# Tytuł raportu

„Analiza danych o piłce nożnej: Ocena wpływu cech zawodników z gry FIFA na typy bukmacherów”

## Wstęp - Motywacja, cele

* Motywacja: Piłka nożna to jedna z najpopularniejszych dyscyplin sportowych na świecie. Analiza danych pozwala lepiej zrozumieć, jakie cechy piłkarzy przyczyniają się do sukcesu drużyny. Może to być użyteczne dla menedżerów sportowych i trenerów.
* Cele:
  + Przeanalizowanie wpływu różnych zmiennych, np. pracowitości w obronie piłkarzy na tracone bramki i pracowitości w ataku piłkarzy na strzelane bramki oraz wpływ ocen ogólnych z gry FIFA na wyniki meczów.
  + Opracowanie modelu przewidującego, które drużyny mają większe szanse na sukces w lidze.
  + Wyciągnięcie praktycznych wniosków dla analizy piłkarskiej.

## Opis danych - Struktura zbioru

Dane pochodzą z publicznego zbioru danych dostępnego na platformie Kaggle: [European Soccer Database](https://www.kaggle.com/datasets/hugomathien/soccer/data?select=database.sqlite). Baza zawiera szczegółowe informacje o zawodnikach, drużynach, wynikach meczów oraz kursach bukmacherskich.

Dane dostępne w zbiorze zawierają szczegółowe informacje na temat piłkarzy, drużyn, lig, wyników meczów oraz atrybutów związanych z grą. Zbiór został podzielony na kilka kluczowych tabel, które umożliwiają analizę zarówno na poziomie zawodników, jak i całych drużyn czy lig.

Tabele obejmują:

#### 1. Country

* Opis: Informacje o krajach, w których odbywają się rozgrywki piłkarskie.
* Kluczowe zmienne:
  + id: Unikalny identyfikator kraju.
  + name: Nazwa kraju.
* Zastosowanie: Łączenie z tabelami League w celu przypisania lig do konkretnych krajów.

##### 2. League

* Opis: Informacje o ligach piłkarskich.
* Kluczowe zmienne:
  + id: Unikalny identyfikator ligi.
  + country\_id: Powiązanie z tabelą Country w celu określenia, w którym kraju znajduje się liga.
  + name: Nazwa ligi (np. “Spain LIGA BBVA”, ”Poland Ekstraklasa”, “England Premier League”).
* Zastosowanie: Segmentacja danych na poziomie lig, co pozwala na analizę różnic między rozgrywkami.

#### 3. Match

* Opis: Dane o meczach piłkarskich, w tym wyniki i statystyki.
* Kluczowe zmienne:
  + id: Unikalny identyfikator meczu.
  + league\_id: Powiązanie z tabelą League.
  + season: Sezon rozgrywek.
  + date: Data rozegrania meczu.
  + home\_team\_api\_id, away\_team\_api\_id: Identyfikatory drużyn gospodarza i gości (łączone z tabelą Team).
  + home\_team\_goal, away\_team\_goal: Liczba zdobytych bramek przez gospodarzy i gości.
* Zastosowanie: Analiza wyników, przewidywanie wyników meczów, ocena formy drużyn.

#### 4. Player

* Opis: Szczegółowe informacje o zawodnikach.
* Kluczowe zmienne:
  + id: Unikalny identyfikator zawodnika.
  + player\_api\_id: Klucz API zawodnika (łączone z tabelą Match).
  + player\_name: Imię i nazwisko zawodnika.
  + Player\_fifa\_api\_id: Klucz API zawodnika z gry FIFA (łączone z tabelą Player\_Attributes).
  + birthday: Data urodzenia zawodnika.
  + height: Wzrost zawodnika (w cm).
  + weight: Waga zawodnika (w funtach).
* Zastosowanie: Analiza cech demograficznych zawodników, np. wpływ wieku lub wzrostu na wyniki.

#### 5. Player\_Attributes

* Opis: Atrybuty techniczne, fizyczne i mentalne zawodników, zebrane w czasie.
* Kluczowe zmienne:
  + id: Unikalny identyfikator rekordu.
  + player\_fifa\_api\_id: Powiązanie z tabelą Player.
  + date: Data aktualizacji danych.
  + Cechy zawodnika:
    - Techniczne: np. crossing (dośrodkowania), finishing (wykończenie akcji), dribbling (drybling).
    - Fizyczne: np. acceleration (przyspieszenie), stamina (wytrzymałość), strength (siła).
    - Mentalne: np. aggression (agresja), vision (wizja gry), positioning (ustawianie się).
* Zastosowanie: Analiza wpływu umiejętności zawodników na wyniki drużynowe, ocena wartości zawodnika.

#### 6. Team

* Opis: Informacje o drużynach biorących udział w rozgrywkach.
* Kluczowe zmienne:
  + id: Unikalny identyfikator drużyny.
  + team\_api\_id: Klucz API drużyny (łączone z tabelą Match).
  + team\_fifa\_api\_id: Klucz API drużyny z gry FIFA (łączone z tabelą Team\_Attributes).
  + team\_long\_name, team\_short\_name: Pełna i skrócona nazwa drużyny.
* Zastosowanie: Łączenie danych meczowych z drużynami, analiza różnic między drużynami.

#### 7. Team\_Attributes

* Opis: Atrybuty drużyn, związane z ich stylem gry i strategią.
* Kluczowe zmienne:
  + id: Unikalny identyfikator rekordu.
  + team\_fifa\_api\_id: Powiązanie z tabelą Team.
  + date: Data aktualizacji danych.
  + Cechy drużyny:
    - Styl gry: np. buildUpPlaySpeed (szybkość budowania akcji), chanceCreationPassing (tworzenie okazji poprzez podania).
    - Strategia: np. defenceAggression (agresywność obrony), defenceWidth (szerokość gry w obronie).
* Zastosowanie: Analiza wpływu strategii drużynowej na wyniki, budowa modeli przewidujących wyniki na podstawie stylu gry.

## Opis procesu przygotowania danych

1. Eksploracja danych

* Wczytanie zbioru danych i zapoznanie się z jego strukturą oraz zawartością.
* Zidentyfikowano braki danych oraz odrzucono rekordy z nieprawidłowymi wartościami.

1. Czyszczenie danych

* Usunięto brakujące wartości w kluczowych cechach zawodników i kursach bukmacherskich.
* Filtrowano zawodników według dat odpowiadających wersjom FIFA, aby dopasować dane do sezonów rozgrywek.

1. Łączenie tabel

* Połączono dane o zawodnikach, drużynach i wynikach meczów za pomocą kluczy API (player\_api\_id, team\_api\_id).
* Uzupełniono dane o różnice w atrybutach drużyn gospodarzy i gości.

1. Tworzenie nowych zmiennych

* Wyliczenie średnich różnic wartości danej statystyki dla danej grupy piłkarzy, np różnica między drużynami w umiejętności krótkich podań, diff\_short\_passing.
* Wyliczono średnie kursy bukmacherskie na wyniki meczu

1. Walidacja przygotowanych danych

* Przeprowadzenie weryfikacji, czy dane są poprawne i spójne po wszystkich etapach przetwarzania.
* Przetworzone dane zapisano jako nowe tabele w bazie SQLite (Player\_Attributes\_filtered\_fifa, Match\_for\_plots).

## Analiza danych

#### Przyjęte założenia

* Wyniki meczów są w dużej mierze zależne od różnic w kluczowych atrybutach drużyn gospodarzy i gości.
* Zamiast rzeczywistych wyników w analizach uwzględniono typy bukmacherów jako wskaźnik, ponieważ lepiej oddają przewidywania oparte na zebranych danych i analizach statystycznych.
* Tylko 2 z 2283 meczów zostały przewidziane jako remis, co podkreśla trudność w dokładnym przewidywaniu remisów przez modele bukmacherskie.

#### Metodologia

* Analiza korelacji między cechami zawodników a wynikami meczów.
* Grupowanie danych według różnic w atrybutach drużyn gospodarzy i gości.
* Wykorzystanie wizualizacji dla lepszego zrozumienia zależności.

Modelowanie danych

#### Przyjęte założenia

* Różnice w cechach drużyn mają kluczowe znaczenie dla wyników meczów.
* Modele klasyfikacyjne, takie jak Random Forest, skutecznie przewidują wyniki.

#### Metodologia

* Model Random Forest trenowany na różnicach w cechach drużyn (np. diff\_acceleration, diff\_stamina).
* Uwzględniono ważenie klas w celu poprawy wyników na nierównoważnych danych.

### Testy z alternatywnymi algorytmami i Wnioski

#### **1. Random Forest (Zaimplementowany)**

* **Opis**: Random Forest to algorytm oparty na wielu drzewach decyzyjnych, które działają jako zespoły. Model dobrze radzi sobie z nierównoważnymi danymi dzięki wagom klas.
* **Wyniki**:
  + Osiągnął dokładność **85%**.
  + Dobre wyniki w przewidywaniu zarówno wygranych gospodarzy, jak i gości.
* **Wniosek**: Random Forest jest efektywny i daje solidne wyniki przy umiarkowanej złożoności obliczeniowej.

A blue squares with white text

Description automatically generated

#### **2. Gradient Boosting**

* **Opis**: Gradient Boosting buduje model iteracyjnie, minimalizując błędy poprzednich iteracji. Jest bardziej złożony niż Random Forest.
* **Wyniki**:
  + Dokładność również na poziomie **85%**.
  + Nie zaobserwowano istotnych różnic w porównaniu do Random Forest.
* **Wniosek**: Gradient Boosting nie wnosi poprawy w wynikach, a jest bardziej złożony w obliczeniach.

#### **3. XGBoost**

* **Opis**: XGBoost to zoptymalizowana wersja Gradient Boosting, która oferuje dodatkowe funkcje, takie jak regularyzacja i lepsza obsługa braków danych.
* **Wyniki**:
  + Dokładność na poziomie **85%**.
  + Nie zaobserwowano istotnych różnic w porównaniu do Random Forest.
* **Wniosek**: Mimo zaawansowanych mechanizmów, XGBoost nie oferuje lepszej dokładności w tym zadaniu.

#### **4. LightGBM**

* **Opis**: LightGBM to alternatywa dla XGBoost, zoptymalizowana pod kątem szybkości działania i obsługi dużych zbiorów danych.
* **Wyniki**:
  + Osiągnął dokładność **84%**.
  + Nie zaobserwowano istotnych różnic w porównaniu do Random Forest.
* **Wniosek**: Podobnie jak XGBoost, LightGBM nie poprawia wyników w porównaniu do Random Forest.

### Podsumowanie

Żaden z testowanych algorytmów nie przyniósł poprawy w wynikach w porównaniu do Random Forest. Kluczowe wnioski:

1. Random Forest pozostaje najbardziej efektywnym modelem w tym przypadku.
2. Większa złożoność modeli nie przynosi lepszych wyników, co sugeruje, że ograniczeniem może być jakość lub zakres danych.
3. Dalsze prace powinny skupić się na wzbogaceniu danych (np. dodanie dodatkowych cech kontekstowych) zamiast na wyborze bardziej złożonych algorytmów.

## Rezultaty i wnioski

#### Rezultaty

* Najważniejsze różnice w cechach zawodników to różnice w szybkości, wytrzymałości i wizji gry.
* Kursy bukmacherskie znacząco poprawiły skuteczność przewidywań modelu.
* Random Forest osiągnął wysoką dokładność w przewidywaniu wyników, 85%.

#### Wnioski

* Wyniki analizy potwierdzają, że różnice w cechach zawodników mają istotny wpływ na wyniki meczów.
* Rozszerzenie modelu o dodatkowe dane (np. warunki pogodowe) mogłoby zwiększyć dokładność predykcji.

A graph with blue and white bars

Description automatically generated with medium confidence