

Analiza porównawcza algorytmów ML w kontekście predykcji klasy emisyjności CO2 pojazdów

Paweł Nowak,
Wydział Matematyki,
Politechnika Wrocławska



OPIEKUN: dr inż. Adam zagdański

Opis zagadnienia

- Staramy się sklasyfikować pojazd mechaniczny do jednej z kilku klas emisyjności dwutlenku węgla na podstawie wybranych jego cech (typ spalanego paliwa, klasa pojazdu, spalanie itp.), korzystając z metod uczenia maszynowego.
- Informacje opisujące pojazdy zostały pobrane z platformy kaggle, która pobrała te dane z oficjalnej strony rządu kanadyjskiego.



W jakim celu wykonujemy tę analizę?

- ▶ ~~Powyższą analizę wykonujemy w celu predykcji klasy emisyjności.~~ Ponadto analiza pomoże nam odpowiedzieć na następujące **pytania badawcze**:
 - ▶ Jak dobór modelu wpływa na skuteczność klasyfikacji?
 - ▶ Który model dokonuje najdokładniejszych predykcji?
 - ▶ Jak różne strategie uczenia kształtują wydajność modeli klasyfikacyjnych?
 - ▶ Czy większa liczba klas docelowych obniża dokładność predykcji algorytmu?
 - ▶ Czy modele bardziej skomplikowane dokonują klasyfikacji z większą dokładnością niż modele proste?

Po co prognozować klasę emisyjności?

- ▶ Aby na własną rękę oszacować rzeczywistą emisyjność pojazdu. Niektórzy producenci pojazdów manipulowali faktyczną emisyjnością pojazdów (afera Volkswagen). To z kolei umożliwi nam:
 - ▶ Oszacowanie podatku od pojazdów (Vehicle Tax), który planują wprowadzić rządy niektórych państw (np. Wielkiej Brytanii).
 - ▶ Zakupienie niskoemisyjnego pojazdu, który będzie przyjazny dla środowiska naturalnego.

Jak przewidywać klasę emisyjności pojazdu?

- ▶ Mamy trzy klasy emisyjności („mała”, „średnia”, „wysoka”). Dany pojazd przypiszemy do jednej z trzech klas emisyjności, korzystając z poniższych metod klasyfikacyjnych.

- ▶ Drzewo decyzyjne (Decision Tree)
 - ▶ Las losowy (RandomForest)
- } Modele zaawansowane



- ▶ K-Najbliższych sąsiadów (KNN)
 - ▶ Regresja Logistyczna (logistic regression)
 - ▶ Regresja Liniowa (linear regression)
- } Modele proste

Jakie strategie uczenia modeli zostaną zastosowane?

- ▶ Aby zapewnić obiektywność analizy, dogłębnie analizując modele, zastosowano cztery różne **strategie trenowania**:
 1. Brak optymalizacji hiperparametrów i brak wyboru cech (**noFS_untuned**)
 2. Optymalizacja hiperparametrów bez wyboru cech (**noFS_tuned**)
 3. Automatyczny wybór cech bez **strojenia** (**FS_untuned**)
 4. Optymalizacja hiperparametrów z automatycznym wyborem cech (**FS_tuned**)

Jak będziemy oceniać skuteczność modeli? – ~~metryki dokładności.~~

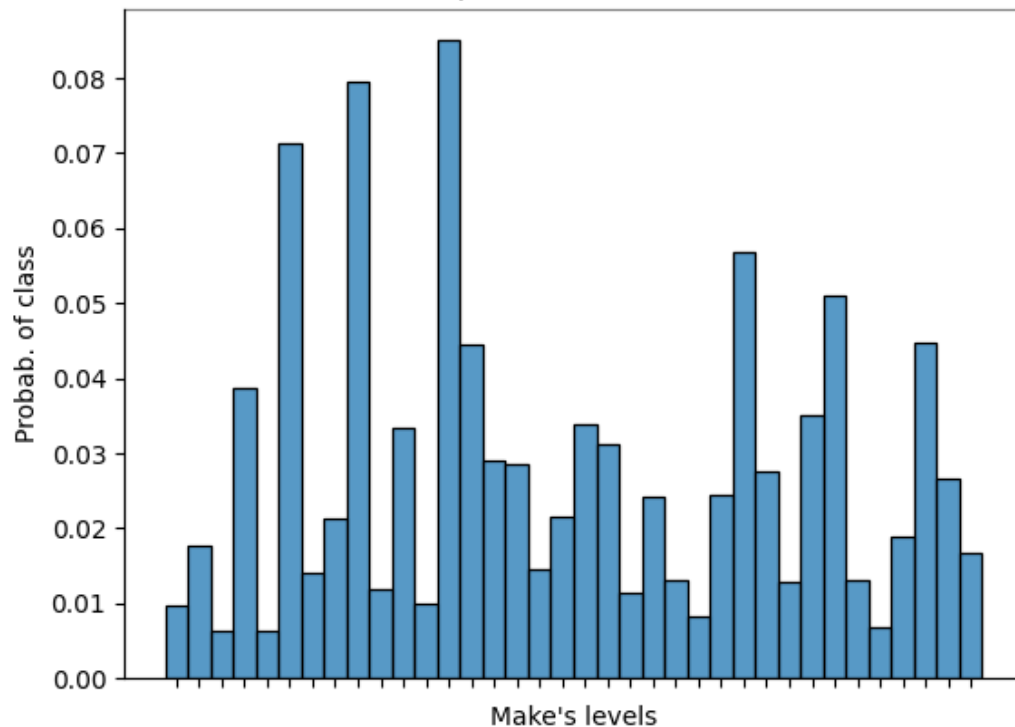
- ▶ W celu zbadania dokładności klasyfikacji modeli, **zastosowane** zostaną trzy wybrane metryki dokładności:
 - ▶ **Confusion matrix**
 - ▶ **accuracy**
 - ▶ **F1**



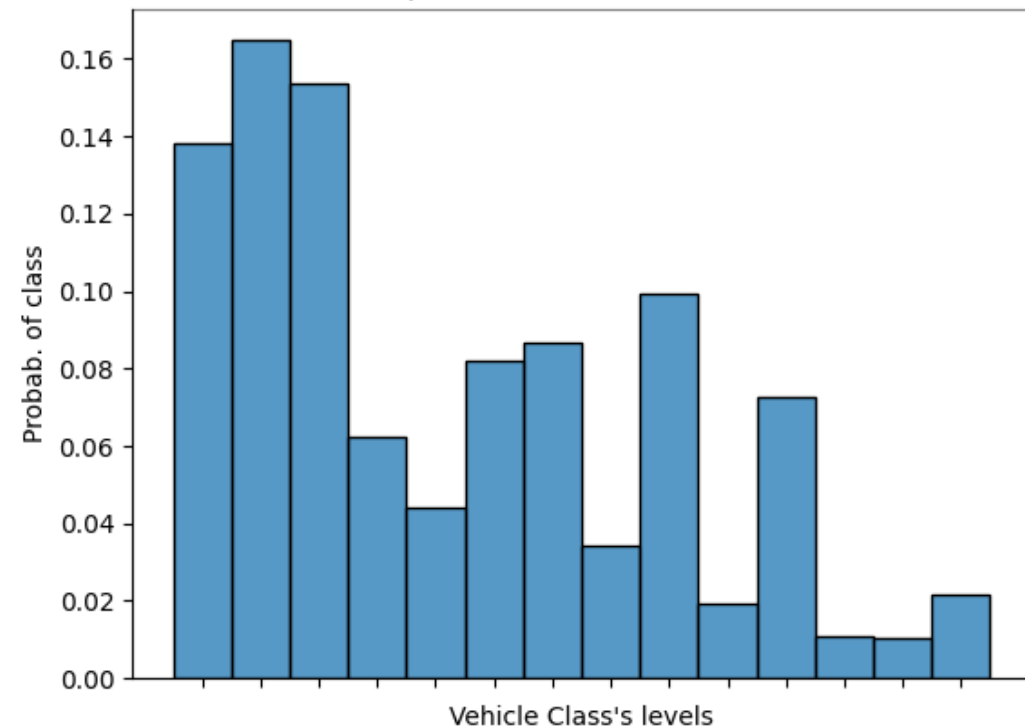
Analiza opisowa danych – zmienne jakościowe

Rozkład **częstotliwości** kategorii zmiennych **Make** oraz **Vehicle Class**

Relatives frequencies of the Make's levels

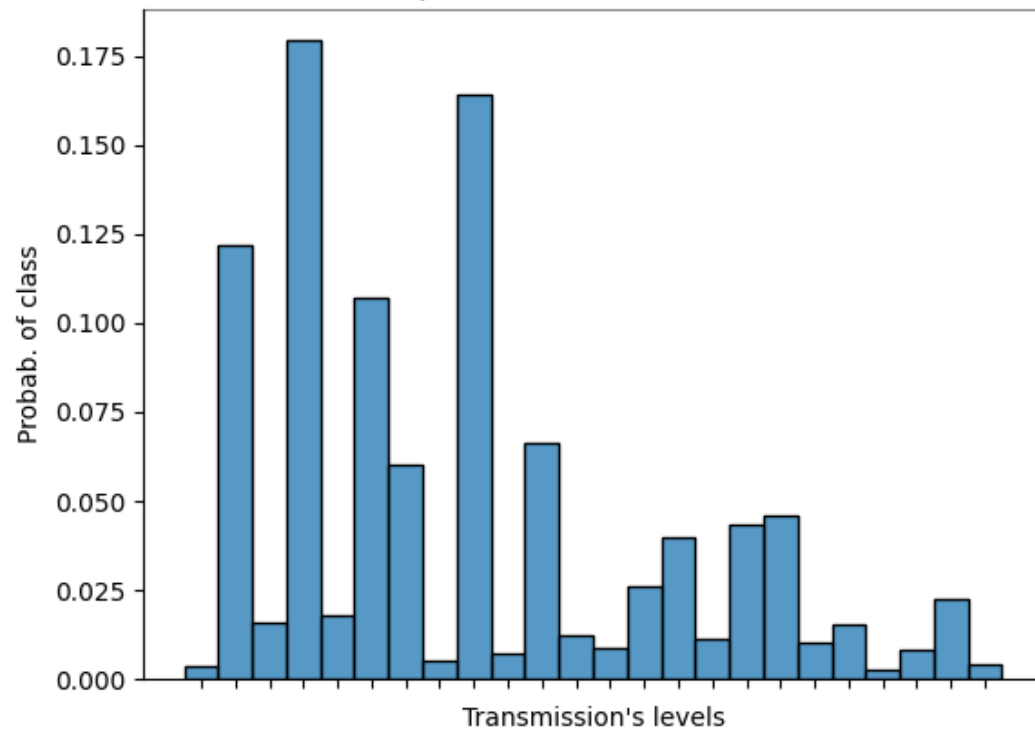


Relatives frequencies of the Vehicle Class's levels

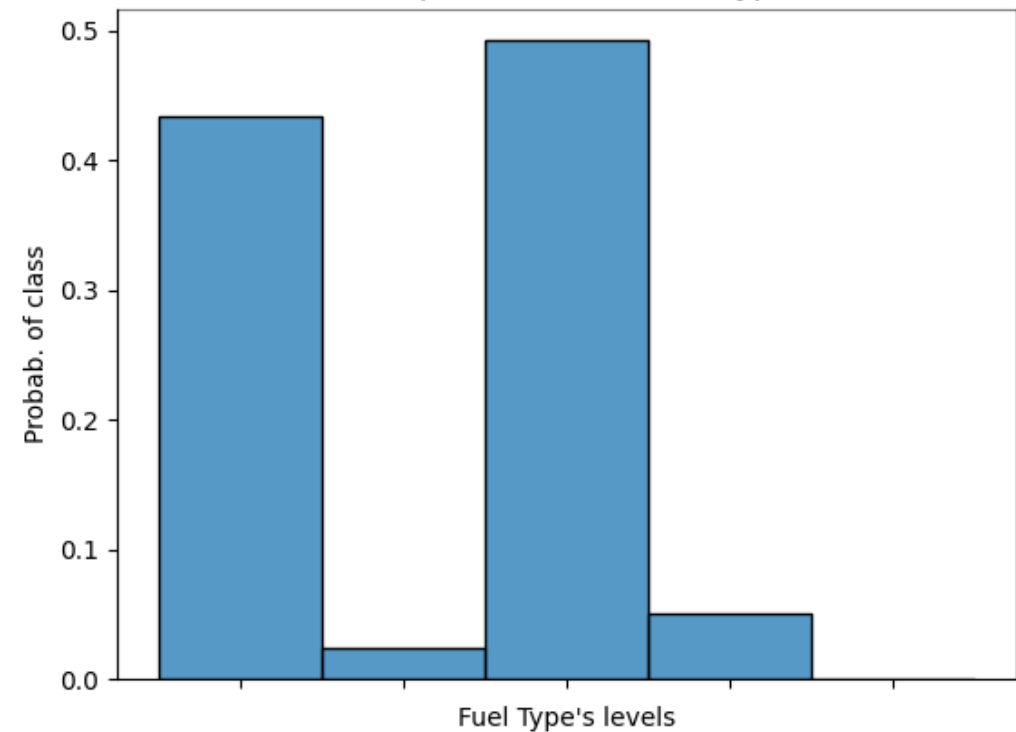


Rozkład **częstotliwości** kategorii zmiennych **Make** oraz **Vehicle Class**

Relatives frequencies of the Transmission's levels



Relatives frequencies of the Fuel Type's levels



Pierwsze wnioski o
cechach
jakościowych.

Pierwsze wnioski o zdolnościach dyskryminacyjnych.

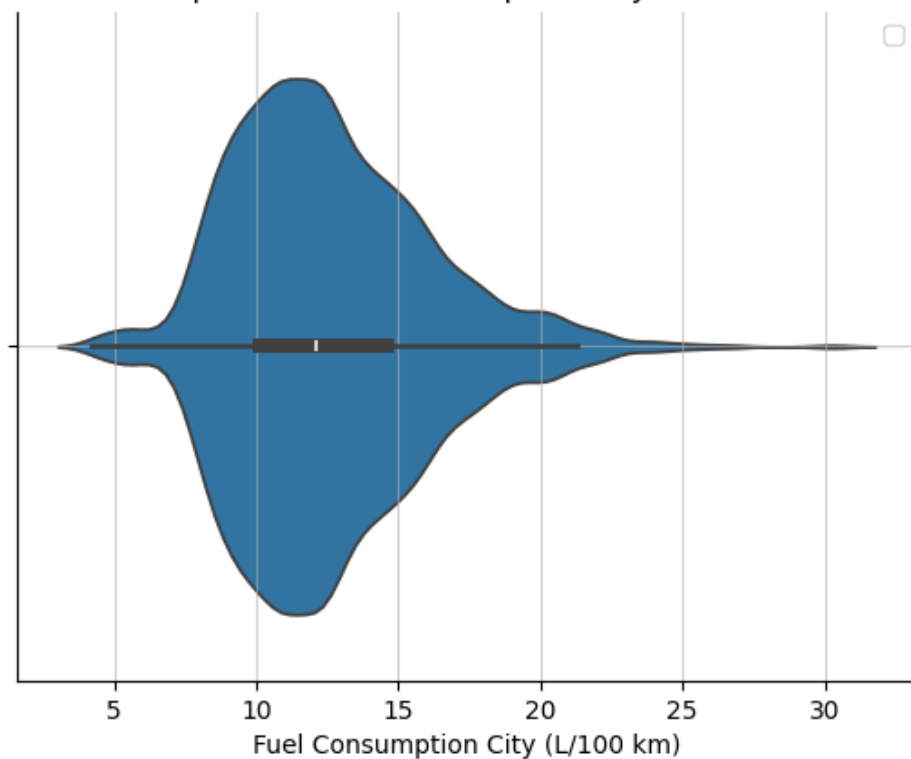
- ▶ Prawie wszystkie cechy jakościowe cechują się dosyć zbalansowanym rozkładem kategorii.
- ▶ Wyjątek stanowi **cecha Fuel Type**, dla której ponadto istnieje klasa występująca dokładnie raz. Z tego powodu ta jedna obserwacja nie będzie uwzględniana w dalszej analizie (~~kategoria jest niedoreprezentowana~~).
- ▶ Zmienna **model**, z uwagi na dużą liczbę unikatowych kategorii, nie będzie brana pod uwagę w dalszej analizie.
- ▶ W celu redukcji kardynalności zmiennych ilościowych, zastosowano metodę kodowania klas rzadkich (z ang. **rare category encoding**)

,

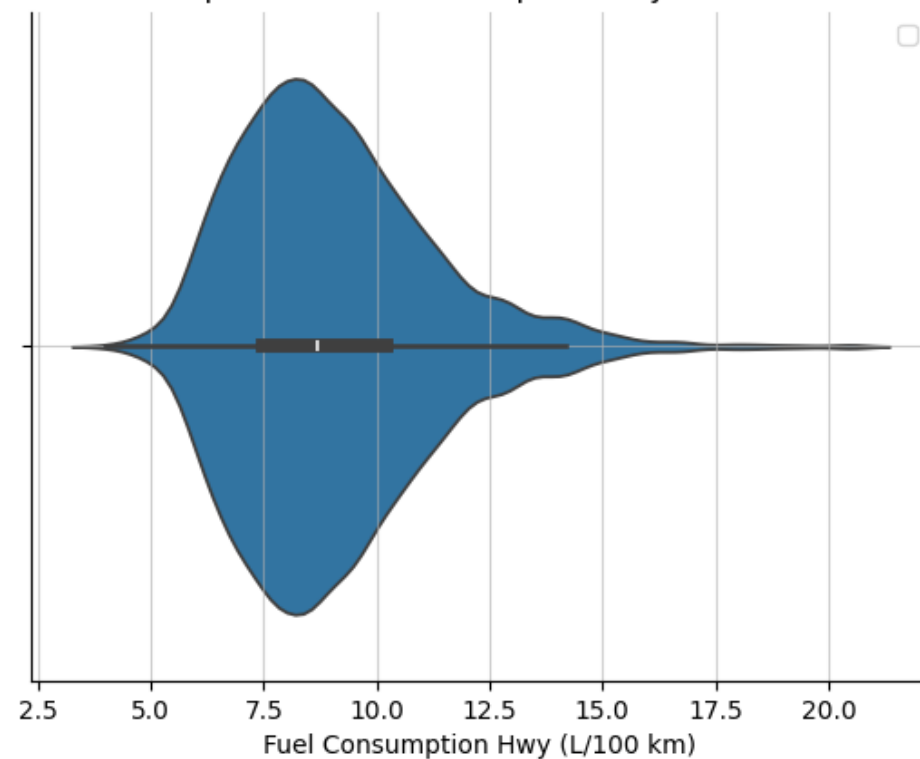
Analiza opisowa danych – zmienne ilościowe

Wykresy skrzypcowe dla zmiennych ciągłych

Violinplot for Fuel Consumption City (L/100 km)

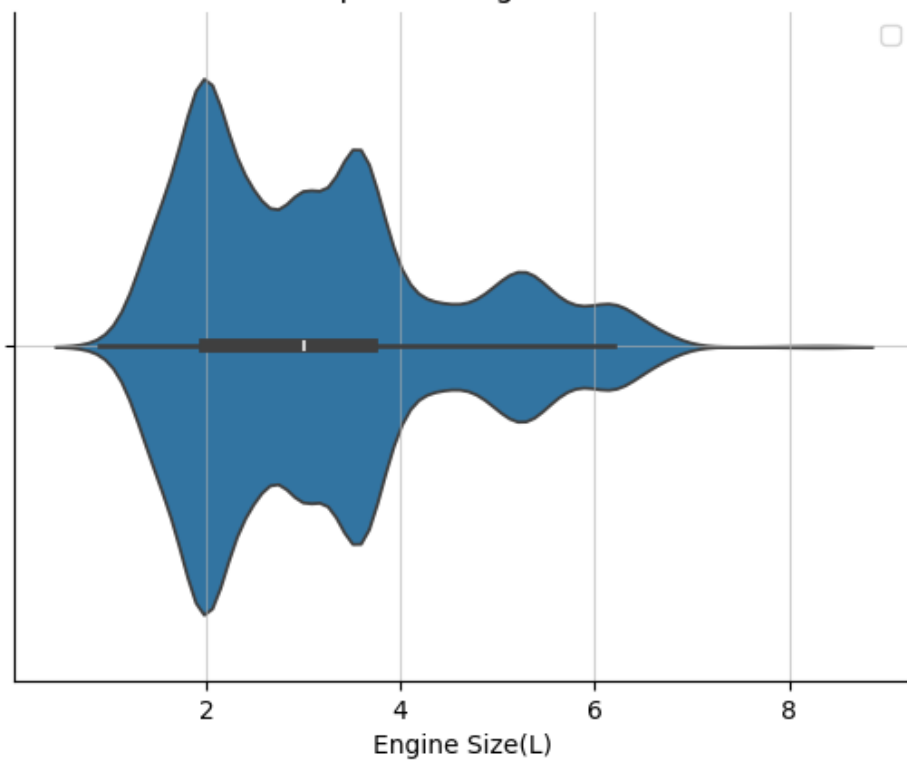


Violinplot for Fuel Consumption Hwy (L/100 km)

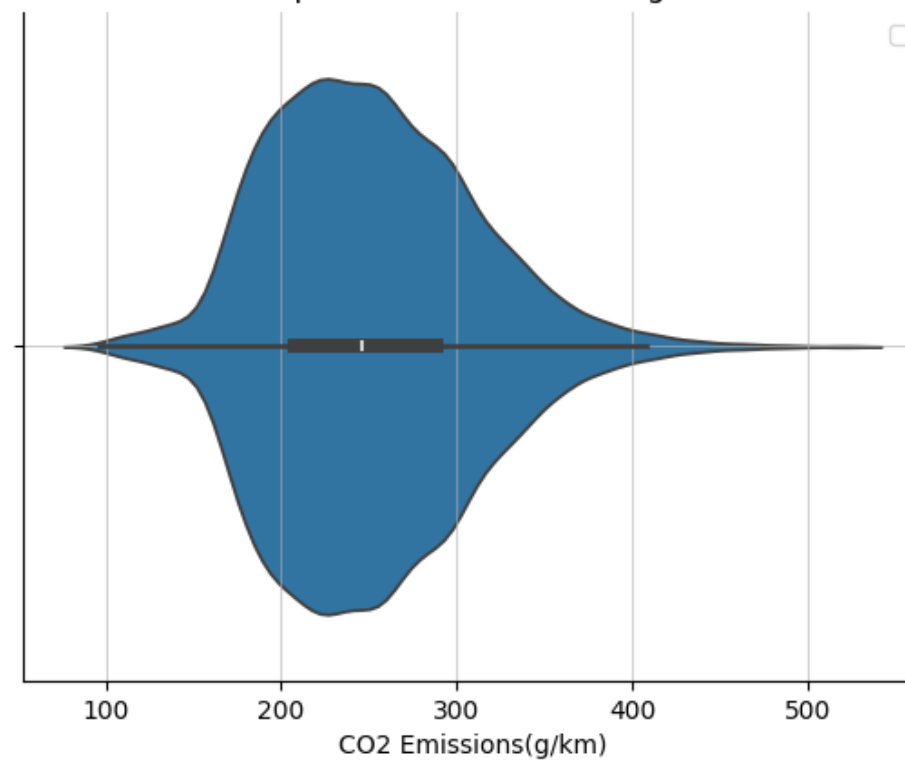



Wykresy skrzypcowe dla zmiennych ciągłych cd.

Violinplot for Engine Size(L)



Violinplot for CO2 Emissions(g/km)





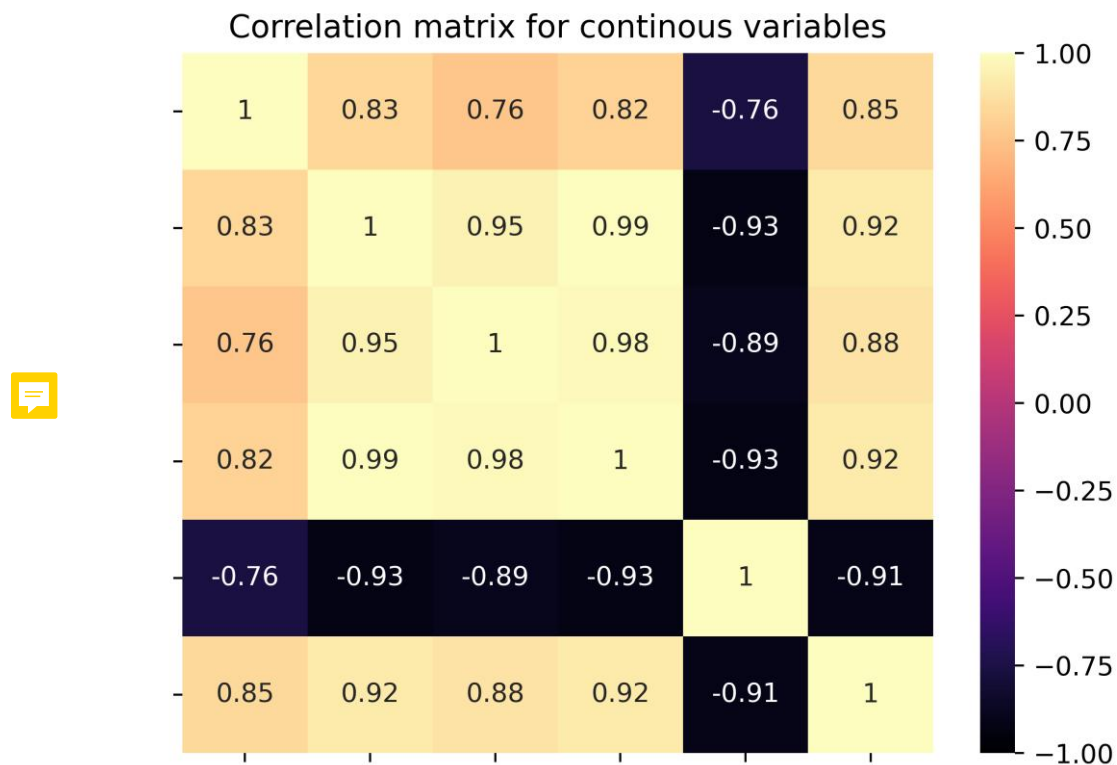
Wnioski o rozkładach cech ilościowych

Wstępne wnioski o zdolnościach dyskryminacyjnych cech ilościowych.

- ▶ Większość wykresów gęstości przypomina ~~tzw. krzywą Gaussa która charakteryzuje rozkład normalny.~~
- ▶ Na tle **krzywych Gaussowskich** wyróżnia się gęstość zmiennej **Fuel Size** , która jest gęstością wielomodalną.
- ▶ Wykresy gęstości **pokazują**, że wariancje **użytych** zmiennych ciągłych są wysokie, ~~co jest pożądaną właściwością.~~

Jak silnie
skorelowane są ze
sobą zmienne
ciągłe?

Macierz korelacji dla zmiennych numerycznych.

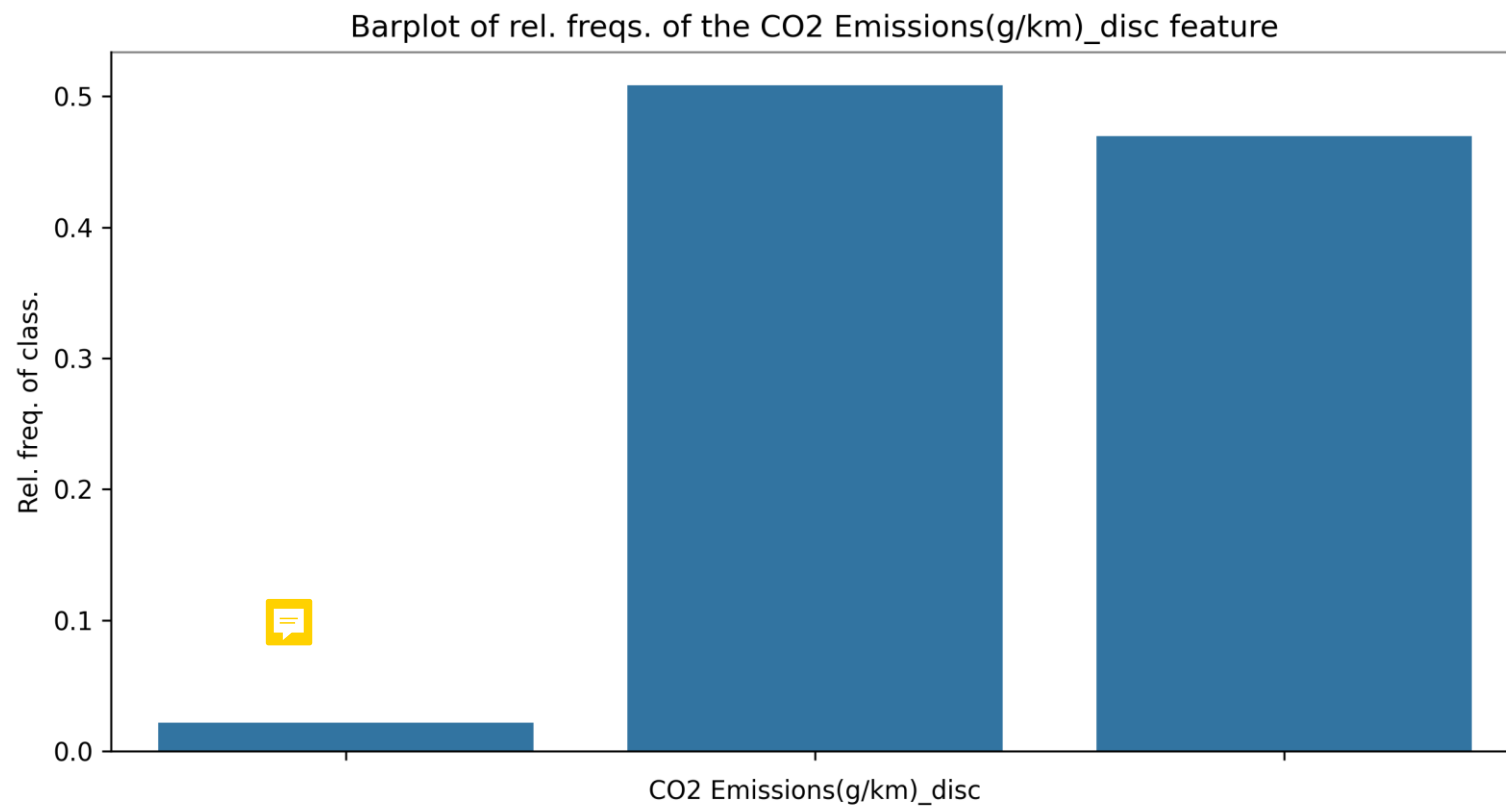


Dyskretyzacja zmiennej celu

Jak dyskretyzować **zmienną**?

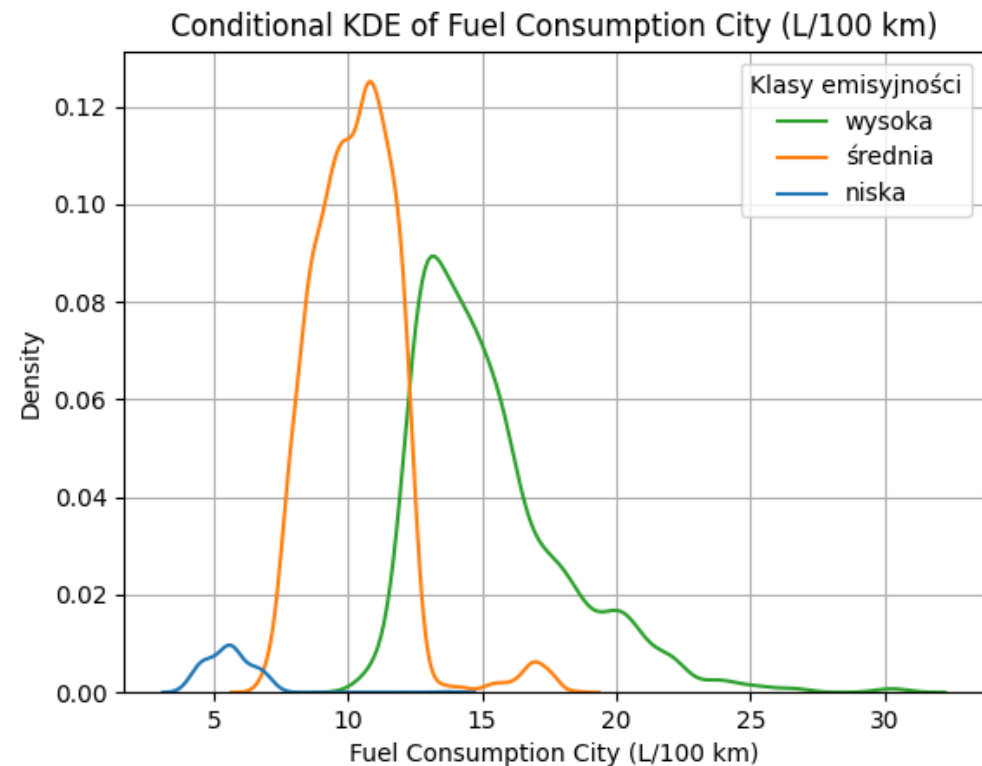
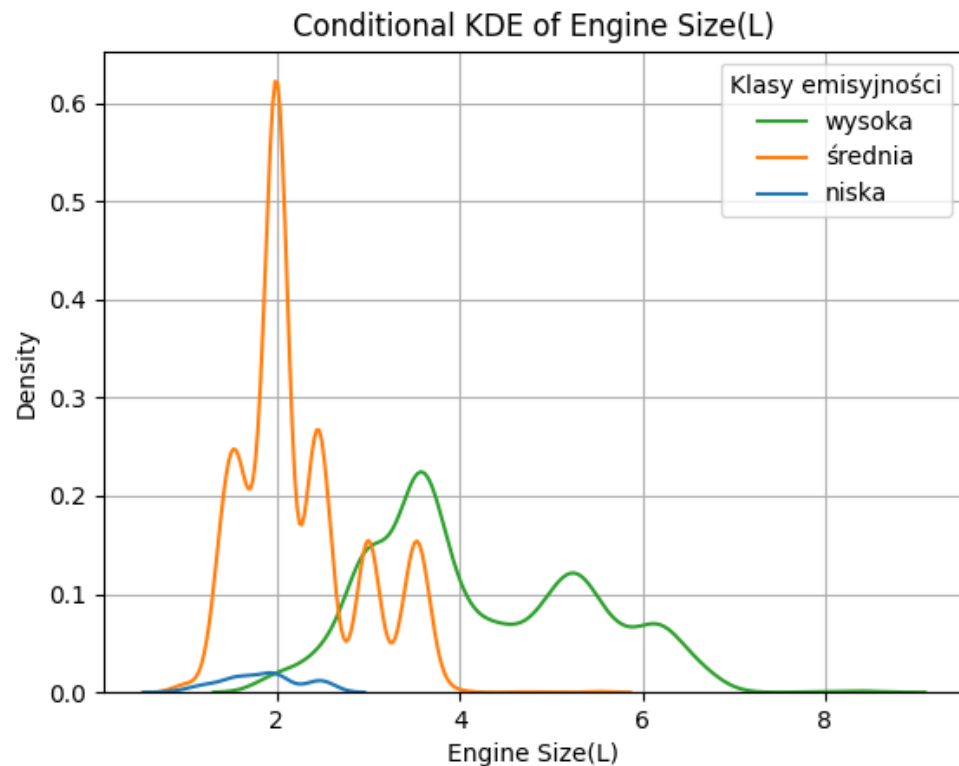
- ▶ W celu budowy modeli klasyfikacyjnych przeprowadzone zostanie przedziałowanie (z ang. binning), zmiennej ciągłej określającej poziom emisji dwutlenku węgla (w g/km).
- ▶ Parametry przedziałowania zostały dobrane tak, aby wynikowe klasy były w dobrym przybliżeniu zgodne z regulacjami kanadyjskiego prawa.
- ▶ W wyniku przedziałowania otrzymaliśmy trzy klasy:
 - ▶ **Niska** – [0;150] (kodowane wewnątrznie jako 0)
 - ▶ **średnia** – (150; 250] (kodowane jako 1)
 - ▶ **wysoka** - powyżej 250 g/km (kodowane jako 2)

Wykres słupkowy klas po dyskretyzacji.



Warunkowe
wykresy gęstości
zmiennych.


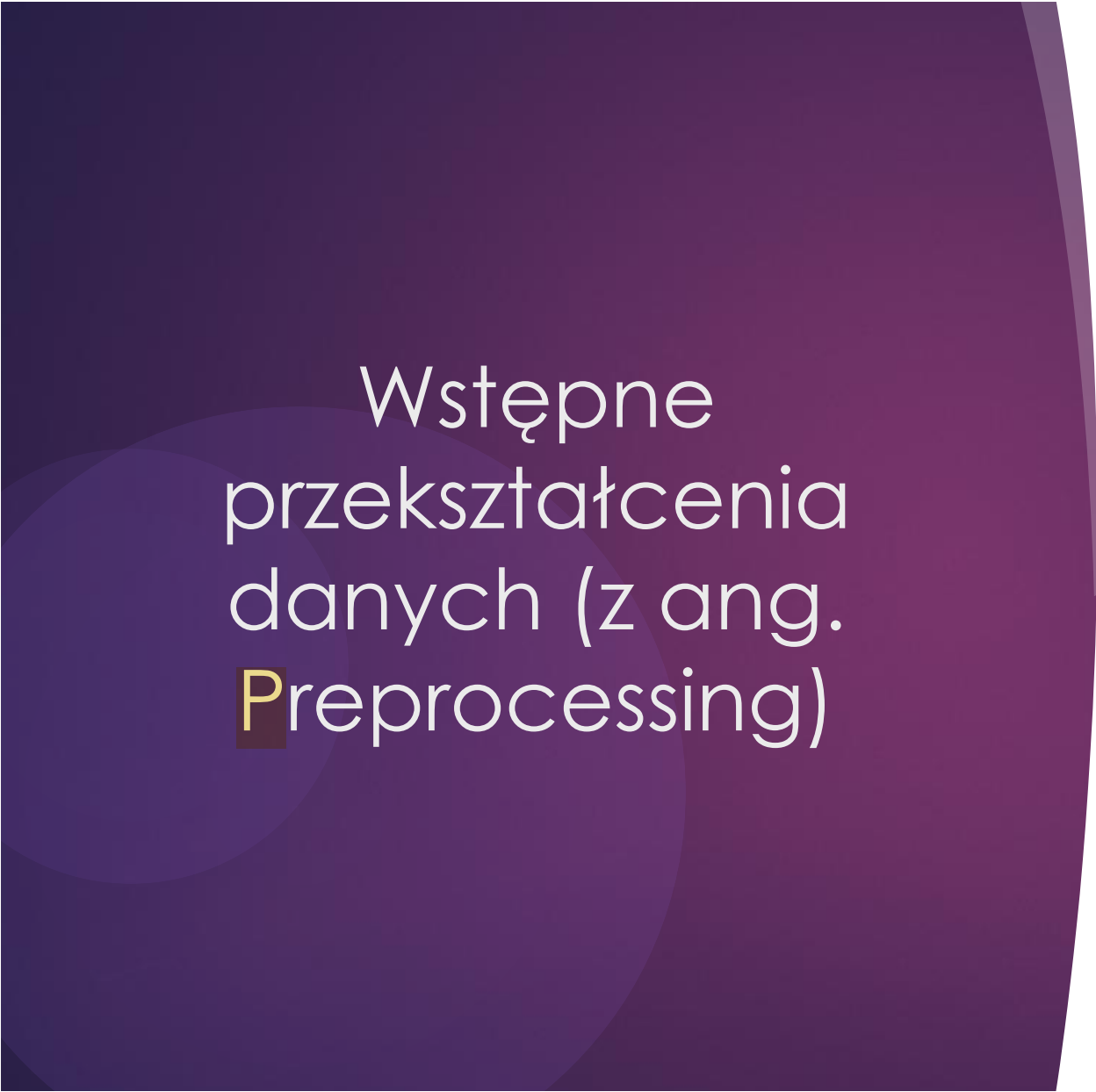
Warunkowy wykres gęstości – zmienna Engine Size



Wnioski o zdolnościach dyskryminacyjnych

Wnioski o zdolnościach dyskryminacyjnych

- ▶ Niemal wszystkie warunkowe wykresy gęstości dobrze odseperowują klasy **emisyjności** – wykresy niebieski, pomarańczowy i zielony nie nachodzą na siebie w znaczącym stopniu.
- ▶ Warto zwrócić uwagę na zmienną **Engine Size**. Wykres gęstości dla klasy 0 (niebieski) jest całkowicie zawarty w wykresie klasy 1 (pomarańczowy), co może prowadzić do błędnej klasyfikacji większości obserwacji klasy 0 jako klasy 1.




Wstępne przekształcenia danych (z ang. Preprocessing)

Wstępne przekształcenia danych

- ▶ Niektóre algorytmy uczenia maszynowego wymagają zastosowania pewnych wstępnych przekształceń danych (z ang. Preprocessing)
- ▶ W naszej analizie wykorzystamy wstępnego:
 - ▶ Standardyzacja (z ang. **Standardization**)
 - ▶ Kodowanie „na gorąco” (z ang. **One hot encoder**)
 - ▶ Analiza składowych głównych (z ang. **Principal components analysis – PCA**) dla zmiennych ciągłych określających poziom spalania paliwa.

Metodologia uczenia.

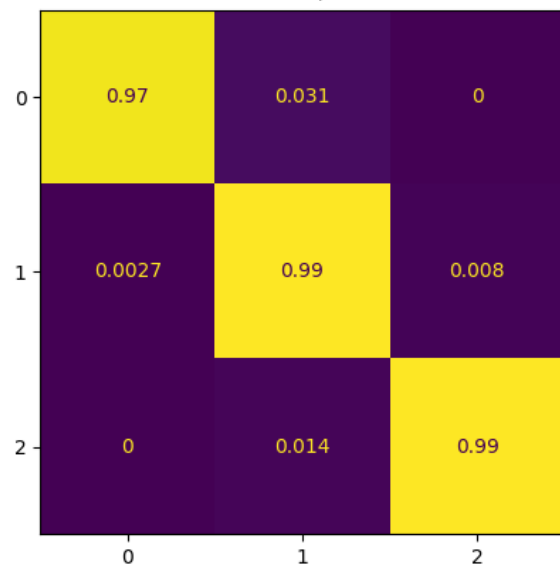
- ▶ Modele są uczone na bazie zbioru danych **45 razy w czterech różnych wersjach.**
- ▶ **Dokonyjemy** podziału zbioru danych na dwa **rozłączne** zbiory:
 - ▶ Zbiór treningowy, który zawiera 80% wszystkich obserwacji
 - ▶ Zbiór testowy, który zawiera 20% wszystkich obserwacji
- ▶ Zbiór treningowy służy do trenowania modeli, natomiast zbiór testowy – do oceny jakości predykcji ~~modeli~~.
- ▶ Rozłączność zbiorów testowego i treningowego jest istotna w celu zapobieganiu tzw. problemu wycieku danych (**data-leakage problem**).



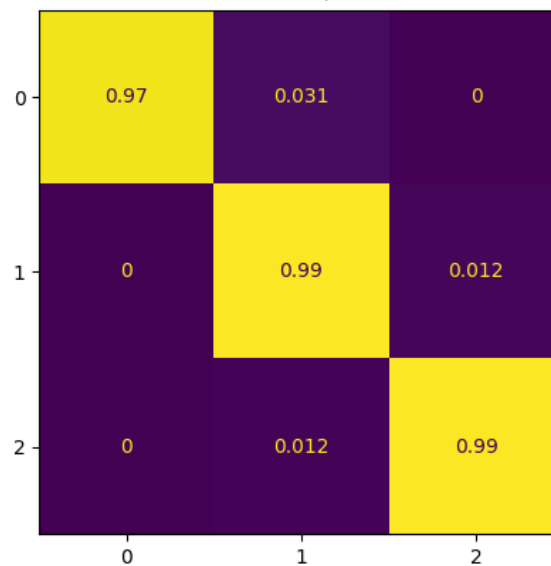
Ocena skuteczności modeli – macierze pomyłek

Macierze pomyłek modeli dla wersji podstawowej (bez optymalizacji i wyboru cech)

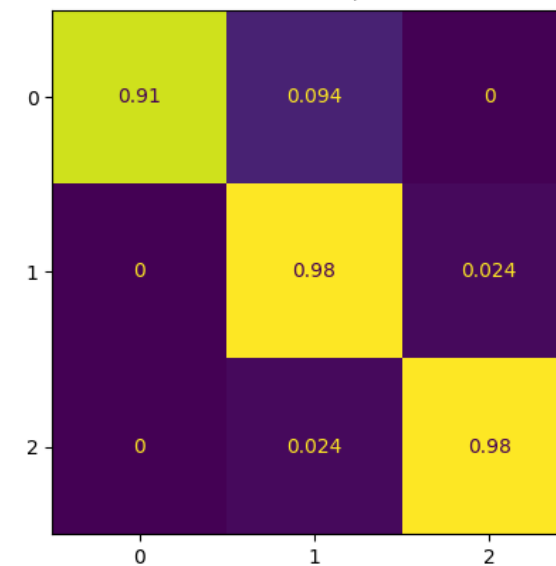
Conf. Matrix, DecTree



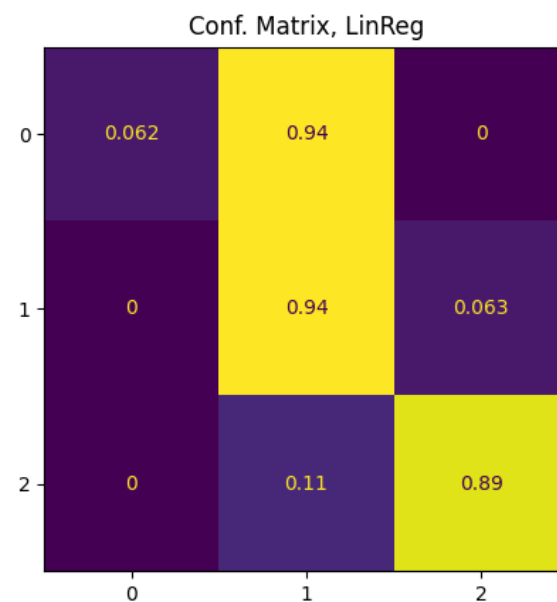
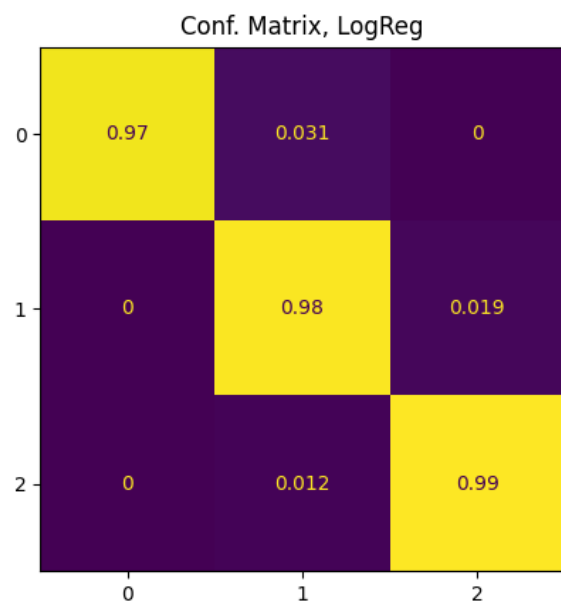
Conf. Matrix, RanFor




Conf. Matrix, KNN

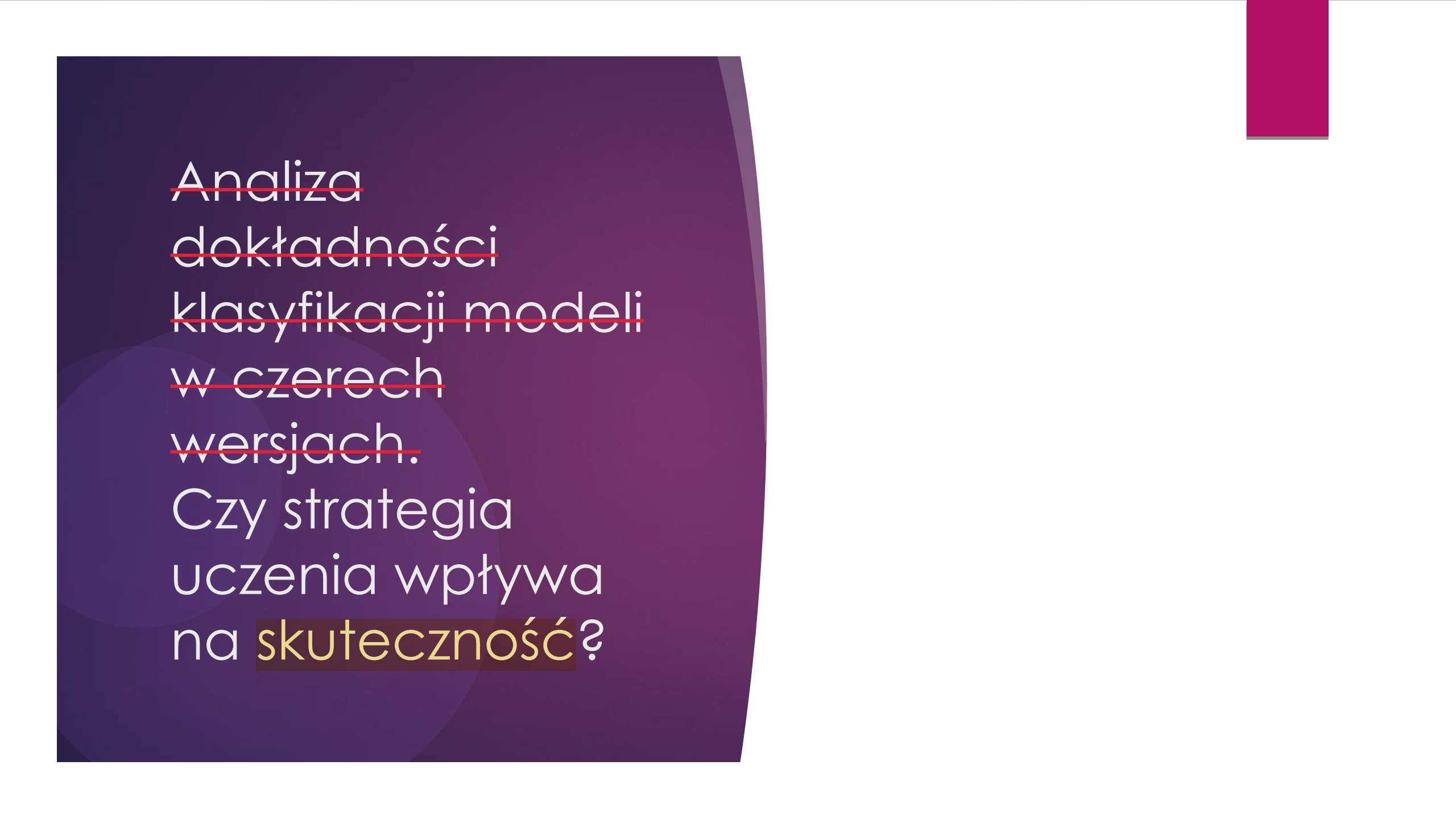


Macierze pomyłek modeli dla wersji podstawowej cd.



Pierwsze wnioski.

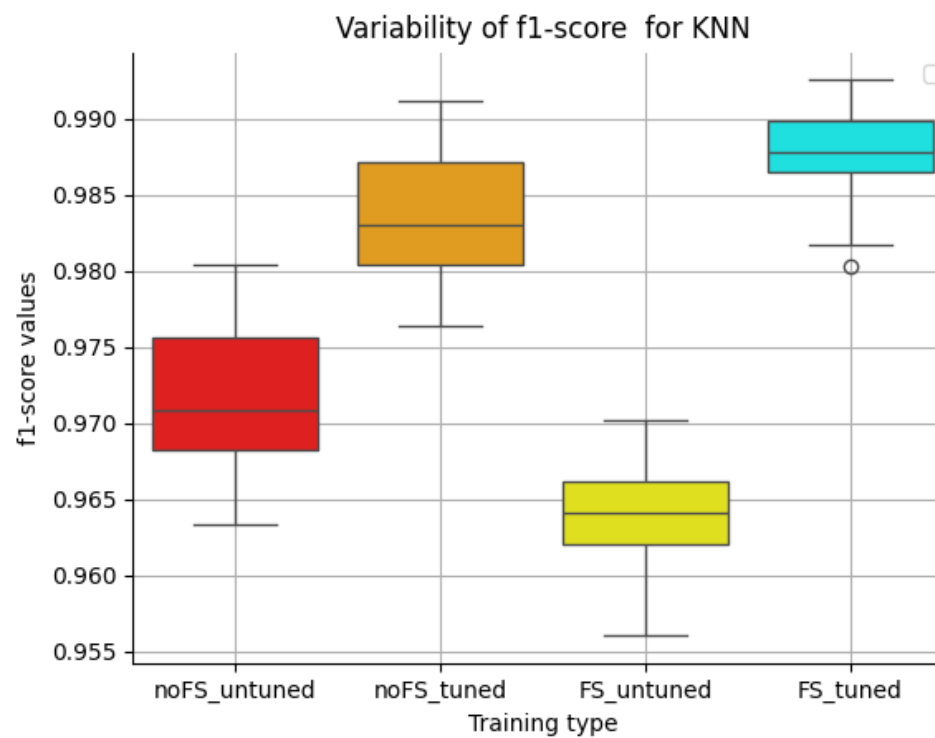
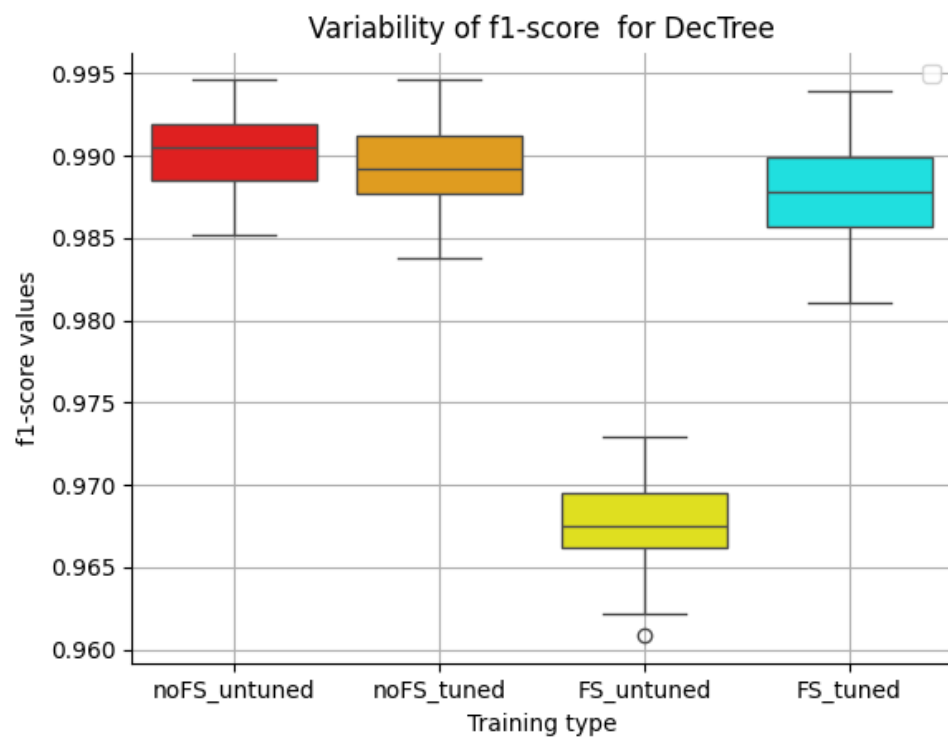
- ▶ Spośród wszystkich klas najmniej dokładnie klasyfikowana jest klasa 0.  Model liniowej regresji niemalże zawsze klasyfikuje ją jako klasa 1 (**masking class problem**), co pokazuje niską nieskuteczność **modelu** w kontekście klasyfikacji.
- ▶ Pozostałe klasy są prognozowane z wysoką skutecznością.
- ▶ **Na ten moment** model prosty (**LogReg**) cechuje się podobną dokładnością klasyfikacji, co modele zaawansowane (**DecTree, RanFor**)



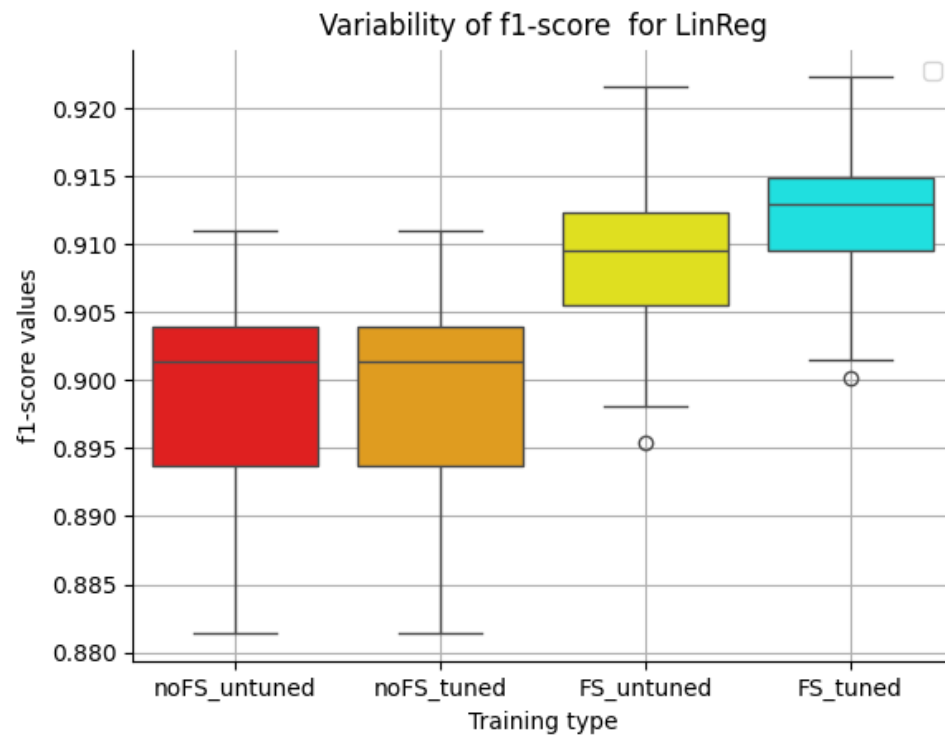
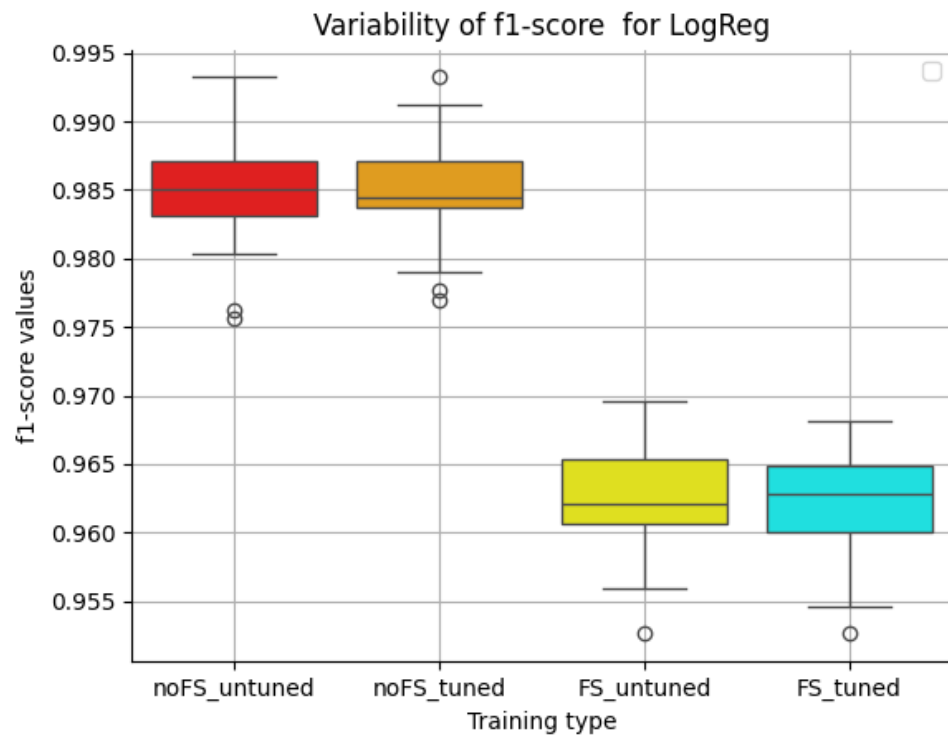
~~Analiza~~
~~dokładności~~
~~klasyfikacji modeli~~
~~w czerech~~
~~wersjach.~~

Czy strategia
uczenia wpływa
na skuteczność?

Skuteczność modeli w różnych wersjach uczenia



Skuteczność modeli w różnych wersjach uczenia

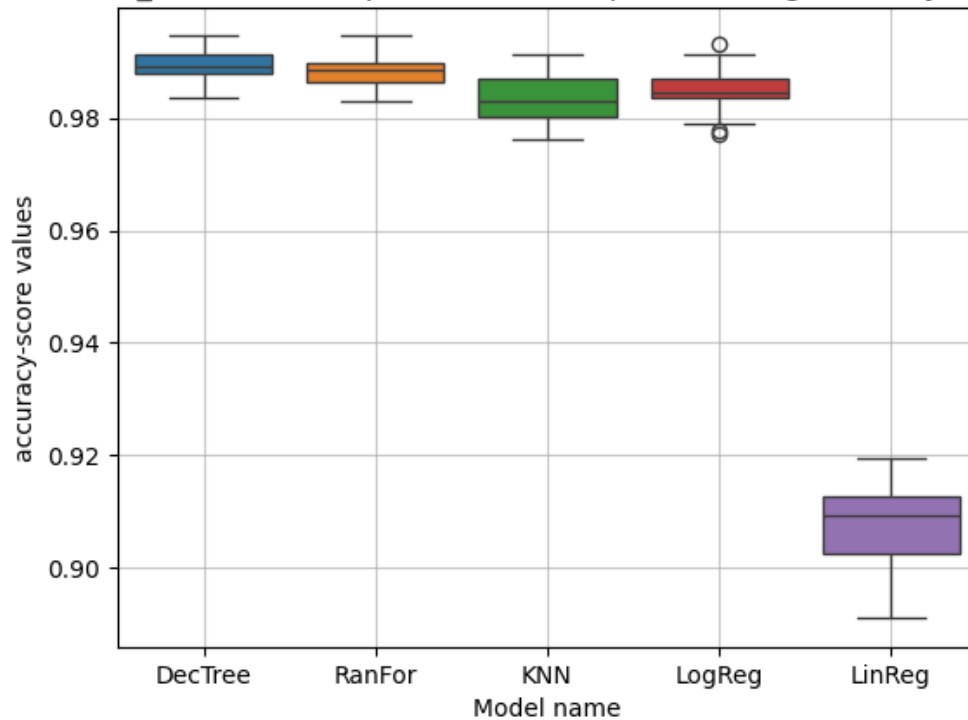


Zbiorcza analiza
dokładności
klasyfikacji modeli.

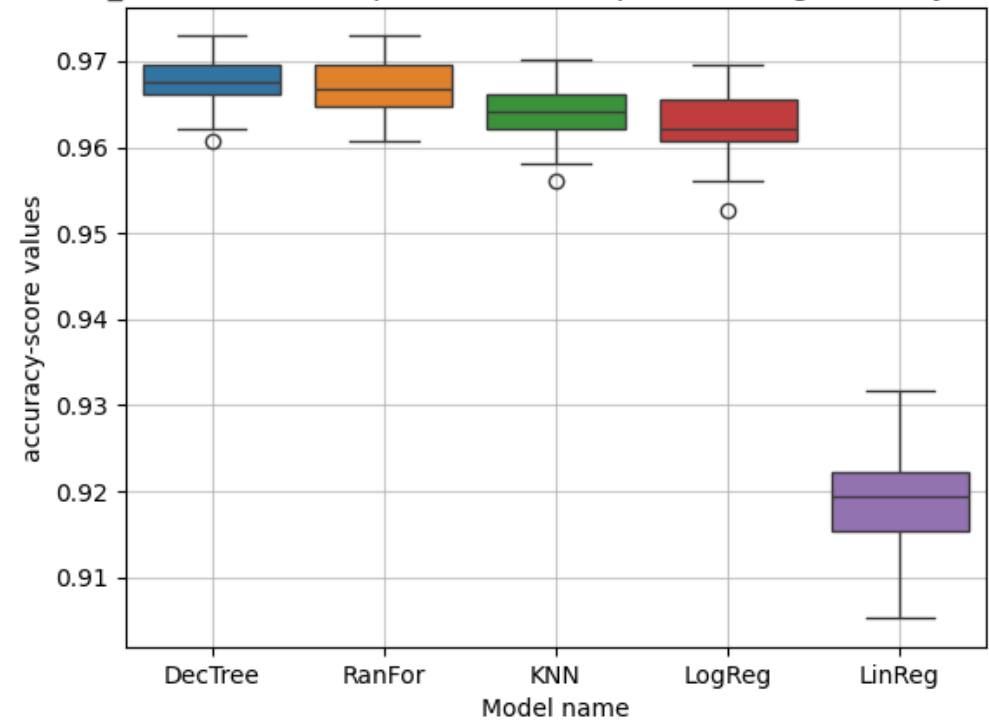
Zbiornicze porównanie wydajności modeli



noFS_tuned: Models performance comparison using accuracy-score

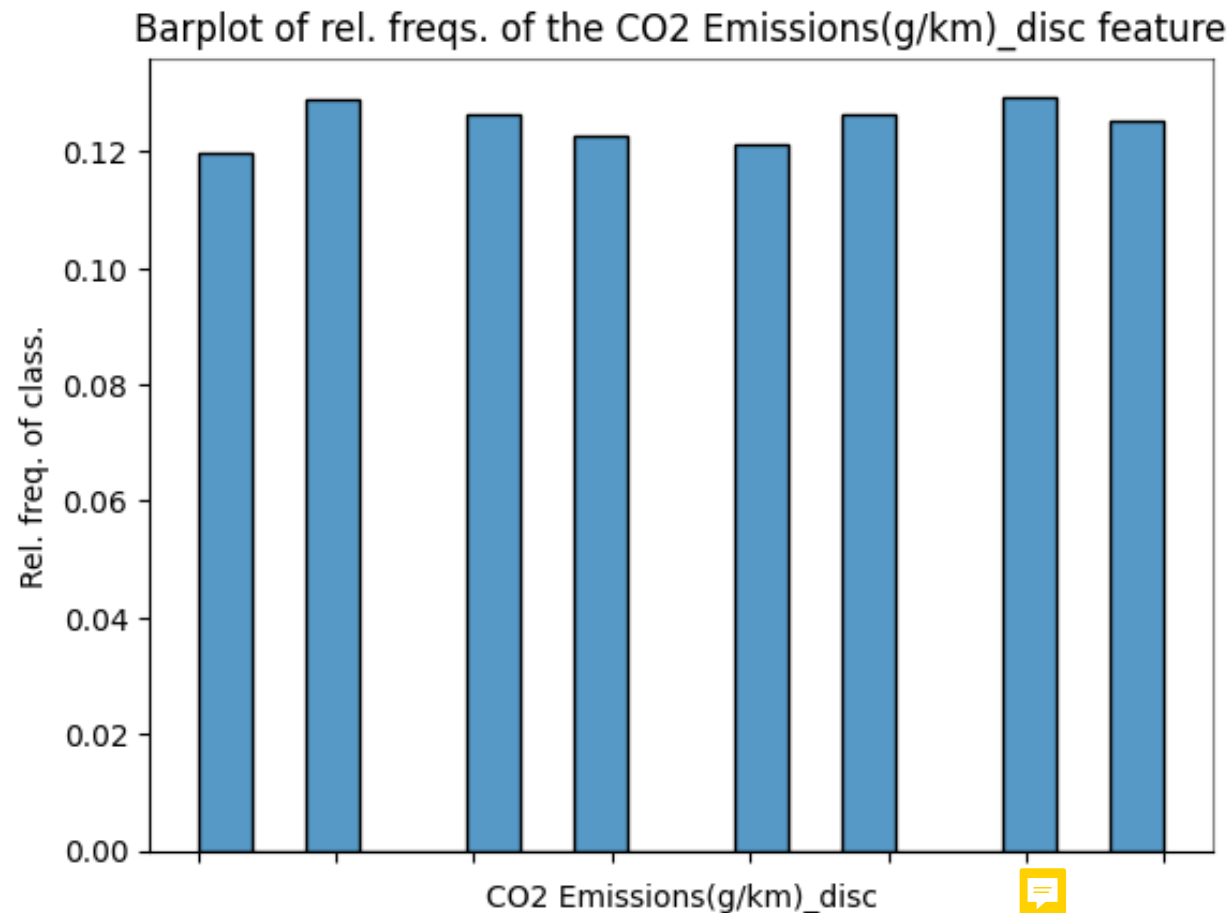



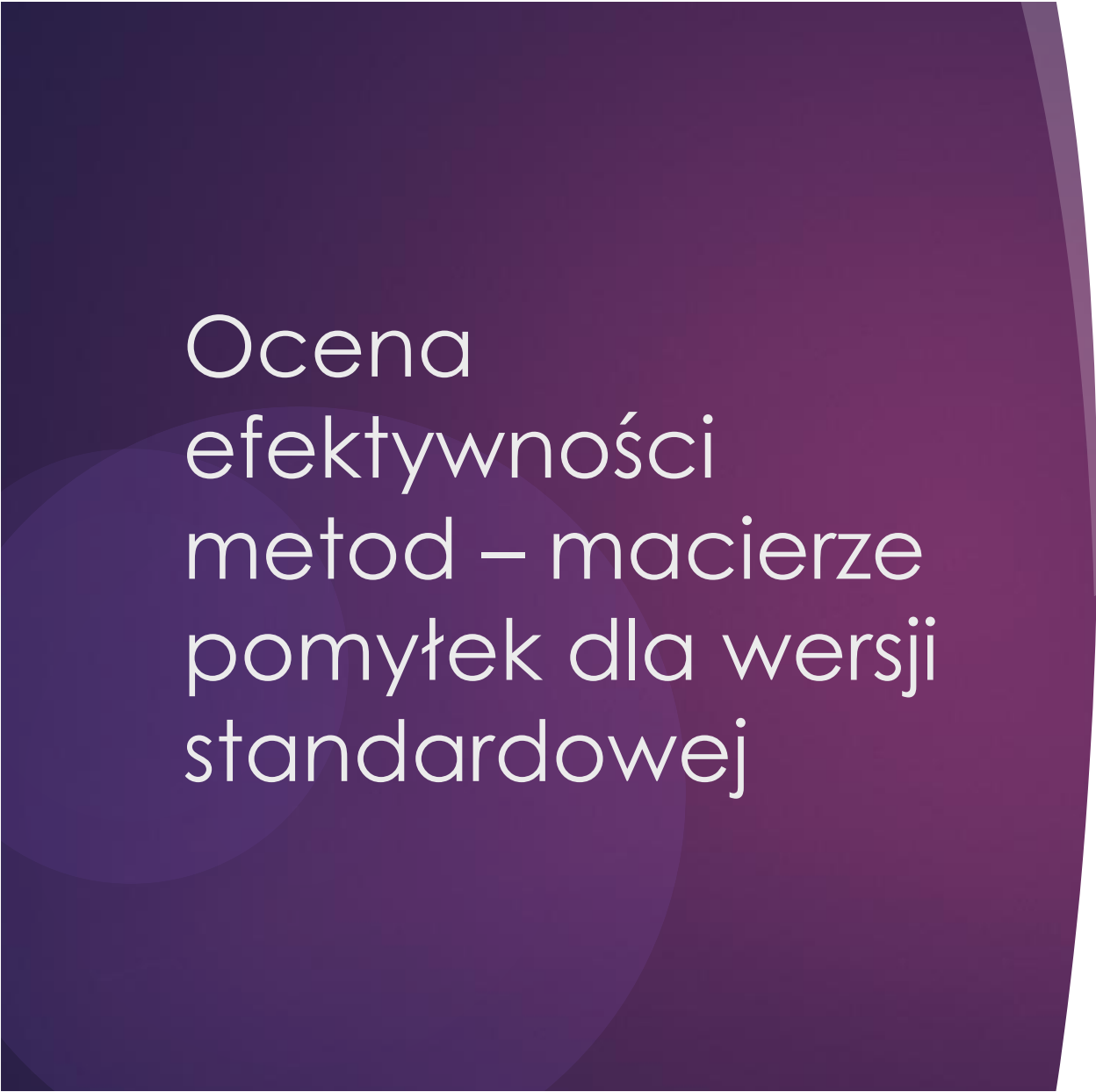
FS_untuned: Models performance comparison using accuracy-score



Czy większa liczba
klas emisyjności
pogorszy jakość
klasyfikacji?

Rozkład klas „nowej” zmiennej docelowej

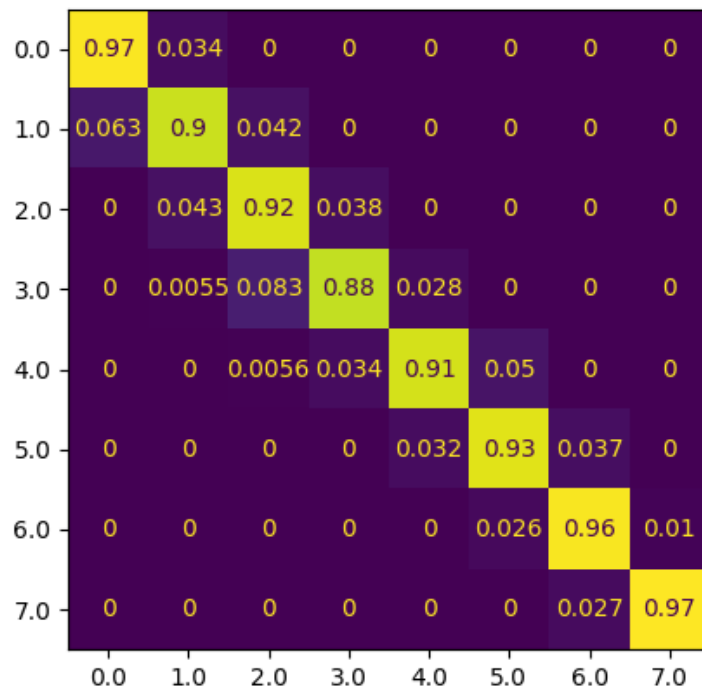




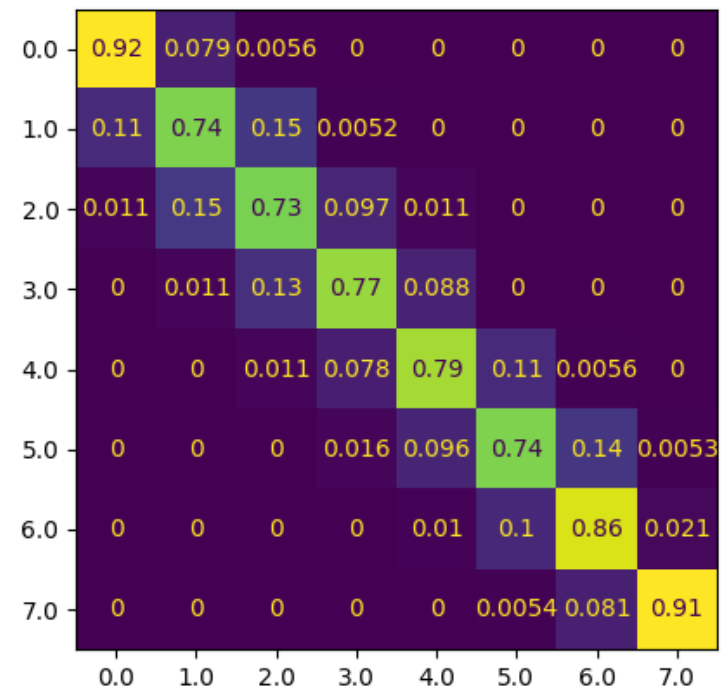
Ocena efektywności metod – macierze pomyłek dla wersji standardowej

Macierz pomyłek dla drzewa decyzyjnego, wersja podstawowa.

Conf. Matrix, DecTree



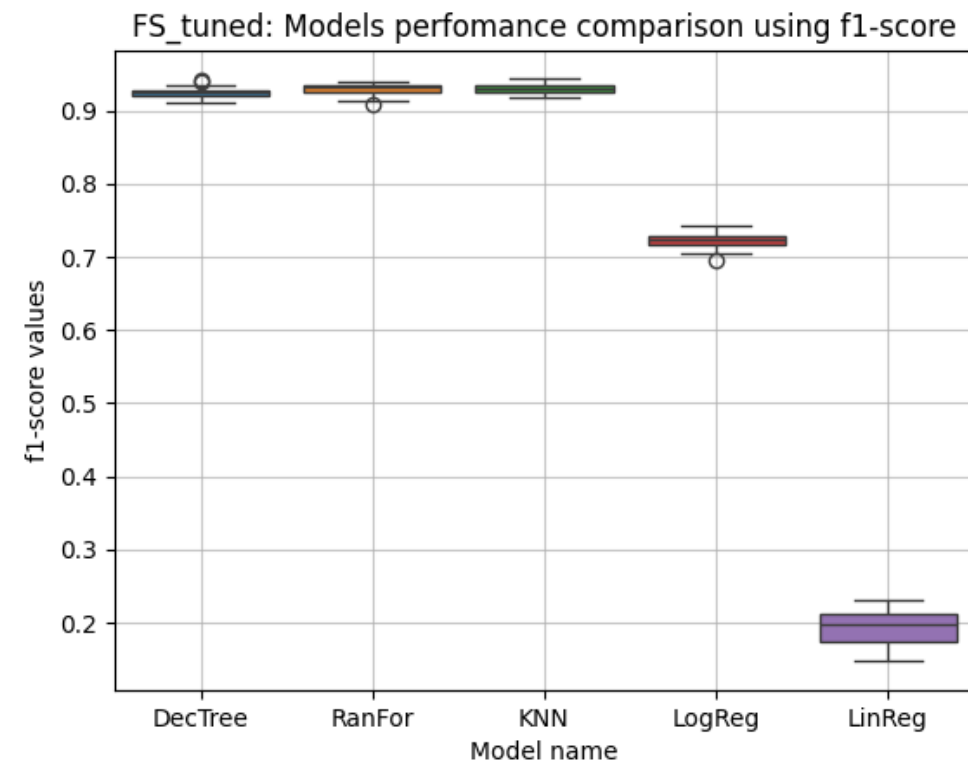
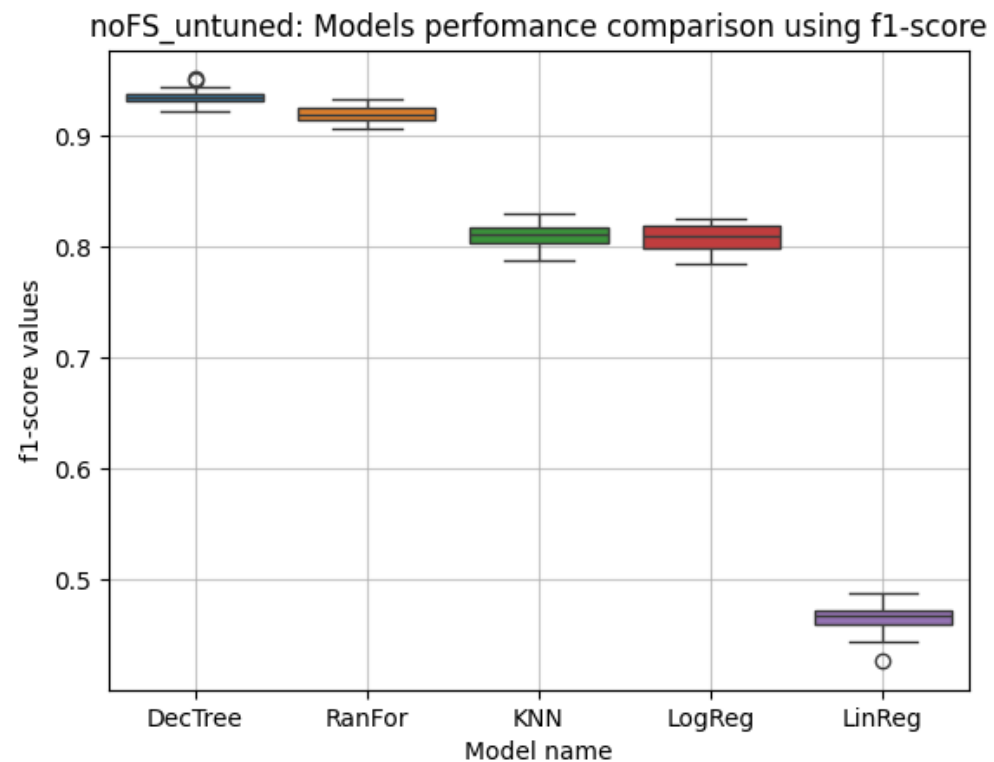
Conf. Matrix, KNN



Ocena

efektywności
metod – ~~zbiorcze~~
wykresy
pudełkowe

Zbiornicze porównanie modeli



Dalsze możliwe
kroki badań nad
efektywnością
modeli.

Co można by dodać do analizy?

- ▶ Naturalnie niniejsza analiza nie uwzględnia wszystkich aspektów badania dokładności modeli.
- ▶ Badania mogłyby zostać rozszerzone poprzez uwzględnienie takich kroków jak:
 - ▶ Optymalizacja hiperparametrów algorytmu **SFS** (sequential feature selector).
 - ▶ Metody syntezy nowych obserwacji (np. metodą **SMOTE**)
 - ▶ Analiza czynnikowa dla zmiennych jakościowych i ilościowych (**FAMD**)
 - ▶ I wiele innych metod ...



Dziękuję za uwagę ~~Waszą!~~

Pytania?

~~Paweł Nowak,~~
~~Wydział Matematyki,~~
~~Politechnika Wrocławska~~