# Raport walidacyjny grupy zajmującej się klasyfikacją rodzaju fasolek, czyli Mateusza Deptucha oraz Zofii Kamińskiej

Anna Ostrowska, Dominika Gimzicka, Norbert Frydrysiak

April 2024

## 1 Projekt

Projekt zrealizowany został przez Mateusza Deptucha i Zofię Kamińską na przedmiot „Wstęp do uczenia maszynowego" na 4 semestrze kierunku Inżynieria i Analiza Danych na Wydziale Matematyki i Nauk Informacyjnych Politechniki Warszawskiej.
Celem projektu było wybranie modelu uczenia maszynowego, który pozwalałby na jak najdokładniejszą klasyfikację rodzaju fasolek z wykorzystaniem zbioru danych ze strony Kaggle (link do danych w ostatnim rozdziale: „Źródła"). W różnych etapach projektu użyte i wypróbowane zostały różne modele, mniej i bardziej zaawansowane i znalezione ich najlepsze do realizacji celu hiperparametry.
Naszym celem, jako walidatorów, było przejrzenie i przeanalizowanie ich kodu i postępów po każdej częsci projektu oraz udzielenie informacji zwrotnej, co naszym zdaniem możnaby było poprawić lub zmienić.
Link do repozytorium na Githubie wraz z naszym feedbackiem można znaleźć w źródłach na końcu raportu walidacynego. W folderze validate na tym repozytorium znajdują się informacje zwrotne przekazane modelarzom.

## 2 Jaki feedback dawaliśmy? Czy wzięto go pod uwagę?

| **Zagadnienie** | **Czy wzięto pod uwagę?** |
|---|---|
| Brak opisu projektu w pliku README.md | Tak |
| Brak wniosków o tym z jakich rozkładów mogą być zmienne, na przykład, że rozkład roundness jest podobny do rozkładu normalnego | Tak |
| Słaba hierarchia plików w projekcie, w późniejszym etapie projektu może to prowadzić do problemów z zarządzaniem plikami oraz ich zrozumieniem czy merge'owaniem branchy itp | Tak |
| Trzeba mieć świadomość, że w waszym projekcie dużo zmiennych ma znaczenie i tych zmiennych nie jest dużo, przez co ja bym nie usuwał za bardzo zmiennych, dopiero na późniejszym etapie przy tworzeniu modelu przetestował, usunięcie których co zmienia. | Tak |
| Jeśli już mamy coś usuwać to korelacje typu 0.9 są BARDZO dużymi kandydatami do usunięcia. | Tak |
| Wypada napisać, że fasolki typu "BOMBAY" są bardzo inne od reszty fasolek, więc nie będzie z nimi problemu. | Tak |
| Wypada po każdym wykresie pisać wnioski, co wnosi dany wykres do analizy, co można z niego wyciągnąć, co jest ciekawe, co jest nieciekawe, co jest zaskakujące, co jest oczywiste itp. (ta, wiem, sam tak nie robię, ale ja jestem wyjątkowy) | Tak |
| Jeśli macie macierze korelacji i nie ma 84 kolumn to wypada wyświetlić macierz korelacji z dokładnymi wartościami napisanymi na niej, a nie tylko kolory. | Tak |
| Polecam skorzystać z modelu, o którym była mowa na wykładzie, który sprawdza, które zmienne są jego zdaniem ważne w kontekście usuwania zmiennych itp. | Tak |

Table 1: Realizacja zagadnień w ramach KM1

Tutaj trzeba przyznać, że Zofia i Mateusz bardzo fajnie wzięli sobie nasz feedback do serca i zrobili bardzo fajną hierarchię plików i opis projektu, że to bardzo pięknie wygląda teraz na ich repozytorium.

| **Zagadnienie** | **Czy wzięto pod uwagę?** |
|---|---|
| Słabo opisane wykresy, które są przy testowaniu różnych modeli, Co to jest frequency? | Tak |
| Fajnie, że są wnioski z EDA i korzystacie z tego w dalszej części projektu. | Tak |
| Zamiast MinMaxScaler można użyć StandardScaler, bo nie ma outlierów lub można spróbować innych transformacji jak BOX COX albo logarytmiczna albo powiedzieć na prezentacji dlaczego uważacie że tak jest wystarczająco dobrze. | Tak |
| Można się zastanowić nad jeszcze innymi modelami jak Naive Bayes, ADA Boost, SGD, QDA, kNN, Neural Network. | Tak |
| Fajnie, że używacie classification_report. | Tak |
| Ja bym nie usuwał za bardzo kolumn. | Tak |
| Ja bym uważał z modelami drzewiastymi, bo one są bardzo podatne na overfitting. | Tak |
| Na zbiorze walidatorów też wychodzą dobre accuracy itp itd. | Tak |

Table 2: Realizacja zagadnień w ramach KM2

| **Zagadnienie** | **Czy wzięto pod uwagę?** |
|---|---|
| brak opisów co dany wykres przedstawia i wniosków płynących z nich | Tak |
| brak wniosków jaki model/modele ostatecznie wybraliście i czemu te są najlepsze, lub chociaż które są lepsze od innych | Tak |
| Za dużo slajdów na prezentacji | Nie |
| Nie powinno być kodu na prezentacjach | Nie |
| dodac cel/motywacje w prezentacji | Tak |
| moze cos dodac o tym, jakie kolumny macie (jakie te cechy fizyczne - przyklady) | Tak |
| w jaki sposob bombay odstaje od reszty? | Tak |
| moze wspomniec, jakie kolumny usunieto (nie wiem) | Tak |
| slajd 11 troche slabo czytelny, jesli tak zostawicie to na pewno warto slownie bardziej wyjasnic, co tam jest | Tak |
| warto dodac jakies podsumowanie: jaki model ostatecznie wypadl najlepiej i z jakimi paramatrami | Tak |
| czego niestandordowego nauczyliście się dzięki temu projektowi??? | Tak |

Table 3: Realizacja zagadnień w ramach KM3

# 3 Sprawdzenie modeli na zbiorze dla walidatorów

Logistic Regression

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| BARBUNYA | 0.93 | 0.92 | 0.93 | 222 |
| BOMBAY | 1.00 | 1.00 | 1.00 | 85 |
| CALI | 0.97 | 0.91 | 0.94 | 276 |
| DERMASON | 0.93 | 0.92 | 0.93 | 604 |
| HOROZ | 0.92 | 0.95 | 0.94 | 319 |
| SEKER | 0.95 | 0.96 | 0.96 | 339 |
| SIRA | 0.86 | 0.88 | 0.87 | 441 |
| accuracy | | | 0.93 | 2286 |
| macro avg | 0.94 | 0.94 | 0.94 | 2286 |
| weighted avg | 0.93 | 0.93 | 0.93 | 2286 |

Figure 1: Rózne metryki na zbiorze dla walidatorów modelu Regresji Liniowej.

SVM

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| BARBUNYA | 0.91 | 0.92 | 0.92 | 222 |
| BOMBAY | 1.00 | 1.00 | 1.00 | 85 |
| CALI | 0.93 | 0.91 | 0.92 | 276 |
| DERMASON | 0.89 | 0.94 | 0.91 | 604 |
| HOROZ | 0.94 | 0.93 | 0.94 | 319 |
| SEKER | 0.97 | 0.92 | 0.95 | 339 |
| SIRA | 0.87 | 0.85 | 0.86 | 441 |
| accuracy | | | 0.92 | 2286 |
| macro avg | 0.93 | 0.93 | 0.93 | 2286 |
| weighted avg | 0.92 | 0.92 | 0.92 | 2286 |

Figure 2: Rózne metryki na zbiorze dla walidatorów modelu Maszyny Wektorów Nośnych.

Random Forest

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| BARBUNYA | 0.94 | 0.92 | 0.93 | 222 |
| BOMBAY | 1.00 | 1.00 | 1.00 | 85 |
| CALI | 0.95 | 0.91 | 0.93 | 276 |
| DERMASON | 0.91 | 0.93 | 0.92 | 604 |
| HOROZ | 0.94 | 0.96 | 0.95 | 319 |
| SEKER | 0.96 | 0.96 | 0.96 | 339 |
| SIRA | 0.87 | 0.87 | 0.87 | 441 |
| accuracy | | | 0.93 | 2286 |
| macro avg | 0.94 | 0.94 | 0.94 | 2286 |
| weighted avg | 0.93 | 0.93 | 0.93 | 2286 |

Figure 3: Rózne metryki na zbiorze dla walidatorów modelu Losowego Lasu Drzew.

Decision Tree

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| BARBUNYA | 0.86 | 0.89 | 0.87 | 222 |
| BOMBAY | 1.00 | 1.00 | 1.00 | 85 |
| CALI | 0.92 | 0.87 | 0.90 | 276 |
| DERMASON | 0.88 | 0.92 | 0.90 | 604 |
| HOROZ | 0.92 | 0.94 | 0.93 | 319 |
| SEKER | 0.95 | 0.91 | 0.93 | 339 |
| SIRA | 0.84 | 0.82 | 0.83 | 441 |
| accuracy | | | 0.90 | 2286 |
| macro avg | 0.91 | 0.91 | 0.91 | 2286 |
| weighted avg | 0.90 | 0.90 | 0.90 | 2286 |

Figure 4: Rózne metryki na zbiorze dla walidatorów modelu Drzewa Decyzyjnego.

Stacking Classifier

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| BARBUNYA | 0.95 | 0.94 | 0.94 | 222 |
| BOMBAY | 1.00 | 1.00 | 1.00 | 85 |
| CALI | 0.96 | 0.93 | 0.94 | 276 |
| DERMASON | 0.91 | 0.95 | 0.93 | 604 |
| HOROZ | 0.93 | 0.96 | 0.94 | 319 |
| SEKER | 0.97 | 0.96 | 0.96 | 339 |
| SIRA | 0.89 | 0.85 | 0.87 | 441 |
| accuracy | | | 0.93 | 2286 |
| macro avg | 0.94 | 0.94 | 0.94 | 2286 |
| weighted avg | 0.93 | 0.93 | 0.93 | 2286 |

Figure 5: Różne metryki na zbiorze dla walidatorów modelu Stacking Classifier złożone z 6 modeli: Regresji Liniowej, Maszyny Wektorów Nośnych, Drzewa Decyzyjnego, Naiwnego Klasyfikatora Bayesowskiego (Gauss), K najbliższych sąsiadów oraz Losowego Lasu Drzew.

# 4 Podsumowanie

Dzięki regularnemu kontaktowi między nami a zespołem walidowanym, udało się nam na każdym etapie pracy nanosić komentarze i poprawki, które liczymy, że usprawniły i ulepszyły efekty ich pracy. Uważamy, że była to owocna współpraca, co pokazuje ilość wziętych pod uwagę naszych uwag. Zespołowi udało się uzyskać bardzo zadowalające modele, które rozwiązują problem z dużym prawdopodobieństwem.

# 5 Źródła

- link do zbioru danych, z których grupa modelarzy korzystała: https://www.kaggle.com/datasets/nimapou bean-dataset-classification/
- link do repozytorium na Githubie grupy modelarzy: https://github.com/DeptuchMateusz/Bean-Classification