Temat:

Eksploracja danych oraz budowa modelu klasyfikacyjnego zdatności wody.

1. Wstęp i opis zbioru danych

Celem projektu było zbudowanie modelu uczenia maszynowego zdolnego do przewidywania zdatności wody do picia (Potability) na podstawie jej parametrów fizykochemicznych.
Do analizy wykorzystano zbiór danych water_potability.csv, zawierający 3276 próbek.
Zmienną celową jest kolumna Potability, przyjmująca wartości:

0 – woda niezdatna do picia,
1 – woda zdatna do picia.

2. Eksploracja danych (EDA)

Wstępna analiza wykazała:

Braki danych: Zidentyfikowano braki w kolumnach ph, Sulfate oraz Trihalomethanes.
Struktura cech: Wszystkie zmienne objaśniające są numeryczne.
Korelacje: Macierz korelacji wykazała niskie powiązania liniowe między poszczególnymi parametrami chemicznymi a zmienną celową.
Rozkład klas: Zbiór jest niezbalansowany z przewagą próbek wody niezdatnej do picia (klasa 0). Nierównowaga ta miała kluczowy wpływ na interpretację wyników modelowania.

3. Przygotowanie danych

Aby umożliwić poprawne działanie algorytmów, wykonano następujące kroki:

Obsługa braków: Brakujące wartości uzupełniono medianą dla każdej kolumny, aby zminimalizować wpływ wartości odstających.
Podział danych: Zbiór podzielono na treningowy (80%) i testowy (20%) z zachowaniem proporcji klas (stratify).
Standaryzacja: Zastosowano StandardScaler w celu przeskalowania cech (średnia=0, odchylenie=1), co jest wymagane m.in. dla Regresji Logistycznej.

4. Budowa modeli predykcyjnych

Jako typ zadania wybrano klasyfikację, ponieważ zmienna celowa jest dyskretna i binarna.
Zaimplementowano dwa modele:

Regresja Logistyczna: Model liniowy.
Drzewo Decyzyjne: Model nieliniowy.

5. Ocena jakości modeli

Modele oceniono na zbiorze testowym przy użyciu metryk Accuracy oraz F1-score. Wyniki (dla zbioru testowego):

Regresja Logistyczna:
Accuracy: 0.8500 (Wynik mylący, wynikający z dominacji klasy 0)
F1-score (dla klasy 1): 0.0741 (Bardzo niski)
Drzewo Decyzyjne:
Accuracy: 0.7930
F1-score (dla klasy 1): 0.3257

6. Wnioski

Choć Regresja Logistyczna osiągnęła wyższą dokładność (Accuracy ~85%), analiza szczegółowa wykazała, że model ten "nauczył się" niemal wyłącznie przewidywać klasę dominującą (woda niezdatna), ignorując wodę zdatną do picia (stąd tragicznie niski F1-score ~0.07). Jest to klasyczna pułapka metryki Accuracy przy niezbalansowanych danych.

Lepszym modelem okazało się Drzewo Decyzyjne, które pomimo niższego ogólnego Accuracy (79%), osiągnęło znacznie wyższy wynik F1-score (0.33) i zdołało poprawnie zidentyfikować ponad 30% próbek wody pitnej (w porównaniu do zaledwie 4% w regresji). Potwierdza to, że nieliniowe podejście Drzewa Decyzyjnego lepiej radzi sobie ze złożoną strukturą danych chemicznych niż proste modele liniowe.