RAPORT

WUM PROJEKT 1 - KLASYFIKACJA

Członkowie zespołu: Zofia Kamińska, Mateusz Deptuch
Walidatorzy: Natalia Choszczyk, Karolina Dunal

I Temat projektu

Przewidywanie rodzaju fasolki na podstawie wymiarów i cech fizycznych

Dane: *Dry Bean Dataset*
Zbiór danych Dry Bean Dataset zawiera informacje na temat 16 różnych cech fizycznych fasolek, takich jak długość, szerokość, powierzchnia, kształt itp. oraz ich etykiet, określających jeden z siedmiu rodzajów fasolki.

(**kaggle:** https://www.kaggle.com/datasets/muratkokludataset/dry-bean-dataset)
cytacja: SKOKLU, M. and OZKAN, I.A., (2020), Multiclass Classification of Dry Beans Using Computer Vision and Machine Learning Techniques. Computers and Electronics in Agriculture, 174, 105507.

II Cel projektu/motywacja:

Projekt ten ma na celu wykorzystanie uczenia maszynowego do klasyfikacji różnych gatunków fasoli na podstawie ich cech fizycznych. Do tego celu wykorzystujemy zbiór danych Dry Bean Dataset, który zawiera informacje na temat 16 różnych cech fizycznych fasolek, takich jak długość, szerokość, powierzchnia, kształt itp. oraz ich etykiet, określających rodzaj fasolki.
Potencjalne zastosowania biznesowe:

- hurtownie żywności
- aplikacja do gotowania, używająca zdjęć składników użytkownika
- kontrola jakości na plantacji fasoli

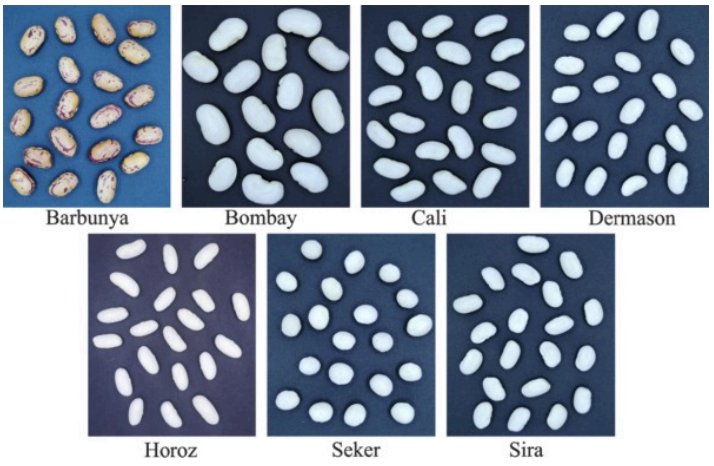
III Proces modelowania

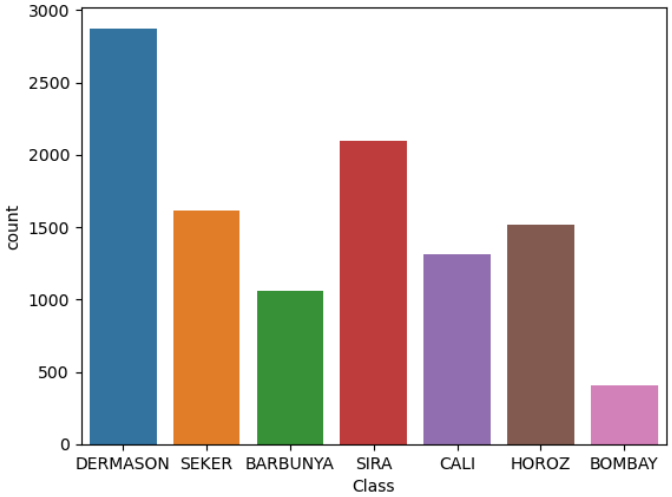
1. EDA:

1.1. Opisy cech

Pierwszym krokiem eksploracji danych jest zapoznanie się z cechami, których w naszym przypadku było 16. Według dokumentacji datasetu, pomiary fasolek dokonane zostały na podstawie zdjęć. Były to głównie cechy fizyczne takie jak pole powierzchni, ilość pikseli, "okrągłość" fasolek i inne.

Na podstawie tych właśnie cech przydzielać będziemy próbki do jednej z siedmiu kategorii: Seker, Barbunya, Bombay, Cali, Dermason, Horoz, Sira (zdjęcia poniżej).

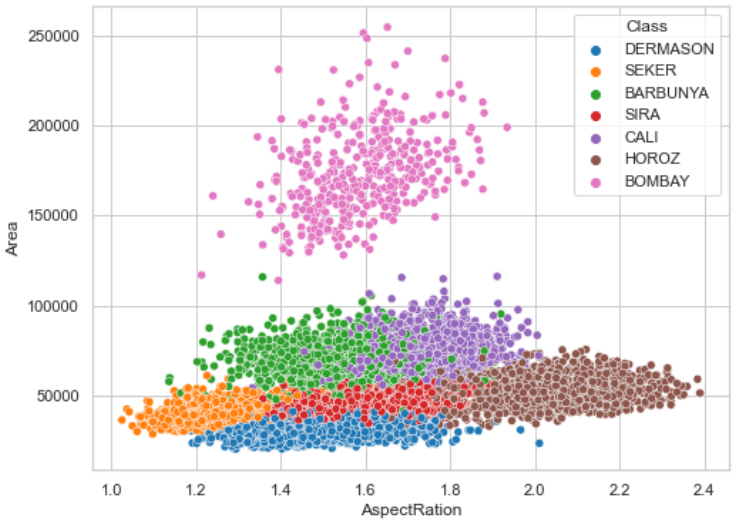




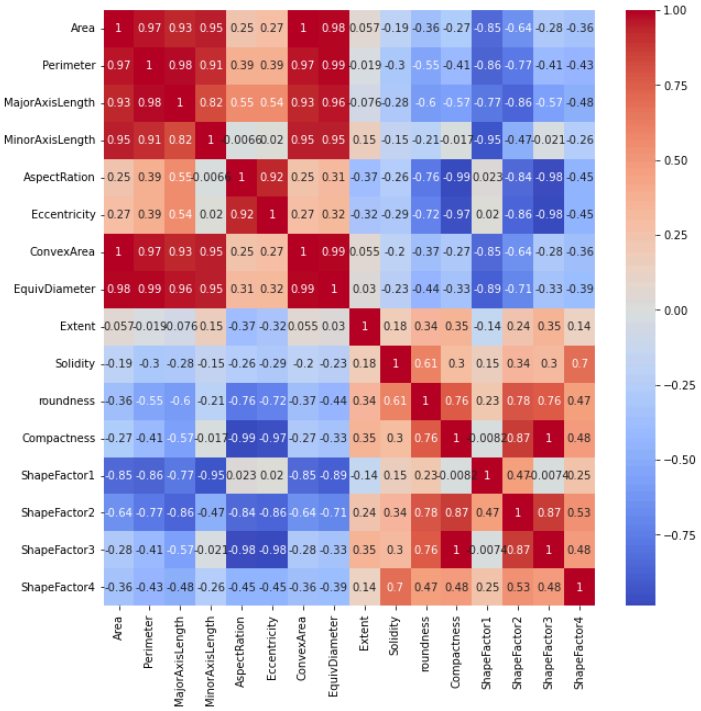
https://ars.els-cdn.com/content/image/1-s2.0-S0168169919311573-gr3.jpg

W naszych danych nie znaleźliśmy żadnych wartości NULL, ani definitywnych outlierów. Próbki nie były rozłożone równomiernie między klasami (wykres powyżej), jednak te, których było najmniej okazały się w późniejszym etapie projektu najłatwiejsze do odróżnienia. W związku z tym zrezygnowaliśmy z oversamplingu danych.

Z dalszej eksploracji wynikło, że fasolka Bombay powinna być najłatwiejsza do oddzielenia od innych, gdyż na wykresach najbardziej "odstaje".

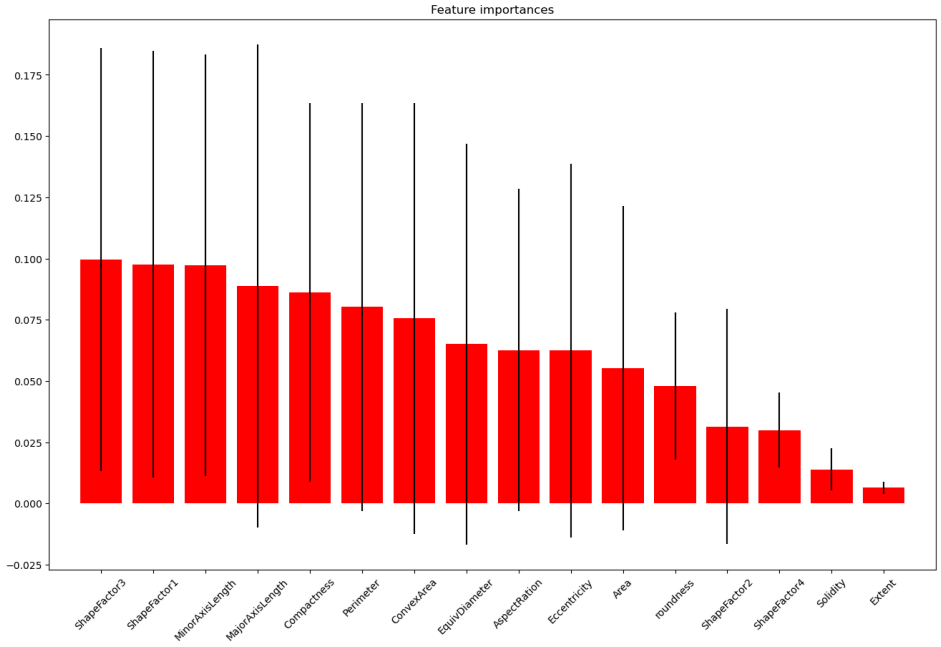


Zauważyliśmy, że niektóre z cech mają bardzo wysoką korelację (równą nawet 1). Zdecydowaliśmy się usunąć niektóre z tych kolumn ('ShapeFactor3','EquivDiameter', 'Area').



2. Feature Engineering + pierwsze modelowanie:

Za pomocą random forest classifier określiliśmy najważniejsze cechy dla przewidywań naszego modelu.



Zdecydowaliśmy się usunąć dwie cechy o najmniejszym wyniku, jednakże z dalszych testów wynikło, że nieusuwanie żadnych z kolumn daje lepsze rezultaty.

Przetestowaliśmy kilka metod normalizacji, z których wyłoniliśmy:

- box-cox, dla regresji logistycznej
- StandardScaler, dla reszty modeli

Sprawdziliśmy również jak wyniki różnią się dla dwóch rodzajów encodingu:

- LabelEncoding - najprostsze rozwiązanie
- OneHotEncoding - mimo nadziei w nim pokładanych wypadał zawsze tak samo, lub gorzej, niż LabelEncoing
- binarne dla one vs. rest

Dla każdej z fasolek przetestowaliśmy osobno wyniki Random Forest z podejściem one vs. all, w celu wychwycenia które z fasolek mogą okazać się najbardziej problematyczne dla naszych modeli. Na podstawie tabeli poniżej widzimy, że najbardziej problematyczna wydaje się Sira, a za to fasolką Bombay nie powinno być dużych problemów.

	BOMBAY	SEKER	DERMASON	HOROZ	SIRA	BARBUNYA	CALI
accuracy	1.00000	0.98863	0.95669	0.98688	0.95188	0.98731	0.98206
f1 score (macro avg)	1.00000	0.97732	0.94446	0.97275	0.92080	0.96301	0.95618
accuracy (selected features)	0.99956	0.97594	0.93526	0.97813	0.92738	0.97550	0.97332
f1 score (selected features)	0.99693	0.95232	0.91718	0.95493	0.88380	0.92987	0.93726

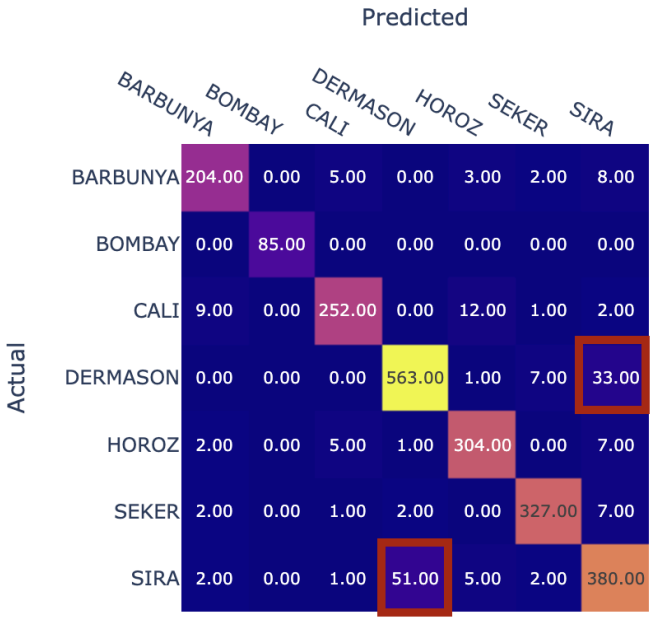
W kolejnym etapie projektu przetestowano 5 modeli uczenia maszynowego (klasyfikacja wieloklasowa):

1. Regresja Logistyczna
2. SVM
3. Decision Tree Classifier
4. **Random Forest**
5. AdaBoost

najlepiej z nich wypadł RandomForest.

	1	2	3	4	5
accuracy	0.92	0.92	0.89	**0.93**	0.74
f1 score (macro avg)	0.93	0.93	0.91	**0.94**	0.69
accuracy (selected features)	0.92	0.92	0.89	**0.92**	-
f1 score (selected features)	0.93	0.93	0.90	**0.93**	-

Na podstawie macierzy (confusion matrix) zauważyliśmy, że każdy z modeli najgorzej radzi sobie z odróżnieniem dwóch z gatunków fasolek: dermason i sira. Rzeczywiście, kiedy porówna się zdjęcia obu tych gatunków, są one niemal identyczne.



Na podstawie macierzy różnic między wariantami z usunięciem kolumn a pełnym zbiorem zdecydowaliśmy, że do dalszej części projektu pozostaniemy z użyciem wszystkich kolumn. Warianty z ograniczoną liczbą kolumn przynosiły najczęściej słabsze rezultaty.

3. KM3

Do tej części projektu dodaliśmy bardziej zaawansowane modele. Lista wszystkich testowanych modeli obejmuje:

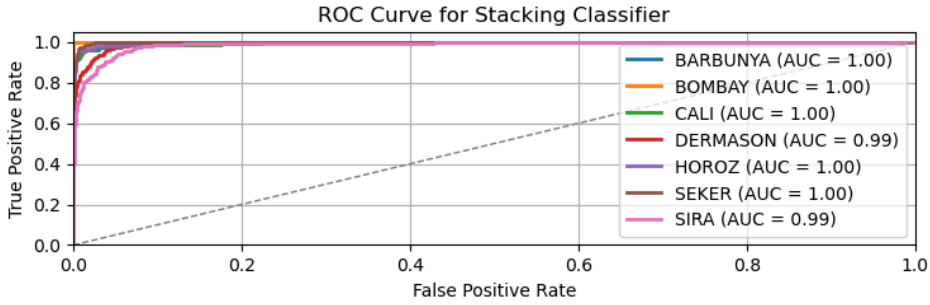
- Regresja Liniowa (ozn. RL)
- Random Forest (RF)
- SVC (SVC)
- Naive Bayes (NB)
- Decision Tree (DT)
- K-Neighbors (KN)
- XGBoost (XGB)
- Stacking (złożony z powyższych, z wyjątkiem XGB)

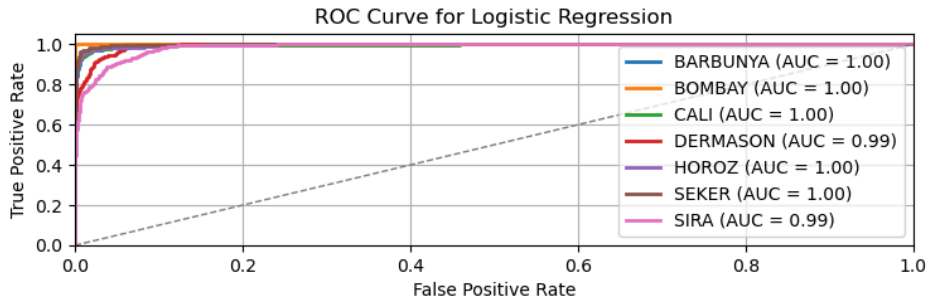
Strojenie parametrów modeli wykonaliśmy przy użyciu funkcji RandomizedSearchCV i dla najlepszych z każdej kategorii przeprowadzaliśmy dalszą analizę. Podczas testowania modeli

stosowaliśmy również kroswalidację. Średnie wartości accuracy z kroswalidacji dla 5 foldów, są przedstawione w tabeli poniżej.

CROSS-VALIDATION RESULTS	RL	RF	SVC	NB	DT	KN	XGB	Stacking
accuracy (przed strojeniem parametrów)	0.92	0.93	0.92	-	0.89	-	-	-
accuracy	**0.923**	0.924	0.928	0.900	0.899	0.924	0.928	**0.930**
cv mean accuracy	**0.925**	0.924	0.930	0.896	0.898	0.926	0.929	**0.930**
cv mean accuracy val	**0.923**	0.914	0.914	0.899	0.881	0.916	0.913	**0.925**

Wyniki wizualizowaliśmy również z uzyciem ROC curve: (wklejone dwa najlepsze modele) Są to oczywiście spojrzenia pod kątem one vs. rest dla każdej fasolki, stąd wybitnie wysokie wyniki AUC.

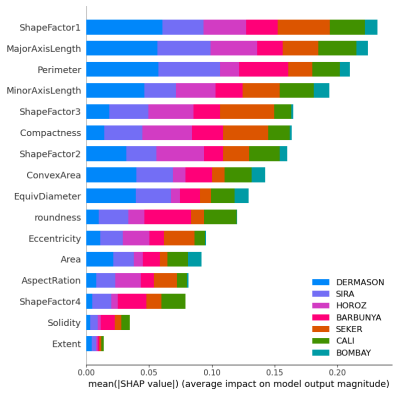




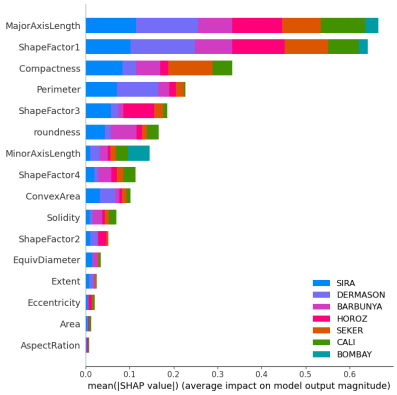
Jako metodę AutoML przetestowaliśmy TPOT Classifier, w rezultacie otrzymując **MLPClassifier(input_matrix, alpha=0.0001, learning_rate_init=0.001)** jako najlepszy pipeline. Jednak z uwagi na dość przeciętny wynik (accuracy 0.93, porównywalny do wcześniejszych rozwiązań) zrezygnowaliśmy z dalszego testowania tego modelu.

3.1 Wizualizacje XAI

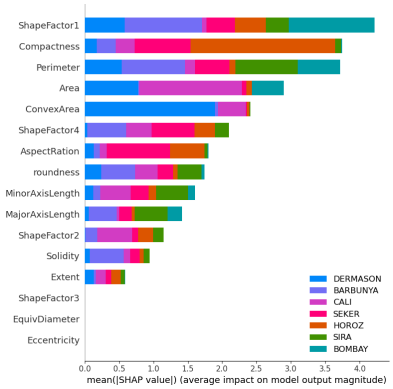
Jako jedna z metod wyjaśnialności skorzystaliśmy z pakietu SHAP, by stworzyć wykresy pokazujące jak dane cechy wpływają na decyzję o klasyfikacji fasolki, w zależności od gatunku, dla różnych modeli.



Random Forest



Decision Tree



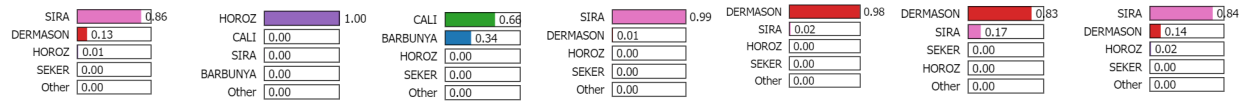
XGBoost

Jak widzimy, różne modele bardzo różnie biorą pod uwagę kolumn. XGBoost w ogóle nie bierze pod uwagę trzech kolumn, co ciekawe, dwie z tych kolumn wyrzucaliśmy wstępnie na etapie EDA na podstawie wysokich korelacji w macierzy korelacji.

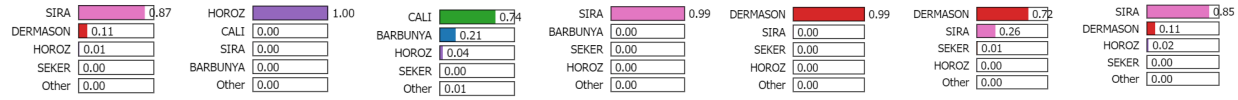
Dla Decision Tree zrobiliśmy również wizualizację drzewiastą (pełne zdjęcie dostępne np. na githubie). Jak widzimy, jest to drzewo mocno rozłożyste.

Korzystaliśmy również z wizualizacji lime, która pozwoliła nam przeanalizować wpływ różnych kolumn na konkretne wyniki danego modelu, a także porównać jak “pewne” swoich decyzji są różne modele (poniżej porównanie 4 modeli dla 7 losowo wybranych obserwacji)

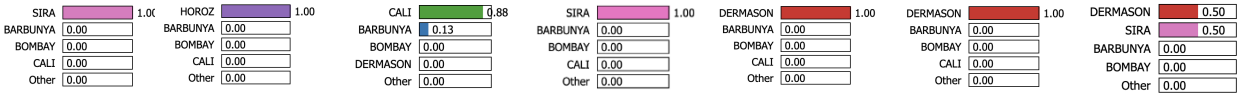
Regresja Logistyczna



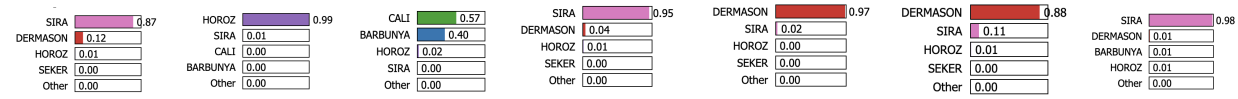
XGBoost



KNeighbors



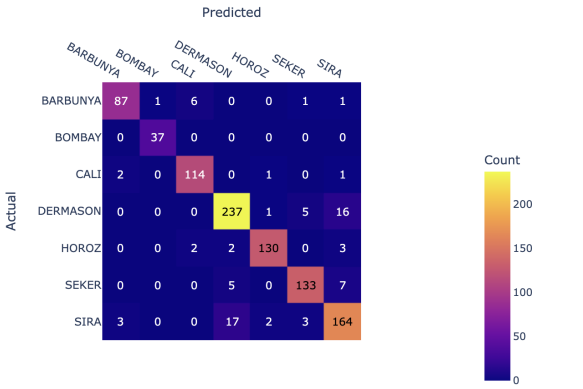
Stacking



"Pewność" stackingu wypada średnio w porównaniu do innych modeli (co oczywiście ma sens jako że wynik to jest pewnego rodzaju średnia innych modeli) co po części skłoniło nas do wyboru tego modelu jako model finalny.

Wybraliśmy Stacking Classifier również ze względu na najwyższy poziom accuracy oraz f1-score (istotny u nas ze względu na niezbalansowanie klas) zarówno na naszym zbiorze, jak i zbiorze walidatorów. Wyniki testu na zbiorze testowym przedstawione są poniżej:

	precision	recall	f1-score	support
BARBUNYA	0.95	0.91	0.93	96
BOMBAY	0.97	1.00	0.99	37
CALI	0.93	0.97	0.95	118
DERMASON	0.91	0.92	0.91	259
HOROZ	0.97	0.95	0.96	137
SEKER	0.94	0.92	0.93	145
SIRA	0.85	0.87	0.86	189
accuracy			0.92	981
macro avg	0.93	0.93	0.93	981
weighted avg	0.92	0.92	0.92	981



Accuracy naszego modelu na danych testowych wyniosło 92% (a właściwie 91,945%), co nie stanowi dużego pogorszenia względem wcześniej używanych danych (acc = 93,0%).

Dokładne parametry finalnego modelu:

- lr = LogisticRegression(C=100, penalty='l1', max_iter=1000,solver='saga', multi_class='multinomial')
- svc = SVC(kernel= 'rbf', gamma= 0.01, C= 10000)
- dt = DecisionTreeClassifier(min_samples_split= 10, max_depth= 15, criterion ='entropy')
- nb = GaussianNB(var_smoothing= 2.848035868435799e-08)
- kn = KNeighborsClassifier(n_neighbors= 9)
- rf = RandomForestClassifier(n_estimators= 200, min_samples_split= 5, max_depth = 25, criterion= "log_loss")

stack = StackingClassifier(estimators=models, final_estimator=LogisticRegression(max_iter=1000,solver='saga', multi_class='multinomial'))

IV Podsumowanie

Udało się nam stworzyć model, który ze skutecznością równą 93% przyporządkowuje gatunek fasolce o danych wymiarach. NIe jest to wynik idealny, jednak satysfakcjonujący, ze względu na uderzające podobieństwo dwóch z gatunków fasolek - Sira i Dermason (gdy usunęliśmy jedną z tych fasolek z danych, accuracy modeli wynosiło ok. 98%). Zatem, gdy klient nie będzie miał styczności z oboma tymi gatunkami fasoli, to nasz model ma tym bardziej zadowalającą skuteczność. Ponadto, w praktyce (np. w hurtowniach żywności) prawdopodobnie bylibyśmy w stanie łatwo otrzymać dodatkowe cechy takie jak waga, lub kolor, co mogłoby również wpłynąć na poprawę dokładności modelu klasyfikacji.

Źródła:

- https://medium.com/@majpsantos/decision-tree-and-logistic-regression-for-dry-bean-classification-369cf78cef20 - zdjęcia fasolek