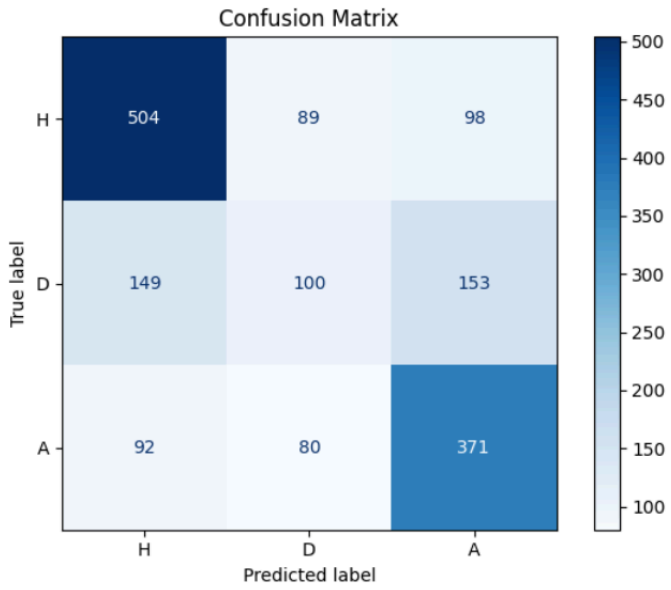
Raport projektu: Predykcja wyników meczów piłkarskich

Bartłomiej Chmiel, Dorian Guz

1. Wyniki testowe i treningowe

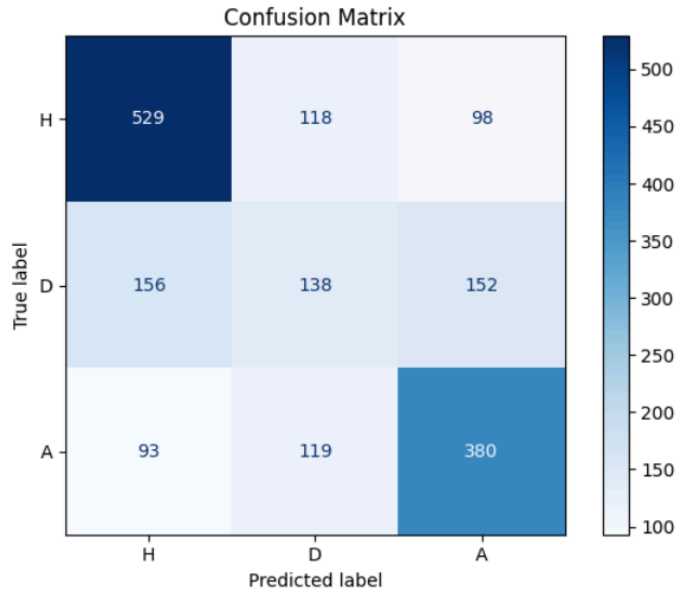
Model został oceniony, przy użyciu metryk takich jak accuracy, F1-score, precision, recall i log-likelihood dla każdej z klas. Wyniki na zbiorze treningowym pokazały, że model dobrze uczy się wzorców w danych, natomiast test na przyszłych meczach pozwolił ocenić jego zdolność do generalizacji. Najlepiej model radził sobie z przewidywaniem zwycięstw gospodarzy, natomiast najwięcej trudności sprawiały remisy, co jest zgodne z oczekiwaniami przy niezbalansowanych danych.

- **Random Forest**



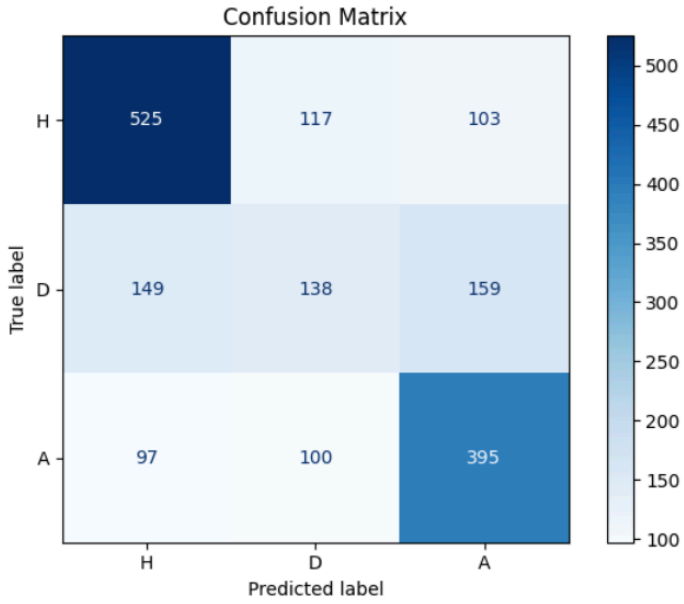
	precision	recall	f1-score	support
A	0.60	0.68	0.64	543
D	0.37	0.25	0.30	402
H	0.68	0.73	0.70	691
accuracy			0.60	1636
macro avg	0.55	0.55	0.55	1636
weighted avg	0.58	0.60	0.58	1636

- **LGBM**



	precision	recall	f1-score	support
A	0.60	0.64	0.62	592
D	0.37	0.31	0.34	446
H	0.68	0.71	0.69	745
accuracy			0.59	1783
macro avg	0.55	0.55	0.55	1783
weighted avg	0.58	0.59	0.58	1783

- **CatBoost**



	precision	recall	f1-score	support
A	0.60	0.67	0.63	592
D	0.39	0.31	0.34	446
H	0.68	0.70	0.69	745
accuracy			0.59	1783
macro avg	0.56	0.56	0.56	1783
weighted avg	0.58	0.59	0.59	1783

2. Uzasadnienie wyboru techniki / modelu

W projekcie jako główny model wybraliśmy **Random Forest** (las losowy). Wybór ten został dokonany ze względu na odporność modelu na przeuczenie, dobrą ogólną wydajność oraz umiejętność radzenia sobie z danymi, które nie mają prostych liniowych zależności. Las losowy sprawdza się dobrze w przypadku zbiorów o niezbalansowanych klasach (co ma miejsce w meczach piłkarskich – wygrane gospodarzy *(Home win - klasa H)* są częstsze niż remisy *(Draw - klasa D)* czy wygrane gości *(Away win - klasa A)*. Random Forest dobrze radzi sobie z dużą liczbą zmiennych wejściowych, a w naszym przypadku było ich dwadzieścia jeden.

3. Strategia podziału danych

Z uwagi na czasowy charakter danych (mecze odbywają się w określonym porządku chronologicznym), zastosowaliśmy podział przy użyciu **TimeSeriesSplit** z 5 foldami. Takie podejście zapobiega wyciekowi informacji, ponieważ model uczy się tylko na danych z przeszłości, a testowany jest na przyszłości — co symuluje rzeczywisty scenariusz predykcyjny.

Ostatni fold został wykorzystany jako zestaw testowy, a wcześniejsze jako treningowy. Przed trenowaniem modelu dane treningowe zostały zbalansowane za pomocą **SMOTE**, co poprawiło reprezentację klas mniejszościowych (remisy i wygrane gości), zmniejszając uprzedzenie modelu wobec zwycięstw gospodarzy.

4. Opis danych wejściowych

- **month** – miesiąc rozegrania meczu (1–12); pozwala uchwycić sezonowość.
- **home_days_since, away_days_since** – liczba dni od ostatniego meczu gospodarzy i gości; informuje o zmęczeniu lub odpoczynku.
- **dow** – dzień tygodnia meczu (0 – poniedziałek, ..., 6 – niedziela).
- **xG_home, xG_away** – oczekiwane gole gospodarzy i gości.
- **xG_home_roll, xG_away_roll** – średnia xG z ostatnich 5 meczów dla gospodarzy i gości.
- **home_roll_xg_against, away_roll_xg_against** – średnie xG stracone w ostatnich 5 meczach.
- **xG_home_ewm, xG_away_ewm** – wykładniczo ważona średnia xG.
- **home_xg_std, away_xg_std** – odchylenie standardowe xG z ostatnich 5 meczów.
- **home_roll_gd, away_roll_gd** – średnia różnica bramek z ostatnich 5 meczów (gole zdobyte minus stracone).
- **avg_goals_home, avg_goals_away** – średnia liczba bramek zdobywanych przez drużyny w historii.
- **lambda_home_for, lambda_home_against, lambda_away_for, lambda_away_against** – średnia liczba goli zdobywanych i traconych przez drużyny, liczona narastająco (expanding mean); aproksymacja parametrów λ rozkładu Poissona.
- **home_roll_pts, away_roll_pts** – suma punktów zdobytych przez drużyny w ostatnich 5 meczach (3 za zwycięstwo, 1 za remis).
- **form_slope_home, form_slope_away** – trend punktowy (nachylenie regresji liniowej punktów z ostatnich 5 meczów).

5. Analiza wyników i dalsze kroki

Model dobrze rozpoznaje zwycięstwa gospodarzy, co jest zgodne z naturalną dominacją tej klasy. Najtrudniejsze okazało się przewidywanie remisów, co wynika zarówno z ich mniejszej liczebności, jak i z tego, że są znacznie cięższe do przewidzenia.

W przyszłości warto rozważyć:

- **Selekcję cech** lub redukcję wymiarowości, by skupić się na najważniejszych zmiennych
- **Rozważenie nowych cech** takich jak wartość składu, kontuzje, trener, stan mentalny drużyny
- **Automatycznie odświeżane** i pobierane dane
- **Interaktywny interfejs** do “wyklikania” drużyn przy predykcji

Podział pracy:

Bartłomiej:

- plan projektu: struktura, technologie, środowisko
- przygotowanie i ustandaryzowanie danych, skrypty fetchujące oraz mergujące, data-loader
- feature-engineering
- Projekt pipeline-u i CLI do łatwiejszej obsługi
- Readme, setup

Dorian:

- analiza eksploracyjna danych (notebook 1.)
- projekt modeli (model.py) Random Forst, LightGBM, CatBoost i StackingClassifier - połączenie poprzednich trzech modeli
- wizualizacja działania modelu i ukazanie wpływu hiperparametrów (notebook 2 i 3)
- symulacja/predykcja meczu (gospodarze - goście, data)
- dokumentacja