Analiza wyników EDA oraz AutoML

Wprowadzenie

Celem analizy było stworzenie modeli predykcyjnych do klasyfikacji jakości powietrza (Air Quality) na podstawie dostarczonych danych o zanieczyszczeniach i czynnikach środowiskowych. W procesie analizy zastosowano dwa podejścia:

1. **Standardowe podejście:** Klasyfikacja wieloklasowa na podstawie nieprzekształconej zmiennej Air Quality (klasy kategoryczne).
2. **Podejście z przekształceniem zmiennej docelowej:** Reprezentacja Air Quality jako zmiennej porządkowej (od 0 dla Good do 3 dla Hazardous).

Wyniki EDA

- **Statystyki numeryczne:** Średnia temperatura wynosiła 25.46°C, a wilgotność 60.07%. Zmienność cech takich jak PM2.5, PM10, i NO2 była istotna.
- **Korelacje:** Silna dodatnia korelacja między PM2.5 a PM10 (0.987). Pozostałe cechy wykazywały niską korelację ze zmienną docelową.
- **Brakujące wartości:** Żadne brakujące wartości nie występowały w zbiorze danych.
- **Rozkład zmiennej Air Quality:**
    - Good: 40% próbek.
    - Moderate: 30% próbek.
    - Poor: 20% próbek.
    - Hazardous: 10% próbek.

Wyniki AutoML (przed przekształceniem zmiennej docelowej)

Najlepsze modele:

1. **MLPClassifier**:
    - **Dokładność na danych testowych:** 40%.

- **Pipeline:** Binarizer + MLPClassifier.
- Model oparty na sieciach neuronowych, wybrany ze względu na możliwość pracy z nieliniowymi zależnościami.

2. **BernoulliNB**:
   - **Dokładność na danych testowych:** ~39.8%.
   - Model probabilistyczny dobrze działający na danych z wartościami binarnymi lub znormalizowanymi.
3. **ExtraTreesClassifier**:
   - **Dokładność na danych testowych:** ~39.7%.
   - Model ensemble z mechanizmem losowych podziałów danych, wybrany za stabilność i szybkość.

Ogólny wniosek:

Żaden z modeli nie osiągnął wysokiej skuteczności, co sugeruje niską korelację cech z Air Quality.

Wyniki AutoML (po przekształceniu zmiennej docelowej)

Najlepsze modele:

1. **ExtraTreesClassifier**:
   - **Dokładność na danych testowych:** 39%.
   - **Pipeline:** ExtraTreesClassifier z parametrami dopasowanymi przez TPOT (np. min_samples_split=20).
   - Model pozostał dominującym wyborem TPOT, jednak wynik nie poprawił się znacząco.
2. **BernoulliNB**:
   - **Dokładność na danych testowych:** ~38.9%.
   - **Pipeline:** MinMaxScaler + BernoulliNB.
   - Wynik podobny do wcześniejszej analizy.
3. **RandomForestClassifier** (nieznacznie gorszy od ExtraTrees):

- **Dokładność:** ~38%.
- Model wskazuje na możliwość dalszej optymalizacji parametrów.

Porównanie wyników przed i po przekształceniu

Kryterium	**Przed przekształceniem**	**Po przekształceniu**
Najlepszy model	MLPClassifier	ExtraTreesClassifier
Dokładność najlepszego	40%	39%
Drugi najlepszy model	BernoulliNB	BernoulliNB
Trzeci najlepszy model	ExtraTreesClassifier	RandomForestClassifier
Ogólny wynik	Porównywalny	Nieco gorszy

Wnioski

1. **Przekształcenie zmiennej docelowej:**
   - Nie przyniosło istotnej poprawy wyników modeli. Modele nadal miały trudności z uchwyceniem złożoności relacji między cechami a klasami Air Quality.
2. **Cechy dominujące:**
   - PM2.5 i PM10 pozostają głównymi predyktorami jakości powietrza. W przyszłych analizach można rozważyć ograniczenie nadmiarowości między tymi zmiennymi.

Podsumowanie

Analiza wykazała, że dane wymagają dalszej pracy nad ich strukturą i wzbogaceniem o nowe cechy. TPOT pozwolił na szybkie wyznaczenie optymalnych modeli, jednak ich skuteczność pozostawia miejsce na usprawnienia.