AutoML dla grzybiarzy

Klaudia Kwoka, Pola Mościcka,
Maciej Wach

Podstawowe informacje

- **Zadanie**: klasyfikacja grzybów na jadalne (0) i trujące (1)
- **Grupa docelowa**: osoby zajmujące się analizą danych w dziedzinie - Biologia i ekologia, branża żywnościowa, entuzjaści grzybobrania
- **Pakiet specjalizuje się w**: klasyfikacji, predykcji prawdopodobieństwa, optymalizacji modeli uczenia maszynowego, ocenie jakości predykcji
- **Dodatkowo**: zwrócenie szczególnej uwagi na miarę czułości

Przykładowe cechy danych

1.cap-diameter (n)

2.cap-shape (k)

3.cap-surface (k)

4.cap-color (k)

5.does-bruise-bleed (k)

6.gill-attachment (k)

7.gill-spacing (k)

8.gill-color (k)

9.stem-height (n)

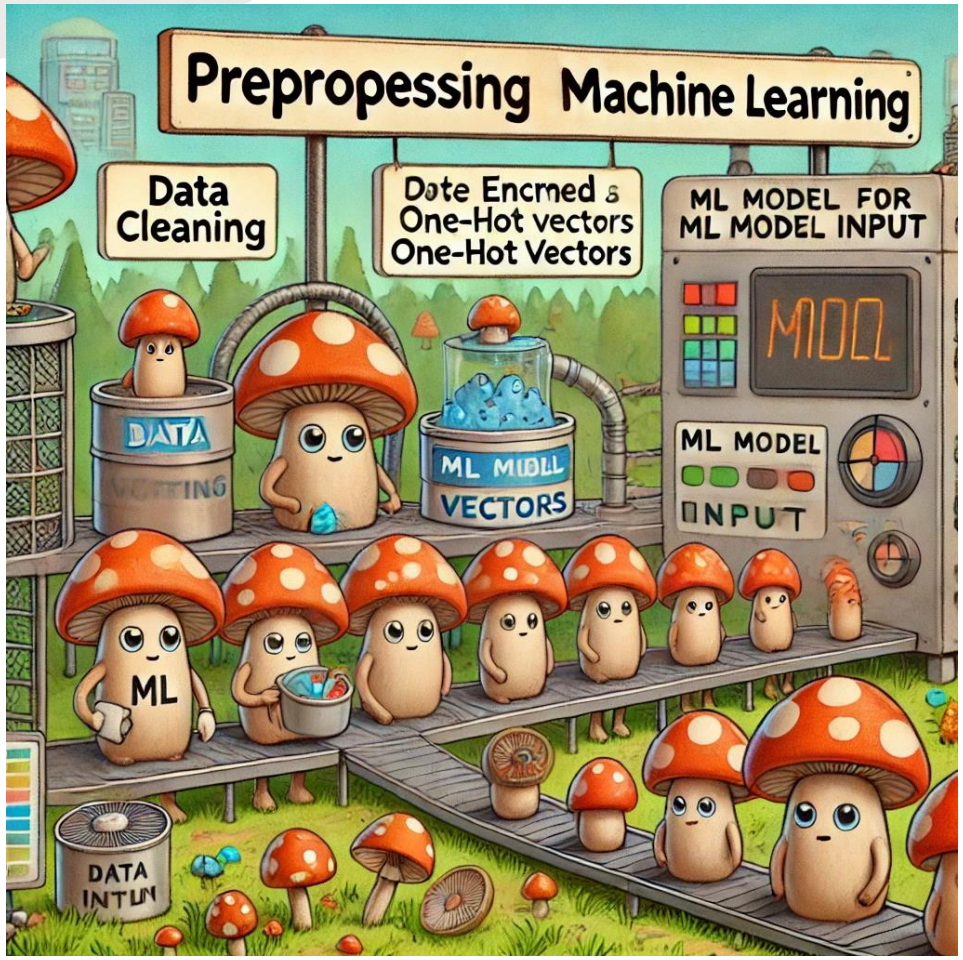
k,- kategoryczna, n - numeryczna

Preprocessing

- **Uzupełnienie brakujących danych** (SimpleImputer,

  'mean' dla zmiennych numerycznych,

  'most-frequent' dla zmiennych kategorycznych)
- **Skalowanie danych numerycznych** (MinMaxScaler na przedział [0,1])
- **Kodowanie zmiennych kategorycznych** (OneHotEncoding)
- **Wybranie istotnych zmiennych** (SelectFromModel z Random Forest)



Algorytmy klasyfikacji

- **Random Forest:** n_estimators: [1, 2000], max_features: [null, 10, 20, "sqrt", "log2"], max_depth: [4, 10]
- **Gradient Boosting:** n_estimators: [100, 500], learning_rate: [0.001, 0.999], subsample: [0.1, 0.9], max_depth: [4, 10]
- **Logistic Regression:** C: [0.0001, 10000], penalty: ["l1", "l2"], solver ="liblinear", max_iter=3000
- **KNeighbors:** n_neighbors: [1,10]
- **Voting Classifier** z powyższych modeli

Selekcja i ewaluacja modeli

- Technika Random Search
- Custom Score = 0.7 roc_auc + 0.3 recall (zbalansowanie tych miar)
- Trzy tryby: szybki, średni, wolny

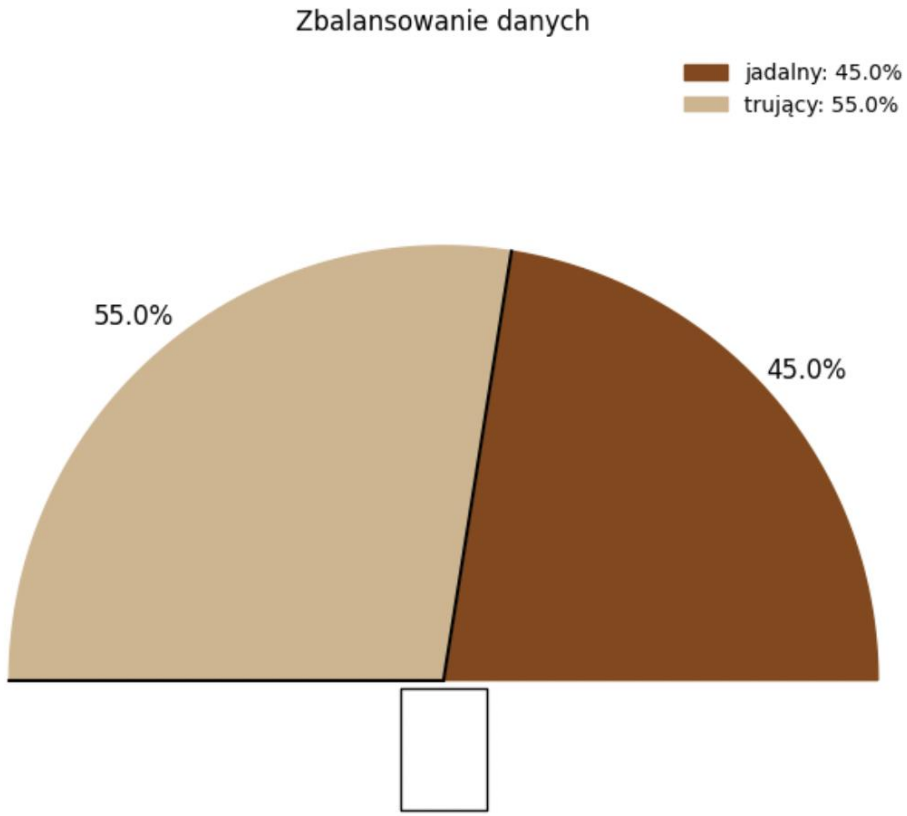
- Ocena finalnego modelu - accuracy, precision, recall, F1, ROC AUC
- Confusion Matrix, krzywa ROC

Raport dla przykładowego zbioru danych

Zbiór danych z kaggle:
prishasawhney/mushroom-dataset

Ogólne informacje o danych

```
Pakiet AutoMushroom dla grzybiarzy
Analizowane są zbiory danych z podziałem na klasy 0 lub 1, gdzie 0 oznacza jadalny grzyb, a 1 trujący.
Analiza danych:
Liczba wierszy: 43228
Liczba kolumn: 8
Liczba zmiennych kategorycznych: 0
Liczba zmiennych numerycznych: 8
Nie ma braków danych.
Balans klas:
```




```
Zbiór jest zbalansowany.
```

Preprocessing

```
Preprocessing składa się z kilku etapów:
Numeryczne dane są wypełniane średnią w przypadku braków, a następnie skalowane do zakresu [0,1] przy użyciu MinMax
Scaler.
Dane kategoryczne są uzupełniane najczęściej występującymi wartościami, a następnie kodowane za pomocą metody one-h
ot encoding.
W trybie treningowym wybierane są istotne cechy za pomocą klasyfikatora Random Forest i SelectFromModel, a dane tes
towe są ograniczane do wybranych cech.
Ważność cech:
Łącznie wybrano 5 cech.
Wybrane cechy:
Index(['cap-diameter', 'gill-attachment', 'gill-color', 'stem-width',
       'stem-color'],
      dtype='object')
```

Modele
+
ewaluacja

Analiza jakości modeli i konfiguracja finalnego komitetu:
1. Miara oceny modeli:
   Do analizy jakości modeli wykorzystano kombinację ważonych miar ROC AUC oraz Recall:
   Custom Score = (Recall: 0.3, ROC AUC: 0.7)

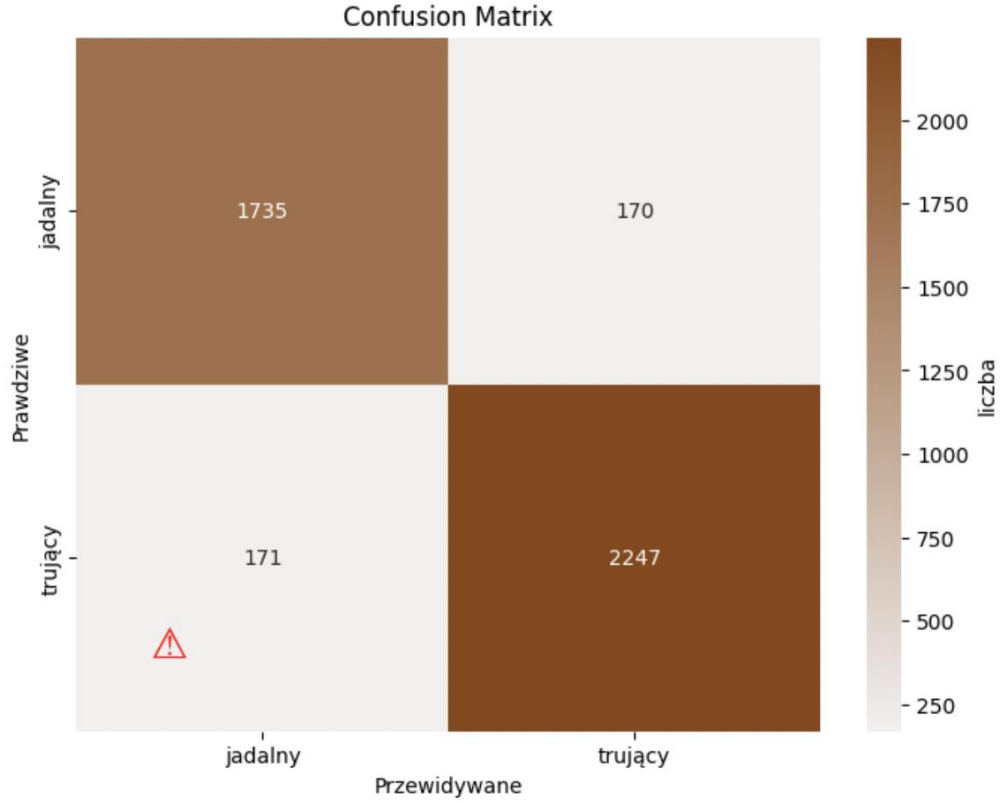
2. Modele użyte w analizie: KNeighborsClassifier, GradientBoostingClassifier, RandomForestClassifier, LogisticRegression
   Dodatkowo komitet VotingClassifier z wyżej wymienionych modeli z optymalnymi parametrami

3. Optymalizacja parametrów:
   Dla każdego z modeli, przy użyciu metody RandomizedSearch, dobrano najlepsze zestawy hiperparametrów.

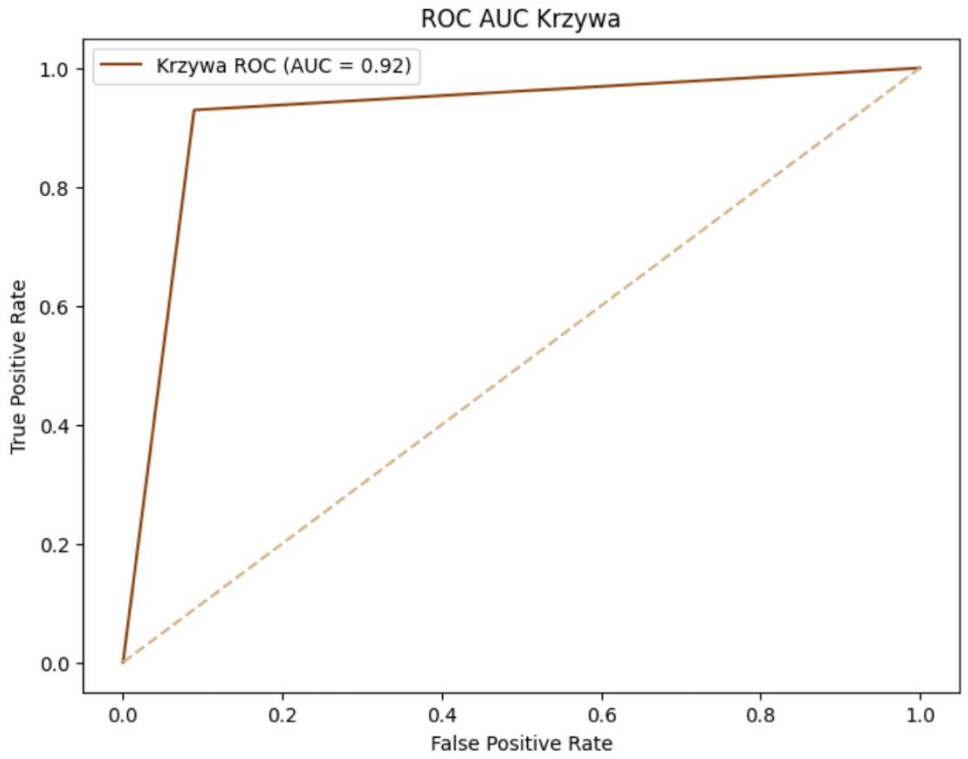
4. Parametry finalnego modelu:
KNeighborsClassifier(n_neighbors=9)

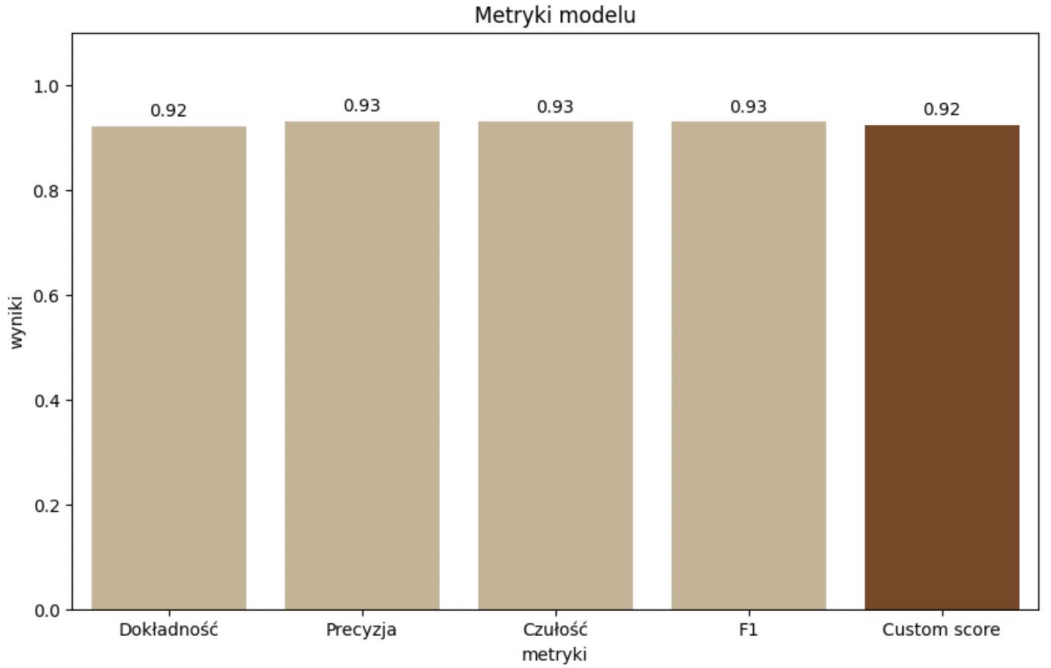
5. Czas trenowania modelu: 293.803111076355 seconds

6. Wynik Custom Score:
   Uzyskana wartość Custom Score dla tego modelu na zbiorze walidacyjnym wynosiła: 0.9227787721449899



Ewaluacja


ROC AUC Krzywa
Krzywa ROC (AUC = 0.92)
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0


Metryki modelu
wyniki
0.92
0.93
0.93
0.93
0.92
Dokładność
Precyzja
Czułość
F1
Custom score
metryki

Analiza wyników

Model osiągnął bardzo wysoką dokładność (>90%).
To oznacza, że jest wyjątkowo skuteczny w klasyfikacji grzybów jako jadalne lub trujące.

Model ma bardzo wysoką czułość (>90%),
co oznacza, że potrafi niemal bezbłędnie wykrywać trujące grzyby.
To kluczowa cecha dla zapewnienia bezpieczeństwa.

Model ma bardzo wysoką precyzję (>90%),
co oznacza, że większość grzybów sklasyfikowanych jako trujące faktycznie jest trująca.

Model osiągnął bardzo wysoki wynik F1 (>90%),
co oznacza, że dobrze równoważy precyzję i czułość.

W przypadku klasyfikacji grzybów kluczowe znaczenie ma czułość (sensitivity/recall),
ponieważ pomyłka w postaci zaklasyfikowania trującego grzyba jako jadalny
może prowadzić do poważnych konsekwencji zdrowotnych.
Dlatego model powinien być zoptymalizowany pod kątem minimalizacji tego ryzyka.

Pierwsi zadowoleni użytkownicy