Dry bean classification
Projekt 1

Bartłomiej Wójcik, Tomasz Żywicki

Politechnika Warszawska, Wydział Matematyki i Nauk Informacyjnych

20.04.2024

1. Opis problemu

W tym raporcie opisany został przebieg projektu 1 ze Wstępu do Uczenia Maszynowego (semestr letni 2024). Naszym zadaniem było przygotowanie modelu klasyfikacji dla podanego zbioru danych. W naszej pracy można było wyróżnić 3 główne etapy:
1. EDA - eksploracja danych
2. inżynieria cech i wstępne modelowanie
3. przygotowanie finalnych modeli

2. Opis zbioru danych

Nasz zespół pracował na zbiorze danych na temat fasoli. Jest on powszechnie dostepny i można go pobrać pod adresem `https://www.kaggle.com/datasets/nimapourmoradi/dry-bean-dataset-classification/`. Nasze zadanie klasyfikacyjne polegało na przewidywaniu typu fasoli.
Do wykonania projektu posługiwaliśmy się językiem Python oraz korzystaliśmy z biblioteki Scikit-learn. Nasz zbiór składał się z:
— 13,611 rekordów
— 17 kolumn

2.1. Opis zmiennych

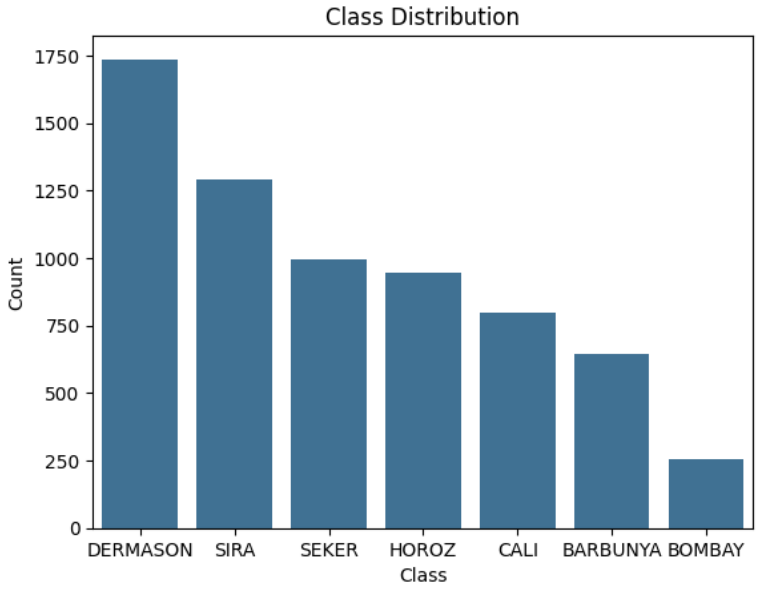
16 z 17 kolumn ma charakter numeryczny, z czego 12 z nich opisuje wymiary danej fasoli, a 4 odpowiadające za opis jej kształtu. Charakter kategoryczny posiada jedynie zmienna celu ('Class') w której zawarta jest informacja o jednym z siedmiu typów fasoli.
Przedstawimy teraz krótko nasze zmienne:

Variable Name	Type	Description
Area	Numeric	The area of a bean zone and the number of pixels within its boundaries.
Perimeter	Numeric	Bean circumference is defined as the length of its border.
MajorAxisLength	Numeric	The distance between the ends of the longest line that can be drawn from a bean.
MinorAxisLength	Numeric	The longest line that can be drawn from the bean while standing perpendicular to the main axis.
AspectRatio	Numeric	Defines the relationship between MajorAxisLength and MinorAxisLength.
Eccentricity	Numeric	Eccentricity of the ellipse having the same moments as the region.
ConvexArea	Numeric	Number of pixels in the smallest convex polygon that can contain the area of a bean seed.
EquivDiameter	Numeric	The diameter of a circle having the same area as a bean seed area.
Extent	Numeric	The ratio of the pixels in the bounding box to the bean area.
Solidity	Numeric	Also known as convexity. The ratio of the pixels in the convex shell to those found in beans.
Roundness	Numeric	Calculated with the formula: (4π * Area) / (Perimeter^2).
Compactness	Numeric	Measures the roundness of an object: Eccentricity divided by AspectRatio.
Shape Factor 1	Numeric	One of the shape factors.
Shape Factor 2	Numeric	One of the shape factors.
Shape Factor 3	Numeric	One of the shape factors.
Shape Factor 4	Numeric	One of the shape factors.
Class	Categorical	One of 7 different bean types.

Rys. 1. Krótki opis danych.

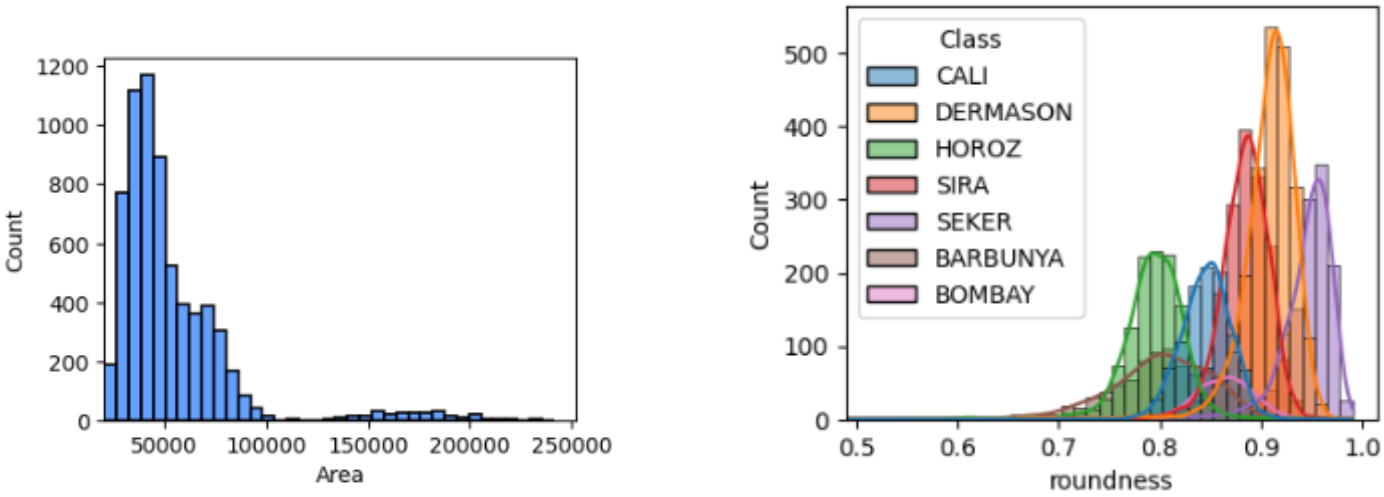
2.2. Wizualizacje i ekploracja danych

Bardzo ważnym elementem naszego zadania projektowego była eksploracja danych, która pozwoliła lepiej zrozumieć zbiór danych na którym pracowaliśmy. Zaczęliśmy więc od naszej kluczowej zmiennej czyli zmiennej celu.



Rys. 2. Rozkład zmmiennej celu. Jak widać, nie jest ona zbalansowana, a róznice w liczności pomiędzy niektórymi typami fasoli są dosyć duże (nawet 7-krotna różnica)

Przeanalizowaliśmy również rozkłady naszych numerycznych zmiennych, które najlepiej obrazują histogramy.

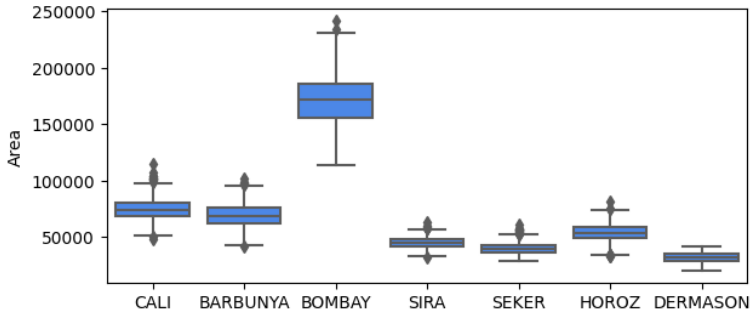


Rys. 3. Przykładowy histogram rozkładu zmiennej bez i z podziałem na typ fasoli

Powyższa analiza pozwoliła nam wysnuć poniższe wnioski:

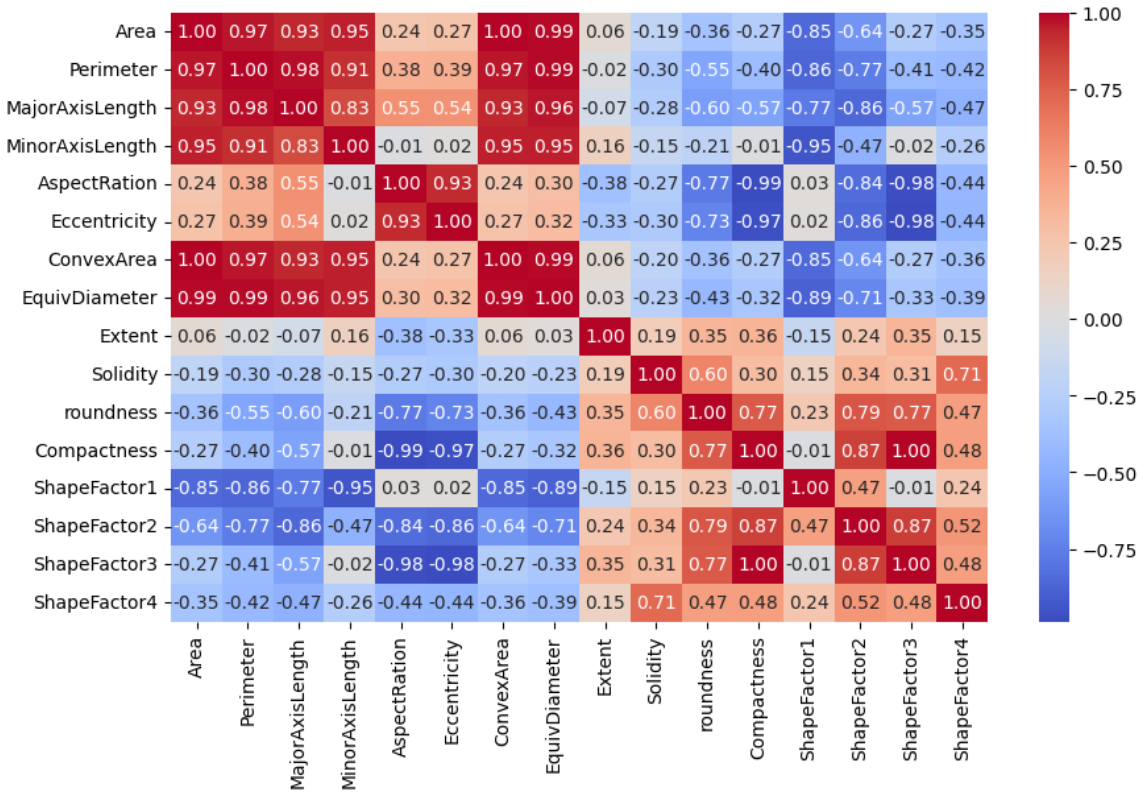
— Fasola typu 'BOMBAY' zdecydowanie wyróżnia się ze względu na atrybuty dotyczące wielkości a fasola typu 'HOROZ' ze względu na niektóre atrybuty dotyczące kształtu

— Zdecydowana większość zmiennych dla każdej z rodzajów fasoli ma rozkład zbliżony do normalnego.



Rys. 4. Fasola typu 'Bombay' jest najliczniejsza w naszym zbiorze, ale wygląda na to, że również najbardziej wyróżnia się ona pod względem kształtu i wielkości, co powinno ułatwić jej identyfikację.

Aby zoptymalizować nasze późniejsze rozwiązanie, sporządziliśmy macierz korelacji, aby dowiedzieć się które zmienne mogą nam sie przydać, a które nie. Mogliśmy również zaobserwować, które zmienne mogą nam przeszkadzać, z powodu zbyt dużej korelacji z innymi zmiennymi.

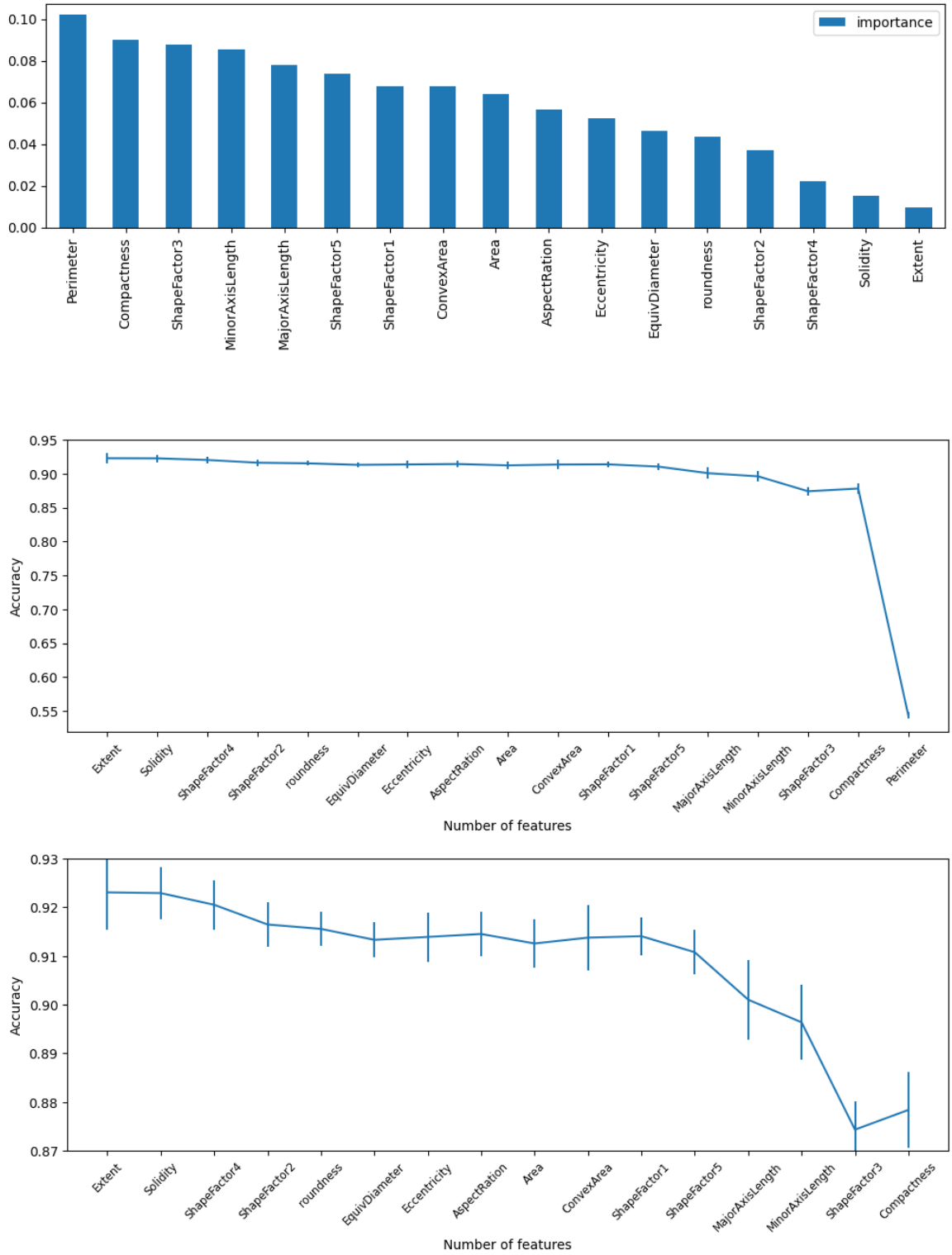


Rys. 5. Macierz korelacji.

3. Inżynieria cech.

Na podstawie eksploracji danych mogliśmy wskazać kilka aspektów, które można było poprawić, aby poprzez preprocessing danych uzyskać bardziej satysfakcjonujące wyniki.
Jako iż nasz zbiór danych nie posiadał zmiennych kategorycznych ani braków danych to w tych aspektach nie musieliśmy podejmować żadnych działań takich jak kodowanie zmiennych czy usuwanie wierszy.
Przeprowadziliśmy również standaryzację naszych danych przy użyciu funkcji *RobustScaler*. Jest on mniej podatny na outliery niż *StandardScaler*.

Sprawdziliśmy również wartości ważności naszych zmiennych przy użyciu *Random Forest Classifier*, aby sprawdzić które kolumny potencjalnie nie mają dla nas dużego znaczenia i warto byłoby je odrzucić przed dalszą fazą modelowania.



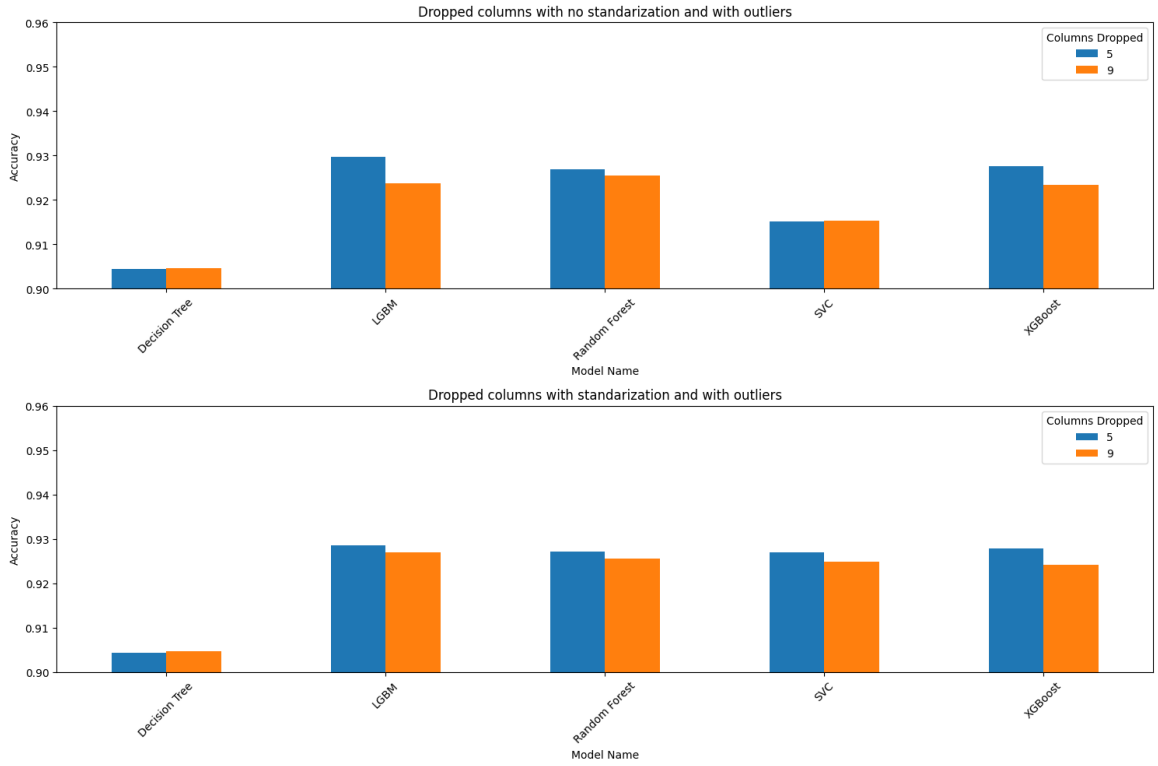
Rys. 6. Wykresy przedstawiają dokładność dla danej ilości zmiennych, zostały one posortowane od najmniej znaczącej (np. dla Solidity nie były już brane zmienne po jej lewej, te mniej znaczące czyli w tym przypadku jedynie Extent

Analiza tych wykresów pozwalała nam wyselekcjonować grupy kolumn, których moglibyśmy nie brać pod uwagę dlatego, że nie będą one miały wpływu na nasze rozwiązanie albo wpływ ten będzie niewielki
Bardzo dobrym przykładem jest tutaj kolumna *ShapeFactor3*, której usunięcie i pozostawienie jedynie dwóch zmiennych i tak powoduje zwiększenie dokładności (jest to najprawdopodobniej związane z tym, że *ShapeFactor3*

jest ściśle skorelowany z *Compactness* co może być problemem).

Wyselekcjonowaliśmy 2 grupy kolumn do usunięcia:
— składająca się z 5 kolumn - ['ShapeFactor3', 'EquivDiameter', 'Eccentricity', 'Area', 'ConvexArea']
— składająca się z 9 kolumn - ['ShapeFactor3', 'EquivDiameter', 'Eccentricity', 'Area', 'ConvexArea', 'Extent', 'ShapeFactor2', 'roundness', 'AspectRation']

Następnie sprawdziliśmy usunięcie której grupy daje lepsze rezultaty. Do weryfikacji skorzystaliśmy z 5 najlepiej sprawdzających się podstawowych modeli przetestowanych na surowych danych (o tym więcej w następnej sekcji tego raportu).
Otrzymaliśmy następujące wyniki:



Rys. 7. Nie ma znacznej różnicy w wynikach pomiędzy zestawami kolumn, więc wybierzemy ten mniej liczny dla uproszczenia modelu

4. Modelowanie

4.1. Wstępne modelowanie

Modelowanie rozpoczęliśmy od sprawdzenia podstawowych modeli oraz kilku bardziej zaawansowanych jednak z podstawowymi parametrami. Przetestowaliśmy je na nie przeprocesowanych danych aby otrzymać pewien punkt wyjścia do naszego późniejszego rozwiązania. Jakość kolejnych tworzonych modeli ocenialiśmy wykorzystując szereg metryk takich jak: accuracy, precision, recall czy f1score. Obliczaliśmy wyniki modeli nie tylko na zbiorze testowym, ale też w kroswalidacji na zbiorze treningowym.
Oto wyniki naszych testów dla nieprzeprocesowanych danych:

	Model Name	Accuracy	Precision	Recall
0	LGBM	0.9286	0.9289	0.9286
1	XGBoost	0.9286	0.9289	0.9286
2	Random Forest	0.9220	0.9222	0.9220
3	SVC	0.9108	0.9114	0.9108
4	Decision Tree	0.9080	0.9091	0.9080
5	K Nearest Neighbors	0.7118	0.7086	0.7118
6	Linear SVC	0.4781	0.4452	0.4781
7	Dummy Classifier	0.2606	0.0679	0.2606

Rys. 8. W naszych późniejszych rozważaniach oraz analizie feature importance bedziemy posługiwali się 4 najlepszymi modelami.

4.2. Tuning hiperparametrów

Po wstepnym modelowaniu, skupiliśy naszą uwagę na 4 najlepiej prosperujących algorytmach(LGBM, XGBoost, RandomForestClassifier oraz SVC). To właśnie dla tych modeli zastosowaliśmy GridSearch z kroswalidacją w celu poszukiwania najlepszych parametrów. Oto wyniki naszych poszukiwań:

Model: Random Forest

	precision	recall	f1-score	support
BARBUNYA	0.93	0.87	0.90	181
BOMBAY	1.00	1.00	1.00	77
CALI	0.94	0.94	0.94	249
DERMASON	0.91	0.96	0.93	491
HOROZ	0.96	0.95	0.96	268
SEKER	0.95	0.96	0.96	295
SIRA	0.87	0.83	0.85	345
accuracy			0.92	1906
macro avg	0.94	0.93	0.93	1906
weighted avg	0.92	0.92	0.92	1906

Accuracy: 0.925

Model: LGBM

	precision	recall	f1-score	support
BARBUNYA	0.91	0.89	0.90	181
BOMBAY	1.00	1.00	1.00	77
CALI	0.96	0.94	0.95	249
DERMASON	0.91	0.95	0.93	491
HOROZ	0.95	0.96	0.95	268
SEKER	0.95	0.96	0.95	295
SIRA	0.89	0.84	0.86	345
accuracy			0.93	1906
macro avg	0.94	0.93	0.94	1906
weighted avg	0.93	0.93	0.93	1906

Accuracy: 0.9271

Model: XGBoost

	precision	recall	f1-score	support
0	0.92	0.90	0.91	181
1	0.99	1.00	0.99	77
2	0.96	0.93	0.95	249
3	0.91	0.95	0.93	491
4	0.96	0.96	0.96	268
5	0.95	0.97	0.96	295
6	0.87	0.83	0.85	345
accuracy			0.93	1906
macro avg	0.94	0.93	0.93	1906
weighted avg	0.93	0.93	0.93	1906

Accuracy: 0.926

Model: SVM

	precision	recall	f1-score	support
BARBUNYA	0.95	0.91	0.93	181
BOMBAY	1.00	1.00	1.00	77
CALI	0.95	0.95	0.95	249
DERMASON	0.90	0.95	0.93	491
HOROZ	0.96	0.95	0.95	268
SEKER	0.96	0.97	0.96	295
SIRA	0.89	0.83	0.86	345
accuracy			0.93	1906
macro avg	0.94	0.94	0.94	1906
weighted avg	0.93	0.93	0.93	1906

Accuracy: 0.9302

Rys. 9. Wygląda na to że naszym docelowym modelem będzie SVC.

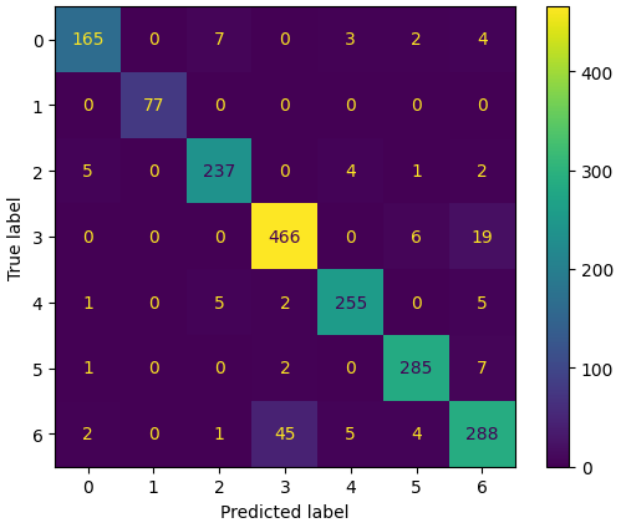
4.3. Finalne modelowanie

Ostatecznie zakończyliśmy z dwoma modelami:
— model 1 - SVC z uzyskanymi wcześniej hiperparametrami
— model 2 - uzyskany metodą stackingu na podstawie naszych najlepszych modeli, z końcowym estymatorem, którym była regresja logistyczna

Model 1:

	precision	recall	f1-score	support
0	0.95	0.91	0.93	181
1	1.00	1.00	1.00	77
2	0.95	0.95	0.95	249
3	0.90	0.95	0.93	491
4	0.96	0.95	0.95	268
5	0.96	0.97	0.96	295
6	0.89	0.83	0.86	345
accuracy			0.93	1906
macro avg	0.94	0.94	0.94	1906
weighted avg	0.93	0.93	0.93	1906

Accuracy: 0.9302

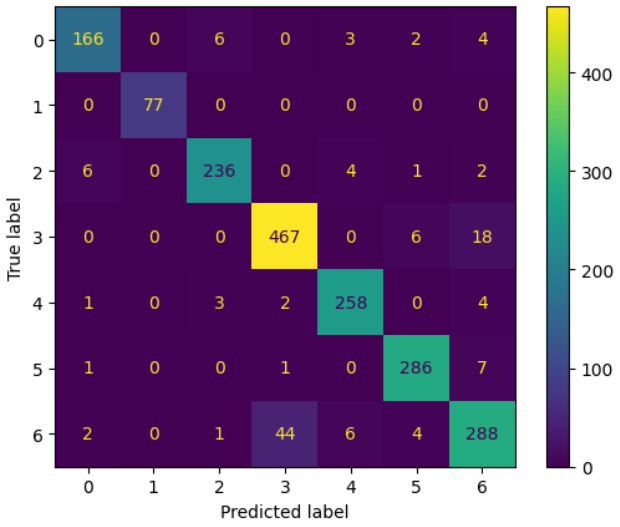


Rys. 10. Wyniki metryk oraz macierz pomyłek dla modelu 1.

Model 2:

	precision	recall	f1-score	support
0	0.94	0.92	0.93	181
1	1.00	1.00	1.00	77
2	0.96	0.95	0.95	249
3	0.91	0.95	0.93	491
4	0.95	0.96	0.96	268
5	0.96	0.97	0.96	295
6	0.89	0.83	0.86	345
accuracy			0.93	1906
macro avg	0.94	0.94	0.94	1906
weighted avg	0.93	0.93	0.93	1906

Accuracy: 0.9328



Rys. 11. Wyniki metryk oraz macierz pomyłek dla modelu 2.

5. Podsumowanie

Finałowe modele bardzo dobrze klasyfikują rodzaje fasoli poza dwoma o nazwach Sira i Dermason. Jest to najprawdopodobniej związane z tym, że rodzaje te mają bardzo zbliżone do siebie charakterystyki i żeby je rozróżnić, przydałyby się jeszcze jakieś zmienne opisujące na przykład kolorystykę.

Jeśli chodzi o zastosowanie biznesowe naszego modelu, mógłby on być przydatny na przykład dla sortowni fasoli, która dzięki niemu mogłaby w prostszy sposób segregować dane rodzaje ziaren, z tym, że tak jak już wspomnieliśmy wcześniej, musiałaby ona dostarczyć również jakieś dodatkowe dane pozwalające w lepszy sposób odróżnić od siebie rodzaje fasól Sira i Dermason. W przypadku fasól innego rodzaju, uważamy, że nasz model poradziłby sobie zdecydowanie wystarczająco dobrze by móc używać go na większą skalę.