AGH
Akademia Górniczo-Hutnicza im. Stanisława Staszica w Krakowie
Wydział Geologii, Geofizyki i Ochrony Środowiska

Projekt na ocenę

Aleksandra Jagiełło

kierunek studiów: Inżynieria i Analiza Danych

EDA

Kraków, 2025

Spis treści

1 Wprowadzenie **1**
1.1 Wprowadzenie do biblioteki Seaborn 1
1.2 Platforma dane.gov.pl 1
1.2.1 Sposoby udostępniania danych 2
1.2.2 Rodzaje danych i sposób ich użycia 2
1.2.3 Rodzaje API 2
1.2.4 Możliwość szybkiego sprawdzenia jakości danych 3

2 Dane **4**
2.1 Pobranie danych i wstępna analiza 4
2.2 Opis kolumn danych 4
2.2.1 Identyfikatory i kody punktów pomiarowych 4
2.2.2 Dane geograficzne i administracyjne 5
2.2.3 Charakterystyka hydrogeologiczna i środowiskowa 6
2.2.4 Dane czasowe 6
2.2.5 Wyniki badań fizykochemicznych (parametry terenowe) 6
2.2.6 Wyniki badań fizykochemicznych (parametry laboratoryjne) . 7

3 Metodologia przygotowania i eksploracji danych **10**
3.1 Inżynieria Cech (Feature Engineering) 10
3.1.1 Usunięcie nieistotnych kolumn 10
3.1.2 Obsługa brakujących wartości 10
3.1.3 Ekstrakcja cech czasowych 11
3.1.4 Kodowanie zmiennych kategorycznych (One-Hot Encoding) . . 11
3.2 Eksploracyjna Analiza Danych (EDA) 11
3.2.1 Badanie rozkładu wartości cech (Histogramy) 11
3.2.2 Badanie wartości odstających (Wykresy pudełkowe) 12
3.2.3 Analiza korelacji (Mapa ciepła) 13

4 Wybór zmiennej docelowej i cech predykcyjnych **16**
4.1 Wybór zmiennej docelowej (TARGET) 16
4.1.1 Uzasadnienie wyboru zmiennej docelowej 16
4.2 Wybór podzbioru zmiennych cech (FEATURES) 16
4.2.1 Uzasadnienie wyboru cech 17

5 Podsumowanie **19**

1 Wprowadzenie

Niniejszy raport przedstawia kompleksową analizę danych pochodzących z monitoringu jakości wód podziemnych w Polsce w 2024 roku. Projekt koncentruje się na etapach eksploracyjnej analizy danych (EDA) oraz inżynierii cech (FE), które są kluczowe w procesie przygotowania danych do budowy modeli uczenia maszynowego.

Analiza obejmuje zapoznanie się z biblioteką **Seaborn** do wizualizacji danych, charakterystykę platformy dane.gov.pl jako źródła danych publicznych, wybór i pobranie odpowiedniego zestawu danych, a następnie przeprowadzenie wstępnych etapów potoku uczenia maszynowego. Szczegółowe omówienie obejmuje proces inżynierii cech, w którym dane surowe są transformowane i wzbogacane, oraz eksploracyjną analizę danych, mającą na celu zrozumienie struktury i charakterystyki zbioru.

Dodatkowo, raport zawiera uzasadnienie wyboru zmiennej docelowej (TARGET) oraz zmiennych objaśniających (FEATURES) w kontekście problemu uczenia nadzorowanego, co stanowi podstawę do dalszych prac nad budową modeli predykcyjnych jakości wód podziemnych. Całość analizy ma na celu nie tylko przedstawienie wyników, ale również wyjaśnienie zastosowanych metod i ich znaczenia dla zrozumienia danych środowiskowych.

1.1 Wprowadzenie do biblioteki Seaborn

Seaborn to biblioteka do wizualizacji danych w języku Python, bazująca na bibliotece Matplotlib. Została zaprojektowana w celu ułatwienia tworzenia atrakcyjnych i informatywnych wykresów statystycznych. Seaborn oferuje wysokopoziomowy interfejs do rysowania różnorodnych wykresów, takich jak histogramy, wykresy pudełkowe, wykresy skrzypcowe, mapy ciepła czy wykresy punktowe. Jest szczególnie przydatna w eksploracyjnej analizie danych, umożliwiając szybką i efektywną wizualizację zależności między zmiennymi, rozkładów danych oraz identyfikację wzorców i anomalii.

Kluczowe cechy biblioteki Seaborn obejmują:

- **Wbudowane zestawy danych:** Ułatwiają eksperymentowanie i naukę.
- **Integracja z pandas:** Bezproblemowa współpraca z obiektami DataFrame.
- **Estetyczne domyślne style:** Wykresy prezentują się profesjonalnie od razu po wygenerowaniu.
- **Funkcje do wizualizacji danych statystycznych:** Umożliwiają czytelne i intuicyjne przedstawienie zależności między zmiennymi.

1.2 Platforma dane.gov.pl

Portal **dane.gov.pl** stanowi centralne miejsce udostępniania otwartych danych publicznych w Polsce. Jego głównym celem jest zapewnienie dostępu do zasobów in-

formacyjnych sektora publicznego, co ma na celu promowanie transparentności, innowacyjności oraz rozwój usług opartych na danych.

1.2.1 Sposoby udostępniania danych

Dane na portalu są udostępniane głównie w trzech formach:

- **Pobieranie plików:** Najpopularniejsza metoda, umożliwiająca użytkownikom pobieranie zestawów danych w różnych formatach, takich jak CSV, XLSX, JSON, XML, KML czy SHP. Jest to najprostsza forma dostępu, idealna do jednorazowej analizy lub wykorzystania danych w trybie offline.
- **API (Application Programming Interface):** Niektóre zestawy danych są dostępne poprzez API, co pozwala na automatyczny i programowy dostęp do danych. API umożliwia dynamiczne pobieranie aktualnych danych bez konieczności ręcznego ściągania plików.
- **Wizualizacje:** Część danych jest prezentowana w formie interaktywnych wykresów, map i tabel, co ułatwia wstępne zapoznanie się z zawartością zbioru bez konieczności jego pobierania.

1.2.2 Rodzaje danych i sposób ich użycia

Na portalu dostępny jest bardzo szeroki zakres danych, obejmujący między innymi:

- **Dane statystyczne:** Demografia, gospodarka, edukacja, zdrowie.
- **Dane przestrzenne:** Mapy, dane geolokalizacyjne.
- **Dane budżetowe i finansowe:** Wydatki publiczne, dotacje.
- **Dane środowiskowe:** Jakość powietrza, wód, monitoring przyrody.
- **Dane publiczne:** Informacje o instytucjach, rejestry.

Dane te mogą być używane do wielu celów, takich jak analizy badawcze, tworzenie aplikacji i usług, dziennikarstwo danych, edukacja czy wspieranie decyzji biznesowych.

1.2.3 Rodzaje API

Na portalu dane.gov.pl nie istnieje jeden uniwersalny rodzaj API. Udostępniane interfejsy API są zazwyczaj specyficzne dla danego zestawu danych lub dostawcy. Często są to **RESTful API**, które zwracają dane w formatach takich jak JSON lub XML. Czasami dostępne są również API bazujące na standardach geoprzestrzennych, np. WMS (Web Map Service) czy WFS (Web Feature Service) dla danych przestrzennych.

1.2.4 Możliwość szybkiego sprawdzenia jakości danych

Możliwość szybkiego sprawdzenia jakości danych na portalu jest ograniczona i w dużej mierze zależy od dostawcy danych. Portal dane.gov.pl nie oferuje wbudowanego narzędzia do automatycznej oceny jakości danych (np. do wykrywania brakujących wartości, duplikatów czy niespójności). Weryfikacja jakości danych zazwyczaj wymaga **samodzielnej analizy eksploracyjnej** po pobraniu zestawu danych.

2 Dane

Do analizy wybrano zestaw danych o nazwie **"2024 - Wyniki badań wskaźników fizykochemicznych monitoringu jakości wód podziemnych - monitoring operacyjny"**. Zestaw ten jest istotny ze względu na jego znaczenie dla monitoringu środowiskowego i zarządzania zasobami wodnymi. Wody podziemne stanowią kluczowy element ekosystemu oraz są źródłem wody pitnej, dlatego ich jakość jest niezwykle ważna.

2.1 Pobranie danych i wstępna analiza

Dane zostały pobrane bezpośrednio z portalu dane.gov.pl. Zgodnie z opisem na stronie, plik zawiera wyniki badań fizykochemicznych wód podziemnych z 2024 roku, w ramach monitoringu operacyjnego. Opis wskazuje, że dane obejmują różne parametry chemiczne i fizyczne wody, mierzone w konkretnych punktach pomiarowych.

Wstępne zapoznanie się z danymi za pomocą biblioteki `pandas` pozwala na szybką ocenę struktury pliku. Poniżej przedstawiono kod użyty do wczytania danych oraz wyświetlenia podstawowych informacji.

```
import pandas as pd

# Wczytanie danych
df = pd.read_csv('2024_Wyniki_badań_wskaźników_fizykochemicznych_
monitoringu_jakości_wód_podziemnych.csv')

print("\nPierwsze 5 wierszy danych:")
display(df.head())
print("\nInformacje o kolumnach i typach danych:")
display(df.info())
print("\nStatystyki opisowe danych numerycznych:")
display(df.describe())
```

2.2 Opis kolumn danych

Zestaw danych zawiera kompleksowe informacje dotyczące monitoringu jakości wód podziemnych, obejmujące zarówno dane identyfikacyjne punktów pomiarowych, ich lokalizację, charakterystykę hydrogeologiczną, jak i szczegółowe wyniki badań fizykochemicznych. Kolumny można podzielić na następujące kategorie:

2.2.1 Identyfikatory i kody punktów pomiarowych

Kolumny te służą do jednoznacznej identyfikacji punktów pomiarowych w różnych systemach klasyfikacyjnych i bazach danych. Są to głównie zmienne kategoryczne (nominalne), choć mogą zawierać wartości numeryczne pełniące rolę etykiet.

- **L.p.:** Numer porządkowy rekordu w zestawie danych. Typowo zmienna numeryczna (całkowita).
- **Numer JCWPd (wg podziału na 161 części):** Numer Jednolitej Części Wód Podziemnych (JCWPd). Kategoryczna/numeryczna.
- **Kod UE JCWPd (wg podziału na 161 części):** Kategoryczna (tekstowa).
- **Identyfikator UE punktu pomiarowego (wg podziału na 161 części):** Kategoryczna (tekstowa).
- **Numer JCWPd (wg podziału na 172 części):** Kategoryczna/numeryczna.
- **Kod UE JCWPd (wg podziału na 172 części):** Kategoryczna (tekstowa).
- **Identyfikator UE punktu pomiarowego (wg podziału JCWPd na 172 części):** Kategoryczna (tekstowa).
- **Numer JCWPd (wg podziału na 174 części):** Kategoryczna/numeryczna.
- **Kod UE JCWPd (wg podziału na 174 części):** Kategoryczna (tekstowa).
- **Identyfikator UE punktu pomiarowego (wg podziału JCWPd na 174 części):** Kategoryczna (tekstowa).
- **Numer punktu pomiarowego wg ID Monitoring:** Kategoryczna/numeryczna.
- **Numer punktu pomiarowego wg MONBADA:** Kategoryczna/numeryczna.
- **Numer punktu pomiarowego wg SOH/SOBWP:** Kategoryczna/numeryczna.
- **Numer punktu pomiarowego wg CBDH:** Kategoryczna/numeryczna.

2.2.2 Dane geograficzne i administracyjne

Kolumny te dostarczają informacji o lokalizacji punktów pomiarowych, co jest kluczowe dla analiz przestrzennych i kontekstowych.

- **PUWG 1992 X:** Współrzędna X punktu pomiarowego w układzie PUWG 1992. Zmienna numeryczna (ciągła).
- **PUWG 1992 Y:** Współrzędna Y punktu pomiarowego w układzie PUWG 1992. Zmienna numeryczna (ciągła).
- **Województwo:** Nazwa województwa, w którym znajduje się punkt pomiarowy. Kategoryczna (nominalna).

- **Powiat:** Nazwa powiatu. Kategoryczna (nominalna).
- **Gmina:** Nazwa gminy. Kategoryczna (nominalna).
- **Miejscowość:** Nazwa miejscowości. Kategoryczna (nominalna).
- **Nazwa dorzecza:** Nazwa dorzecza. Kategoryczna (nominalna).
- **RZGW:** Regionalny Zarząd Gospodarki Wodnej. Kategoryczna (nominalna).

2.2.3 Charakterystyka hydrogeologiczna i środowiskowa

Kolumny te opisują specyfikę warstwy wodonośnej i otoczenia punktu pomiarowego.

- **Stratygrafia:** Opis warstwy geologicznej, z której pobierane są próbki. Kategoryczna (nominalna).
- **Głębokość do stropu warstwy wodonośnej [m p.p.t.]:** Głębokość od powierzchni terenu do górnej granicy warstwy wodonośnej. Zmienna numeryczna (ciągła).
- **Przedział ujętej warstwy wodonośnej [m p.p.t.]:** Zakres głębokości, z którego pobierana jest woda. Zmienna numeryczna (ciągła).
- **Zwierciadło wody:** Poziom zwierciadła wody. Kategoryczna (nominalna).
- **Typ ośrodka wodonośnego:** np. szczelinowy, porowy. Kategoryczna (nominalna).
- **Rodzaj punktu pomiarowego:** Np. piezometr, studnia. Kategoryczna (nominalna).
- **Użytkowanie terenu:** Np. rolnicze, leśne, miejskie. Kategoryczna (nominalna).

2.2.4 Dane czasowe

- **Rok badań:** Rok, w którym przeprowadzono badania. Zmienna numeryczna (całkowita).
- **Data poboru próbki:** Dokładna data poboru próbki. Zmienna czasowa (datetime).

2.2.5 Wyniki badań fizykochemicznych (parametry terenowe)

Kolumny te zawierają wyniki pomiarów wykonanych bezpośrednio w terenie. Są to zmienne numeryczne (ciągłe).

- **Przewodność elektrolityczna właściwa w 20°C - wartość terenowa [µS/cm]:** Pomiar zdolności wody do przewodzenia prądu elektrycznego. Jest to wskaźnik ogólnej zawartości rozpuszczonych soli.
- **Odczyn pH - wartość terenowa:** Miara kwasowości lub zasadowości wody.
- **Temperatura - wartość terenowa [°C]:** Temperatura wody.
- **Tlen rozpuszczony - wartość terenowa [mgO2/l]:** Ilość tlenu rozpuszczonego w wodzie. Kluczowy wskaźnik stanu ekologicznego.

2.2.6 Wyniki badań fizykochemicznych (parametry laboratoryjne)

Jest to rdzeń danych, zawierający wyniki szczegółowych analiz chemicznych wykonanych w laboratorium. Większość to zmienne numeryczne (ciągłe), często z małymi wartościami (stężenia śladowe).

- **Przewodność elektrolityczna właściwa w 20°C - wartość laboratoryjna [µS/cm]:** Wartość przewodności uzyskana w laboratorium.
- **Odczyn pH - wartość laboratoryjna:** Wartość pH uzyskana w laboratorium.
- **Ogólny węgiel organiczny [mgC/l]:** Suma wszystkich związków organicznych zawierających węgiel. Wskaźnik zanieczyszczenia organicznego.
- **Amonowy jon [mgNH4/l]:** Stężenie jonów amonowych, wskaźnik zanieczyszczenia organicznego, często związany z procesami rozkładu.
- **Antymon [mgSb/l]:** Stężenie pierwiastka śladowego.
- **Arsen [mgAs/l]:** Stężenie pierwiastka śladowego, toksyczny.
- **Azotany [mgNO3/l]:** Stężenie azotanów, wskaźnik zanieczyszczenia nawozami lub ściekami.
- **Azotyny [mgNO2/l]:** Stężenie azotynów, również wskaźnik zanieczyszczenia, często przejściowy etap cyklu azotowego.
- **Bar [mgBa/l]:** Stężenie pierwiastka śladowego.
- **Beryl [mgBe/l]:** Stężenie pierwiastka śladowego.
- **Bor [mgB/l]:** Stężenie pierwiastka śladowego.
- **Chlorki [mgCl/l]:** Stężenie chlorków, często związane z zanieczyszczeniami antropogenicznymi lub naturalnym zasoleniem.
- **Chrom [mgCr/l]:** Stężenie pierwiastka śladowego.

- **Cyjanki wolne [mgCN/l]:** Stężenie cyjanków, toksyczne.
- **Cyna [mgSn/l]:** Stężenie pierwiastka śladowego.
- **Cynk [mgZn/l]:** Stężenie pierwiastka śladowego.
- **Fluorki [mgF/l]:** Stężenie fluorków.
- **Fosforany [mgPO4/l]:** Stężenie fosforanów, wskaźnik zanieczyszczenia nawozami lub detergentami.
- **Glin [mgAl/l]:** Stężenie pierwiastka śladowego.
- **Kadm [mgCd/l]:** Stężenie pierwiastka śladowego, toksyczny.
- **Kobalt [mgCo/l]:** Stężenie pierwiastka śladowego.
- **Magnez [mgMg/l]:** Stężenie magnezu, ważny składnik wody.
- **Mangan [mgMn/l]:** Stężenie manganu.
- **Miedź [mgCu/l]:** Stężenie pierwiastka śladowego.
- **Molibden [mgMo/l]:** Stężenie pierwiastka śladowego.
- **Nikiel [mgNi/l]:** Stężenie pierwiastka śladowego.
- **Ołów [mgPb/l]:** Stężenie pierwiastka śladowego, toksyczny.
- **Potas [mgK/l]:** Stężenie potasu.
- **Rtęć [mgHg/l]:** Stężenie pierwiastka śladowego, toksyczny.
- **Selen [mgSe/l]:** Stężenie pierwiastka śladowego.
- **Siarczany [mgSO4/l]:** Stężenie siarczanów.
- **Sód [mgNa/l]:** Stężenie sodu.
- **Srebro [mgAg/l]:** Stężenie pierwiastka śladowego.
- **Tal [mgTl/l]:** Stężenie pierwiastka śladowego.
- **Tytan [mgTi/l]:** Stężenie pierwiastka śladowego.
- **Uran [mgU/l]:** Stężenie pierwiastka śladowego, potencjalnie radioaktywny.
- **Wanad [mgV/l]:** Stężenie pierwiastka śladowego.
- **Wapń [mgCa/l]:** Stężenie wapnia, ważny składnik wody.
- **Wodorowęglany [mgHCO3/l]:** Stężenie wodorowęglanów, wpływają na twardość i buforowanie pH.

- **Żelazo [mgFe/l]:** Stężenie żelaza.
- **Węglany CO32- [mgCO32-/l]:** Stężenie węglanów.

3 Metodologia przygotowania i eksploracji danych

Przygotowanie danych do modelowania uczenia maszynowego oraz ich wstępna eksploracja są kluczowymi etapami każdego projektu analitycznego. Celem tych działań jest przekształcenie surowych danych w format zrozumiały dla algorytmów oraz uzyskanie głębszego wglądu w ich strukturę i zależności.

3.1 Inżynieria Cech (Feature Engineering)

Inżynieria cech to proces tworzenia nowych zmiennych lub przekształcania istniejących w celu poprawy wydajności modeli uczenia maszynowego.

3.1.1 Usunięcie nieistotnych kolumn

W pierwszym kroku usunięto kolumny, które nie niosą ze sobą wartości predykcyjnej lub stanowią jedynie identyfikatory. Kolumny te, takie jak 'L.p.', 'Numer JCWPd', 'Kod UE JCWPd', 'Identyfikator UE punktu pomiarowego' (dla wszystkich podziałów), 'Numer punktu pomiarowego wg ID Monitoring', 'Numer punktu pomiarowego wg MONBADA', 'Numer punktu pomiarowego wg SOH/SOBWP' oraz 'Numer punktu pomiarowego wg CBDH', zostały wyeliminowane w celu zredukowania wymiarowości danych i uniknięcia szumu informacyjnego. Współrzędne geograficzne (PUWG 1992 X, PUWG 1992 Y) zostały celowo zachowane, ponieważ mogą zawierać cenne informacje przestrzenne o rozkładzie mierzonych parametrów.

3.1.2 Obsługa brakujących wartości

Przed dalszymi operacjami dokonano uzupełnienia brakujących wartości w zbiorze danych. Obserwowane braki dotyczyły głównie kolumn numerycznych, takich jak 'Głębokość do stropu warstwy wodonośnej [m p.p.t.]' oraz 'Węglany CO32-[mgCO32-/l]'.

Dla kolumn numerycznych brakujące wartości zostały uzupełnione **medianą** danej kolumny. Mediana, jako miara odporna na wartości odstające, jest preferowana nad średnią, co zapewnia stabilne uzupełnienie danych. W przypadku kolumn kategorycznych (tekstowych), brakujące wartości zostały uzupełnione **modą** (najczęściej występującą wartością), co jest standardową praktyką dla tego typu danych. Ta operacja jest fundamentalna dla zapewnienia kompletności danych, ponieważ wiele algorytmów uczenia maszynowego nie radzi sobie z brakującymi wartościami.

Wnioski: Uzupełnienie brakujących wartości jest kluczowym krokiem w przygotowaniu danych. Jak widać w logu programu, udało się uzupełnić wszystkie braki, co jest niezbędne dla dalszych analiz i modelowania.

3.1.3 Ekstrakcja cech czasowych

Z kolumny 'Data poboru próbki' wyodrębniono nowe, bardziej użyteczne cechy czasowe: 'Rok poboru próbki', 'Miesiąc poboru próbki' oraz 'Dzień poboru próbki'. Konwersja oryginalnej kolumny daty na format `datetime` umożliwia łatwe wyciąganie tych komponentów. Cechy te mogą pomóc w uchwyceniu sezonowości, trendów długoterminowych lub innych wzorców czasowych, które mogą wpływać na jakość wody. Po ekstrakcji, oryginalna kolumna 'Data poboru próbki' została usunięta.

3.1.4 Kodowanie zmiennych kategorycznych (One-Hot Encoding)

Zmienne kategoryczne (nominalne), takie jak 'Województwo', 'Powiat', 'Gmina', 'Miejscowość', 'Nazwa dorzecza', 'RZGW', 'Stratygrafia', 'Zwierciadło wody', 'Typ ośrodka wodonośnego', 'Rodzaj punktu pomiarowego' oraz 'Użytkowanie terenu', zostały przekształcone w format numeryczny przy użyciu metody **One-Hot Encoding**. Technika ta polega na utworzeniu nowych kolumn binarnych dla każdej unikalnej kategorii, gdzie wartość 1 oznacza obecność danej kategorii, a 0 jej brak. Jest to niezbędne, ponieważ algorytmy uczenia maszynowego operują na danych numerycznych. Dodatkowo, zastosowano parametr `drop_first=True`, aby zapobiec problemowi współliniowości (ang. *multicollinearity*).

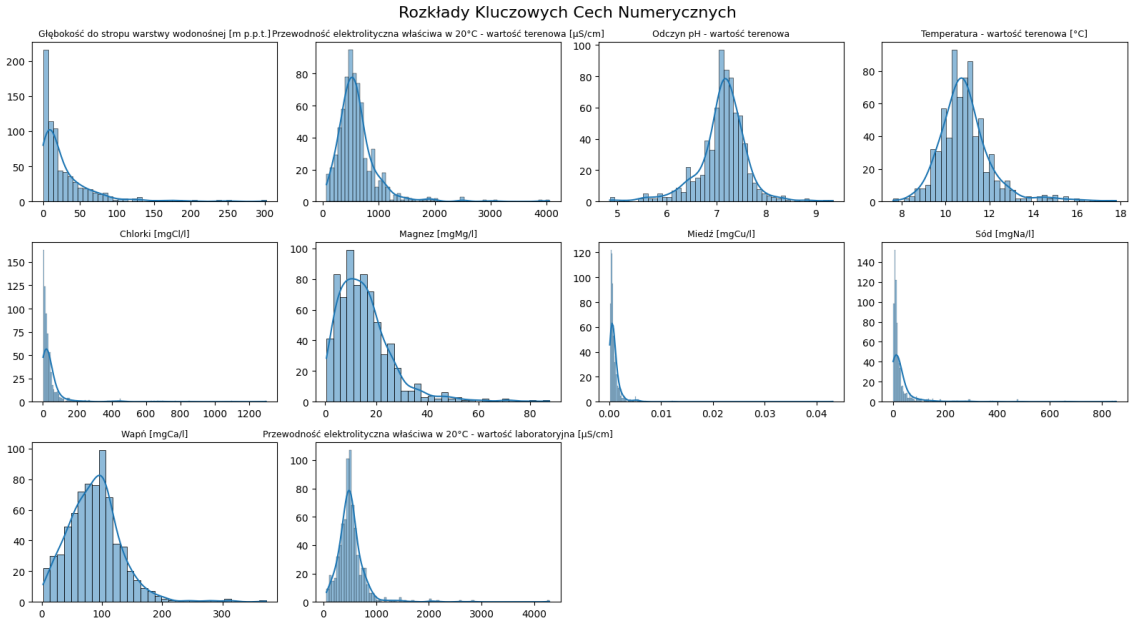
Wnioski: One-Hot Encoding znacząco zwiększył liczbę kolumn w zbiorze danych (do 5114), co jest typowe dla tej metody przy dużej liczbie unikalnych kategorii w zmiennych kategorycznych. Jest to niezbędny krok, ale może zwiększyć złożoność obliczeniową przyszłych modeli.

3.2 Eksploracyjna Analiza Danych (EDA)

Eksploracyjna Analiza Danych (EDA) jest procesem analizowania zbiorów danych w celu podsumowania ich głównych cech, często przy użyciu metod wizualnych. Pozwala to na zrozumienie struktury danych, identyfikację wzorców, wykrywanie anomalii oraz testowanie hipotez.

3.2.1 Badanie rozkładu wartości cech (Histogramy)

Rozkłady kluczowych cech numerycznych zostały zwizualizowane za pomocą **histogramów**. Histogramy umożliwiają ocenę kształtu rozkładu danych (np. normalny, skośny, dwumodalny), co jest istotne dla wyboru odpowiednich metod statystycznych i algorytmów uczenia maszynowego.

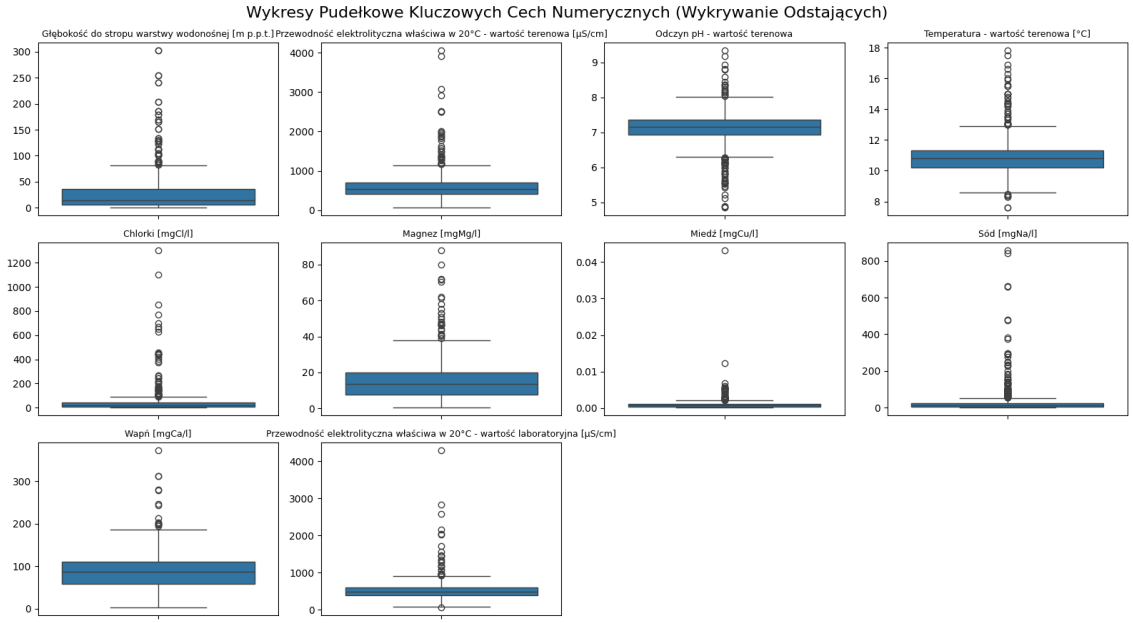


Rysunek 1: Rozkłady kluczowych cech numerycznych (histogramy).

Wnioski z histogramów: Wiele z analizowanych rozkładów, takich jak 'Przewodność elektrolityczna właściwa w 20°C - wartość terenowa', 'Chlorki [mgCl/l]', 'Magnez [mgMg/l]' czy 'Przewodność elektrolityczna właściwa w 20°C - wartość laboratoryjna', wykazuje silną **skośność prawostronną**. Oznacza to, że większość wartości jest skoncentrowana w niższym zakresie, z długim "ogonem"rozciągającym się w kierunku wyższych wartości. Taka charakterystyka może wymagać zastosowania transformacji danych (np. logarytmicznej) przed modelowaniem, aby poprawić wydajność niektórych algorytmów. Rozkłady 'Odczyn pH - wartość terenowa' i 'Temperatura - wartość terenowa [°C]' wydają się być bardziej zbliżone do rozkładu normalnego lub lekko skośnego.

3.2.2 Badanie wartości odstających (Wykresy pudełkowe)

Wizualizacja **wykresów pudełkowych (boxplotów)** dla wybranych cech numerycznych umożliwiła identyfikację wartości odstających. Wartości odstające to obserwacje, które znacznie odbiegają od pozostałych danych i mogą negatywnie wpływać na wydajność modeli.

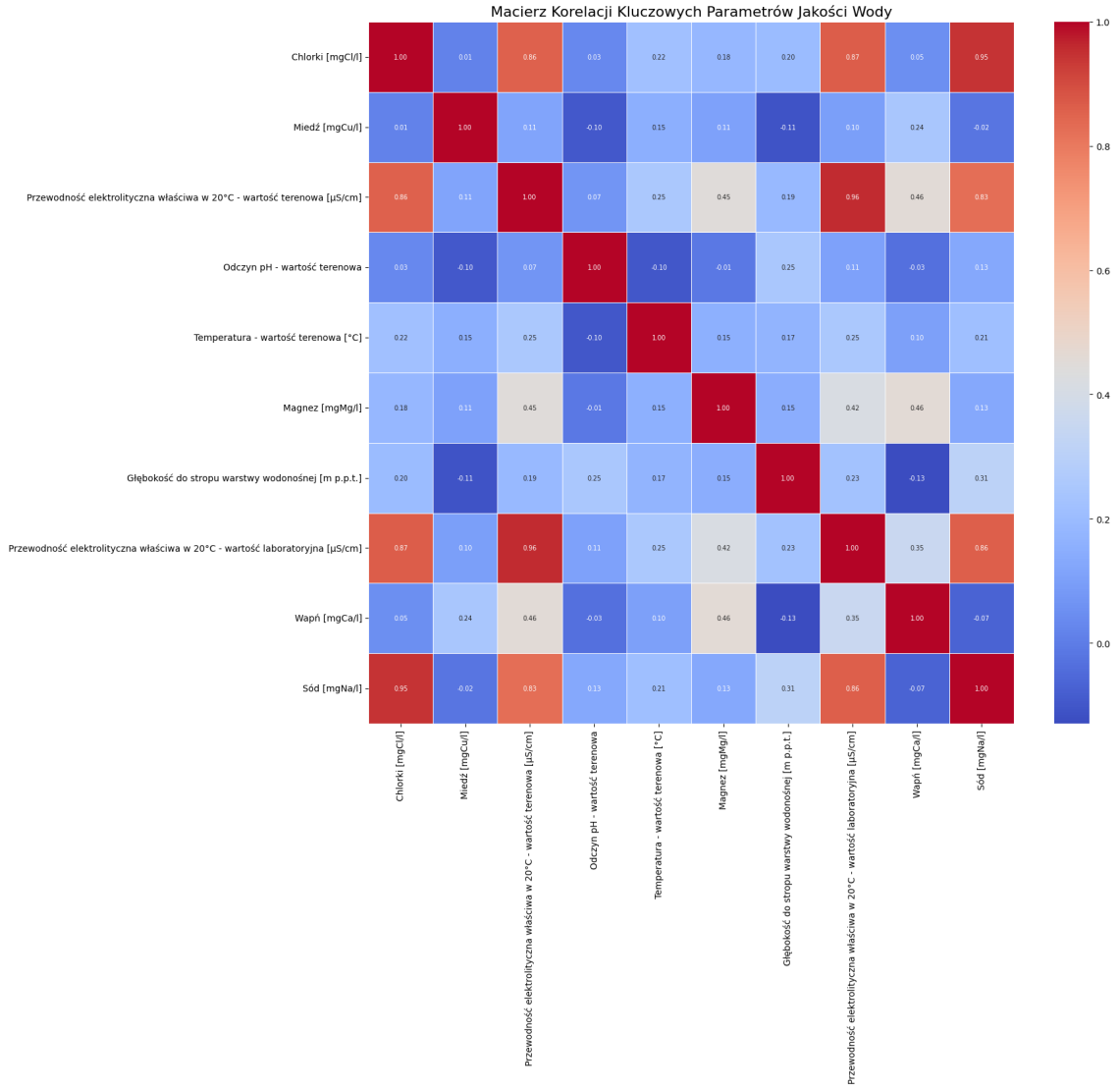


Rysunek 2: Wykresy pudełkowe kluczowych cech numerycznych (wykrywanie odstających).

Wnioski z wykresów pudełkowych: Wykresy pudełkowe jasno wskazują na obecność **licznych wartości odstających** w większości analizowanych kolumn numerycznych. Jest to szczególnie widoczne dla 'Głębokość do stropu warstwy wodonośnej [m p.p.t.]', 'Przewodność elektrolityczna właściwa w 20°C - wartość terenowa [µS/cm]', 'Chlorki [mgCl/l]', 'Magnez [mgMg/l]', 'Sód [mgNa/l]', 'Wapń [mgCa/l]' oraz 'Przewodność elektrolityczna właściwa w 20°C - wartość laboratoryjna [µS/cm]'. Wartości odstające mogą być wynikiem błędów pomiarowych, nietypowych zdarzeń lub rzeczywistych, ekstremalnych obserwacji. Ich obecność wymaga dalszej uwagi i może prowadzić do konieczności zastosowania metod ich obsługi (np. winsoryzacji, transformacji, usuwania) w zależności od wybranego algorytmu modelowania.

3.2.3 Analiza korelacji (Mapa ciepła)

Macierz korelacji między cechami numerycznymi została zwizualizowana za pomocą **mapy ciepła**. Mapa ciepła pozwala na graficzne przedstawienie siły i kierunku liniowych zależności między parametrycznymi zmiennymi. Silne korelacje między cechami mogą wskazywać na redundancję informacji, co jest istotne przy selekcji cech i unikaniu problemów z multikolinearnością w modelach regresji.



Rysunek 3: Macierz korelacji kluczowych parametrów jakości wody.

Wnioski z mapy ciepła korelacji: Mapa ciepła ujawnia kilka interesujących zależności.

- Stwierdzono **bardzo silną pozytywną korelację** (0.96) pomiędzy 'Przewodność elektrolityczna właściwa w 20°C - wartość terenowa [µS/cm]' a zmienną docelową 'Przewodność elektrolityczna właściwa w 20°C - wartość laboratoryjna [µS/cm]'. Jest to zgodne z oczekiwaniami, ponieważ są to pomiary tego samego parametru w różnych warunkach (teren vs. laboratorium).

- Występuje **silna pozytywna korelacja** zmiennej docelowej z 'Chlorki [mgCl/l]' (0.87), 'Sód [mgNa/l]' (0.86),. To potwierdza, że stężenia jonów rozpuszczonych są kluczowymi determinantami przewodności wody.

- Zauważalna jest również **silna pozytywna korelacja** między 'Chlorki [mgCl/l]' a 'Sód [mgNa/l]' (0.95), co jest typowe dla wód zawierających chlorek sodu.

- 'Odczyn pH - wartość terenowa' oraz 'Temperatura - wartość terenowa [°C]' wykazują stosunkowo **słabe korelacje** z większością pozostałych cech, w tym ze zmienną docelową. Oznacza to, że ich wpływ na przewodność, choć ważny, nie jest liniowo dominujący w porównaniu do stężeń jonów.

Analiza korelacji potwierdza, że parametry jonowe są kluczowymi predyktorami przewodności, a także wskazuje na potencjalną multikolinearność między niektórymi cechami (np. chlorki i sód), co należy wziąć pod uwagę przy wyborze modelu.

4 Wybór zmiennej docelowej i cech predykcyjnych

Niniejsza sekcja przedstawia wybór zmiennej docelowej (TARGET) dla zadania uczenia nadzorowanego oraz uzasadnienie wyboru podzbioru zmiennych objaśniających (FEATURES), które mogą być wykorzystane do budowy modelu.

4.1 Wybór zmiennej docelowej (TARGET)

Dla celów modelowania uczenia nadzorowanego, jako zmienną docelową wybrano kolumnę **'Przewodność elektrolityczna właściwa w 20°C - wartość laboratoryjna [µS/cm]'**.

4.1.1 Uzasadnienie wyboru zmiennej docelowej

Wybór tej zmiennej podyktowany jest kilkoma kluczowymi względami:

- **Ważność w kontekście jakości wody:** Przewodność elektrolityczna jest fundamentalnym i powszechnie stosowanym wskaźnikiem jakości wody. Odzwierciedla ona ogólną zawartość rozpuszczonych soli i jonów, co ma bezpośrednie przełożenie na ocenę przydatności wody do różnych celów. Prognozowanie tej wartości ma istotne zastosowanie praktyczne w monitoringu środowiska, zarządzaniu zasobami wodnymi oraz ocenie ryzyka zanieczyszczeń.
- **Charakterystyka numeryczna:** Zmienna ta jest zmienną ciągłą, co kwalifikuje problem jako zadanie **regresji**. Regresja jest typowym zastosowaniem uczenia maszynowego, pozwalającym na przewidywanie wartości numerycznych na podstawie dostępnych cech.
- **Dostępność danych:** Kolumna 'Przewodność elektrolityczna właściwa w 20°C - wartość laboratoryjna [µS/cm]' jest stosunkowo dobrze wypełniona i zawiera dane w odpowiednim formacie, co jest kluczowe dla efektywnego trenowania modelu.
- **Potencjalne zależności:** Spodziewana jest silna zależność między tą zmienną a innymi parametrami fizykochemicznymi wody oraz czynnikami środowiskowymi. Oznacza to, że pozostałe dostępne cechy mogą skutecznie posłużyć do jej predykcji. Analiza korelacji (Sekcja 3.2.3) dodatkowo potwierdza silne liniowe zależności tej zmiennej z innymi parametrami chemicznymi.

4.2 Wybór podzbioru zmiennych cech (FEATURES)

Wybór zmiennych objaśniających (FEATURES), które zostaną użyte do prognozowania zmiennej docelowej ('Przewodność elektrolityczna właściwa w 20°C - wartość laboratoryjna [µS/cm]'), oparto na wiedzy dziedzinowej z zakresu hydrogeologii i

chemii wody, a także na wstępnej analizie danych (EDA), bez wykorzystywania zaawansowanych algorytmów selekcji cech.

4.2.1 Uzasadnienie wyboru cech

Wybrano następujące kategorie cech, które na podstawie wiedzy eksperckiej są najbardziej prawdopodobne do wpływania na przewodność elektrolityczną wody:

- **Parametry fizykochemiczne (wartości terenowe i laboratoryjne):**
  - **'Odczyn pH - wartość terenowa', 'Temperatura - wartość terenowa [°C]', 'Tlen rozpuszczony - wartość terenowa [mgO_2/l]':** Te parametry mierzone w terenie stanowią podstawowe wskaźniki fizykochemiczne. Temperatura ma bezpośredni wpływ na przewodność, a pH wpływa na rozpuszczalność jonów. Tlen rozpuszczony pośrednio wskazuje na procesy biologiczne i chemiczne.
  - **'Ogólny węgiel organiczny [mgC/l]', 'Amonowy jon [mgNH_4/l]', 'Azotany [mgNO_3/l]', 'Azotyny [mgNO_2/l]':** Jony te oraz zawartość węgla organicznego są kluczowymi składnikami chemicznymi wód podziemnych, których stężenie bezpośrednio wpływa na przewodność.
  - **Jony główne: 'Chlorki [mgCl/l]', 'Siarczany [mgSO_4/l]', 'Sód [mgNa/l]', 'Wapń [mgCa/l]', 'Magnez [mgMg/l]', 'Potas [mgK/l]', 'Węglany CO_3^{2-} [mgCO_3^{2-}/l]', 'Wodorowęglany [mgHCO_3/l]':** Te jony są głównymi składnikami rozpuszczonymi w wodzie i w znacznym stopniu decydują o jej przewodności elektrolitycznej. Korelacje potwierdzają ich silny związek ze zmienną docelową.
  - **Metale ciężkie i pierwiastki śladowe (np. 'Antymon [mgSb/l]', 'Arsen [mgAs/l]', 'Bar [mgBa/l]', 'Beryl [mgBe/l]', 'Bor [mgB/l]', 'Chrom [mgCr/l]', 'Cyjanki wolne [mgCN/l]', 'Cyna [mgSn/l]', 'Cynk [mgZn/l]', 'Fluorki [mgF/l]', 'Fosforany [mgPO4/l]', 'Glin [mgAl/l]', 'Kadm [mgCd/l]', 'Kobalt [mgCo/l]', 'Mangan [mgMn/l]', 'Miedź [mgCu/l]', 'Molibden [mgMo/l]', 'Nikiel [mgNi/l]', 'Ołów [mgPb/l]', 'Rtęć [mgHg/l]', 'Selen [mgSe/l]', 'Srebro [mgAg/l]', 'Tal [mgTl/l]', 'Tytan [mgTi/l]', 'Uran [mgU/l]', 'Wanad [mgV/l]', 'Żelazo [mgFe/l]'):** Ich obecność, nawet w śladowych ilościach, może wpływać na przewodność i są ważnymi wskaźnikami jakości wody, odzwierciedlającymi specyfikę geologiczną lub potencjalne zanieczyszczenia.
  - **'Przewodność elektrolityczna właściwa w 20°C - wartość terenowa [µS/cm]':** Wartość terenowa tego parametru jest silnym predyktorem wartości laboratoryjnej, często używanym do weryfikacji pomiarów i stanowiącym bardzo silny wskaźnik.

- **Cechy geoprzestrzenne i lokalizacyjne:**
  - **'PUWG 1992 X', 'PUWG 1992 Y':** Współrzędne geograficzne punktów pomiarowych umożliwiają modelowi uwzględnienie wpływu czynników przestrzennych, takich jak geologia regionu, obecność źródeł zanieczyszczeń czy bliskość zbiorników wodnych.
  - **Zakodowane zmienne kategoryczne (np. 'Województwo_X', 'Powiat_X', 'Gmina_X', 'Miejscowość_X', 'Nazwa dorzecza_X', 'RZGW_X'):** Różnice w jakości wody często korelują z konkretnymi regionami administracyjnymi lub hydrologicznymi, ze względu na zróżnicowane uwarunkowania geologiczne i działalność antropogeniczną.
- **Cechy hydrogeologiczne:**
  - **'Głębokość do stropu warstwy wodonośnej [m p.p.t.]':** Głębokość, z której pobierana jest próbka, może wpływać na skład chemiczny wody poprzez kontakt z różnymi warstwami geologicznymi i minerałami.
  - **'Przedział ujętej warstwy wodonośnej [m p.p.t.]':** Zakres, z którego pobierana jest woda, oraz specyfika samej warstwy wodonośnej (np. jej typ litologiczny) mają bezpośredni wpływ na skład chemiczny wody.
  - **Zakodowane zmienne kategoryczne charakteryzujące ujęcie (np. 'Stratygrafia_X', 'Zwierciadło wody_X', 'Typ ośrodka wodonośnego_X', 'Rodzaj punktu pomiarowego_X', 'Użytkowanie terenu_X'):** Te cechy opisują środowisko hydrogeologiczne i sposób ujęcia wody, co jest kluczowe dla zrozumienia jej parametrów fizykochemicznych.
- **Cechy czasowe:**
  - **'Rok poboru próbki', 'Miesiąc poboru próbki', 'Dzień poboru próbki':** Cechy czasowe mogą uchwycić sezonowe wahania jakości wody, wpływ czynników meteorologicznych (opady, temperatura powietrza) oraz długoterminowe trendy wynikające ze zmian klimatycznych lub działalności człowieka.

Wybrany zestaw cech stanowi kompleksową bazę informacji, która w oparciu o wiedzę dziedzinową jest najbardziej znacząca do modelowania przewodności elektrolitycznej wody.

5 Podsumowanie

Niniejszy raport szczegółowo przedstawił proces przygotowania i eksploracji danych dotyczących jakości wód podziemnych. Omówiono każdy etap, od wstępnej obróbki danych po inżynierię cech i eksploracyjną analizę danych (EDA), podkreślając niezbędność wykonanych kroków dla skutecznego modelowania uczenia maszynowego.