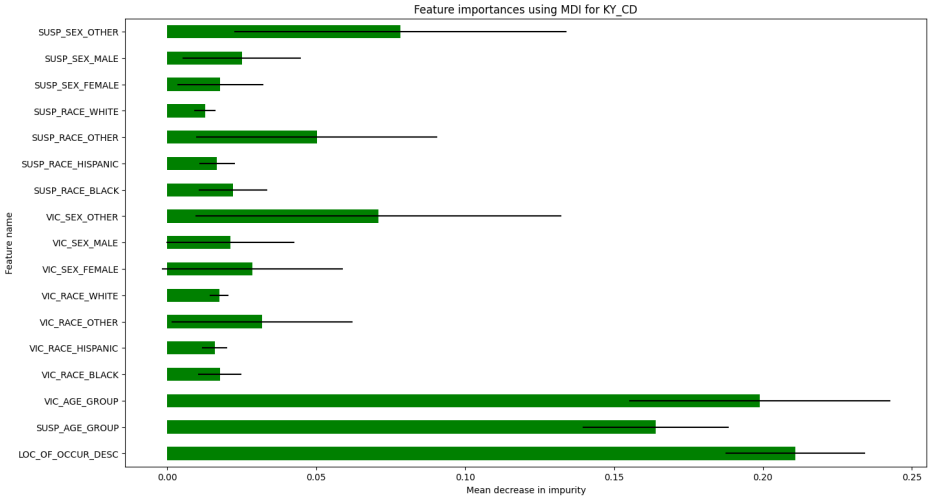
Rysunek 18. Feature importances (permutation), w klasyfikacji key-code

Nazwa metryki	Wartość	Zmiana wartości	Klasyfikacja
Accuracy	29,605%	0	key-code
Accuracy	50,9%	0	law-breaking-level

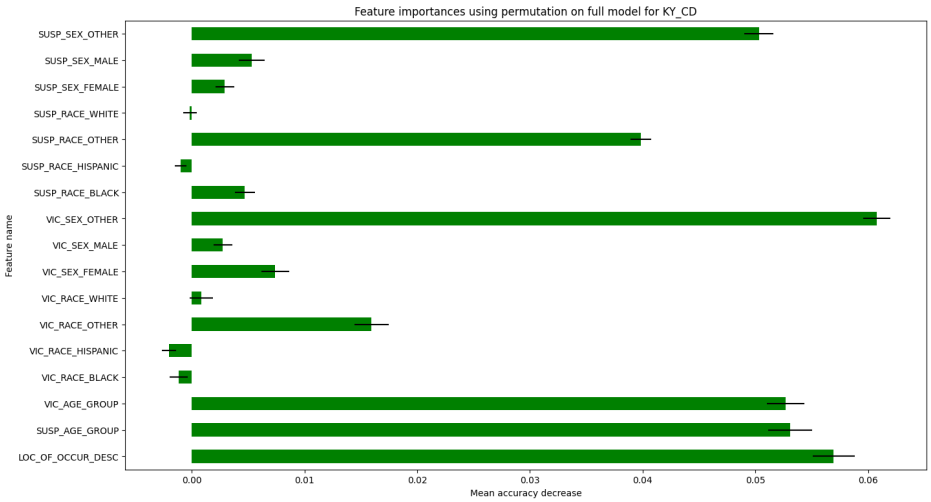
Tabela 1. Wyniki dla eksperymentu nr 1

Warto zwrócić uwagę na dużą rozbieżność dotyczącą przydatności cech według algorytmów (wykresy 17 i 18), w kontekście daty zgłoszenia zdarzenia, zostało to uwzględnione w kolejnym eksperymencie.

Eksperyment nr 2 W tym eksperymencie pogrupowano rasy pogrupowano zgodnie z kategoriami ["WHITE", "BLACK", "HISPANIC", "UNKNOWN", "OTHER"], natomiast nie wykorzystano dnia tygodnia i dnia roku.



Rysunek 19. Feature importances (MDI), w klasyfikacji key-code



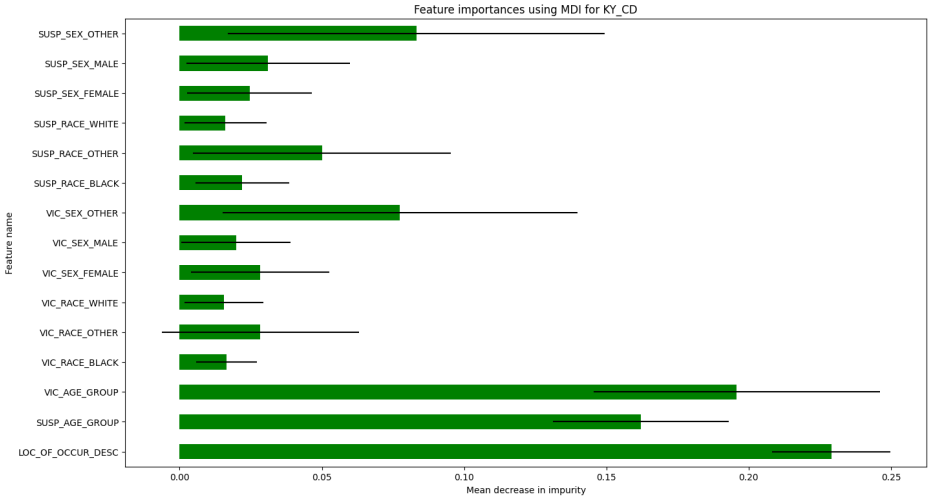
Rysunek 20. Feature importances (permutation), w klasyfikacji key-code

Nazwa metryki	Wartość	Zmiana wartości	Klasyfikacja
Accuracy	34,815%	+5,21	key-code
Accuracy	54,0%	+3,1	law-breaking-level

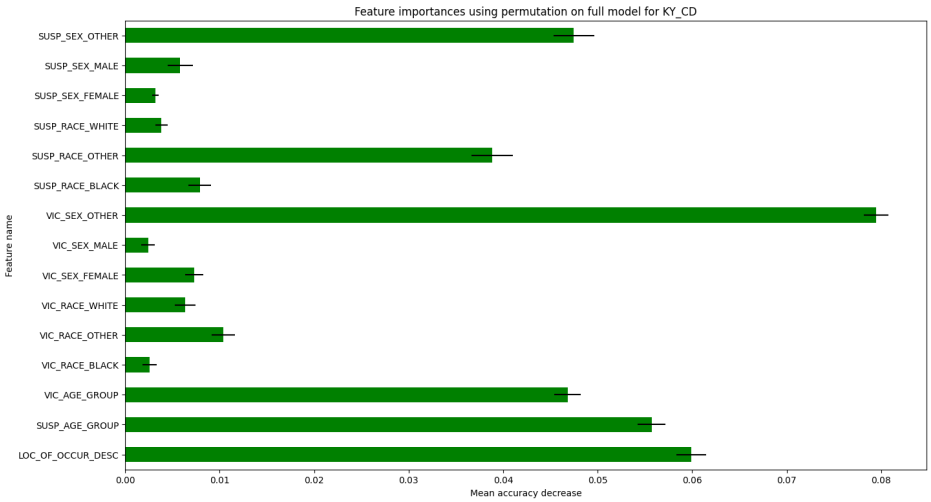
Tabela 2. Wyniki dla eksperymentu nr 2

Na podstawie wzrostów dokładności klasyfikacji dla obu cech, widocznych w tabeli 2, decyzję o nie wykorzystaniu daty należy uznać za korzystną. Po analizie danych na wykresach 19 i 20 zmienione zostaną etykiety ras, tak aby nie wykorzystywać jednej z obecnych - "HISPANIC".

Eksperyment nr 3 W tym eksperymencie pogrupowano rasy pogrupowano zgodnie z kategoriami ["WHITE", "BLACK", "UNKNOWN", "OTHER"].



Rysunek 21. Feature importances (MDI), w klasyfikacji key-code



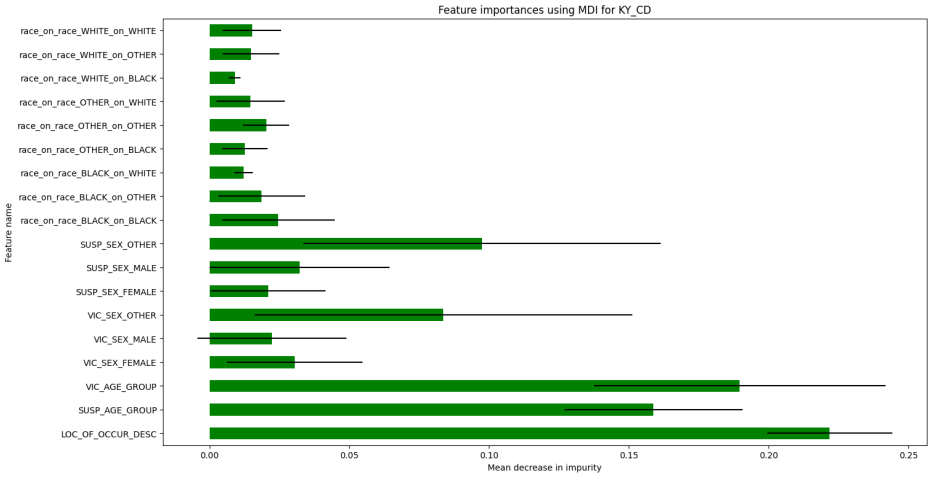
Rysunek 22. Feature importances (permutation), w klasyfikacji key-code

Nazwa metryki	Wartość	Zmiana wartości	Klasyfikacja
Accuracy	35,275%	+0,46	key-code
Accuracy	54,245%	+0,245	law-breaking-level

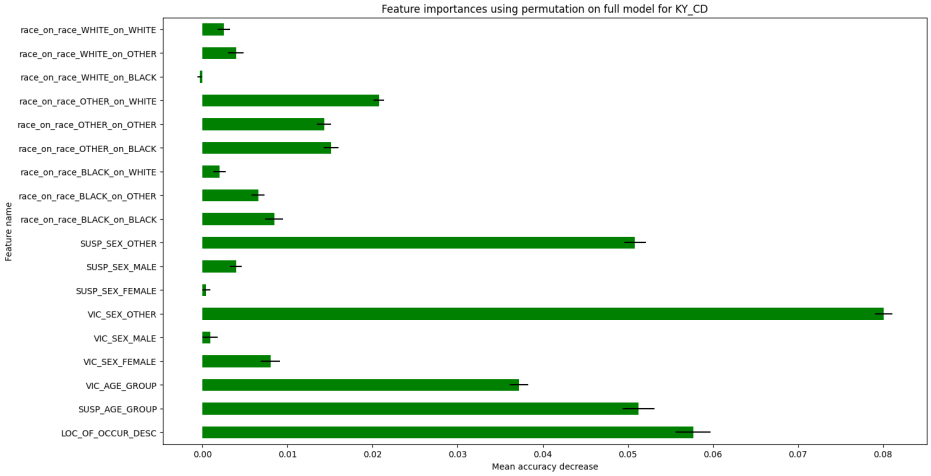
Tabela 3. Wyniki dla eksperymentu nr 3

Na podstawie wzrostów dokładności klasyfikacji dla obu cech, widocznych w tabeli 3, decyzję o zmianie sposobu grupowania ras należy uznać za korzystną.

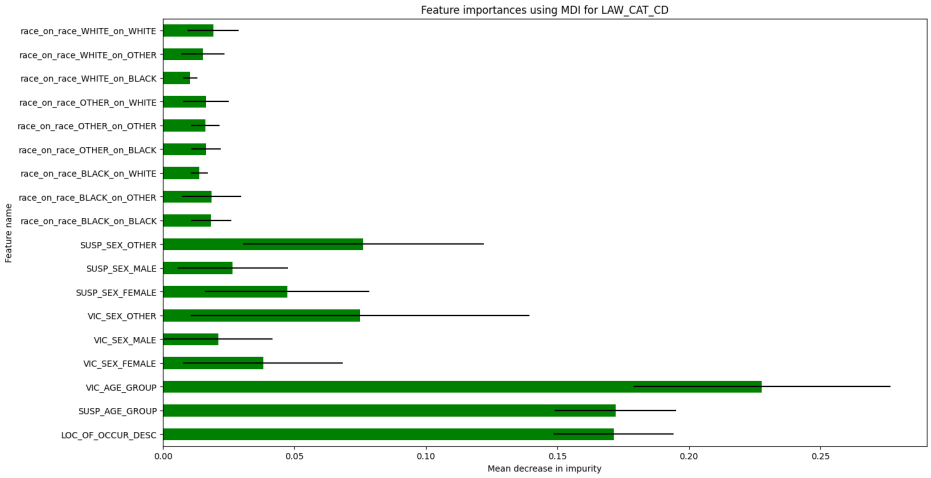
Eksperyment nr 4 W tym eksperymencie, w oparciu o poprzednie eksperymenty zmieniono podejście do przechowywania informacji o rasie podejrzanego i ofiary. Postanowiono wykorzystać zależność pomiędzy ofiarą a podejrzanym, zaprezentowaną na wykresie 15.



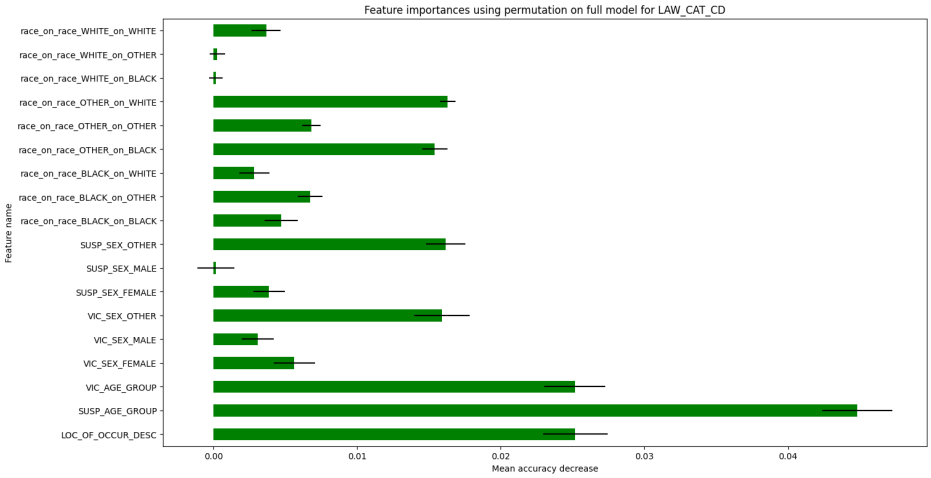
Rysunek 23. Feature importances (MDI), w klasyfikacji key-code



Rysunek 24. Feature importances (permutation), w klasyfikacji key-code



Rysunek 25. Feature importances (MDI), w klasyfikacji law-breaking-level



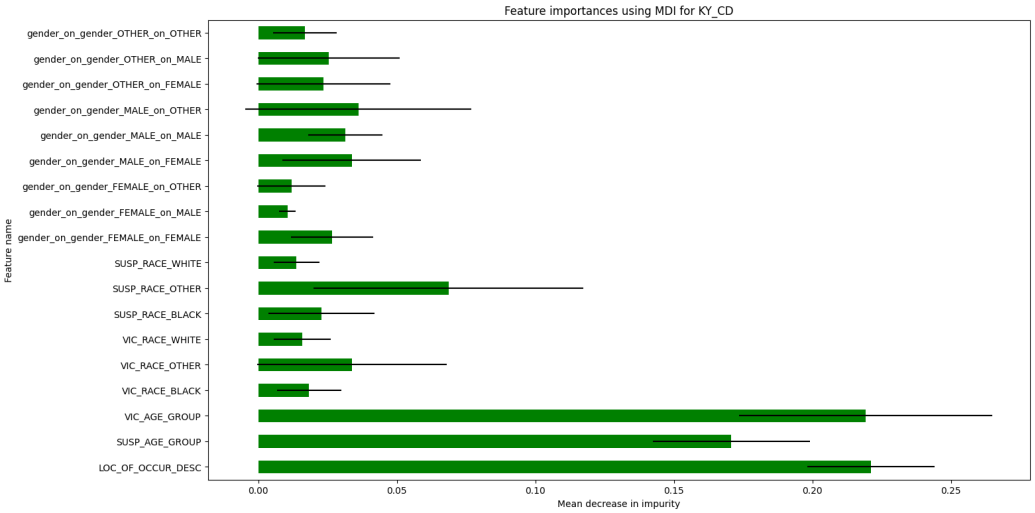
Rysunek 26. Feature importances (permutation), w klasyfikacji law-breaking-level

Nazwa metryki	Wartość	Zmiana wartości	Klasyfikacja
Accuracy	35,0%	-0,275	key-code
Accuracy	53,71%	-0,535	law-breaking-level

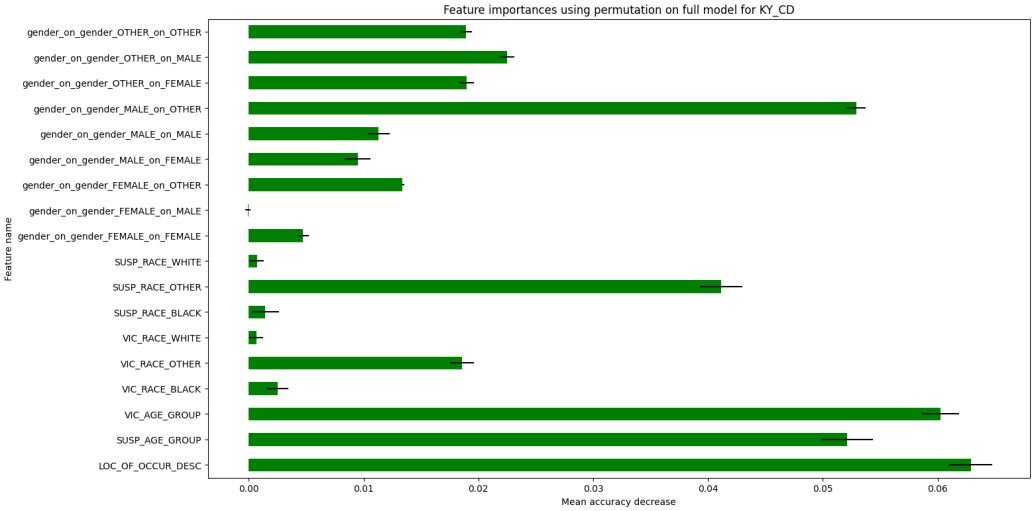
Tabela 4. Wyniki dla eksperymentu nr 4

Na podstawie spadków dokładności klasyfikacji dla obu cech, widocznych w tabeli 4, decyzję o zmianie sposobu grupowania ras należy uznać za niekorzystną. Jednak wykresy 23, 25, 24, 26, skłaniają do pomysłu aby podobnej modyfikacji poddać płeć.

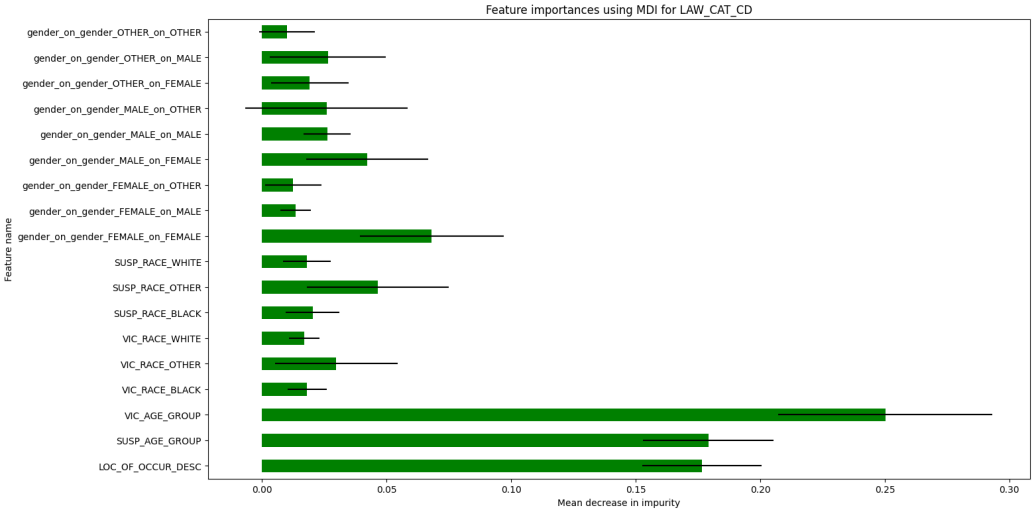
Eksperyment nr 5 W tym eksperymencie, postanowiono nie modyfikować sposobu przechowywanie informacji o rasie podejrzanego i ofiary. Wykorzystano natomiast zależność pomiędzy ofiarą a podejrzanym, zaprezentowaną na wykresie 16 i zmieniomo sposób przechowywania płci.



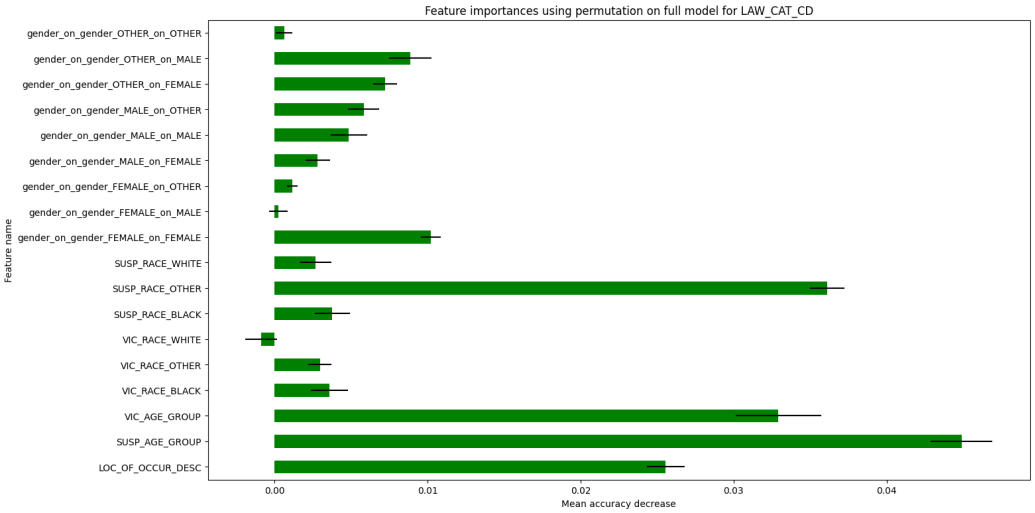
Rysunek 27. Feature importances (MDI), w klasyfikacji key-code



Rysunek 28. Feature importances (permutation), w klasyfikacji key-code



Rysunek 29. Feature importances (MDI), w klasyfikacji law-breaking-level

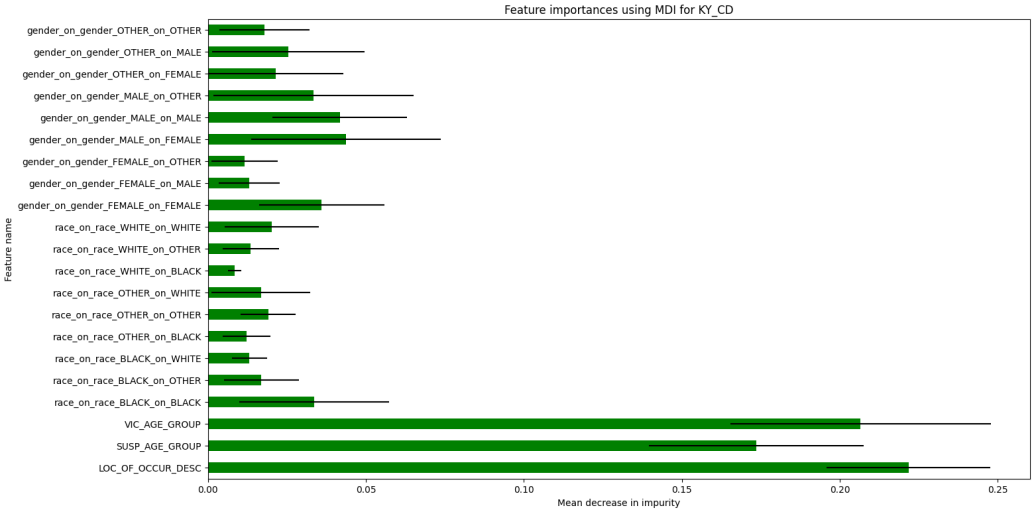


Rysunek 30. Feature importances (permutation), w klasyfikacji law-breaking-level

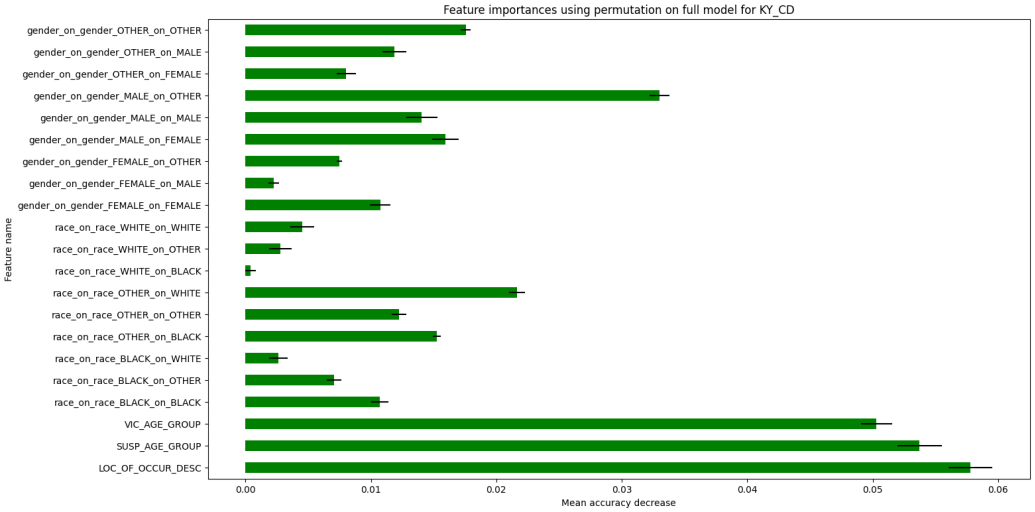
Nazwa metryki	Wartość	Zmiana wartości	Klasyfikacja
Accuracy	35,325%	+0,325	key-code
Accuracy	53,575%	-0,135	law-breaking-level

Tabela 5. Wyniki dla eksperymentu nr 5

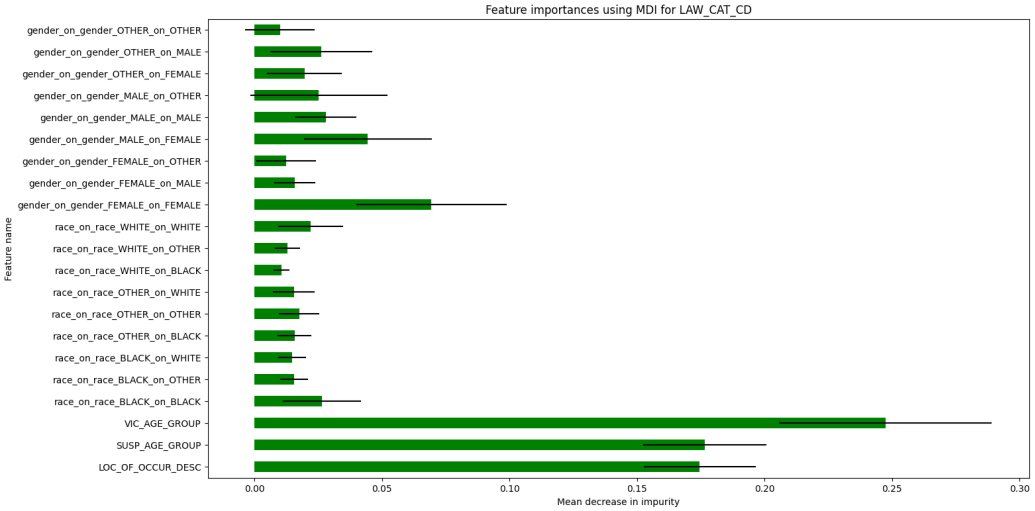
Eksperyment nr 6 W tym eksperymencie, postanowiono sprawdzić zachowanie klasyfikacji przy zmienionym sposobie przechowywania informacji zarówno o rasie jak i o płci.



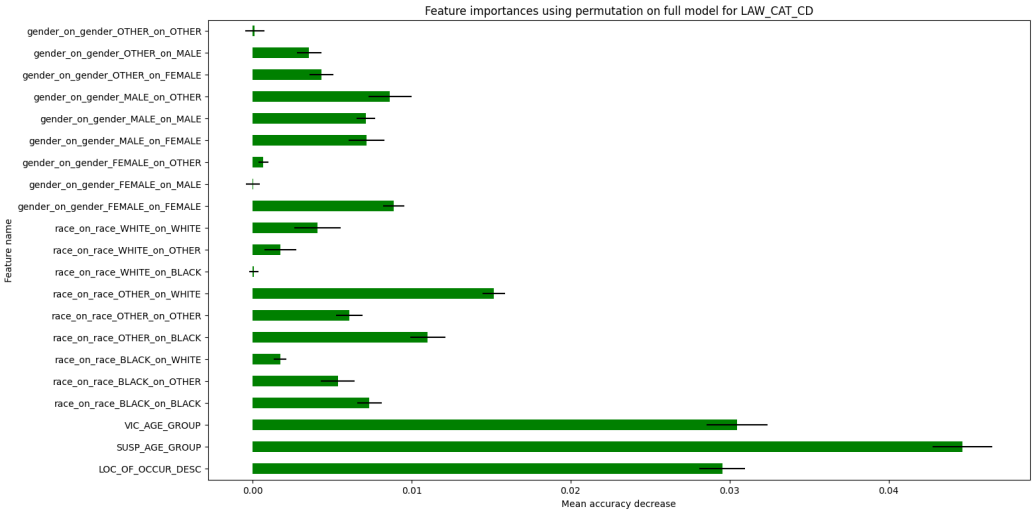
Rysunek 31. Feature importances (MDI), w klasyfikacji key-code



Rysunek 32. Feature importances (permutation), w klasyfikacji key-code



Rysunek 33. Feature importances (MDI), w klasyfikacji law-breaking-level



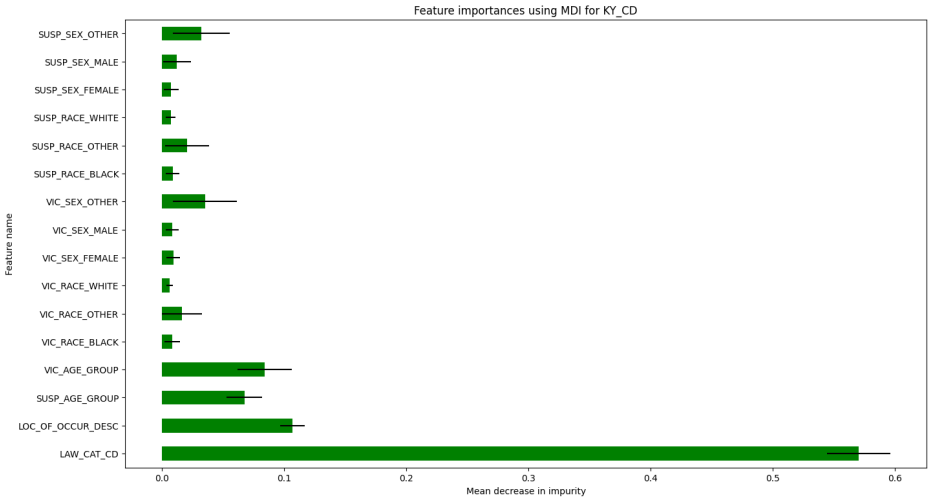
Rysunek 34. Feature importances (permutation), w klasyfikacji law-breaking-level

Nazwa metryki	Wartość	Zmiana wartości	Klasyfikacja
Accuracy	35,11%	-0,215	key-code
Accuracy	54,465%	+0,89	law-breaking-level

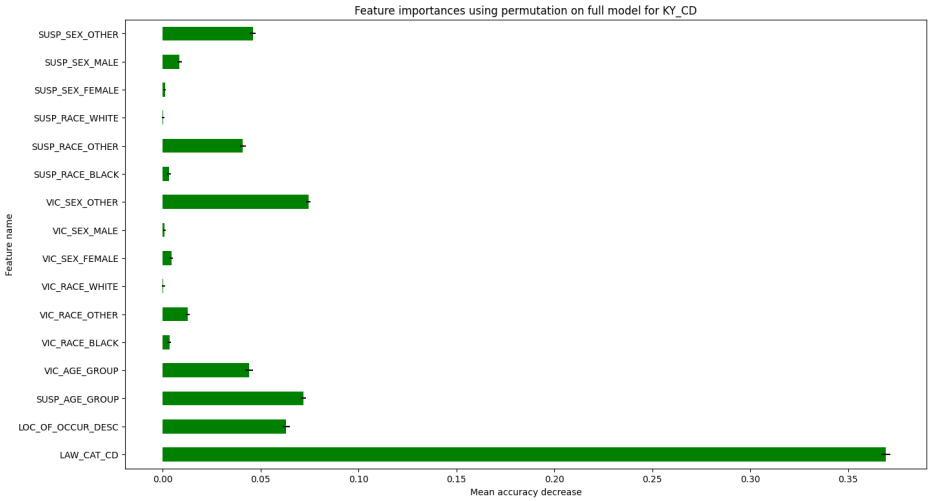
Tabela 6. Wyniki dla eksperymentu nr 6

Na podstawie spadków dokładności klasyfikacji dla obu cech, widocznych w tabelach 4, 5 i 6, biorąc pod uwagę wpływ na wydajność rozwiązania, jaki ma tak duże rozwinięcie zbioru kolumn, pomysł przedstawiony w eksperymentach 4, 5 i 6 nie będzie dalej badany.

Eksperyment nr 7 W tym eksperymencie, postanowiono sprawdzić zachowanie klasyfikacji key-code przy dodaniu do zbioru cech, w oparciu o które jest dokonywana klasyfikacja kolumny law-breaking-level



Rysunek 35. Feature importances (MDI), w klasyfikacji key-code



Rysunek 36. Feature importances (permutation), w klasyfikacji key-code

Nazwa metryki	Wartość	Zmiana wartości	Klasyfikacja
Accuracy	59,39%	+24,28	key-code

Tabela 7. Wyniki dla eksperymentu nr 7

Na podstawie wzrostu dokładności klasyfikacji, widocznej w tabeli 7 oraz wykresów 35 i 36 pomysł aby dodać dodatkową cechę należy uznać za bardzo korzystny.

Eksperyment końcowy W oparciu o przeprowadzone eksperymenty wyodrębniono zbiór cech, dla których klasyfikacja zostanie sprawdzona na całym zbiorze danych, wraz z doborem odpowiednich parametrów. Postanowiono skonfrontować to z wynikami na pełnym zbiorze, niepoddanym imputacji grup wiekowych. Jako parametry inne niż domyślnie dostarczane przez algorytm wybrano te uzyskane we wcześniejszej fazie projektu, tj. podczas eksperymentów z checkpointu nr 3, zostały one przedstawione w tabeli 8. Wspomniany zbiór cech, wykorzystany w końcowej klasyfikacji składał się z:
— LAW_CAT_CD - poziom popełnionego przestępstwa - tylko przy klasyfikacji key-code;
— LOC_OF_OCCUR_DESC - opis otoczenia zdarzenia;
— SUSP_AGE_GROUP - Grupa wiekowa podejrzanego
— SUSP_RACE - Rasa podejrzanego
— SUSP_SEX - Płeć podejrzanego
— VIC_AGE_GROUP - Grupa wiekowa ofiary
— VIC_RACE - Rasa ofiary
— VIC_SEX - Płeć ofiary
Sposób grupowania tych cech został opisany w odpowiedniej podsekcji, warto wspomnieć, że rasy zostały pogrupowane zgodnie ze zbiorem ["WHITE", "BLACK", "HISPANIC", "UNKNOWN", "OTHER"].

Dodatkowo dla scenariuszy z największą dokładnością klasyfikacji, spośród tych wykorzystujących niedomyślne parametry, tj. scenariusz nr 6 i 8, przygotowano macierze pomyłek, zamieszczone w tabelach 10 i 11. Przygotowano je tylko dla klasyfikacji poziomu wykroczenia, ponieważ tutaj klasyfikowana jest mała liczba klas, dzięki czemu statystyka ta jest bardziej czytelna.

Nazwa parametru	Wartość parametru
min_samples_leaf	100
max_depth	15
n_estimators	100
max_samples	0.99

Tabela 8. Parametry klasyfikatora

ID	Parametry domyślne	Accuracy	Imputacja grup wiekowych	Klasyfikacja
1	Tak	52,254%	Tak	key-code
2	Tak	58,33%	Tak	law-breaking-level
3	Tak	60,378%	Nie	key-code
4	Tak	52,694%	Nie	law-breaking-level
5	Nie	52,071%	Tak	key-code
6	Nie	58,312%	Tak	law-breaking-level
7	Nie	60,146%	Nie	key-code
8	Nie	52,822%	Nie	law-breaking-level

Tabela 9. Wyniki klasyfikacji

VIOLATION	MISDEMEANOR	FELONY
36768	138446	16298
27891	718781	82192
11945	338206	104672

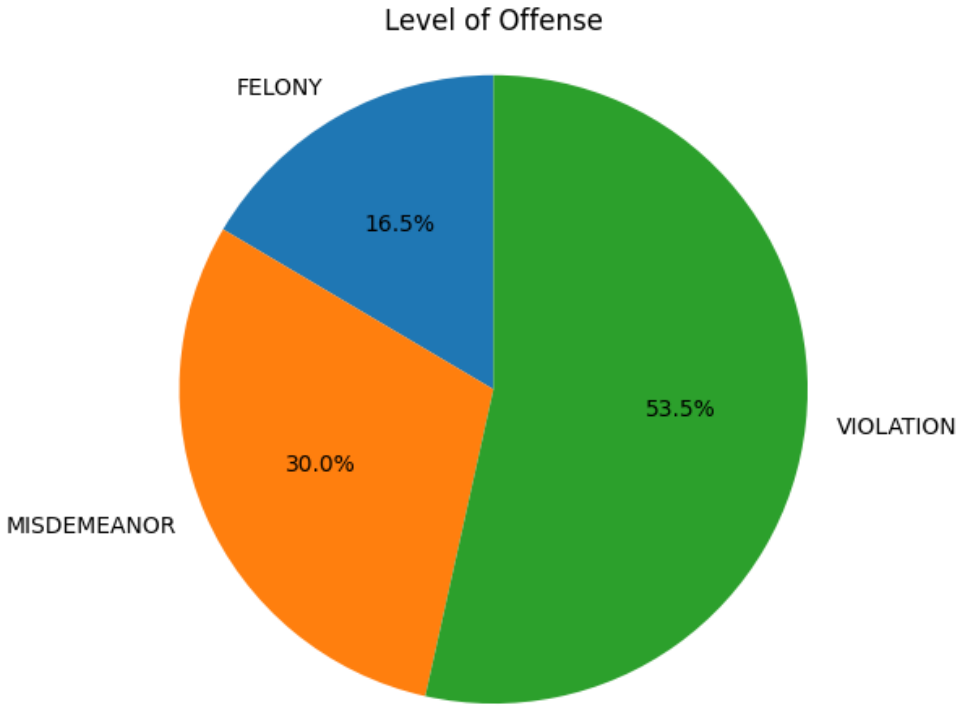
Tabela 10. Macierz pomyłek scenariusza 6

VIOLATION	MISDEMEANOR	FELONY
23955	86421	3272
18145	221943	9316
7840	106419	13202

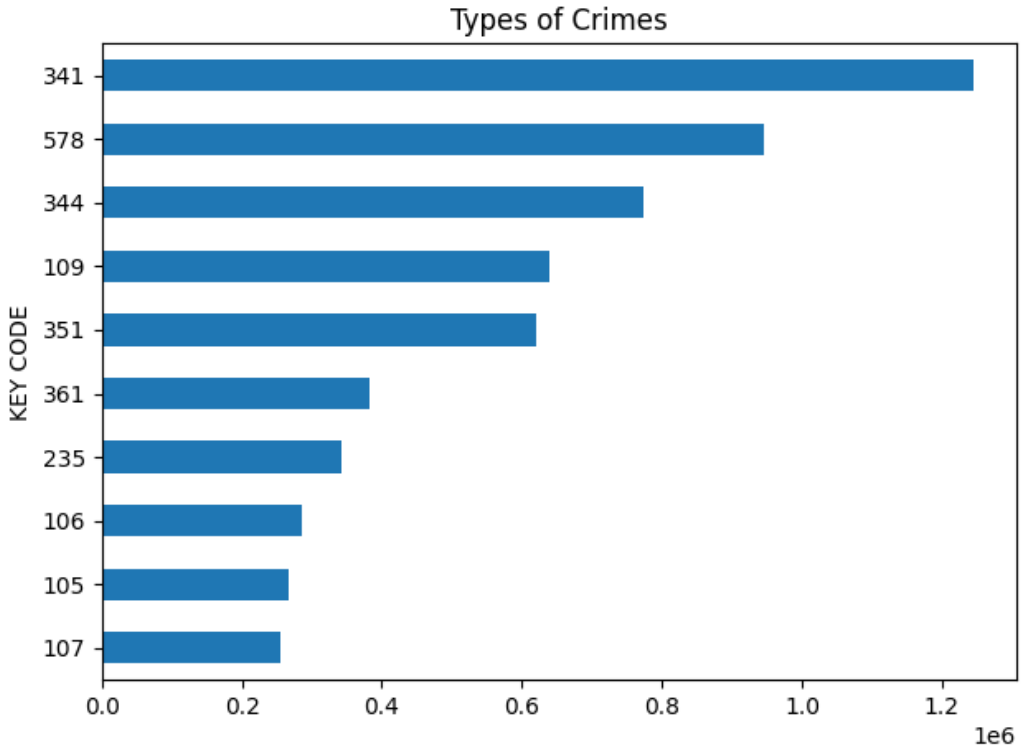
Tabela 11. Macierz pomyłek scenariusza 8

5.1.4. Dyskusja i wnioski

Analizując wyniki końcowe przedstawione w tabeli 9, można zauważyć, że ustawienie parametrów klasyfikatora, wartościami wybranymi w checkpoincie nr 3, nie wpłynęło w sposób znaczący na wyniki, może to być spowodowane zbyt małym zbiorem testowym, na podstawie którego były ustalane wartości parametrów klasyfikatora. Warto zauważyć, że imputacja grup wiekowych poprzez przypisanie najczęściej występującej wartości, wpływa na wzrost dokładności klasyfikacji poziomu wykroczenia w sposób korzystny, natomiast wyniki klasyfikacji kodu przestępstwa, przy zastosowaniu tej metody jest niższy. Najwyższe wyniki dokładności jakie udało się uzyskać podczas klasyfikacji to kolejno dla klasyfikacji kodu przestępstwa i poziomu wykroczenia: 60% i 58%. Aby odpowiedzieć na pytanie czy to dużo, postanowiono zobrazować rozkład danych klasyfikowanych na dwóch wykresach.



Rysunek 37. Rozkład procentowy wartości poziomu wykroczenia.



Rysunek 38. Histogram 10 najczęstszych kodów przestępstwa.

Na wykresie 37 można zaobserwować, że najczęściej występującym zdarzeniem jest 'violation', które stanowi 53,5% przypadków, klasyfikator osiągnął dokładność na poziomie 58%, można zatem stwierdzić, że nie klasyfikuje on w sposób losowy, a jego dokładność jest wyższa niż w przypadku, gdyby przyporządkowywał do każdego klasyfikowanego przypadku najczęściej występujący poziom. Z wykresu 38 można wywnioskować, że kod 341 jest najczęściej występującym kodem przestępstwa, dodatkowe obliczenia pozwoliły wskazać, że występuje on w 16,87% przypadków, na podstawie dokładności klasyfikacji na poziomie 60%, można uznać uzyskane wyniki klasyfikacji za bardzo dobre, gdyż nie zawężają się one do klasyfikacji tylko jednego kodu przestępstwa.

5.2. Regresja godziny wystąpienia zdarzenia

5.2.1. Opis

Regresja godziny wystąpienia zdarzenia jest zadaniem regresji i sprowadza się do przewidywania, w której sekundzie dnia wydarzyło się przestępstwo, na podstawie takich danych jak cechy podejrzanego i ofiary, otoczenia zdarzenia czy też pory roku i dnia tygodnia.

Oryginalnie drugim celem była regresja czasu trwania przestępstwa, jednakże ze względu na bardzo nierównomierny rozkład (patrz charakterystyka zbioru) tej wartości, okazało się to zbyt trudne. Aby pozostać w temacie regresji oraz informacji na temat chwili wystąpienia zdarzenia, cel ten został przeformułowany do obecnej postaci.

Aby zrealizować ten cel wykorzystane zostaną algorytmy/modele przystosowane do zadania regresji. Głównie są to drzewa i lasy losowe zaimplementowane w popularnej bibliotece XGBoost. Wykonane zostało również wstępne przetwarzanie danych w oparciu o wcześniejszą analizę statystyczną. Wykorzystano takie metody jak redukcja wymiarów poprzez selekcję i ekstrakcję cech, oraz podstawowe techniki imputacji danych.

5.2.2. Przygotowanie danych

Wstępne przetwarzanie danych dla tego celu składa się z dwóch etapów. Pierwszy z nich dotyczy czyszczenia zbioru (m. in. imputacji) odrzucenia zbędnych danych, selekcji i ekstrakcji cech. Drugi dotyczy kodowania kolumn. Pierwszy etap można opisać w kontekście poszczególnych grup kolumn:

1. Identyfikator
   - Nie został wykorzystany - nie wnosi żadnej informacji
2. Typ zdarzenia
   - Opisy słowne zostały odrzucone jako redundantne
   - Kod klasyfikacji wewnętrznej (PD_CD) został odrzucony ze względu na zbyt dużą liczbę unikalnych wartości
3. Czy się udało
   - Imputacja najczęstszą wartością („udało się”)
4. Data i czas zdarzenia
   - Odrzucenie daty i czasu końca zdarzenia - dużo braków i nie wydaje się sensowne w kontekście estymacji godziny początku zdarzenia
   - Wyekstrahowanie dnia roku i dnia tygodnia oraz imputacja średnią
   - Wyekstrahowanie sekundy dnia i imputacja średnią
5. Data i czas zgłoszenia
   - Odrzucone jako zbędne w kontekście tego celu
6. Otoczenie zdarzenia
   - Wykorzystane obie kolumny
   - Brakujące dane nie są imputowane ze względu na znacząco liczne braki
7. Lokalizacja zdarzenia
   - Odrzucona jako zbędna w kontekście tego celu
8. Cechy podejrzanego i ofiary
   - Grupy wiekowe są ograniczone do wartości: <18, 18-24, 25-44, 45-64, 64+, NaN (nieznane)

— Rasy UNKNOWN i OTHER są oznaczone jako NaN
— Płcie są ograniczone do wartości: F (kobieta), M (mężczyzna), NaN (nieznane)

Niezastosowanie imputacji i pozostawienie wartości NaN w niektórych kolumnach, jako „nieznane" zostało wykonane ze względu na charakterystykę wykorzystanego algorytmu (XGBoost), który potrafi „radzić" sobie z brakującymi wartościami. Zastosowanie prostej imputacji (np. wartość najczęstsza) w przypadku silnie wybrakowanych danych mogłoby znacząco obciążyć zbiór.

Po drugim etapie wstępnego przetwarzania danych zbiór składa się z następujących kolumn, zakodowanych w następujący sposób:
— CMPLNT_DAY_OF_YEAR - wartość całkowita
— CMPLNT_DAY_OF_WEEK - wartość całkowita
— KY_CD - one hot encoding
— LAW_CAT_CD - one hot encoding
— CRM_ATPT_CPTD_CD - one hot encoding (bez kolumny „ATTEMPTED")
— LOC_OF_OCCUR_DESC - one hot encoding (bez kolumny NaN)
— PREM_TYP_DESC - one hot encoding (bez kolumny NaN)
— VIC_SEX - one hot encoding (bez kolumny NaN)
— SUSP_SEX - one hot encoding (bez kolumny NaN)
— VIC_RACE - one hot encoding (bez kolumny NaN)
— SUSP_RACE - one hot encoding (bez kolumny NaN)
— VIC_AGE_GROUP - ordinal encoding (wartości całkowite od 0 do 4)
— SUSP_AGE_GROUP - ordinal encoding (wartości całkowite od 0 do 4)

Zbiór danych został podzielony na zbiory uczący (70%) i testowy (30%).

5.2.3. Przetwarzanie i analiza danych

Przetwarzanie i analiza danych została przeprowadzona z wykorzystaniem biblioteki XGBoost i algorytmów Gradient Boosted Tree oraz Gradient Boosted Random Forest. W celu optymalizacji hiperparametrów wykorzytano bibliotekę „optuna", która przeszukuje określoną przestrzeń hiperparametrów z wykorzystaniem specjalizowanych algorytmów, w celu minimalizacji zadanej wartości.

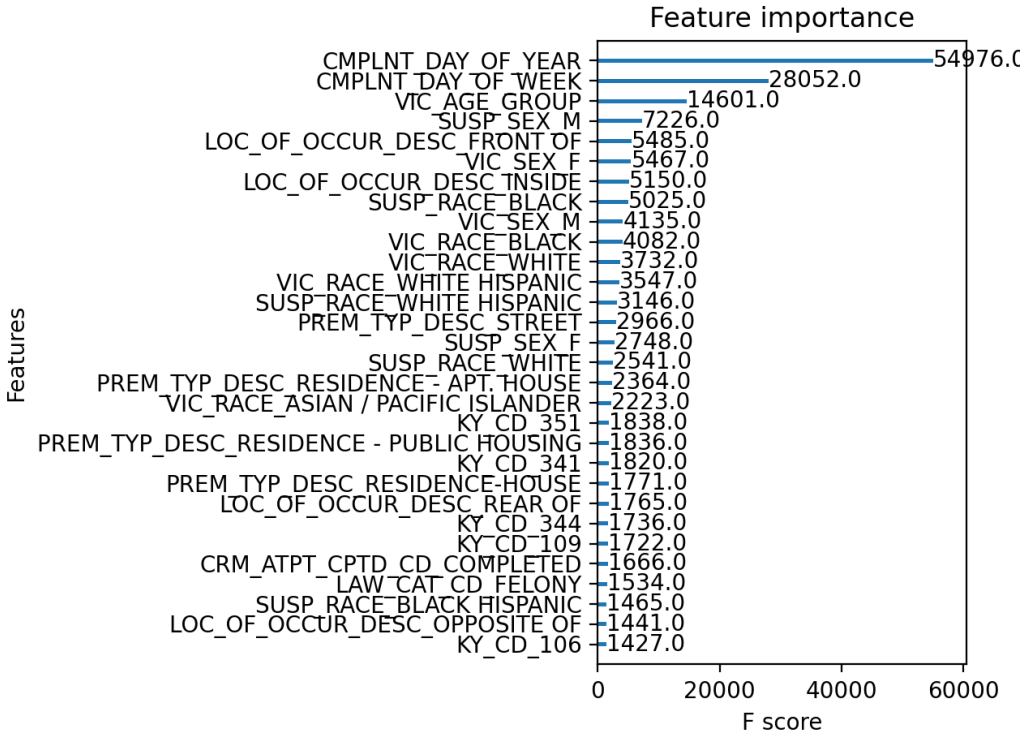
Przeprowadzonych zostało kilka eksperymentów, gdzie każdy jest pojedynczym uruchomieniem algorytmu „optuna" w celu znalezienia optymalnych wartości hiperparametrów dla modelu Gradient Boosted Tree (Forest) z biblioteki XGBoost. Przeszukiwane zakresy hiperparametrów za każdym razem są następujące:
— eta - [0.1 - 1]
— max_depth - [3, 10]
— min_child_weight - [1, 10]
— subsample - [0.6, 1]
— colsample_bynode - [0.6, 1]
— num_parallel_tree - [1, 10]

Pozostałe hiperparametry trenigu są następujące:
— tree_method - gpu_hist
— objective - reg:squarederror
— eval_metric - RMSE, MAE

— num_boost_rounds - 100
— early_stopping_rounds - 10
Optymalizowaną metryką przez bibliotekę „optuna” jest MAE na zbiorze testowym (pełniącego w tym przypadku rolę zbioru walidacyjnego).

Eksperymenty różnią się między sobą jedynie odrzuceniem niektórych ze zdefiniowanego w poprzedniej sekcji zbioru kolumn.

Eksperyment 1. W tym przypadku wykorzystane zostały wszystkie dostępne kolumny. Znalezione optymalne wartości parametrów:
— eta - 0.79
— max_depth - 10
— min_child_weight - 4.87
— subsample - 0.9
— colsample_bynode - 0.85
— num_parallel_tree - 9
Wartości metryk:
— RMSE (zbiór uczący) - 22831
— MAE (zbiór uczący) - 18647
— RMSE (zbiór testowy) - 23120
— MAE (zbiór testowy) - 18887



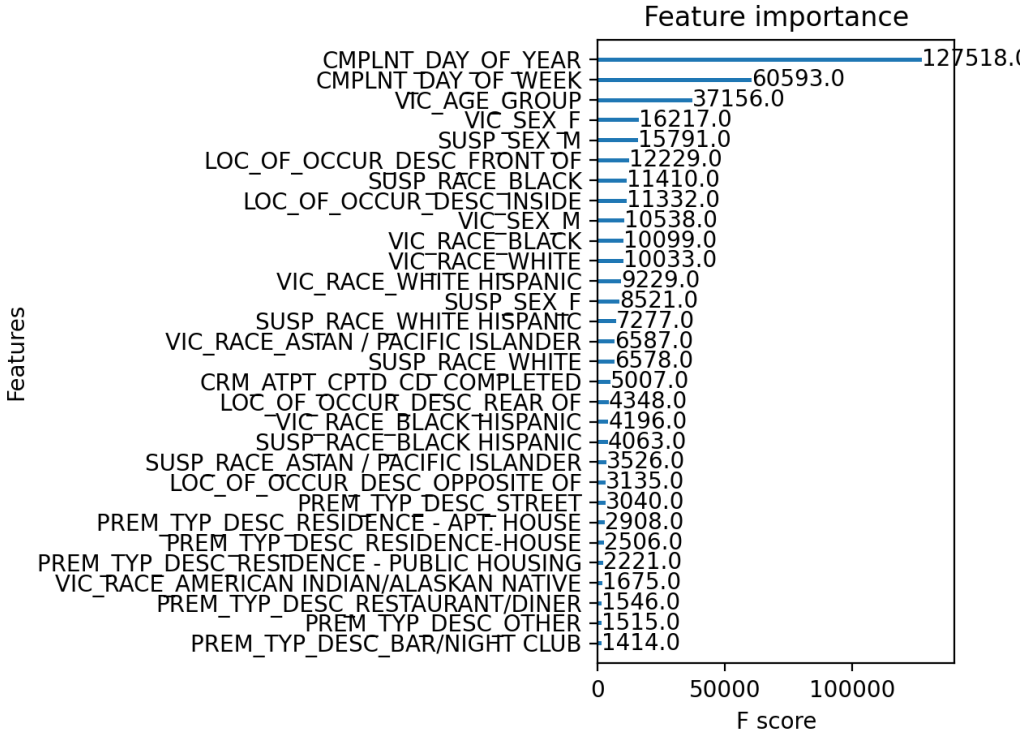
Rysunek 39. Wykres znaczenia poszczególnych cech dla modelu wyuczonego w eksperymencie 1.

Eksperyment 2. W tym eksperymencie nie wykorzystano informacji na temat typu zdarzenia. Znalezione optymalne wartości parametrów:

— eta - 0.57
— max_depth - 10
— min_child_weight - 3.47
— subsample - 0.83
— colsample_bynode - 0.88
— num_parallel_tree - 9

Wartości metryk:

— RMSE (zbiór uczący) - 23118
— MAE (zbiór uczący) - 18993
— RMSE (zbiór testowy) - 23450
— MAE (zbiór testowy) - 19282



Rysunek 40. Wykres znaczenia poszczególnych cech dla modelu wyuczonego w eksperymencie 2.

5.2.4. Dyskusja i wnioski

W eksperymencie 1. wykorzystane zostały wszystkie dostępne dane, które wydają się przydatne dla tego zadani, zakodowane w optymalny sposób tak, aby nie stracić żadnej informacji. Podsumowując dostępne cechy można opisać realizowane zadanie, jako regresja godziny zdarzenia na podstawie cech jego typu, cech jego uczestników, otoczenia i pory roku oraz dnia tygodnia.

Po przeprowadzeniu wszystkich niezbędnych obliczeń znalezione zostały optymalne hiperparametry, dla których model osiągnął wartość średniego błędu bezwzględnego równą 18887 sekund, na zbiorze testowym oraz 18647 sekund na zbiorze uczącym. Pierwszym spostrzeżeniem jest, że wartości te bardzo są do siebie zbliżone (obie wynoszą trochę ponad 5 godzin), co ozna-

cza, że model ma niską wariancję i nie został przetrenowany. Z drugiej strony oznacza to, że najprawdopodobniej można osiągnąć jeszcze trochę lepsze wyniki, zbliżając się do momentu, w którym dojdzie do przetrenowania. Być może potrzebny jest dłuższy trening. Obserwując nieprzedstawione tu metryki podczas uczenia widać jednak, że wartości te zmieniają się po 100 rundach już bardzo nieznacznie.

Jeżeli chodzi o interpretację wartości średniego błędu bezwzględnego, to oznacza on, że średnio udaje się przewidzieć godzinę zdarzenia z dokładnością do około 5 godzin. Innymi słowy, można ocenić porę dnia - czy jest ranek, popołudnie, wieczór czy noc. Wydaje się, że informacja ta może być jak najbardziej przydatna. Z jednej strony trzeba jednak przyznać, że nie jest to bardzo spektakularny wynik i gdyby udało się estymować z dokładnością do godziny byłoby to znacznie ciekawsze osiągnięcie. Z drugiej strony natomiast, osiągnięcie tak dużej dokładności przy takim zbiorze dostępnych cech może okazać się zupełnie niemożliwe.

Rysunek 39 pokazuje znaczenie poszczególnych cech dla wyuczonego modelu. Widać tutaj, że najbardziej istotna jest pora roku i pora tygodnia, oraz cechy podejrzanego i ofiary. Zwłaszcza pora roku wydaje się być logiczna, ponieważ zależnie od pory roku o danej porze dnia jest już ciemno lub jeszcze nie. Duże znaczenie ma też czy zdarzenie miało miejsce na zewnątrz czy wewnątrz np. budynku.

Drugi przeprowadzony eksperyment daje bardzo nieznacznie gorsze wyniki. Tym razem dla zbioru testowego MAE 19282 sekund a dla zbioru uczącego 18993. Ponownie nie mamy tu do czynienia z przetrenowaniem i ponownie jakość regresji praktycznie przestała się poprawiać. Okazuje się więc, że rezygnacja z nawet mniej znaczących kolumn pogarsza jakość regresji i sytuacja taka miałaby miejsce prawdopodobnie dla dowolnych innych cech. Wynika to z faktu, że wyuczony model znajduje bardzo złożone i subtelne zależności, tak więc praktycznie każda cecha może okazać się przydatna. Dodatkowo należy zwrócić uwagę, że algorytmy lasu losowego nie są tak bardzo podatne na obecność nadmiarowych i nieprzydatnych kolumn. Na rysunku 40 te same cechy okazują się być istotne, jak w przypadku pierwszego eksperymentu.

Wnioski z przeprowadzonych badań są następujące:
— Dla wielu kombinacji hiperparametrów model osiąga bardzo podobne wyniki
— Osiągnięta dokładność zawsze znajduje się w okolicach 5 godzin co pozwala ocenić porę dnia zdarzenia
— Najistotniejszymi cechami okazują się pora roku i dzień tygodnia
— Rezygnacja z danych o typie zdarzenia powoduje obniżenie jakości regresji

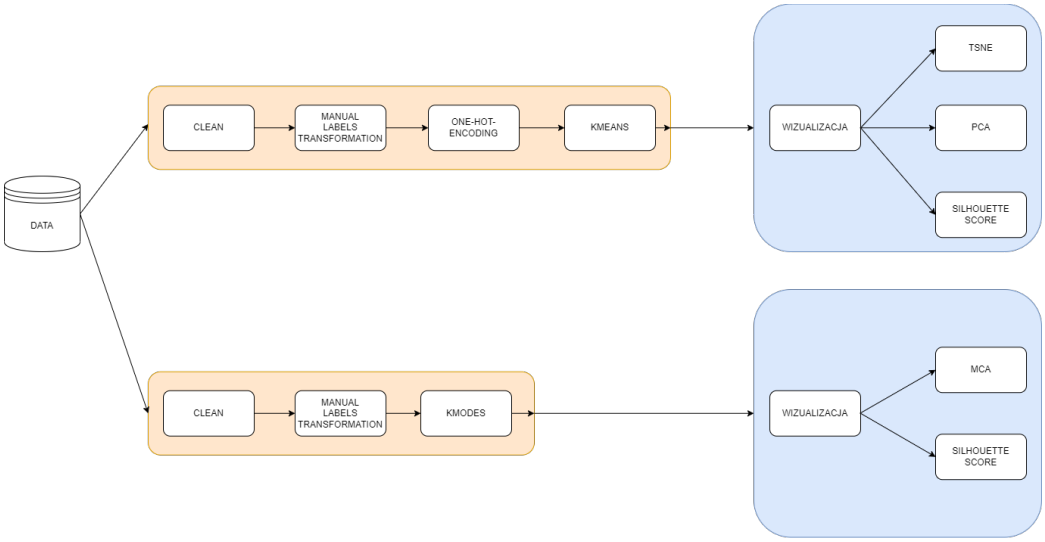
5.3. Grupowanie przestępstw na podstawie podzbiorów cech

5.3.1. Opis

Oryginalnie spośród cech podejrzanego i ofiary miał zostać wybrany kilkakrotnie podzbiorów podzbiór cech, które utworzą przedmiot eksperymentu - etykietę, a dla każdej z nich (która może być złożeniem kilku atrybutów) miała być przeprowadzona seria eksperymentów, mająca na celu automatyczne pogrupowanie zdarzeń zgodnie z tą etykietą, wykorzystując wszystkie pozostałe atrybuty (poza tworzącymi etykietę). Koncept ten został porzucony ze względu na wysoką złożoność obliczeniową przy wyborze permutacji najlepszego zestawu przetworzonych danych.

W ramach tego celu zostało przeprowadzone grupowanie zdarzeń na podstawie etykiety związanej z poziomem przestępstwa. Zostały do tego wykorzystane cechy osoby podejrzanej, poszkodowanej oraz informacje o czasie wystąpienia danego zdarzenia.

Po przeprowadzeniu grupowania z wykorzystaniem kilku różnych algorytmów (kmeans oraz kmodes), zmierzona została jakość grupowania za pomocą metryk, a same dane zwizualizowano, aby ocenić jakość klasteryzacji.



Rysunek 41. Schemat realizacji celu - pipeline

5.3.2. Przygotowanie danych

Przez rozpoczęciem eksperymentów z pełnego zestawu kolumn zostały wybrane te opisujące zarówno podejrzanego, jak i ofiarę, które składały się z informacji o płci, grupie wiekowej, oraz rasie. Informacje o grupie wiekowej zostały zmodyfikowane, aby reprezentowały dane z przedziału 0 do 1.

Początkowa wartość grupy wiekowej	Wartość po transformacji
<18	0
18-24	0.25
25-44	0.5
45-64	0.75
65+	1

Tabela 12. Wartości kolumny zawierającej dane o grupie wiekowej po przekształceniu

Ponadto, w analizie brana była pod uwagę pora dnia zdarzenia, która została wyekstrahowana z godziny zgłoszenia, a następnie zamieniona na dane kategoryczne.

Początkowa wartość godziny	Wartość po transformacji
od 1 do 5	BEFORE_DAWN
5 do 12	MORNING
12 do 17	AFTERNOON
17 do 20	EVENING
20 do 23	NIGHT
0	MIDNIGHT

Tabela 13. Wartości kolumny zawierającej dane o wystąpieniu zdarzenia po przekształceniu

W celu przeprowadzenia eksperymentów wzięte pod uwagę rekordy, nie mogły posiadać wybrakowanych danych w ramach opisanych kolumn. Dodatkowo, brane pod uwagę były wyłącznie rekordu w których wartość płci oznaczała kobietę, lub mężczyznę. Pozwoliło to na wyszczególnienie **2301826** rekordów, które zostały poddane analizie. Usunięta została również kolumna zawierając kod przestępstwa, gdyż stanowiła ona etykiety do późniejszej weryfikacji.

5.3.3. Przetwarzanie i analiza danych

Każdy z algorytmów wymagał dodatkowego przetworzenia rekordów, które będą stanowiły dane eksperymentu. W przypadku algorytmu KMeans dane kategorialne zostały poddane operacji one hot encodingu, aby poprawnie przeprowadzić klasteryzację. Liczba klastrów została ustawiona zgodnie z liczbą etykiet poziomów przestępstw na 3.

Po uzyskanych wynikach klasteryzacji rozpatrywany był każdy przepadek dopasowania danej etykiety do klastra, a następnie wybrany został wariant z największą wartością accuracy.

LABEL 0	LABEL 1	LABEL 2	Accuracy
FELONY	VIOLATION	MISDEMEANOR	36.837

Tabela 14. Najlepsza permutacja dla eksperymentu KMeans

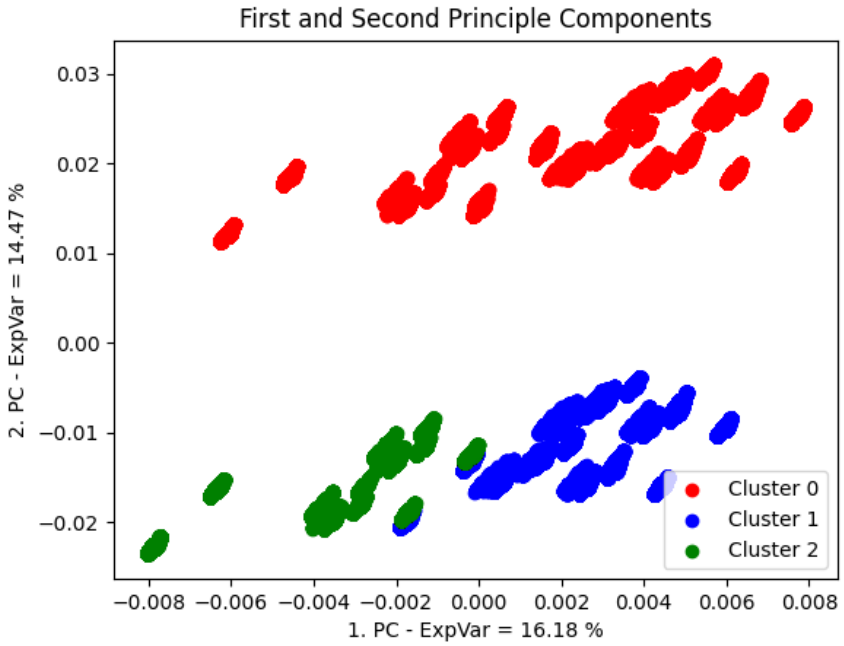
	FELONY	VIOLATION	MISDEMEANOR
DATASET	562577	714884	1024365
CLUSTER	863971	697736	740119

Tabela 15. Liczebność próbek danych etykiet dla zbioru danych i klastrów

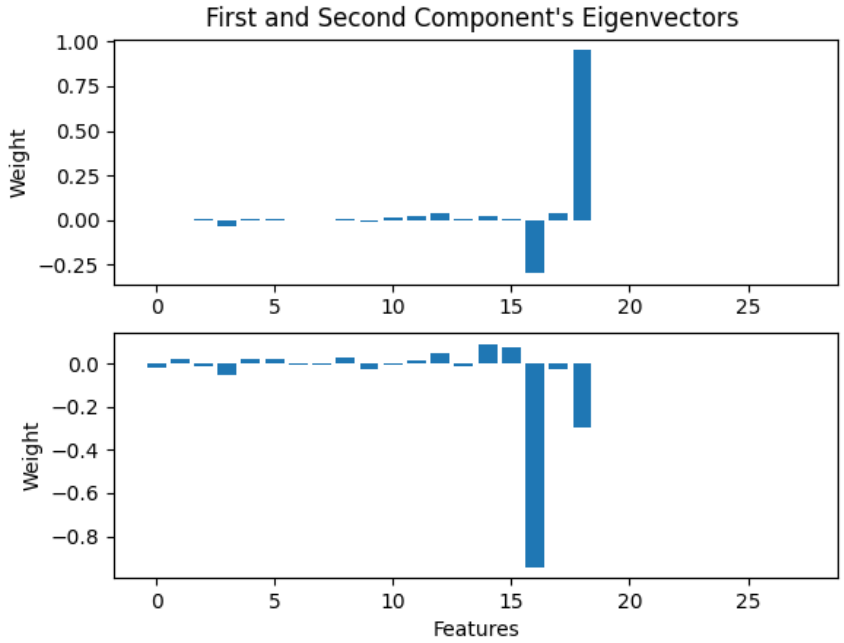
	VIOLATION	MISDEMEANOR	FELONY
accuracy	0.597510	0.524523	0.614722
precision	0.348382	0.452642	0.312340
recall	0.340026	0.327041	0.479673
sensitivity	0.340026	0.327041	0.479673
specificity	0.713501	0.682879	0.658405

Tabela 16. Wartości poszczególnych metryk dla uzyskanych klastrów - KMeans

W celu wizualizacji otrzymanych klastrów w wyniku algorytmu KMeans została zastosowana metoda redukcji wymiarów PCA (*Principal component analysis*), której wynikiem były dane dwuwymiarowe.

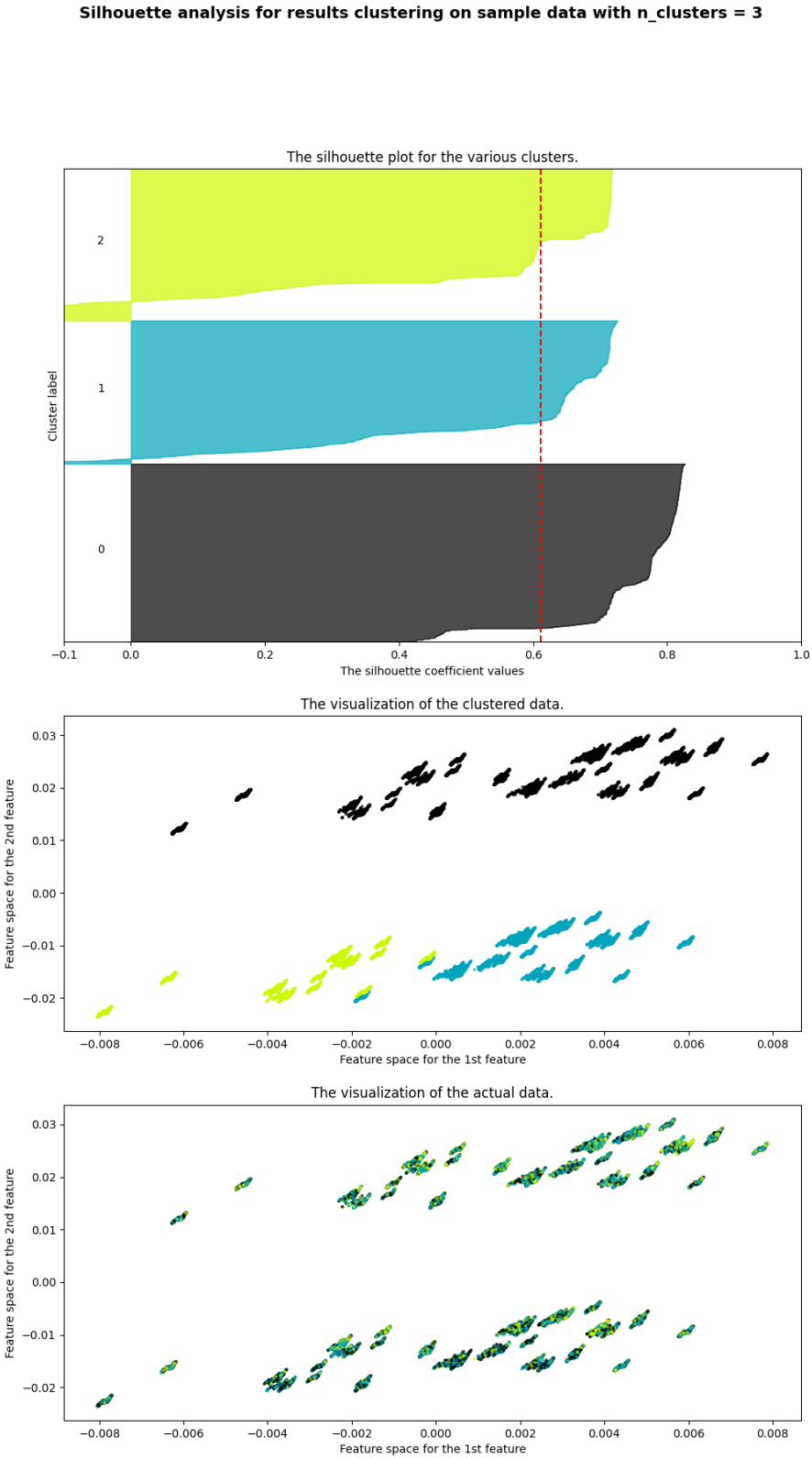


Rysunek 42. Wykres przedstawiający wyniki klasteryzacji po zastosowaniu PCA

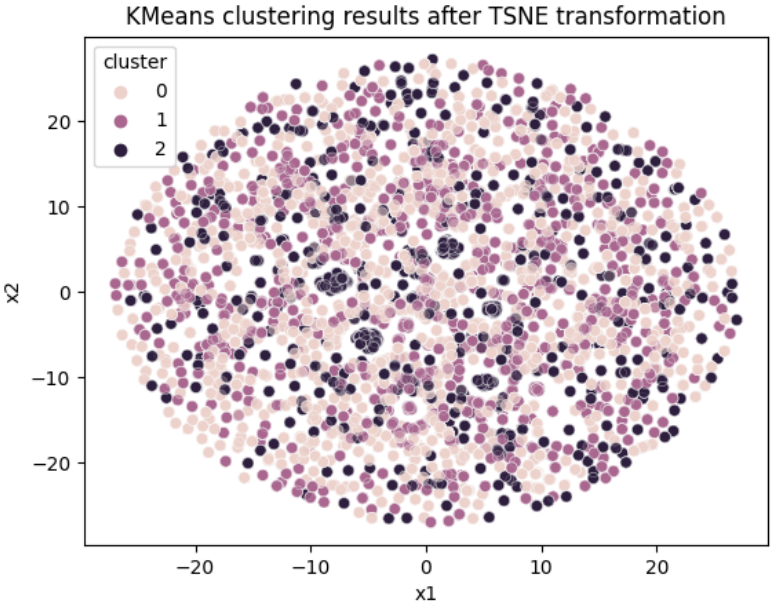


Rysunek 43. Wagi cech wektorów własnych PCA

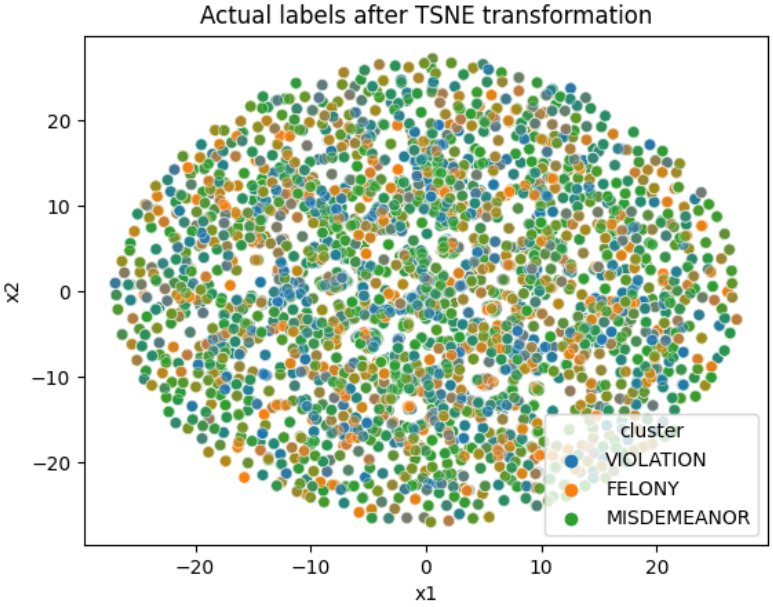
Ponadto została wyznaczona średnia wartość silhouette score, które w przypadku eksperymentu KMeans wynosiła **0.61**



Rysunek 44. Silhouette score poszczególnych próbek oraz porównanie wizualne danych z etykietą po klastrowaniu oraz faktyczną dla algorytmu KMeans



Rysunek 45. Wizualizacja danych przetworzonych za pomocą algorytmu TSNE - dane po klasteryzacji



Rysunek 46. Wizualizacja danych przetworzonych za pomocą algorytmu TSNE - dane faktyczne

Drugi eksperyment przeprowadzono z wykorzystaniem algorytmu KModes, który jest silnie nastawiony na wykorzystanie danych kategorialnych i wydawał się w przypadku analizowanego zbioru obiecującym wyborem.

Ponieważ algorytm ten bazuje na danych kategorialnych kolumny nie musiały zostać poddane one-hot encodingowi.

LABEL 0	LABEL 1	LABEL 2	Accuracy
MISDEMEANOR	VIOLATION	FELONY	52.32

Tabela 17. Najlepsza permutacja dla eksperymentu KModes

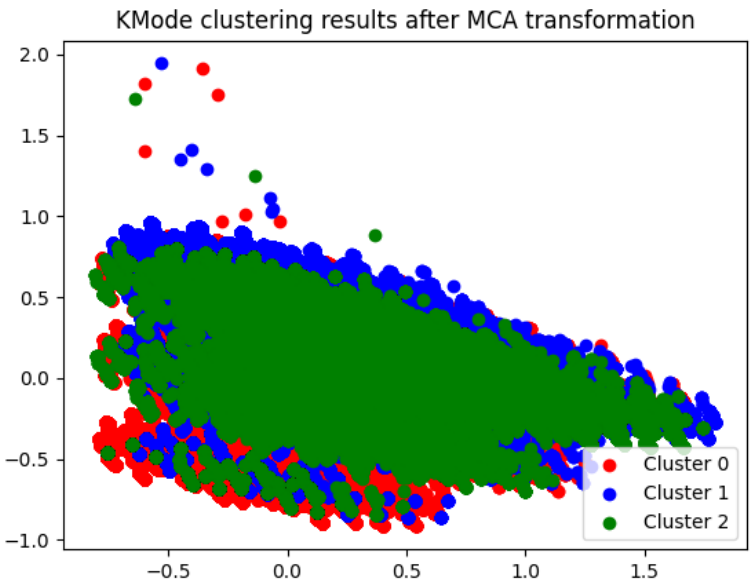
	FELONY	VIOLATION	MISDEMEANOR
DATASET	562577	714884	1024365
CLUSTER	211447	504967	1585412

Tabela 18. Liczebność próbek danych etykiet dla zbioru danych i klastrów

	VIOLATION	MISDEMEANOR	FELONY
accuracy	0.689722	0.610955	0.745811
precision	0.500672	0.540637	0.446746
recall	0.353656	0.836745	0.167911
sensitivity	0.353656	0.836745	0.167911
specificity	0.841113	0.429900	0.932739

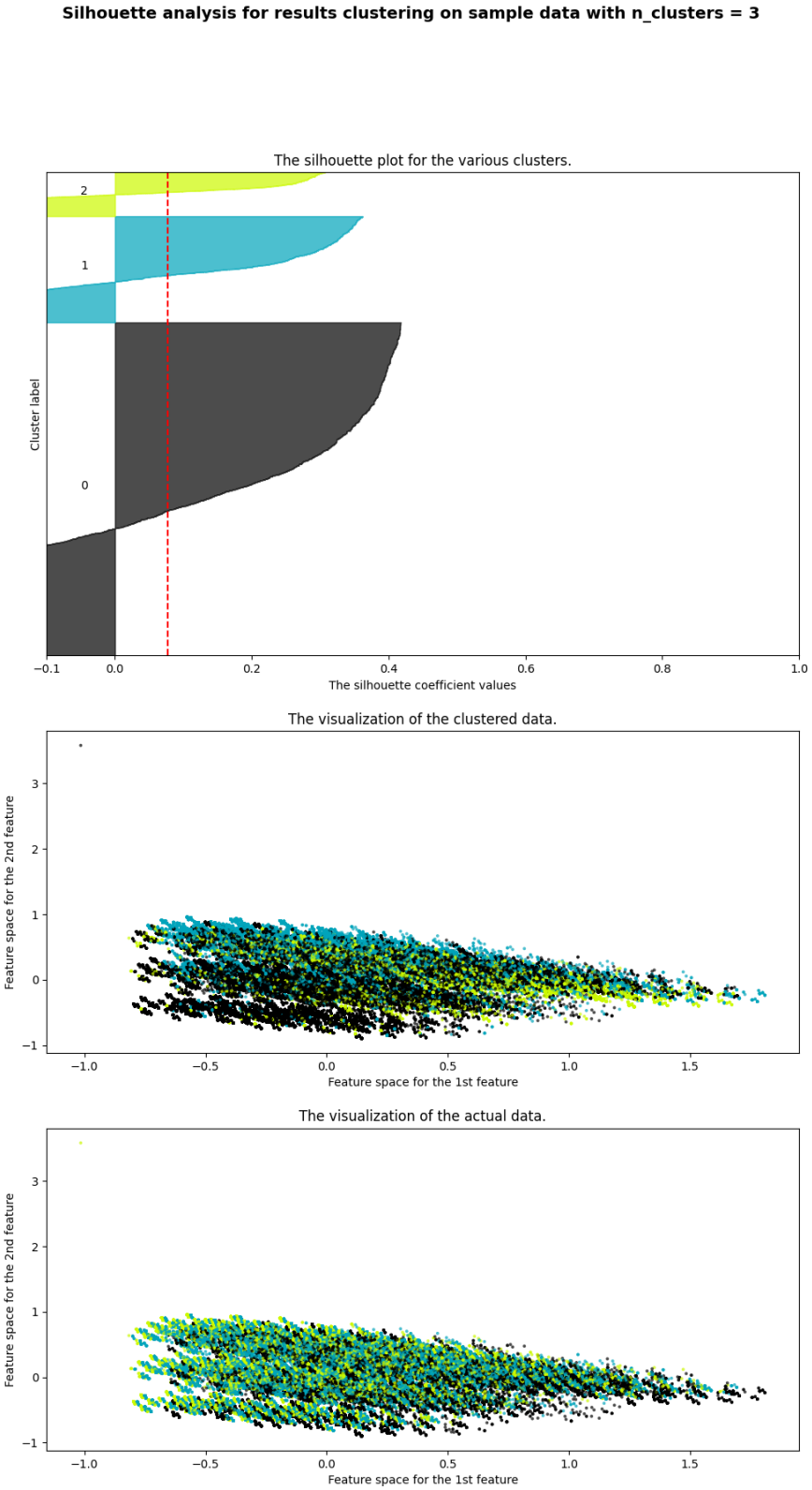
Tabela 19. Wartości poszczególnych metryk dla uzyskanych klastrów - KModes

Również w tym przypadku wizualizacja danych wymagała redukcji wymiarowości. Zamiana danych na zgodne z PCA jest błędne, dlatego zastosowana została metoda MCA (ang. *Multiple correspondence analysis*) przystosowana do danych kategorialnych.



Rysunek 47. Wykres przedstawiający wyniki klasteryzacji po zastosowaniu MCA

Dla eksperymentu z algorytmem KModes również wyznaczono wartość silhouette score, które wynosiła **0.07**.



Rysunek 48. Silhouette score poszczególnych próbek oraz porównanie wizualne danych z etykietą po klastrowaniu oraz faktyczną dla algorytmu Kmodes

5.3.4. Dyskusja i wnioski

W przypadku algorytmu KMeans można zauważyć w tabeli 16 niższą wartość accuracy dla algorytmu KModes. Przeprowadzenie redukcji wymiarów za pomocą metody PCA zwróciło wartość wariancji oscylującą około 15%, co jest dalekie od optymalnych wartości referencyjnych 70% - 80%. Rozkład liczby próbek uzyskanej przez klastrowanie nie odpowiada wartościom występującym w faktycznym zbiorze danych. Ponadto, w przypadku każdej z etykiet miary precision oraz recall nie przekroczyły 50% próbek, a większość oscylowała wokół 35%.

Algorytm KModes poradził sobie lepiej w przypadku rozważanego problemu. Accuracy otrzymane dla tego eksperymentu przekroczyło 50%. Wyniki MCA pozwalają też na zaobserwowanie ułożenia danych, które są bardziej skupione i łatwiejsze do klasteryzacji. Metoda utworzyła klastry, z których jeden zawsze był znacznie większy. Przy założeniu, że ten klaster otrzymałby etykietę najliczniejszej grupy danych (MISDEMEANOR) przełożyłoby się to na wysokie wartości metryk precision oraz recall kosztem jakości klasteryzacji dla najmniejszego klastra (w tym przypadku FELONY).

Wnioski:

— Zastosowanie algorytmu KMeans bez odpowiedniego przygotowania danych za pomocą one-hot encodingu prowadzi do uzyskania losowych wyników (najczęściej accuracy 1/n)
— Algorytm KMeans nie pozwala na uzyskanie stabilnych wyników w przypadku danych kategorialnych
— Algorytm KModes oferuje lepsze wyniki dla danych kategorialnych

6. Ewolucja projektu

Od momentu kiedy rozpoczęte zostały prace nad tym projektem nastąpiło bardzo wiele zmian dotyczących zarówno realizowanych celów, sposobów ich realizacji jak również zmian związanych ze świadomością i wiedzą uczestników projektu.

Pierwszym znaczącym błędem, który został popełniony, było niedokładne przeanalizowanie danych pod kątem statystycznym. Należało dokładnie zbadać jaki jest rozkład wartości we wszystkich kolumnach i gdzie to tylko możliwe dokonać wizualizacji, które zawsze pomagają zrozumieć dane ludzkiemu odbiorcy. W ten sposób znacznie łatwiej byłoby sformułować sensowne cele. Ponieważ zabrakło tej analizy od samego początku, dwa cele zostały sformułowane nie do końca rozsądnie i uległy modyfikacjom.

Drugim ważnym krokiem, który został podjęty w trakcie realizacji projektu a nie został zrealizowany wcześniej było skupienie się na temacie wstępnego przetwarzania danych: ich oczyszczania, imputacji, kodowania, selekcji i ekstrakcji cech, itp. W przypadku danych tabelarycznych, z jakimi mamy tu do czynienia, często najbardziej istotnym etapem jest właśnie ten, kiedy wybiera się które kolumny zostaną użyte i w jakiej formie zostaną podane do algorytmu. Zabrakło nie tylko poświęcenie większej ilości czasu temu zagadnieniu ale również wiedzy i świadomości uczestników projektu. Należało na ile to możliwe zapoznać się lub przypomnieć sobie dokładnie możliwe podejścia do tego tematu. Z powodu, że aspekt ten nie był wzięty pod uwagę od samego początku wiele pierwszych eksperymentów okazało się zupełnie zbędnych.

Ostatnim tematem, który ulegał istotnym zmianom jest sposób organizacji pracy z kodem źródłowym i etapami przetwarzania danych. Na początku zespół próbował uwspólnić ile tylko możliwe etapów pracy i fragmentów kodu. Zwłaszcza dotyczyło to wstępnego przetwarzania danych. Okazało się to jednak problematyczne i prowadziło do zamieszania, ponieważ na potrzeby każdego z celów inne cechy były wykorzystywane. Dodatkowo sposób kodowania i podejście do imputacji danych jest silnie zależne od wykorzystywanych algorytmów. Z tego powodu, najprostszym rozwiązaniem okazało się zrezygnowanie z nadmiernej generalizacji i rozdzielenie wstępnego przetwarzania danych dla każdego celu z osobna.

7. Opis kodu źródłowego

Biorąc pod uwagę doświadczenia nabyte w czasie realizacji checkpoint numer 1, 2, 3 zdecydowaliśmy się zmienić podejście od strony programistycznej. Pierwotnym podejściem było stworzenie jednego programu ze wspólnym preprocessingiem danych oraz różnymi modułami odpowiadającymi za realizacji założonych celów. Niestety pomimo szczerych chęci takie podejście absolutnie się nie sprawdziło a co więcej mogło pogarszać uzyskiwane wyniki, gdyż próby tworzenia polimorficznego kodu w tym przypadku bardzo komplikowały program i generowały nowe problemy. W przypadku ostatniego checkpointa zdecydowaliśmy się na nieskomplikowane założenie a mianowicie każdy cel to osobny zbiór plików, które zawierają kod w języku Python. W przypadku

celu nr 1 jest to katalog "random_forest", dla celu nr 2 został stworzony katalog "clustering" natomiast kod odpowiedzialny za cel nr 3 jest umieszczony w katalogu "time_regression". Pliki w tych 3 katalogach wykorzystują kod znajdujący się w "shared". Co więcej naszę nieudane podejścia do analizy danych zostały umieszczone w "old_approach" natomiast dodatkowe skrypty do statystycznej analizy danych, który były wykorzystywane w ramach Checkpoint 1 są umieszczone w "scripts".

Literatura

[1] https://data.cityofnewyork.us/Public-Safety/NYPD-Complaint-Data-Historic/qgea-i56i