Feature engineering + EDA

Uczniowie uchodźcy z Ukrainy w podziale na typy szkół, klasy i powiaty – stan na 31.03.2025

Analiza danych i przygotowanie zbioru do modelowania

Roksana Jandura, Inżynieria i Analiza Danych

Spis treści

1. Zapoznanie z biblioteką Seaborn

Seaborn to biblioteka wizualizacji danych dla języka Python, oparta na bibliotece Matplotlib. Umożliwia tworzenie estetycznych i złożonych wykresów za pomocą prostych poleceń. Została zaprojektowana z myślą o analizie statystycznej i eksploracyjnej wizualizacji danych.

Biblioteka integruje się bezpośrednio z obiektami typu DataFrame z biblioteki pandas, co ułatwia tworzenie wykresów opartych na kolumnach danych. Seaborn oferuje m.in. wykresy: słupkowe, rozrzutu, pudełkowe, liniowe, macierze korelacji czy wykresy gęstości. Umożliwia również szybkie tworzenie wykresów z podziałem na kategorie (np. hue, col, row) oraz automatyczne dobieranie palet kolorystycznych.

Biblioteka jest często wykorzystywana podczas eksploracyjnej analizy danych (EDA) w celu identyfikacji zależności, rozkładów i anomalii w danych.

2. Charakterystyka portalu Dane.gov.pl

2.1. Opis portalu

Portal Dane.gov.pl jest oficjalną platformą udostępniania danych publicznych w Polsce. Został uruchomiony w ramach działań Ministerstwa Cyfryzacji i służy jako centralne repozytorium otwartych danych udostępnianych przez instytucje publiczne. Celem portalu jest zwiększenie przejrzystości administracji publicznej oraz umożliwienie ponownego wykorzystania danych w analizach, projektach badawczych i aplikacjach.

2.2. Rodzaje danych dostępnych na portalu

Na portalu udostępniane są dane z różnych dziedzin, takich jak:

- edukacja, kultura i sport,
- energia
- kwestie międzynarodowe
- ludność i społeczeństwo
- nauka i technologia
- regiony i miasta
- rolnictwo, rybołówstwo, leśnictwo i żywność
- rząd i sektor publiczny
- sprawiedliwość, ustrój sądów i bezpieczeństwo publiczne
- transport,
- środowisko,
- zdrowie,
- gospodarka i finanse,

Dane są publikowane w postaci zbiorów, często aktualizowanych, i mogą pochodzić od urzędów centralnych, samorządów, instytucji publicznych czy GUS.

Udostępnione dane mogą być wykorzystywane do prowadzenia analiz statystycznych, modelowania predykcyjnego, tworzenia aplikacji internetowych, raportów, wizualizacji danych oraz wspierania procesów decyzyjnych w administracji publicznej i sektorze prywatnym. Dzięki otwartemu dostępowi możliwe jest ponowne wykorzystywanie danych w różnych projektach badawczych, społecznych i komercyjnych.

2.3. Format i sposób pobierania danych

Zbiory danych dostępne są w różnych formatach, m.in. CSV, XLSX, JSON, XML. Każdy zbiór można pobrać ręcznie jako plik, ale część z nich dostępna jest także przez interfejs API, umożliwiający automatyczne pobieranie danych.

2.4. Interfejsy API dostępne na portalu

Portal Dane.gov.pl umożliwia dostęp do danych publicznych nie tylko poprzez bezpośrednie pobieranie plików, ale również za pośrednictwem interfejsów API (Application Programming Interface). Głównym i oficjalnie wspieranym interfejsem API na portalu jest:

- RESTful API: Interfejs oparty na architekturze REST, wykorzystujący protokół HTTP. Umożliwia on pobieranie danych w formacie JSON. API to pozwala na dostęp do metadanych, listy zbiorów oraz pobieranie danych w sposób automatyczny. Dostęp do API nie wymaga uwierzytelnienia, jednak istnieją ograniczenia dotyczące liczby zapytań w określonym czasie.

2.5. Ocena jakości danych

Dla każdego zbioru danych na portalu podany jest poziom otwartości w postaci gwiazdek (od 1 do 5), co umożliwia szybką ocenę ich formatu oraz przydatności do przetwarzania. W opisach zbiorów dostępne są również informacje o dacie aktualizacji, formacie pliku, języku oraz warunkach wykorzystania danych.

Na portalu istnieje możliwość podejrzenia początkowych wierszy danych bez konieczności pobierania całego pliku. Ponadto, dla niektórych zbiorów dostępne są funkcje umożliwiające wykonanie podstawowych wizualizacji, takich jak histogramy czy wykresy rozkładu, na podstawie wybranych kolumn danych.

Portal nie oferuje natomiast zaawansowanego, automatycznego narzędzia do kompleksowej analizy jakości danych, takiej jak wykrywanie braków lub niespójności. W większości przypadków pełna ocena jakości danych wymaga ręcznej weryfikacji metadanych oraz własnej analizy pobranych zasobów.

3. Wybór zestawu danych

3.1. Wybrany zbiór danych

W ramach realizacji projektu wybrano zbiór danych pt. „Uczniowie uchodźcy z Ukrainy w podziale na typy szkół, klasy i powiaty – stan na 31.03.2025”, opublikowany na portalu Dane.gov.pl przez Ministerstwo Edukacji Narodowej. Zestawienie zawiera informacje o liczbie oddziałów oraz liczbie uczniów uchodźców z Ukrainy przypisanych do oddziałów podstawowych lub przygotowawczych w szkołach lub placówkach oświatowych. W zestawieniu uwzględniono uczniów zarejestrowanych od dnia 24 lutego 2022 roku, posiadających aktywne przypisanie do oddziału w dniu aktualizacji raportu, tj. 31 marca 2025 roku.

3.2. Uzasadnienie wyboru zbioru

Zbiór danych został wybrany ze względu na aktualność, szczegółowość oraz społeczne znaczenie prezentowanych informacji. Pozwala na analizę rozmieszczenia uczniów-uchodźców w polskim systemie oświaty z podziałem na lokalizację oraz typy szkół. Struktura danych umożliwia wykonanie eksploracyjnej analizy danych (EDA), inżynierii cech (Feature Engineering). Dodatkowo, dane mają przejrzystą strukturę tabelaryczną i są dostępne w formacie umożliwiającym łatwe przetwarzanie.

4. Wykonanie pierwszego etapu pipeline'u ML

4.1. Pobranie danych i zapoznanie się z ich opisem

Dane zostały pobrane w formacie CSV z portalu Dane.gov.pl, co umożliwiło ich bezpośrednie wczytanie do środowiska analitycznego w języku Python. Plik posiada przejrzystą strukturę tabelaryczną, sprzyjającą dalszemu przetwarzaniu i analizie.
W ramach zapoznania się ze strukturą zbioru danych przeprowadzono wstępną inspekcję zawartości pliku. Każdy wiersz odpowiada jednej unikalnej kombinacji lokalizacji (województwo, powiat), rodzaju szkoły, klasy oraz typu oddziału i zawiera informacje o liczbie oddziałów oraz liczbie uczniów uchodźców z Ukrainy przypisanych do tej struktury organizacyjnej. Struktura danych została przedstawiona na rysunku Fig. 4.1. Zbiór danych zawiera zarówno zmienne kategoryczne, jak i liczbowe. Kluczowe kolumny to m.in.:

- Województwo, Powiat – lokalizacja placówki,
- Typ podmiotu, Rodzaj placówki, Publiczność – charakterystyka jednostki edukacyjnej,
- Typ oddziału, Klasa – poziom edukacyjny,
- Liczba oddziałów, Liczba uczniów pobyt legalny – dane liczbowe, z których ostatnia stanowi potencjalną zmienną docelową (target) do predykcji w ramach uczenia nadzorowanego.

Tak przygotowany zbiór stanowi punkt wyjścia do dalszych etapów procesu analitycznego, w tym inżynierii cech i eksploracyjnej analizy danych.

	idTerytWojewodztwo	Województwo	idTerytPowiat	Powiat	idRodzajPlacowki	Rodzaj Placowki	idPublicznosc	Publiczność	idTypPodmiotu	Typ Podmiotu
0	2.0	DOLNOŚLĄSKIE	201.0	bolesławiecki	1.0	samodzielna	1.0	publiczna	1.0	Przedszkole
1	2.0	DOLNOŚLĄSKIE	201.0	bolesławiecki	1.0	samodzielna	1.0	publiczna	3.0	Szkoła podstawowa
2	2.0	DOLNOŚLĄSKIE	201.0	bolesławiecki	1.0	samodzielna	1.0	publiczna	3.0	Szkoła podstawowa
3	2.0	DOLNOŚLĄSKIE	201.0	bolesławiecki	1.0	samodzielna	1.0	publiczna	3.0	Szkoła podstawowa
4	2.0	DOLNOŚLĄSKIE	201.0	bolesławiecki	1.0	samodzielna	1.0	publiczna	3.0	Szkoła podstawowa

Typ Podmiotu	idTypOddzialu	Typ Oddziału	idKlasa	Klasa	Liczba oddziałów	Liczba uczniów pobyt legalny
Przedszkole	10.0	Podstawowe	24	Wychowanie przedszkolne	32.0	50.0
Szkoła podstawowa	10.0	Podstawowe	4	I	13.0	32.0
Szkoła podstawowa	10.0	Podstawowe	5	II	12.0	30.0
Szkoła podstawowa	10.0	Podstawowe	6	III	19.0	45.0
Szkoła podstawowa	10.0	Podstawowe	7	IV	15.0	35.0

Fig 4.1. Przykładowe rekordy zbioru danych przedstawiające kombinację cech lokalizacji, rodzaju szkoły, klasy oraz liczby uczniów z Ukrainy.

4.2. Klasyfikacja typu problemu ML i potencjalne zastosowania danych

Zbiór danych umożliwia zastosowanie podejścia opartego na uczeniu maszynowym, w szczególności w ramach uczenia nadzorowanego (supervised learning). Celem modelowania może być przewidywanie liczby uczniów uchodźców z Ukrainy przypisanych do określonej kombinacji cech jednostki edukacyjnej, takich jak lokalizacja, typ placówki, publiczność, rodzaj oddziału czy klasa.

Z uwagi na fakt, że dane zawierają zarówno zmienne kategoryczne, jak i numeryczne, możliwe jest zastosowanie szerokiego zakresu algorytmów regresyjnych oraz klasycznych metod przetwarzania danych tablicowych. Konieczne będzie odpowiednie przygotowanie danych, w tym kodowanie zmiennych jakościowych, skalowanie oraz weryfikacja spójności informacji.

Zbiór może znaleźć zastosowanie w praktyce m.in. w:

- planowaniu zasobów edukacyjnych w jednostkach samorządowych,
- analizie rozmieszczenia uczniów-uchodźców w systemie oświaty,
- wsparciu w podejmowaniu decyzji dotyczących otwierania nowych oddziałów lub zwiększania liczby klas,
- modelowaniu scenariuszy zmian demograficznych w kontekście edukacji.

4.3. Inżynieria cech i eksploracyjna analiza danych

Na początkowym etapie dokonano sprawdzenia podstawowych informacji o zbiorze danych. Zawiera on 13 130 rekordów (wierszy) oraz 16 kolumn, z których większość stanowi zmienne kategoryczne.

Poniżej przedstawiono typy danych oraz ich klasyfikację:

- Zmienne jakościowe (kategoryczne):
  - Tekstowe (object): Województwo, Powiat, Rodzaj Placówki, Publiczność, Typ Podmiotu, Typ Oddziału, Klasa – są to zmienne opisowe
  - Identyfikatory (float64): idTerytWojewodztwo, idTerytPowiat, idRodzajPlacowki, idPublicznosc, idTypPodmiotu, idTypOddzialu –pełnią funkcję identyfikatorów kategorii, a więc traktowane są jako zmienne jakościowe.
- Zmienne ilościowe (float64):
  - Liczba oddziałów
  - Liczba uczniów pobyt legalny

Warto zauważyć, że dla każdej zmiennej kategorycznej dostępny jest odpowiadający jej identyfikator liczbowy. Struktura danych została przedstawiona na rysunku Fig. 4.2, prezentującym typy zmiennych dla każdej kolumny.

```
Rozmiar zbioru danych:
Liczba wierszy (rekordów): 13130
Liczba kolumn: 16

Typy danych w kolumnach:
idTerytWojewodztwo              float64
Województwo                      object
idTerytPowiat                   float64
Powiat                           object
idRodzajPlacowki                float64
Rodzaj Placowki                  object
idPublicznosc                   float64
Publiczność                      object
idTypPodmiotu                   float64
Typ Podmiotu                     object
idTypOddzialu                   float64
Typ Oddziału                     object
idKlasa                          object
Klasa                            object
Liczba oddziałów                float64
Liczba uczniów pobyt legalny    float64
dtype: object
```

Fig 4.2. Rozmiar zbioru danych i typy danych w kolumnach

Sprawdzono obecność brakujących wartości. Jak pokazano na rysunku Fig. 4.3, każda kolumna – z wyjątkiem Klasa, która zawiera 76 braków – zawiera dokładnie 16 brakujących rekordów. Braki te dotyczą zarówno zmiennych kategorycznych, jak i liczbowych.

```
Braki danych:
idTerytWojewodztwo              16
Województwo                     16
idTerytPowiat                   16
Powiat                          16
idRodzajPlacowki                16
Rodzaj Placowki                 16
idPublicznosc                   16
Publiczność                     16
idTypPodmiotu                   16
Typ Podmiotu                    16
idTypOddzialu                   16
Typ Oddziału                    16
idKlasa                         16
Klasa                           76
Liczba oddziałów                16
Liczba uczniów pobyt legalny    16
dtype: int64
```

Fig 4.3. Braki danych

W kolejnym kroku wyświetlono rekordy zawierające brakujące wartości we wszystkich kolumnach, z wyjątkiem kolumny Klasa. Okazało się, że są to ostatnie 16 wierszy zbioru, które były całkowicie puste – brakowało w nich wartości we wszystkich kolumnach. Ich obecność wynikała z błędnego formatowania pliku źródłowego, a nie z rzeczywistego braku danych. Zidentyfikowane braki należą do kategorii MCAR

(Missing Completely At Random), ponieważ ich wystąpienie nie było powiązane z żadną zmienną w zbiorze. W związku z tym podjęto decyzję o usunięciu tych rekordów ze zbioru danych.

Po usunięciu 16 pustych wierszy końcowych, ponownie sprawdzono kompletność danych. Jak przedstawiono na rysunku Fig. 4.4, braki danych występują już tylko w jednej kolumnie – Klasa, która zawiera 60 nieuzupełnionych wartości. Pozostałe kolumny są kompletne.

```
idTerytWojewodztwo              0
Województwo                     0
idTerytPowiat                   0
Powiat                          0
idRodzajPlacowki                0
Rodzaj Placowki                 0
idPublicznosc                   0
Publiczność                     0
idTypPodmiotu                   0
Typ Podmiotu                    0
idTypOddzialu                   0
Typ Oddziału                    0
idKlasa                         0
Klasa                          60
Liczba oddziałów                0
Liczba uczniów pobyt legalny    0
dtype: int64
```

Fig 4.4. Braki danych po usunięciu 16 końcowych wierszy

W celu lepszego zrozumienia charakteru braków w kolumnie Klasa, przeanalizowano ich powiązania z innymi zmiennymi kategorycznymi. Jak pokazano na rysunku Fig. 4.5, wszystkie 60 brakujących wartości występują w wierszach, w których Typ Oddziału przyjmuje wartość „Oddział przygotowawczy". W tego typu oddziałach uczniowie nie są przypisani do konkretnej klasy, co wyjaśnia brak danych w tej kolumnie.

Dodatkowo, wartości brakujące w kolumnie Klasa pojawiają się głównie w szkołach podstawowych i liceach ogólnokształcących, co jest zgodne z charakterem tych placówek – to właśnie one najczęściej prowadzą oddziały przygotowawcze.

Warto również podkreślić, że w całym zbiorze danych znajduje się dokładnie 60 rekordów dotyczących oddziałów przygotowawczych i wszystkie zawierają brak w kolumnie Klasa, co jednoznacznie wskazuje, że braki te nie są przypadkowe, lecz wynikają ze specyfiki struktury organizacyjnej tych jednostek. Z tego względu uznaje się je za braki typu MNAR (Missing Not At Random), ponieważ ich obecność jest bezpośrednio związana z wartością innej zmiennej – Typu Oddziału – i wynika z reguł funkcjonowania systemu edukacji.

Oddziały przygotowawcze to specjalne klasy przeznaczone dla uczniów cudzoziemskich – w tym przypadku uchodźców z Ukrainy – którzy rozpoczynają edukację w polskim systemie oświaty bez znajomości języka polskiego. Ich celem jest umożliwienie uczniom nauki języka polskiego oraz adaptacji do systemu edukacyjnego i kultury, a nie realizacja podstawy programowej w ramach konkretnej klasy. W związku z tym, brak przypisania do poziomu edukacyjnego (Klasa) w takich przypadkach jest uzasadniony merytorycznie i nie powinien być traktowany jako błąd danych.

```
Typ Podmiotu:
Typ Podmiotu
Szkoła podstawowa             34
Liceum ogólnokształcące       17
Technikum                      8
Branżowa szkoła I stopnia      1
Name: count, dtype: int64

Typ Oddziału:
Typ Oddziału
Oddział przygotowawczy    60
Name: count, dtype: int64

Ilość rekordów z oddziałem przygotowawczym: 60
```

Fig 4.5. Rozkład liczby rekordów według typu podmiotu i typu oddziału w przypadkach braków w kolumnie Klasa

W celu uporządkowania braków w kolumnie Klasa podjęto decyzję o ich uzupełnieniu wartością „Wstępna". Taka wartość występuje już w zbiorze danych i odnosi się do podobnych przypadków – uczniów, którzy nie zostali jeszcze formalnie przypisani do klasy zgodnie z polskim systemem edukacji.

Dzięki temu uniknięto obecności wartości brakujących i zachowano spójność semantyczną zbioru. Dodatkowo, dla zgodności ze strukturą zbioru, wszystkim tym rekordom przypisano również wartość idKlasa = 3, tak aby nie pozostawiać braków w tej kolumnie numerycznej.

Kolumna „Typ Podmiotu" zawierała 16 unikalnych wartości opisujących różne typy jednostek oświatowych, m.in. szkoły podstawowe, przedszkola, szkoły artystyczne, technika, licea czy szkoły specjalne (Rysunek Fig. 4.6). Wiele z tych wartości było jednak do siebie zbliżonych pod względem funkcji edukacyjnej (np. liceum, technikum, szkoła branżowa), co sprawiało, że naturalnym krokiem było ich pogrupowanie w szersze kategorie. Takie podejście pozwala uprościć strukturę danych, zwiększyć przejrzystość analiz i ograniczyć nadmiarowość zmiennych kategorycznych, która mogłaby utrudniać proces kodowania i modelowania danych.

W celu uproszczenia tej zmiennej, zastosowano podejście polegające na grupowaniu podobnych kategorii w cztery główne grupy:

- przedszkolne – obejmujące przedszkola i punkty przedszkolne,
- podstawowe – szkoły podstawowe,
- ponadpodstawowe – technika, licea, szkoły policealne, branżowe i specjalne,
- artystyczne – szkoły plastyczne, muzyczne i baletowe.

Dla ułatwienia dalszego przetwarzania utworzono również nową kolumnę numeryczną „idGrupaPodmiotu", będącą zakodowaną wersją grupy podmiotu. Zabieg ten umożliwia dalsze modelowanie i analizę zmiennej bez ryzyka przeuczenia modelu wynikającego z nadmiaru unikalnych etykiet.

```
Unikalne wartości:
=== Typ Podmiotu ===
['Przedszkole' 'Szkoła podstawowa' 'Liceum ogólnokształcące'
 'Punkt przedszkolny' 'Szkoła policealna' 'Technikum'
 'Branżowa szkoła I stopnia' 'Szkoła specjalna przysposabiająca do pracy'
 'Branżowa szkoła II stopnia' 'Liceum sztuk plastycznych'
 'Ogólnokształcąca szkoła muzyczna I stopnia'
 'Ogólnokształcąca szkoła muzyczna II stopnia'
 'Ogólnokształcąca szkoła baletowa' 'Bednarska Szkoła Realna'
 'Zespół wychowania przedszkolnego' 'Poznańska szkoła chóralna']
Liczba unikalnych wartości: 16
```

Fig 4.6. Lista unikalnych typów podmiotów edukacyjnych występujących w zbiorze danych.

Podczas analizy unikalnych wartości w kolumnie Klasa (rysunek fig 4.7) zauważono występowanie zarówno klas oznaczonych cyframi rzymskimi (np. I, II, III), jak i wartości semestralnych (np. sem. I, sem. II, … sem. VIII). Te ostatnie odnoszą się głównie do szkół ponadpodstawowych – takich jak szkoły policealne, licea ogólnokształcące czy szkoły branżowe – w których program nauczania jest podzielony na semestry.

Występowanie wartości semestralnych w szkołach podstawowych było marginalne i ograniczało się do pojedynczych obserwacji, co może świadczyć o błędnym przypisaniu lub specyfice rejestracji danych.

W celu uproszczenia analizy i ograniczenia liczby unikalnych wartości w kolumnie Klasa, zdecydowano się na zastąpienie wszystkich wartości semestralnych jedną wspólną etykietą: semestralna. Zabieg ten pozwolił na:

- redukcję kardynalności zmiennej,
- zachowanie spójności interpretacyjnej pomiędzy różnymi typami szkół,

```
Unikalne wartości:
=== Klasa ===
['Wychowanie przedszkolne' 'I' 'II' 'III' 'IV' 'V' 'VI' 'VII' 'VIII'
 'sem. I' 'sem. II' 'sem. III' 'sem. IV' 'Wstępna' 'sem. VI' 'sem. VIII'
 'sem. V' 'sem. VII' 'IX']
Liczba unikalnych wartości: 19
```

Fig 4.7. Lista unikalnych klas występujących w zbiorze danych.

Dodatkowo zauważono (Fig 4.8) , że oznaczenie np. klasy I występuje zarówno w szkołach podstawowych, jak i ponadpodstawowych (np. w liceum, technikum czy szkole policealnej), co mogło prowadzić do niejednoznaczności interpretacyjnej. W związku z tym podjęto decyzję o rozróżnieniu tych przypadków i wprowadzeniu nowych etykiet w formacie:

- I podstaw., II podstaw., … dla szkół podstawowych,
- I ponadpodst., II ponadpodst., … dla szkół ponadpodstawowych,
- analogicznie: I artyst., II artyst., … dla szkół artystycznych.

Zmodyfikowana kolumna Klasa zachowuje teraz jednolity format i lepiej odzwierciedla strukturalne różnice pomiędzy etapami edukacyjnymi w szkołach podstawowych i ponadpodstawowych.

Grupa Podmiotu	artystyczne	podstawowe	ponadpodstawowe
Klasa			
I	60	735	1380
II	54	758	1320
III	48	755	1183
IV	27	759	411
IX	3	0	0
V	22	743	54
VI	16	742	0
VII	11	744	0
VIII	11	737	0
Wstępna	0	34	33
Wychowanie przedszkolne	0	457	0
semestralna	0	16	918

Fig 4.8. Liczba rekordów w poszczególnych klasach z podziałem na grupy typu szkoły.

Na rysunku Fig. 4.9 przedstawiono rozkład liczby uczniów uchodźców z Ukrainy, którzy posiadają legalny pobyt w Polsce, w podziale na uproszczone grupy typów placówek edukacyjnych. Do wizualizacji zastosowano wykres pudełkowy (boxplot) ze skalą logarytmiczną na osi Y, co umożliwiło lepsze zobrazowanie rozpiętości wartości oraz licznych obserwacji odstających (outliers).

Analiza wykresu pozwala stwierdzić, że największe liczebności uczniów odnotowano w szkołach podstawowych i przedszkolnych – to tam występują najwyższe mediany i szeroki rozkład wartości. W przypadku szkół ponadpodstawowych widoczna jest również duża zmienność, jednak mediana liczby uczniów jest niższa niż w szkołach podstawowych. Szkoły artystyczne charakteryzują się najniższą liczbą uczniów – zarówno pod względem mediany, jak i maksymalnych wartości.

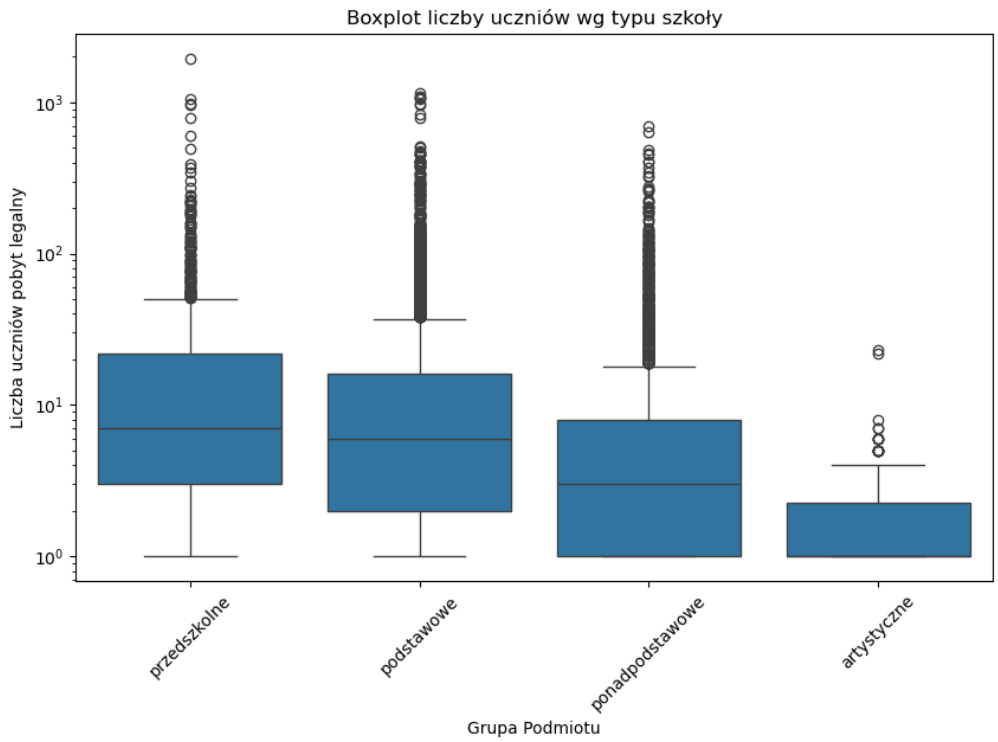

Fig 4.9. Boxplot liczby uczniów z legalnym pobytem w Polsce w podziale na grupy typu szkoły (skala logarytmiczna).

Podczas analizy zauważono obecność pojedynczych bardzo wysokich wartości – przekraczających 1000 uczniów. W związku z tym podjęto decyzję o ich dalszym zbadaniu, aby ustalić, czy są to obserwacje odstające wynikające z błędów w danych, przypadkowych anomalii, czy może mają one uzasadnienie merytoryczne (np. duże zespoły szkolne w aglomeracjach). To działanie miało na celu poprawną interpretację danych oraz rozważenie ewentualnego ich przekształcenia lub wykluczenia z modelowania.

Na podstawie analizy tych rekordów (rysunek Fig. 4.10), można wyciągnąć kilka istotnych wniosków. Po pierwsze, nie są to wartości przypadkowe ani wyniki błędów w danych – wszystkie te obserwacje dotyczą dużych ośrodków miejskich, takich jak Warszawa i Kraków, które ze względu na swoją wielkość, dostępność infrastruktury edukacyjnej oraz status administracyjny naturalnie przyciągają większe grupy uczniów, w tym uchodźców z Ukrainy.

Po drugie, wysokie liczby uczniów pojawiają się głównie w przedszkolach i szkołach podstawowych, co może wskazywać na przewagę młodszych dzieci w populacji uchodźców, a także na lepsze przygotowanie tych placówek do przyjęcia większych grup. W związku z tym wartości te nie są anomaliami.

Rekordy, gdzie Liczba uczniów przekracza wartość 1000

Województwo	Powiat	Typ Podmiotu	Typ Oddziału	Klasa	Liczba uczniów pobyt legalny
MAŁOPOLSKIE	m. Kraków	Przedszkole	Podstawowe	Wychowanie przedszkolne	1060.0
MAZOWIECKIE	m. st. Warszawa	Przedszkole	Podstawowe	Wychowanie przedszkolne	1931.0
MAZOWIECKIE	m. st. Warszawa	Szkoła podstawowa	Podstawowe	III podstaw.	1052.0
MAZOWIECKIE	m. st. Warszawa	Szkoła podstawowa	Podstawowe	IV podstaw.	1074.0
MAZOWIECKIE	m. st. Warszawa	Szkoła podstawowa	Podstawowe	V podstaw.	1088.0
MAZOWIECKIE	m. st. Warszawa	Szkoła podstawowa	Podstawowe	VI podstaw.	1146.0
MAZOWIECKIE	m. st. Warszawa	Szkoła podstawowa	Podstawowe	VII podstaw.	1092.0

Fig 4.10. Rekordy, w których liczba uczniów uchodźców z Ukrainy przekracza 1000 – analiza przypadków o najwyższej liczebności.

Na rysunku Fig. 4.11 oraz 4.12 przedstawiono sumaryczną liczbę uczniów uchodźców z Ukrainy z legalnym pobytem w Polsce w podziale na klasy. Dane zostały zaprezentowane za pomocą wykresu słupkowego z błędami standardowymi, co pozwala na ocenę nie tylko skali wartości, ale i ich zmienności.

Największą liczbę uczniów odnotowano w kategorii „Wychowanie przedszkolne" (33 700) oraz w pierwszej klasie szkół ponadpodstawowych (17 264), co potwierdza obserwację, że najmłodsze dzieci oraz uczniowie rozpoczynający naukę na nowych etapach są najczęściej obejmowani systemem edukacyjnym.

W przypadku klas szkoły podstawowej, liczba uczniów utrzymuje się na relatywnie wysokim poziomie aż do klasy VIII, co wskazuje na dobrą kontynuację nauki przez dzieci uchodźcze w polskim systemie oświaty. Warto zaznaczyć, że największe wartości pojawiają się nie na samym początku, lecz w środkowych klasach szkoły podstawowej (IV–VII), co może być efektem intensyfikacji zapisów dzieci uchodźczych w kolejnych latach po wybuchu wojny oraz elastycznego podejścia szkół do przypisywania uczniów do odpowiedniego poziomu edukacji.

Liczba uczniów w klasach ponadpodstawowych wyraźnie maleje wraz z kolejnymi poziomami nauczania: z 12 168 w klasie II, przez 7 157 w klasie III. Może to wskazywać na większe trudności starszej młodzieży w kontynuowaniu nauki w nowym systemie edukacyjnym, zarówno ze względów językowych, jak i adaptacyjnych.

Dane dotyczące klas artystycznych pokazują, że tego typu placówki mają bardzo ograniczoną liczbę uczniów uchodźców – sumarycznie poniżej 600 osób na wszystkie poziomy klas, co może być skutkiem zarówno niskiej dostępności tych szkół, jak i wymagań rekrutacyjnych.

Warto również zwrócić uwagę na oddziały przygotowawcze („Wstępna" – 3 949 uczniów), które stanowią istotny element wsparcia dla uczniów cudzoziemskich. Ich obecność potwierdza, że system edukacyjny w Polsce reaguje na potrzeby integracyjne poprzez tworzenie klas umożliwiających płynne przejście do regularnych form nauczania.

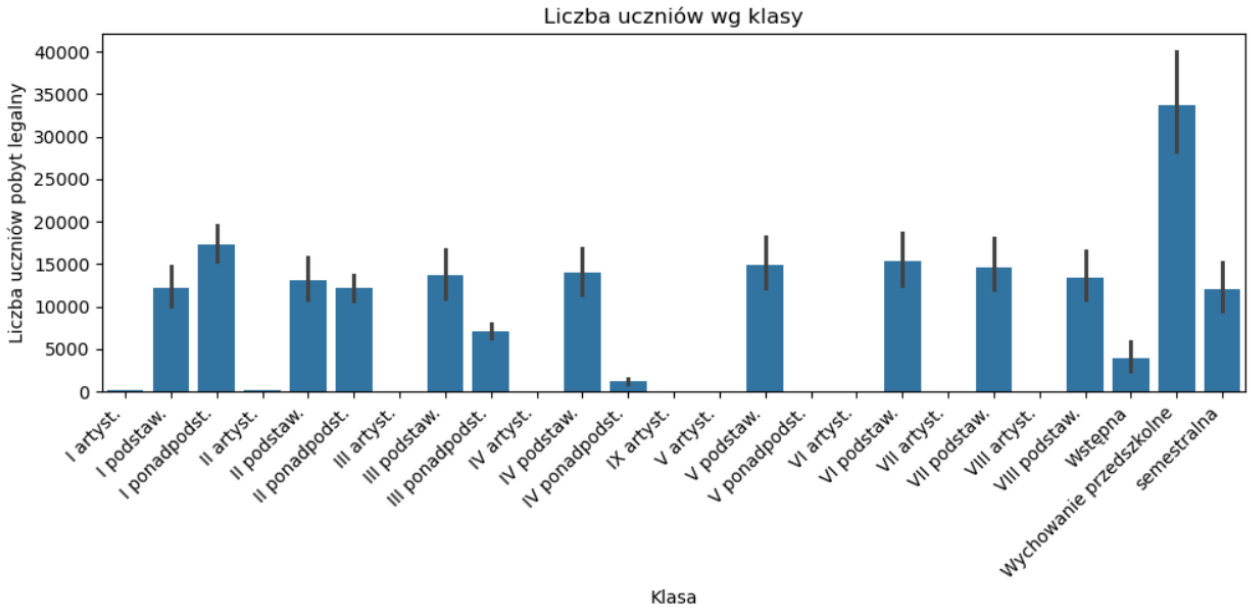

Fig 4.11. Łączna liczba uczniów uchodźców z Ukrainy w poszczególnych klasach na wykresie.

```
Klasa
Wychowanie przedszkolne    33700.0
I ponadpodst.              17264.0
VI podstaw.                15379.0
V podstaw.                 14971.0
VII podstaw.               14630.0
IV podstaw.                13951.0
III podstaw.               13700.0
VIII podstaw.              13340.0
II podstaw.                13133.0
I podstaw.                 12211.0
II ponadpodst.             12168.0
semestralna                12092.0
III ponadpodst.             7157.0
Wstępna                     3949.0
IV ponadpodst.              1193.0
I artyst.                    162.0
II artyst.                   154.0
III artyst.                   90.0
V ponadpodst.                 78.0
IV artyst.                    36.0
V artyst.                     29.0
VI artyst.                    27.0
VII artyst.                   16.0
VIII artyst.                  15.0
IX artyst.                     6.0
Name: Liczba uczniów pobyt legalny, dtype: float64
```

Fig 4.12 . Łączna liczba uczniów uchodźców z Ukrainy w poszczególnych klasach

Na rysunku Fig. 4.13 przedstawiono łączną liczbę uczniów uchodźców z Ukrainy z legalnym pobytem w Polsce, zarejestrowanych w szkołach według województw. Wykres słupkowy poziomy wybrano ze względu na jego czytelność przy prezentacji danych dla wielu kategorii (województw) – ułatwia on bezpośrednie porównanie liczebności pomiędzy regionami.

Z wykresu wynika, że zdecydowanie największa liczba uczniów przypada na województwo mazowieckie, co jest zgodne z jego największym zaludnieniem i obecnością stolicy – Warszawy – będącej głównym punktem recepcyjnym i osiedleńczym dla uchodźców. Kolejne miejsca zajmują województwa: śląskie, dolnośląskie i wielkopolskie, czyli silnie zurbanizowane i rozwinięte regiony, w których dostęp do edukacji i infrastruktury miejskiej jest wysoki.

Województwa o najmniejszej liczbie uczniów uchodźców to: podlaskie, warmińsko-mazurskie i świętokrzyskie. Są to obszary o mniejszym zagęszczeniu ludności, mniejszej liczbie dużych ośrodków miejskich oraz ograniczonych możliwościach przyjęcia dużej liczby migrantów.

Zróżnicowanie przestrzenne pokazuje, że rozmieszczenie uczniów-uchodźców silnie koreluje z ogólną demografią oraz dostępnością usług edukacyjnych i socjalnych w regionie.

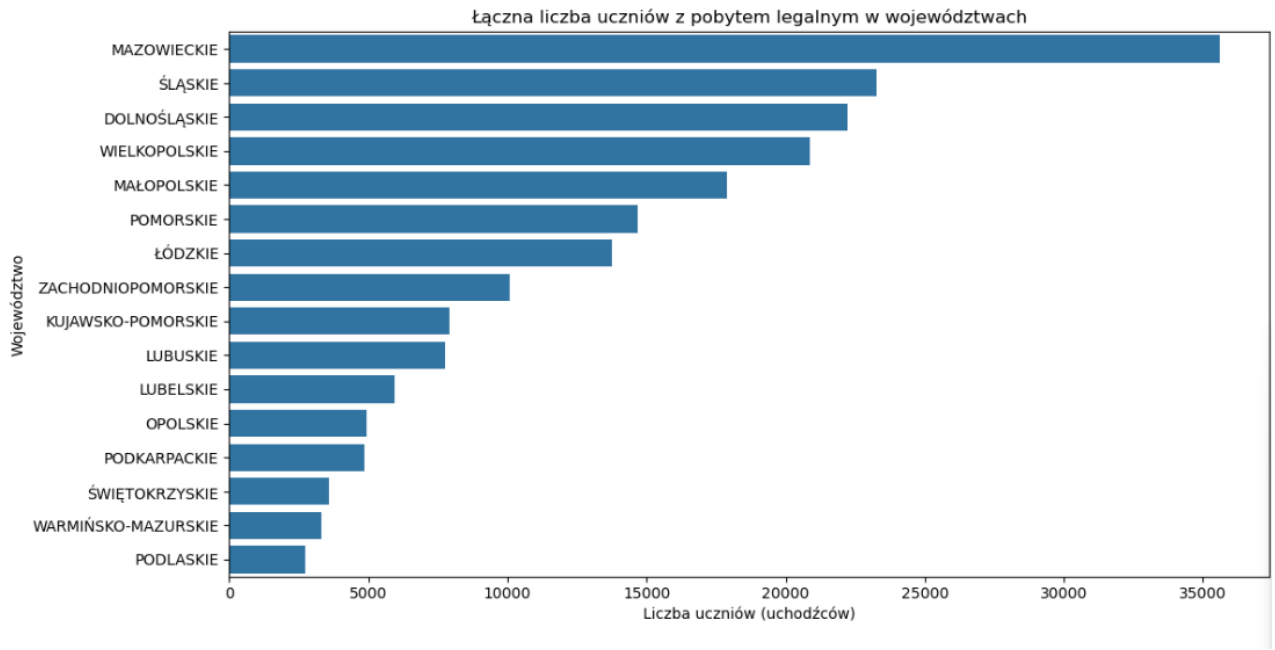

Fig 4.13. Łączna liczba uczniów uchodźców z Ukrainy z legalnym pobytem w podziale na województwa.

W celu uproszczenia struktury danych oraz zwiększenia przejrzystości analiz i modelowania, zdecydowano się na agregację danych według dwóch kluczowych wymiarów: powiatu oraz uproszczonej grupy typu placówki edukacyjnej (przedszkolne, podstawowe, ponadpodstawowe, artystyczne). Celem tej operacji było uzyskanie skondensowanego widoku danych, który lepiej odzwierciedla lokalne potrzeby edukacyjne w poszczególnych regionach.

Zastosowano funkcję groupby() w bibliotece Pandas, sumując wartości liczby uczniów oraz liczby oddziałów w każdej z grup. Pozwoliło to uzyskać nową ramkę danych, w której każda obserwacja reprezentuje łączną liczbę uczniów i oddziałów danego typu szkoły w konkretnym powiecie. Podgląd zagregowanych danych widnieje na Fig 4.14.

Takie podejście:

- redukuje szum i zmienność danych (np. różnice między poszczególnymi klasami czy pojedynczymi szkołami),
- zwiększa stabilność predykcji,
- ułatwia dalsze etapy modelowania.

Dodatkowo, do agregowanych danych dołączono kolumny opisujące lokalizację (Województwo, idTerytWojewodztwo, idTerytPowiat) oraz cechy strukturalne placówek, takie jak Publiczność, idPublicznosc, idGrupaPodmiotu. Umożliwia to uwzględnienie czynników kontekstowych przy budowie modelu predykcyjnego (np. różnice między regionami lub sektorami publicznym i niepublicznym). Dzięki agregacji danych po powiecie i grupie typu szkoły liczba rekordów zmniejszyła się z 13 114 do 1 159, co znacząco uprościło strukturę zbioru.

	Województwo	idTerytWojewodztwo	Powiat	idTerytPowiat	Publiczność	idPublicznosc	Grupa Podmiotu	idGrupaPodmiotu	Liczba oddziałów	Liczba uczniów pobyt legalny
0	KUJAWSKO-POMORSKIE	4.0	aleksandrowski	401.0	publiczna	1.0	podstawowe	1	34.0	66.0
1	KUJAWSKO-POMORSKIE	4.0	aleksandrowski	401.0	publiczna	1.0	ponadpodstawowe	2	16.0	23.0
2	KUJAWSKO-POMORSKIE	4.0	aleksandrowski	401.0	publiczna	1.0	przedszkolne	0	14.0	20.0
3	PODLASKIE	20.0	augustowski	2001.0	publiczna	1.0	podstawowe	1	64.0	128.0
4	PODLASKIE	20.0	augustowski	2001.0	publiczna	1.0	ponadpodstawowe	2	21.0	50.0

Fig 4.14. Przykładowe 4 początkowych wierszy nowej ramki danych (zagregowane dane – liczba uczniów i oddziałów według powiatu i grupy typu placówki edukacyjnej)

Na rysunku Fig. 4.15 przedstawiono średnią liczbę uczniów uchodźców z Ukrainy przypadających na jeden oddział w zależności od grupy typu szkoły, z dodatkową informacją o całkowitej liczbie oddziałów w każdej kategorii. Wybrano taki wykres, aby jednocześnie przedstawić wartość średnią oraz kontekst liczebności próbki w każdej grupie – co pozwala na pełniejszą interpretację wyników.

Z wykresu wynika, że największą średnią liczbą uczniów na oddział charakteryzują się szkoły ponadpodstawowe. Jednocześnie liczba oddziałów w tej grupie (ok. 20 000) jest istotnie mniejsza niż w przypadku szkół podstawowych (ponad 56 000 oddziałów), co oznacza większą koncentrację uczniów w pojedynczych klasach na poziomie ponadpodstawowym.

Szkoły podstawowe mają nieco niższy wskaźnik uczniów na oddział, ale znacznie większą liczbę oddziałów, co może świadczyć o bardziej rozproszonym systemie nauczania i większym rozłożeniu dzieci w mniejszych grupach. W przedszkolach średnia liczba uczniów na oddział jest najniższa, co może być efektem ograniczeń organizacyjnych, jak maksymalna liczba dzieci w grupie, a także większej liczby placówek.

Szkoły artystyczne cechują się najmniejszą liczbą oddziałów (zaledwie 301), jednak średnia liczba uczniów na oddział jest stosunkowo wysoka, co może wynikać ze specyfiki organizacyjnej tych szkół oraz ograniczonego dostępu do tego typu edukacji.

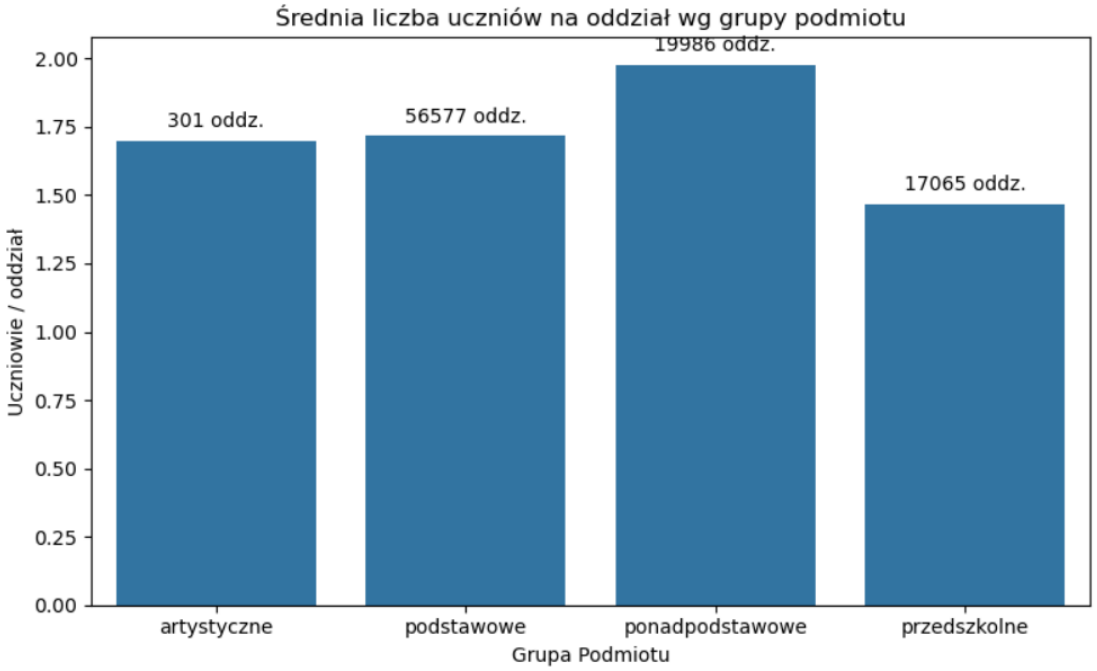

Fig 4.15. Średnia liczba uczniów na oddział w podziale na grupy typów szkół wraz z liczbą oddziałów w każdej grupie

4.4. Wybór zmiennej docelowej (TARGET) do modelowania ML

Jako zmienną docelową (TARGET) wybrano kolumnę „Liczba uczniów pobyt legalny", która zawiera informacje o liczbie uczniów-uchodźców z Ukrainy posiadających legalny pobyt na danym poziomie edukacyjnym, w konkretnej placówce lub grupie placówek (po agregacji). Zmienna ta ma charakter ilościowy (ciągły), co czyni ją odpowiednią do zastosowania w regresji – jednej z podstawowych metod w uczeniu nadzorowanym. Jej wartość jest bezpośrednio związana z celem analizy, jakim jest prognozowanie zapotrzebowania na miejsca w szkołach w zależności od lokalizacji i typu placówki. Ponadto, zmienna ta jest kompletna (brak wartości brakujących) oraz dobrze zróżnicowana, co stanowi ważną cechę w kontekście budowy skutecznego modelu predykcyjnego. Dodatkowo, po dokonaniu agregacji danych względem powiatu oraz grupy typu placówki edukacyjnej, zmienna ta zyskała na znaczeniu – odzwierciedla bowiem lokalne, skumulowane potrzeby edukacyjne, co sprzyja stabilniejszemu modelowaniu i lepszej interpretacji wyników.

4.5. Wybór zmiennych objaśniających (FEATURES) do predykcji zmiennej docelowej

Wybrane zmienne: idTerytWojewodztwo, idPublicznosc, idGrupaPodmiotu oraz Liczba oddziałów zostały uwzględnione jako zmienne objaśniające (FEATURES), ponieważ zawierają kluczowe informacje o lokalizacji placówki, jej charakterze (publiczna/niepubliczna), typie szkoły (np. podstawowa, ponadpodstawowa) oraz skali działania mierzonej liczbą oddziałów. Wszystkie te czynniki mogą realnie wpływać na liczbę uczniów – np. większa liczba oddziałów zwykle wiąże się z większą liczbą dzieci, a lokalizacja (województwo) odzwierciedla zagęszczenie ludności i poziom urbanizacji.

Zmienna idPublicznosc została uwzględniona jako jedna ze zmiennych objaśniających, jednak ze względu na jej kategoryczny charakter, warto rozważyć zastosowanie kodowania typu one-hot encoding. Zabieg ten pozwala na reprezentację każdej kategorii jako oddzielnej zmiennej binarnej, dzięki czemu model nie będzie traktować wartości liczbowych tej zmiennej jako ciągłych lub uporządkowanych. Zapobiega to nadinterpretacji.

Zmienna idTerytPowiat nie została uwzględniona w zbiorze cech, ponieważ zawiera aż 371 unikalnych wartości. Tak wysoka kardynalność w przypadku identyfikatora może prowadzić do przeuczenia modelu (overfittingu), zwłaszcza że sama wartość identyfikatora nie niesie informacji opisowej, a jedynie służy jednoznacznej identyfikacji jednostki administracyjnej.