# Zaawansowane metody uczenia maszynowego

Projekt 2

Mikołaj Małkiński

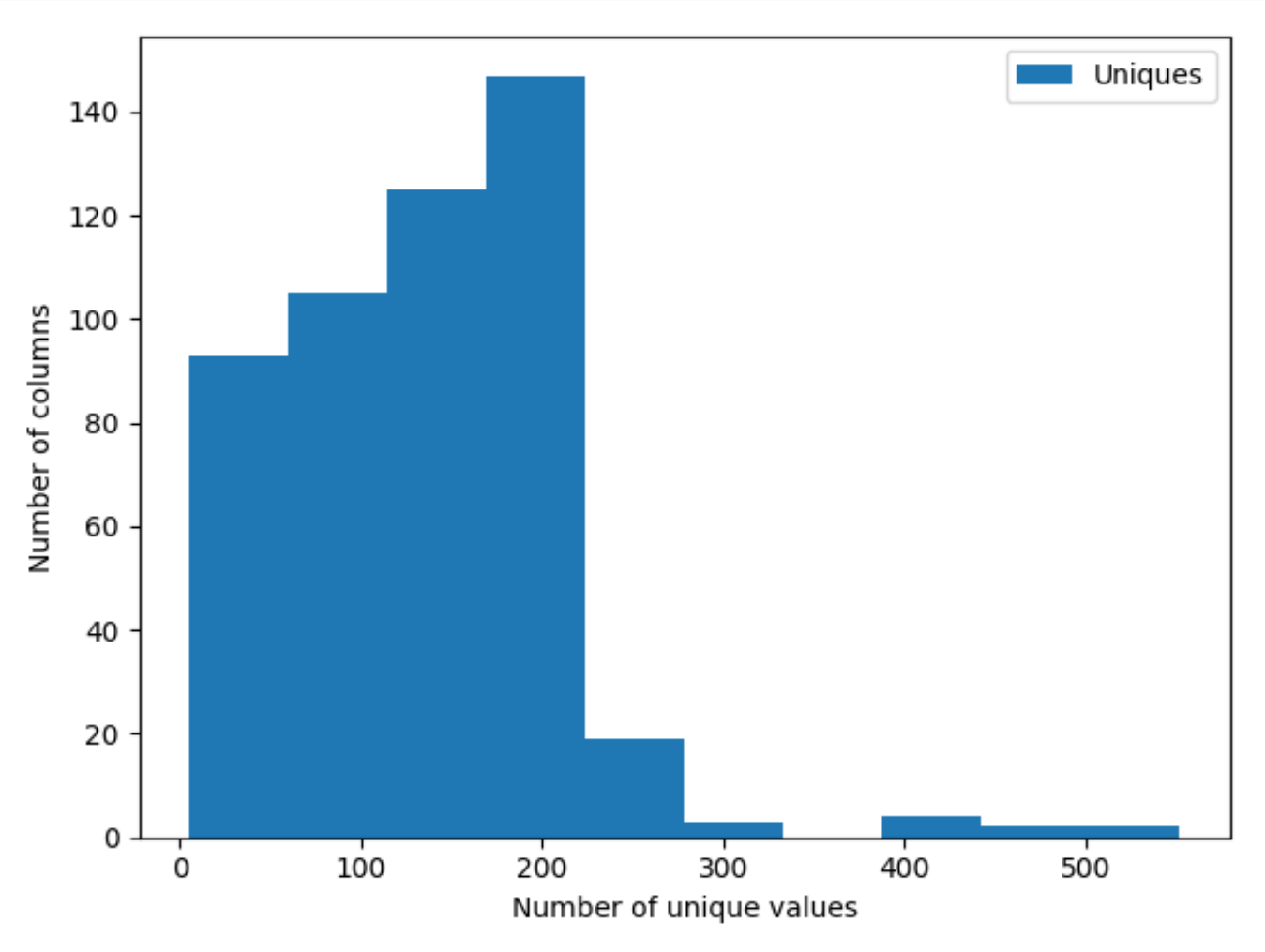
19 maja 2019

Politechnika Warszawska
Wydział Matematyki i Nauk Informacyjnych

# Spis treści

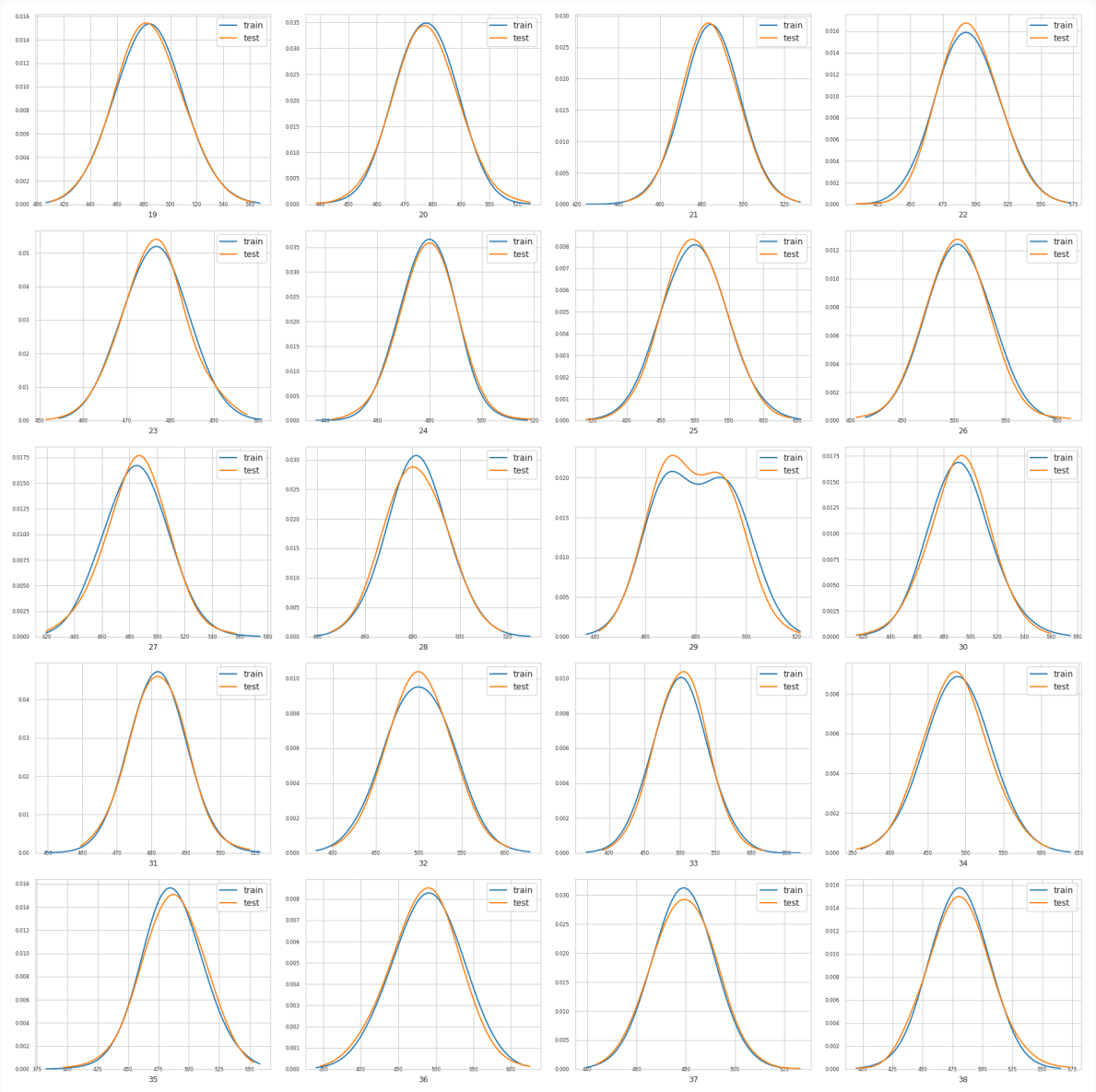
# Wstępna analiza danych

# Unikalność danych



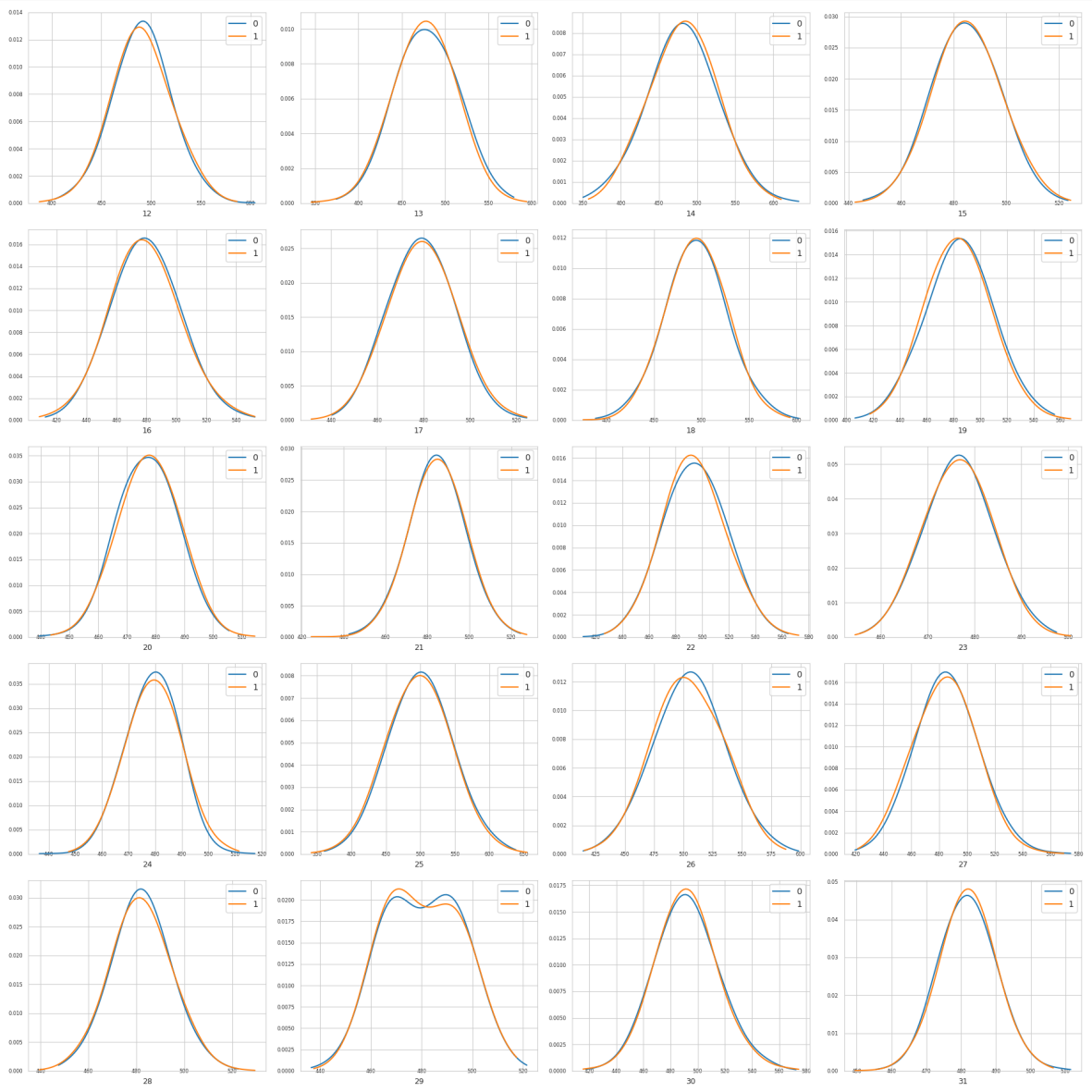

**Rysunek 1:** Unikalne wartości

# Rozkład cech z podziałem na zbiór




**Rysunek 2:** Rozkład cech z podziałem na zbiór

# Rozkład cech z podziałem na klasę



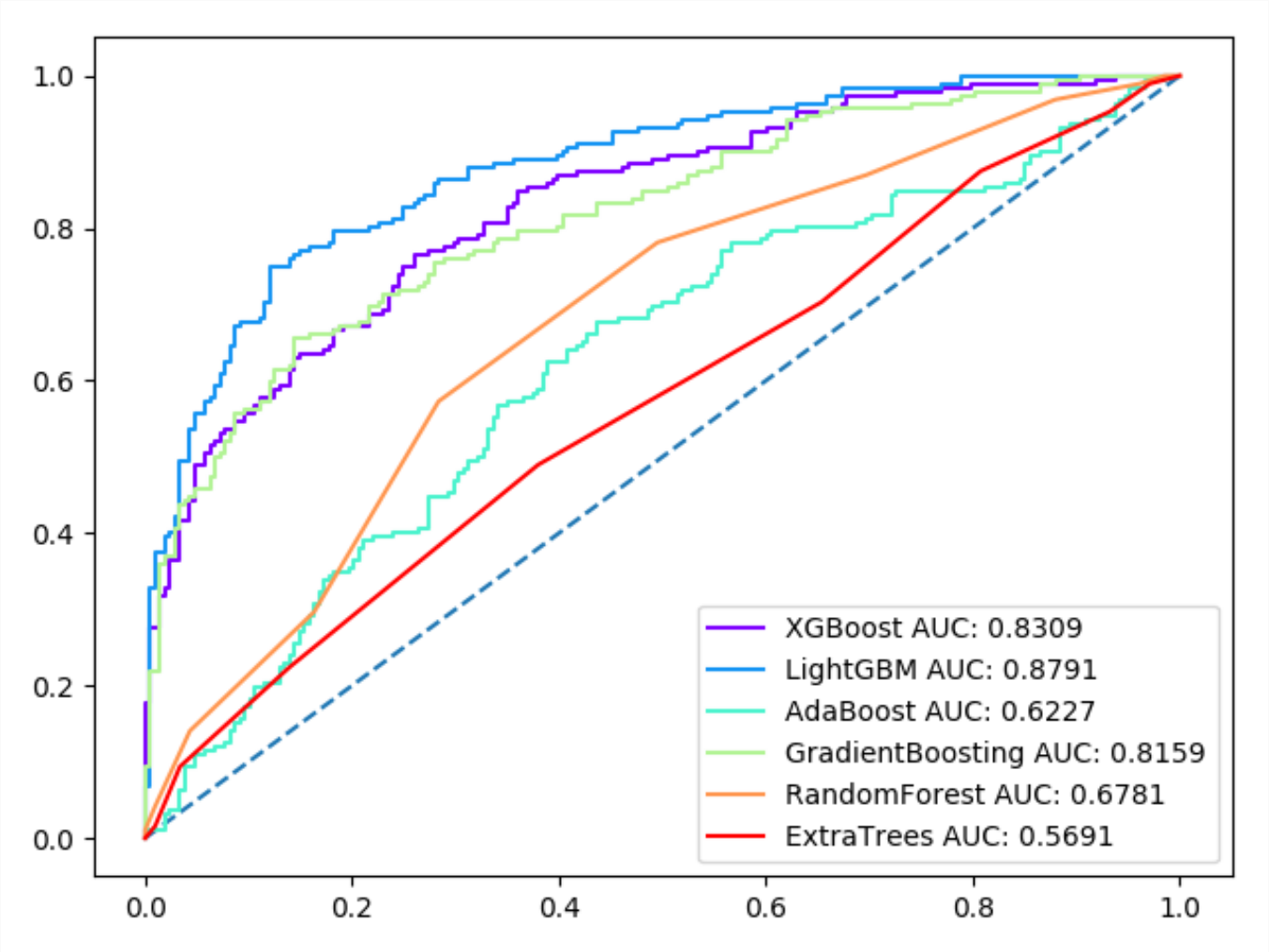

**Rysunek 3:** Rozkład cech z podziałem na klasę

# Podstawowe modele w domyślnych konfiguracjach

# Wykorzystane klasyfikatory

- XGBoost
- LightGBM
- AdaBoost
- GradientBoosting
- RandomForest
- ExtraTrees

# Krzywe ROC




**Rysunek 4:** Porównanie krzywych ROC podstawowych modeli klasyfikacyjnych

# Metryki

| | AUC | Dokładność | Zbalansowana dokładność |
|---|---|---|---|
| XGBoost | 0.8309 | 0.7400 | 0.7390 |
| LightGBM | **0.8791** | **0.8025** | **0.8023** |
| AdaBoost | 0.6227 | 0.6125 | 0.6138 |
| GradientBoosting | 0.8159 | 0.7275 | 0.7272 |
| RandomForest | 0.6781 | 0.6475 | 0.6446 |
| ExtraTrees | 0.5691 | 0.5575 | 0.5549 |

**Tablica 1:** Metryki podstawowych modeli klasyfikacyjnych

# Metryki

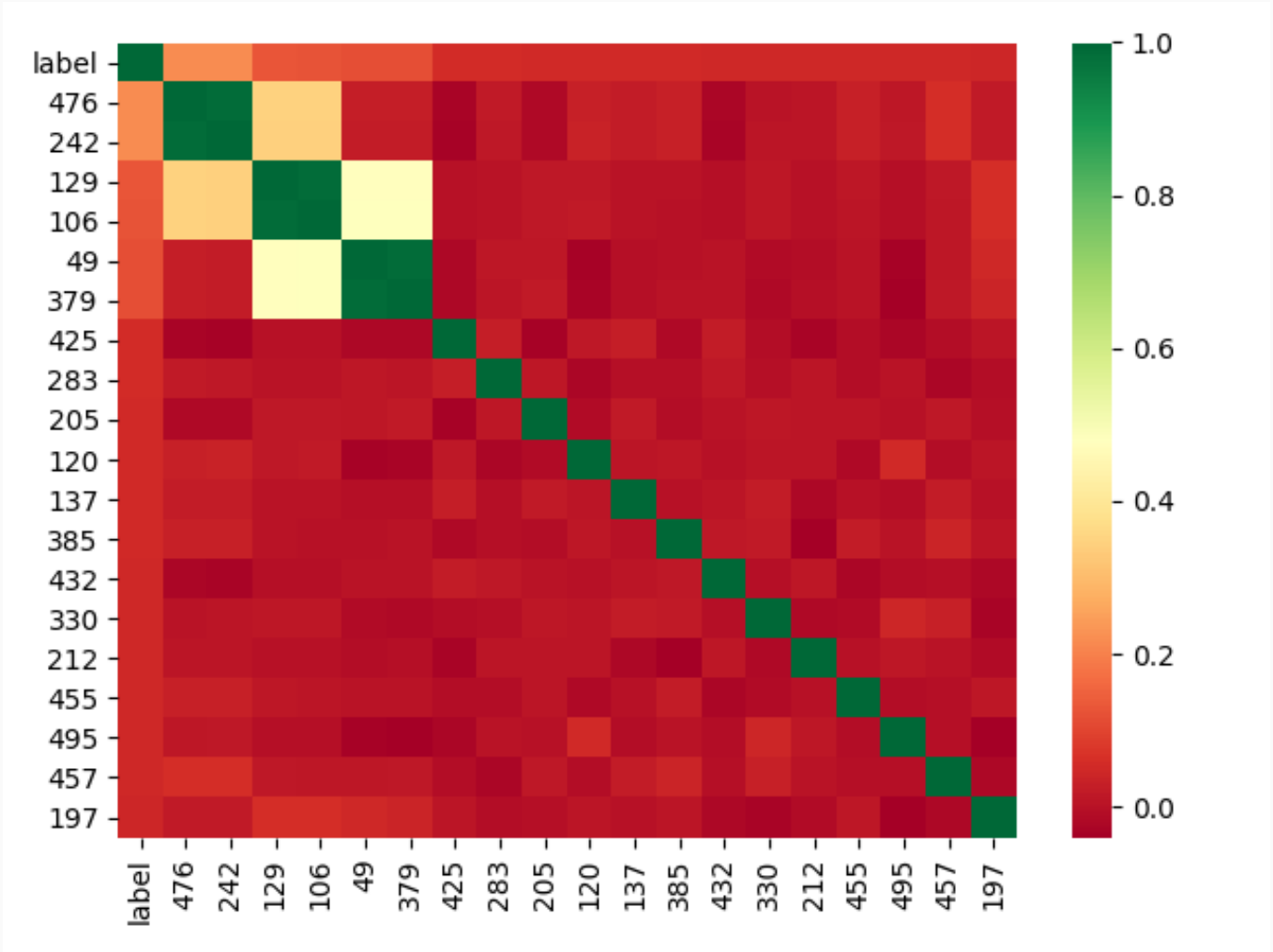
| | F1 | Precyzja | Czułość |
|---|---|---|---|
| XGBoost | 0.7249 | 0.7366 | 0.7135 |
| LightGBM | **0.7948** | **0.7927** | **0.7969** |
| AdaBoost | 0.6154 | 0.5877 | 0.6458 |
| GradientBoosting | 0.7169 | 0.7150 | 0.7188 |
| RandomForest | 0.6094 | 0.6509 | 0.5729 |
| ExtraTrees | 0.5151 | 0.5434 | 0.4896 |

**Tablica 2:** Metryki podstawowych modeli klasyfikacyjnych

# Wybór cech

# Macierz korelacji

Ważne cechy: 476, 242, 129, 106, 49, 379.

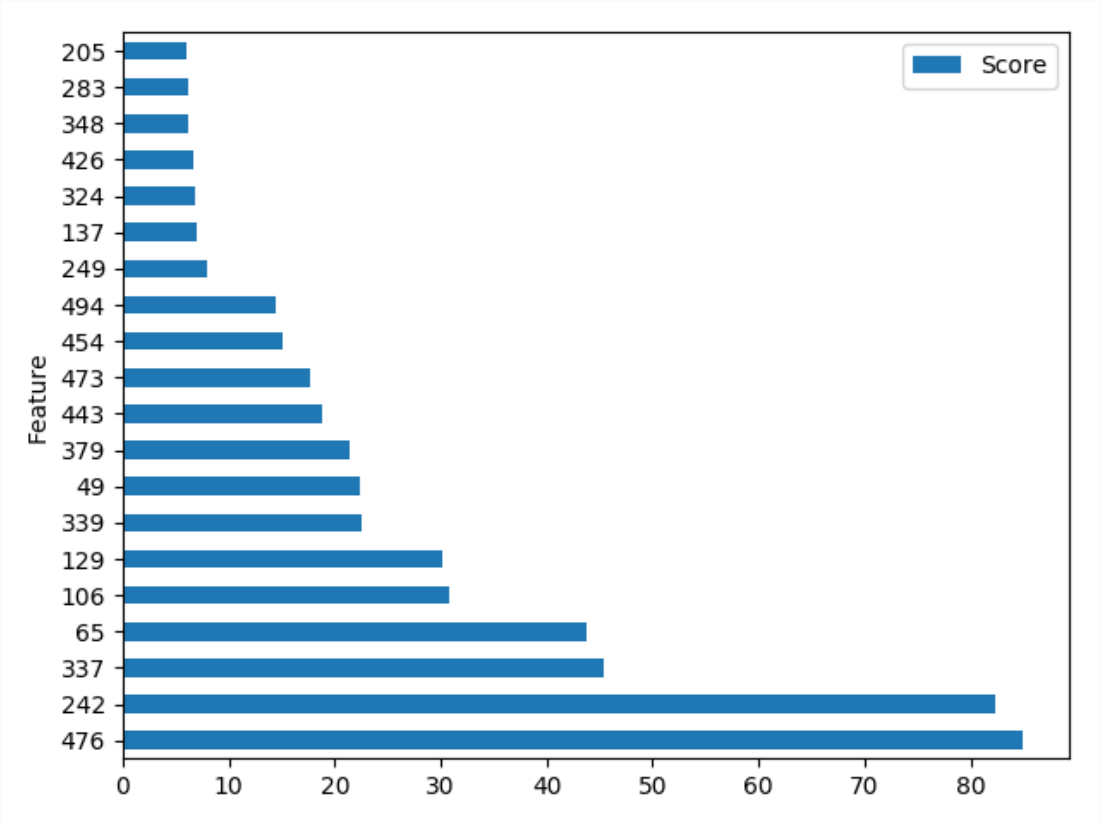



**Rysunek 5:** Macierz korelacji ograniczona do 20 największych wartości

# Analiza jednowymiarowa

Ponownie wyróżnione cechy: 476, 242, 106, 129, 49, 379.

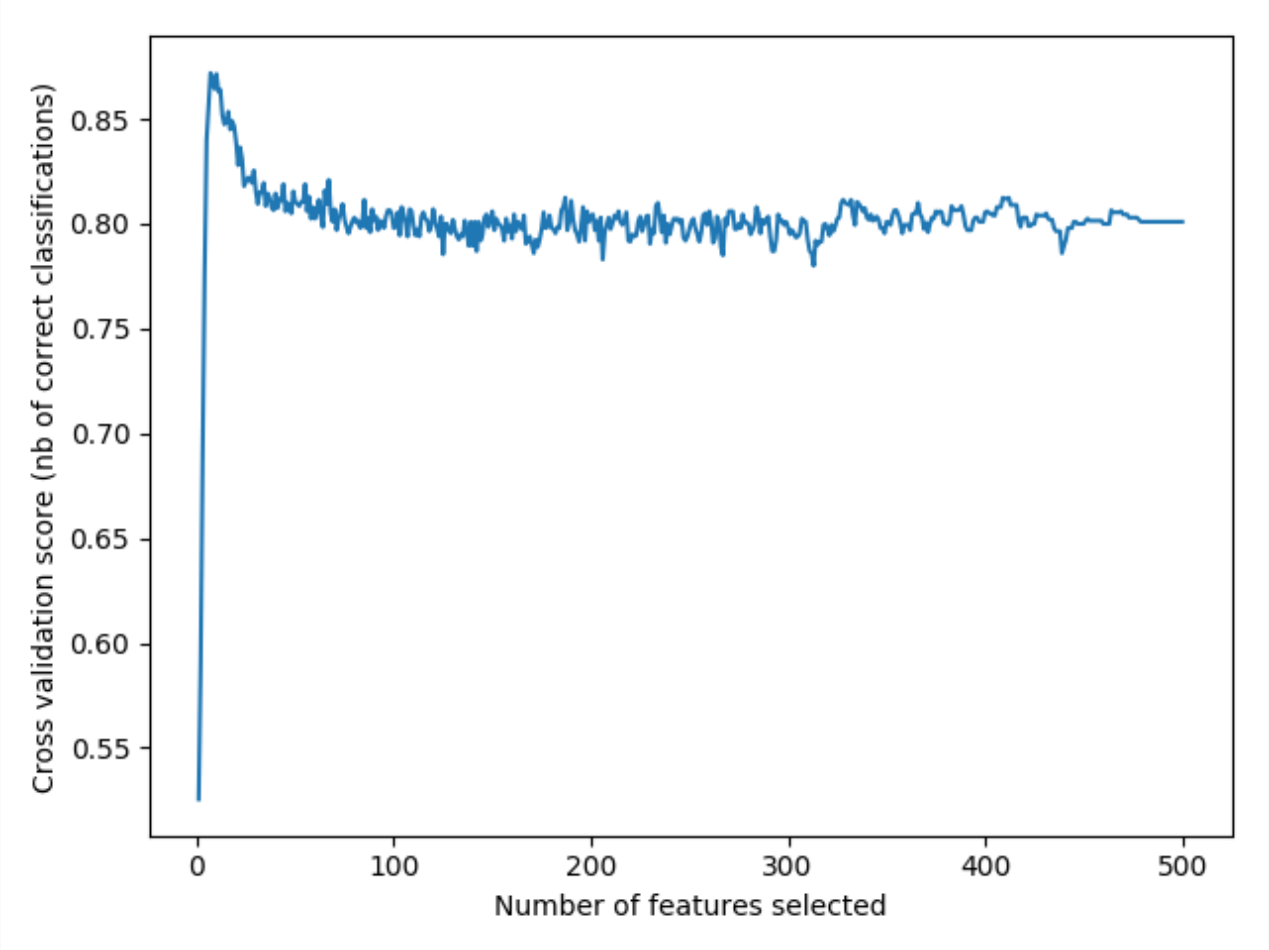
Nowe cechy: 337, 339, 443, 473, 454.




**Rysunek 6:** 20 największych wartości otrzymanych za pomocą testu ANOVA

# Rekurencyjna eliminacja cech - LightGBM

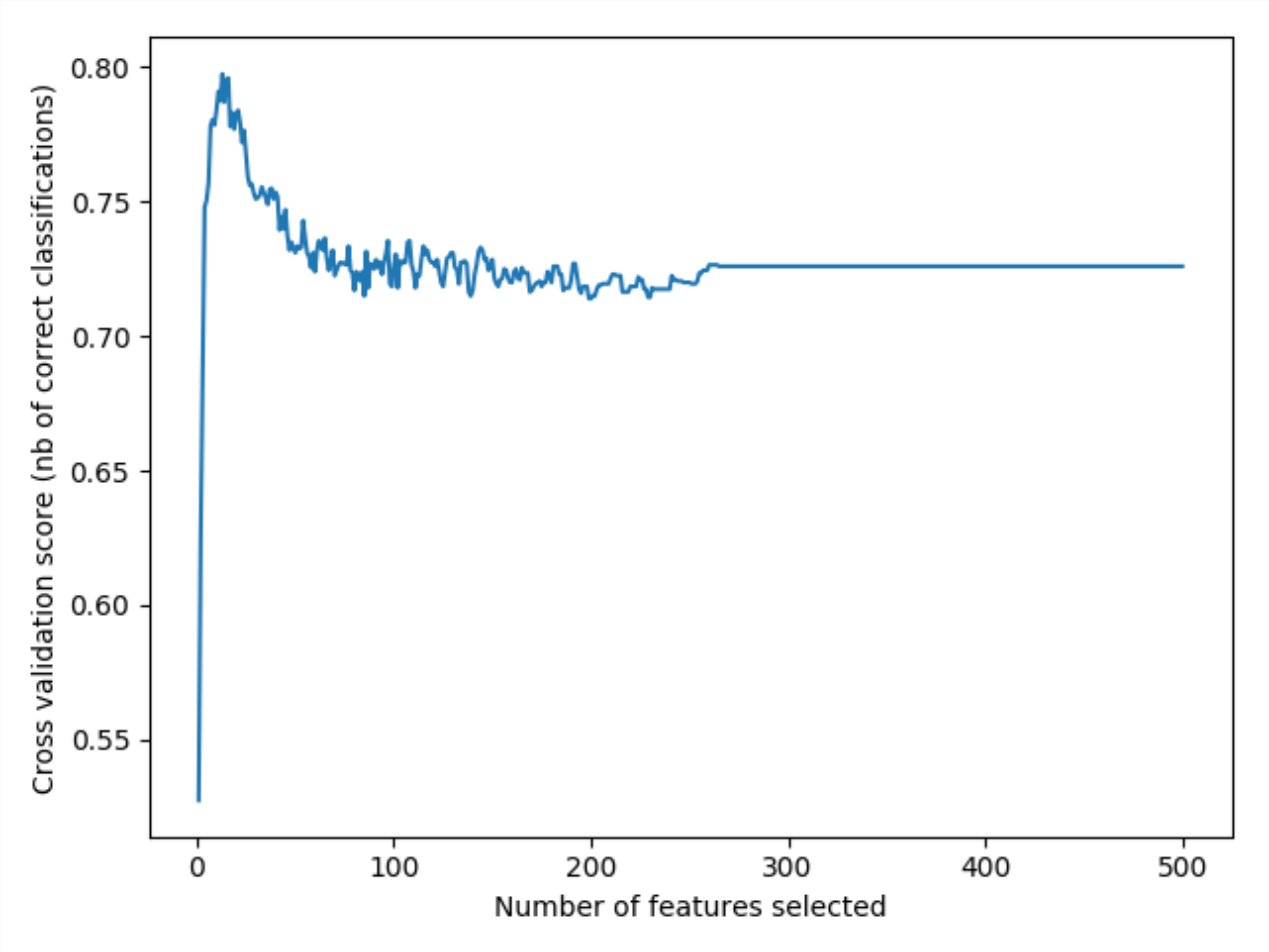
Wybrane cechy: **106**, 154, 319, **337**, **379**, **443**, **454**.




**Rysunek 7:** Zbalansowana dokładność w zależności od liczby zmiennych

# Rekurencyjna eliminacja cech - XGBoost

Wybrane cechy: 29, 49, **106**, 154, 242, 282, 319, **339**, **379**, **443**, 452, **473**, 476.




**Rysunek 8:** Zbalansowana dokładność w zależności od liczby zmiennych

# Dobór hiperparametrów

# Dobór hiperparametrów

- Grid search
- 5-krotna kroswalidacja
- Maksymalizacja zbalansowanej dokładności
- Trening tylko na cechach wybranych przez dany model
- LightGBM - 87.2%
- XGBoost - 85.9%

Dziękuję za uwagę