Politechnika Wrocławska
Wydział Informatyki i Telekomunikacji

Kierunek: **Informatyka algorytmiczna (INA)**
Specjalność: —

PRACA DYPLOMOWA MAGISTERSKA

Przewidywanie zdarzeń w piłce nożnej

Predicting events in football

Radosław Wojtczak

Opiekun pracy
dr hab. inż. Stanisław Saganowski

Słowa kluczowe: sztuczna inteligencja, uczenie głębokie, przewidywanie zdarzeń, modele optymalizacyjne

WROCŁAW, 2024

Streszczenie

Obiektem badań niniejszej pracy jest sprawdzenie możliwości przewidywania wybranych zdarzeń w meczach piłkarskich. Eksperyment obejmuje prognozowanie różnorodnych zdarzeń, począwszy od podstawowych, takich jak wynik meczu czy zwycięstwo drużyny, aż po bardziej szczegółowe aspekty, takie jak liczba rzutów rożnych danej drużyny.

Praca ta składa się z obszernego opisu procesu projektowania, jak i wdrażania modelu predykcyjnego dla elementów piłkarskiego widowiska, w szczególności skupiając się na wykorzystaniu technik maszynowego głebokiego uczenia (Deep Learning, DL). Dodatkowo zawiera liczne przykłady zastosowań modelu, szczegóły implementacyjne oraz porównanie otrzymanych wyników z rzeczywistymi rezultatami oraz innymi powszechnie stosowanymi rozwiązaniami dostępnymi w literaturze naukowej.

Dodatkowym aspektem pracy jest opracowanie metodologii umożliwiającej rywalizację w zakładach bukmacherskich poprzez wykrywanie rozbieżności między ustalonym prawdopodobieństwem wystąpienia zdarzenia, a kursem oferowanym przez wybranych bukmacherów. Przedstawiono także definicje umożliwiające czytelnikowi zapoznanie się z nomenklaturą używaną w zakładach bukmacherskich.

Słowa kluczowe: sztuczna inteligencja, uczenie głębokie, przewidywanie zdarzeń, modele optymalizacyjne

Abstract

The main goal of the research is to investigate the possibility of predicting selected events in soccer matches. The experiment involves the predicition of a variety of events, ranging from basic ones, such as the result of a match or a given team's victory, to more detailed aspects, such as the number of corner kicks of a given team.

Presented thesis consists of comprehensive description of the process of designing, as well as implementing, a predicitve model for elements of the soccer game, particularly focusing on the usage of machine learning techniques, such as deep learning (Deep Learning, DL). Additionally, thesis includes numerous examples of the model's application, implementation details and a comparison of the results obtained with actual results of games and other commonly used solutions available in the scientific literature.

An additional aspect of the work is the development of a methodology that enables competitive betting by detecting discrepancies between the established probability of an event and the odds offered by the selected bookmakers. All needed definitions are provided in order for the reader to become familiar with the nomenclature used in betting's world.

Keywords: artificial intelligence, deep learning, event prediction, optimization models

Spis treści

Spis rysunków

Spis tabel

Spis listingów

Skróty

SC (ang. *Scoring Chances*)
BP (ang. *Ball Possesion*)
SOG (ang. *Shots on Goals*)
CK (ang. *Corner Kicks*)
FK (ang. *Free Kicks*)
YC (ang. *Yellow Cards*)
RC (ang. *Red Cards*)
OFF (ang. *Offsides*)
O/U (ang. *Over/Under*)
ML (ang. *Moneyline*)

Rozdział 1

Wprowadzenie

1.1. Popularność piłki nożnej

Piłka nożna - od wielu dekad nieprzerwanie najpopularniejszy sport na świecie, który trenowany, jak i obserwowany, jest przez miliardy ludzi z różnych zakątków globu. Wpływ wspomnianego sportu na społeczeństwa jest tak ogromny, że czasem przekracza granice czysto sportowego zainteresowania, stając się istotnym elementem kultury i życia społecznego. Przykłady tego wpływu można odnaleźć w różnych aspektach codzienności - od piłkarzy jako idoli wśród młodzieży kultywujący w nich takie idee jak pracowitość po infrastrukture stadionową, która może służyć jako oblegane miejsca turystyczne. Zdarzają się również tak ekstremalne zdarzenia jak ogłoszenie dni wolnych od pracy po osiągnięciu sukcesów przez narodowe reprezentacje, umożliwiając w ten sposób obywatelom celebrowanie triumfu swojej drużyny, czy osobiste poświęcenia fanów, którzy nie wahają się zaciągać długów, aby móc uczestniczyć w wydarzeniach piłkarskich na żywo. Najświeższym wydarzeniem dostarczających ogrom tego typu historii są mistrzostwa świata rozgrywane w Katarze w 2022 roku. To właśnie po szokującym zwycięstwie w meczu rozgrywanym w ramach przytoczonych rozgrywek w grupie C, do której również należała Polska, Arabii Saudyjskiej nad Argentyną władzę owego kraju ogłosiły dzień wolny od pracy. Mecz ten według wielu ekspertów określany jest jako największa sensacja w całej, niemalże stuletniej, historii mistrzostw świata. Mimo pierwszej porażki, również podczas tego wydarzenia wielu Argentyńczyków decydowało się zaciągać pożyczki na ogromne sumy aby ujrzeć swój zespół w drodze ku zwycięstwu, na czele której stał piłkarz uznawany za jednego z najwybitniejszych w całej historii futbolu — Lionel Messi.

Wpływ piłki nożnej na społeczeństwo nie kończy się na wydarzeniach narodowych czy ogólnokrajowych. Indywidualne osiągnięcia najlepszych piłkarzy także mają ogromne znaczenie. Ikony tego sportu, tacy jak wspomiany wcześniej Lionel Messi czy Cristiano Ronaldo, nie tylko osiągają sukcesy na boisku, ale stają się również symbolem aspiracji i marzeń dla milionów ludzi na całym świecie. To właśnie dlatego, obaj ci piłkarze zajmują czołowe pozycje w rankingu najpopularniejszych kont na platformach społecznościowych, co jest wyraźnym dowodem na ich międzynarodową rozpoznawalność i wpływ na kulturę popularną.

Ze względu na swoją popularność przemysł piłkarski przyciąga do siebie wielkie pieniądze. Kluby oraz narodowe federacje inwestują miliony w pozyskanie, bądź wytrenowanie utalentowanych piłkarzy, prowadząc transakcje transferowe na rekordowe sumy lub zakupując najwyższej klasy sprzęt ułatwiający rozwój zawodnika od najmłodszych lat. Rozwój infrastruktury stadionowej, w tym modernizacje obiektów oraz budowa nowych aren sportowych, przyciąga inwestorów oraz generuje zyski z biletów, sprzedaży koncesji, reklam i innych dochodów związanych z organizacją wydarzeń sportowych. Oprócz inwestycji w kluby i infrastrukturę, piłka nożna staje się również istotnym sektorem dla przemysłu hazardowego. Rozwój zakładów bukmacherskich

oraz internetowych platform do obstawiania wyników meczów generuje ogromne zyski dla firm z wspomnianej branży. Ludzie, poszukując emocji i adrenaliny związanych z obstawianiem wyników meczów, podejmują ryzyko finansowe, co przekłada się na znaczący dochód dla firm bukmacherskich.

Fenomen piłki nożnej jako zakładu hazardowego nie tylko wynika z komercyjnych korzyści, ale także z naturalnej ludzkiej potrzeby rywalizacji i emocji towarzyszących oglądaniu meczów. Obstawianie wyników zdarzeń staje się często sposobem na dodatkową stymulację emocjonalną podczas oglądania spotkań, co sprawia, że ów aspekt biznesu piłkarskiego ma potencjał dalszego wzrostu, rozwoju, a co najważniejsze z perspektywy przedstawianego badania — eksploatowania przez sprytnego gracza.

1.2. Znaczenie przewidywania zdarzeń

Przewidywanie zdarzeń w meczach piłkarskich ma duże znaczenie dla wielu grup zawodowych, takich jak trenerzy, analitycy sportowi, wspomniani wcześniej bukmacherzy, czy sami piłkarze, ale również dla wszystkich osób, które interesują się rozgrywkami sportowymi. Zastosowanie zaawansowanych technik analizy danych, uczenia maszynowego oraz modeli predykcyjnych pozwala na wykorzystanie ogromnej ilości informacji zgromadzonych podczas meczów do prognozowania wyników, taktyk zespołów oraz indywidualnych osiągnięć graczy, co w znaczy sposób może ułatwić pracę trenerów przy opracowywaniu taktyk i piłkarzy, przy wyznaczaniu swoich słabych stron. Ponadto specjalnie przeszkolone modele mogą pomagać zarządcom klubów optymalizować swoje wydatki transferowe oraz bukmacherom przy ustalaniu kursów zakładów. Znaczenie przewidywania wyników meczów piłkarskich nie ogranicza się tylko do aspektów sportowych i hazardowych. Sport ma ogromny wpływ na społeczeństwo, kulturę oraz życie codzienne, dlatego też precyzyjne prognozy mogą mieć znaczenie również dla mediów sportowych, które korzystają z nich do tworzenia analiz oraz komentarzy dotyczących meczów. Ponadto, prognozy te mogą wpływać na zachowania kibiców, ich emocje oraz zaangażowanie w wydarzenia sportowe, co może przyczynić się do budowania więzi społecznych i wspólnotowych. Przewidywanie zdarzeń w meczach piłkarskich, mimo swojej wymagającej i nieprzewidywalnej natury, staje się coraz bardziej istotne z punktu widzenia różnych grup społecznych, mniej lub bardziej związanych z piłką nożną.

1.3. Cel i zakres pracy

Głównym celem niniejszej pracy jest stworzenie modelu, który bazując na starannie dobranych danych, będzie w stanie prognozować z jak najwyższą możliwą do uzyskania dokładnością wybrane zdarzenia mogące zajść w trakcie meczu piłkarskiego. W odróżnieniu od większości prac naukowych zajmujących się opisywanym tematem, ów badanie ma za zadanie spojrzeć na mecz piłkarskich z szerszej perspektywy niż same wyniki oraz liczba zdobytych bramek. Utworzony model ma również na celu przewidywać liczbę rzutów rożnych, wolnych, spalonych, strzałów na bramkę, czy otrzymanych kartek w obu kolorach przez zespoły uczestniczący w widowisku.

Drugi rozdział skupia się na przeglądzie literatury związanej z tematem przewidywania wyników meczów. Zawarte jest w nim omówienie różnych podejść do prognozowania rezultatów meczów, począwszy od metod opartych na analizie statystycznej, przez modele matematyczne, aż po wykorzystanie sztucznej inteligencji i uczenia maszynowego. Przedstawia przegląd aktualnych osiągnięć naukowych, zawierając najważniejsze prace naukowe, uszeregowane w sposób chronologiczny, które swoją innowacyjnością wywarły duży wpływ na obszar przewidywania

zdarzeń nie tylko w piłce nożnej, ale i w innych sportach.

W trzecim rozdziale zostaną przedstawione wszystkie aspekty dotyczące danych, na których bazuje cała metodologia nauczania. W nim przedstawione zostaną dokładne dane, które zostały wyselekcjonowane na potrzeby badania, ich źródła, oraz sposób, w jaki zostały zorganizowane. Ze względu na użycie relacyjnych baz danych wyróżniono poszczególne tabele oraz kolumny, które wchodzą w jej skład jak i połączenia między encjami.

Czwarty rozdział skupia się na modelu oraz wszystkich aspektach z nim związanych. Przedstawia system rankingowy utworzony na potrzeby mierzenia siły drużyn, który wykorzystywany jest w celu przewidywania wyników meczów. Ponadto przedstawione zostały metodologie oraz mechanizmy wykorzystywane w celu predykcji pozostałych zdarzeń takich jak rzuty rożne czy wolne. W celu potwierdzenia poprawności wypracowanych schematów przedstawione zostaną obrazowe przykłady, na podstawie których czytelnik będzie mógł zweryfikować poprawność zaimplementowanych metod.

Piąty rozdział skupia się na szczegółach implementacyjnych. W tej sekcji zostaną przedstawione fragmenty kodów źródłowych wraz z ich opisami, pozwalające czytelnikowi na głębsze zrozumienie procesów oraz mechanizmów wdrożonych w ramach przeprowadzonego badania.

Szósty rozdział przedstawia wyniki intenstywnych testów prowadzonych w celu dokładnego zweryfikowania możliwości zaimplementowanego modelu. Testy te głównie będą opierały się na przewidywaniu wybranych zdarzeń z przyszłości na podstawie zebranych wcześniej danych, porównania z testami przeprowadzonymi w innych pracach naukowych, czy na porównaniu z obszernym rynkiem bukmacherskim. To również w tej części czytelnik zostanie wprowadzony w podstawowe pojęcia stosowane w zakładach bukmacherskich oraz zostaną przedstawione zaimplementowane metodologie, mające na celu układanie optymalnych zakładów nastawionych na jak największy profit gracza.

Ostatni, siódmy rozdział, stanowi podsumowanie zrealizowanych w pracy elementów. Przedstawione zostaną aspekty pracy, które udało się zrealizować, możliwe rozszerzenia modelu oraz sposoby jego wykorzystania w branży piłkarskiej.

Rozdział 2

Przegląd literatury

2.1. Początki

Piłka nożna jako sport z małą liczbą bramek, a co za tym idzie, z dużo liczbą remisów (remisy średnio stanowią od $\frac{1}{4}$ do $\frac{1}{3}$ wyników, w zależności od specyfiki ligii [6]), szczególnie na poziomie profesjonalnym, nie jest sportem łatwym jeśli chodzi o przewidywanie wyników. Większość popularnych sportów, takich jak koszykówka, hokej na lodzie czy siatkówka, nawet w sezonach regularnych wprowadza mechanizm dogrywek (ang. tiebreaker) w celu wyeliminowania remisów z puli możliwych wyników. Istnienie nadprogramowego wyniku starcia okazuje się być większym utrudnieniem niż mogłoby się pierwotnie wydawać, co z pewnością przykuło uwagę wielu badaczy, którzy podjęli rękawicę i rozpoczęli badania w tym obszarze.

Lata 30' ubiegłego wieku to moment, w którym pojawiają się pierwsze udokumentowane schematy próbujące przewidywać wyniki spotkań. Na łamach czasopisma *Lincoln Journal Star* publikowano system utworzony przez P.B. Williamsona. W swoim dziele brał pod uwagę takie czynniki jak trudność terminarza, liczbę rozgrywanych meczów w najbliższym okresie co w połączeniu z matematycznymi metodami takimi jak metoda najmniejszej sumy kwadratów (ang. least squares [1]) oczarowało czytelników gazety. Ideę wykorzystania metody *least square* po śmierci Williamsona kontynuował Stefani [41], który wykorzystał ją do ustalenia ocen drużyn, które mogły być aktualizowane co tydzień, aby przewidzieć wyniki meczów na podstawie różnicy w ocenach między tymi drużynami. Takie podejście dało początek systemom rankingowym, które wykorzystywane są również w nowoczesnych pracach naukowych.

Pierwsze poważniejsze modele służace do przewidywań wyników opierały się głównie na narzędziach wywodzących się wprost ze statystyki. Jedną z pionierskich prac poruszających temat przewidywania zdarzeń w piłce nożnej była praca autorstwa *M. J. Moroney'a* , który w 1951 roku wykorzystał *rozkład Poissona* do przewidywania liczby bramek zdobytych przez daną drużynę w meczu piłkarskim. Niezadowolony z rezultatów swojego badania, zasugerował, iż użycie zmodyfikowanego rozkładu Poisson'a (co w rzeczywistości oznaczało wykorzystanie ujemnego rozkładu dwumianowego, znanego również jako *Rozkład Pascal'a*), doprowadzi do znacznie lepszego dopasowania, co zostało zweryfikowane między innymi przez *M.J Maher'a* [31]. Podobna metodologia została użyta w pracy *Reep'a i Benjamina*, którzy poszli o krok dalej i wykorzystali zaimplementowany model do przewidywania wyników pozostałych sportów, których związek z piłką nożna był taki, iż do rozgrywania meczów używały piłki [39]. Jednym z głównych wniosków płynących z ów badania było przeświadczenie o tym, że takowe sporty są zbyt losowe, aby móc efektywnie przewidywać ich wyniki, co szybko zostało zdementowane w pracy [22]. Faktem jest, iż wystepowanie takich sytuacji, jak czerwona kartka, która znacznie osłabia szansę drużyny, która ją otrzymała, na zwycięstwo, czy nieprawidłowości w sędziowaniu, często prowadzą do wypaczania wyników, jednakże porównania predykcji ekspertów z faktycznymi

wynikami pokazują, iż balans między przysłowiowym szczęciem a pechem, predzej czy później musi wyjść na zero i ostatecznie lepsze drużyny wygrywają częściej [22]. Dixon i Coles [13] zmodyfikowali w swojej pracy model Maher'a [31], aby mógł radzić sobie z niekompletnymi danymi i danymi z różnych rogrywek oraz aby umożliwić czasowe zmiany w wydajności zespołów. Był to jeden z pierwszych modelów, który w swoich rozważaniach próbował brać pod uwagę zagadnienie *formy* drużyny. Idea badania formy rozwinięta została również w pracy autorstwa Babooty i Kaur'a [4], którzy przy pomocy następujących formuł:

$$Form_j^A = Form_{j-1}^A + \alpha Form_{j-1}^B$$
$$Form_j^B = Form_{j-1}^B - \alpha Form_{j-1}^A$$

a w przypadku remisów:

$$Form_j^A = Form_{j-1}^A - \alpha(Form_{j-1}^A - Form_{j-1}^B)$$
$$Form_j^B = Form_{j-1}^B - \alpha(Form_{j-1}^B - Form_{j-1}^A)$$

określili sposób wyznaczania formy drużyn A i B. Należy zauważyć, iż pojęcie formy w ów pracy jest zależne od obu drużyn biorących udział w spotkaniu (branie pod uwage jedynie osiągnięc jednej drużyny zostało określone mianiem *serii* i bazuje na fenomienie gorącej ręki[1]). Testy wykazały, iż hiperparametr α o wartości równej 0.33 skutkuje otrzymaniem najdokładniejszych wyników. Należy zauważyć, iż pojęcia formy, serii, czy liczby meczów branych pod uwagę, może różnić się w zależności od intencji autora badania.

2.2. Analiza wykraczająca poza statystyki meczowe

Istnieją dane związane z widowiskiem piłkarskim, których nie da się bezpośrednio pozyskać z czystych pomeczowych statystyk przeszłych spotkań. Mowa tutaj o pogodzie panującej danego dnia, wymiarach boiska[2], które mogą wpływać na sposób gry drużyn, czy dostępność, jak i samopoczucie zawodników poszczególnych ekip biorących udział w wydarzeniu. W jednej z prac badacze podjęli się wyzywania przewidywania wyników meczów biorąc pod uwagę jedynie warunki atmosferyczne panujące w trakcie spotkania, otrzymując zaskakującą dokładność predykcji na poziomie 50 procent [34]. Inne badania brały pod uwagę średni wiek zawodników [20] czy ich wartość rynkową [28]. Pojęcie **Mądrość tłumu** (ang. **Wisdom of the crowd**) odnosi się do zjawiska, w którym zbiorowa mądrość grupy osób okazuje się być trafniejsza niż wiedza poszczególnych ekspertów [45]. Koncepcja ta opiera się na założeniu, iż agregacja różnych punktów widzenia finalnie może prowadzić do lepszych decyzji i bardziej precyzyjnych prognoz. W przypadku sportów za opinię tłumu można uznać komentarze głoszone przez kibiców na trybunach bądź w mediach społecznościowych. W pracy [43] Wunderlich'a i Mement'a zastosowano analizę semantyczną wpisów opublikowanych na platformie Twitter[3] dotyczących spotkań rozgrywanych w ramach najwyższej klasy rozgrywkowej w Anglii. Doszukiwali się w nich informacji odnośnie nastawienia kibiców do nadchodzącego meczu, informacji o statusie kluczowych zawodników czy obawach o wybranych piłkarzy drużyny przeciwnej. Badacze z

[1] **Gorąca ręka** to pojęcie wywodzące się z koszykówki opisujące zjawisko, wcześniej uważane za poznawcze uprzedzenie społeczne, polegające na tym, że osoba, która doświadczyła udanego wyniku, ma większą szansę na sukces w kolejnych próbach.

[2] Standardem odnośnie wymarów boiska uchwalonym przez Fédération Internationale de Football Association (FIFA) jest 105x68, jednakże przepisy pozwalają na marginesy błędu. Ostatecznie dozwolone wymiary boiska wahają się od 100 do 110 metrów długości i od 64 do 75 metrów szerokości

[3] Obecnie platforma ta nosi nazwę **X**

Turcji [16], czy chociażby z Anglii [28] zastosowali inne podejście, w którym w swoich pracach wykorzystali informacje pozyskane z mediów społecznościowych jako cechy trenowanych modelów [16].

Wzmożony wzrost wykorzystania zaawansowanej technologii w sporcie pozwolił na prężny rozwój takich firm jak **OPTA**[4] . Firmy te zajmują się zbieraniem szczegółowych danych z meczów piłkarskich, obejmujących różnorodne aspekty gry, takie jak podania, strzały, odbiory poszczególnych piłkarzy, czy liczba interwencji bramkarskich. Każdy mecz jest analizowany przez zespół ekspertów, który rejestruje tysiące indywidualnych zdarzeń. Informacje na temat osiągnięć poszczególnych zawodników mogą znacznie wpłynąć na poprawienie skuteczności modelów, ze względów na fakt, iż finalnie to oni odpowiadają za sukces danej drużyny na polu gry [24]. Na podstawie zebranych informacji o zawodnikach podjęto próby przedstawienia ich umiejętności w różnych aspektach meczu przy pomocy ocen. Najpopularniejszym przykładem tego procesu jest seria gier **EA FC**[5] autorstwa studia **Electronic Arts**, w której każdy z piłkarzy otrzymuje osobną ocenę w 6 kategoriach:

- Pace (szybkość) — ocenia przyspieszenia i maksymalnej prędkości zawodnika
- Shooting (strzały) — przedstawia dokładność, siłę i technikę strzałów na bramkę
- Passing (podania) — dotyczy precyzji i siły podań zawodnika
- Dribbling (drybling) — prezentuje umiejętność prowadzenia piłki przez piłkarza
- Defending (obrona) — ocenia zdolność zawodnika do przechwytywania piłki, blokowania strzałów i skutecznego bronienia
- Physical fizyczność — odnosi się do siły fizycznej oraz wytrzymałości zawodnika

Rys. 2.1: Przykładowa karta zawodnika prezentująca oceny jego poszczególnych atrybutów. W lewym górnym rogu znajduje się ocena ogólna piłkarza wyliczana przez twórców gry przy pomocy autorskiego algorytmu. Źródło: `https://www.ea.com/games/ea-sports-fc/ratings`, ostatnia data wizyty: 19.05.2024

Wszystkie parametry zawodnika przyjmują wartości z zakresu od 1 do 99, gdzie ocena 99 nadawana jest zawodnikom, którzy zostali uznani za najlepszych względem danej własności, a 1 nada-

[4] `https://optaplayerstats.statsperform.com/pl_PL/soccer`

[5] Seria gier znana jest ze swojej poprzedniej nazwy FIFA, która obowiązywała do 2024 roku

wana tym, którzy zostali uznani za najgorszych. Badania, w których wykorzystano powyzsze atrybuty zawodników do predykcji wyników prezentują obiecujące rezultaty, wskazując jakoby ich wykorzystanie mogło znacznie wpłynąć na możliwości przewidywania zdarzeń w meczach piłkarskich [11] [36].

2.3. Systemy rankingowe

Pierwsze próby przewidywań opierały się na analizie przeszłych meczów pod kątem historycznych osiągnięc drużyn w obrębie danych rozgrywek, ale również w starciach bezpośrednich. Nie jest tajemnicą, iż istnieją zestawienia, w których notorycznie jedna z drużyn osiąga zwycięstwa bez względu na datę rozgrywek czy intensywność terminarza, jednakże rozwój piłki nożnej doprowadził do sytuacji, w której istotniejszymi rozgrywkami są te rozgrywane na poziomie międzynarodowym. W Europie w rozgrywkach drużynowych naistotniejszym trofeum jest trofeum **Ligi Mistrzów**, natomiast w rozgrywkach krajowych - puchar **Mistrzostw Świata**. Łatwo zauważyć, iż aktualny sposób predykcji jest nieadekwatny do nowej sytuacji, gdyż drużyny z różnych krajów grają przeciwko sobie zdecydowanie rzadziej, niż jakby grały wspólnie w jednej lidze. Dodatkowym problemem okazuje się również określenie różnicy w sile lig, gdyż piąta drużyna ligii francuskiej nie musi być lepsza niż ósma drużyna ligi angielskiej. Ze względu na potrzebę porównywania siły drużyn z różnych krajów, naukowcy podjęli się próby utworzenia systemu, który pozwoliłby ów cel uzyskać z jak największą dokładnością.

Ranking ELO

Pierwsze prace wykorzystywały znany z szachów ranking **Elo** [17], którego nazwa pochodzi od nazwiska twórcy, amerykańskiego matematyka pochodzenia węgierskiego, *Arpad'a Elo*. W systemie rankingowym Elo każda drużyna zwyczajowo zaczyna z rankingiem 1500. System można podzielić na dwa etapy: etap szacowania: krok E (ang. estimation step) i etap aktualizacji: krok U (ang. update step). Krok E polega na oszacowaniu prawdopodobieństwa z jakim drużyna wygra dany mecz, podczas gdy krok U obejmuje aktualizację rankingu drużyny po określonym meczu.
Krok E:

$$p_{i,j}(t) = \frac{1}{1 + 10 * \frac{-(R_i(t) - R_j(t))}{400}} \tag{2.1}$$

$$p_{j,i}(t) = \frac{1}{1 + 10 * \frac{-(R_j(t) - R_i(t))}{400}} \tag{2.2}$$

$p_{i,j}(t)$ oznacza prawdopodobieństwo zwycięstwa drużyny **i** nad drużyna **j**, w momencie **t** a $R_i(t)$ oznacza aktualny ranking drużyny **i**. Przez zmienną t rozumiemy aktualnie rozpatrywany moment biorąc pod uwage, iż w rozważaniach wszystkie mecze analizujemy chronologicznie względem terminu rozegrania spotkania zaczynając od tych najstarszych, kończąc na najbliższych teraźniejszości.
Krok U:

$$R_i(t+1) = R_i(t) + K[A_i(t) - p_{i,j}(t)] \tag{2.3}$$

$$R_j(t+1) = R_j(t) + K[A_j(t) - p_{j,i}(t)] \tag{2.4}$$

$A_i(t)$ oznacza wynik starcia z perspektywy drużyny **i**. Funkcja ta przyjmuje wartości ze zbioru $\{1, 0.5, 0\}$, gdzie 1 oznacza zwycięstwo drużyny **i**, 0.5 remis, a 0 - porażkę. Współczynnik **K** znany jest jako tempo wzrostu - określa skalę zmian wynikających z rozegrania pojedynczego

spotkania. **K** moze przyjmować wartość stała (w pracy [17] zaproponowana wartość przez autora systemu wynosi 10), lecz w rozgrywkach szachowych stała ta zależy od poziomu zaawansowania gracza (im gracz rozegrał więcej gier lub znajduje się wyżej w rankingu tym stała K używana do obliczenia zmian jego rankingu jest mniejsza).

W kontekście piłki nożnej również prowadzony jest ranking Elo klubów jak i reprezentacji narodowych. Ponadto, idąc śladami szachistów, w football'u również zapronowano wykorzystywanie różnych wartości K, tym razem w zależności od prestiżu rozgrywek (posługując się przykładem, strona eloratings [6] przypisuje $K = 60$ meczom rozgrywanym w ramach mistrzostw świata, $K = 40$ kwalifikacjom do mistrzostw świata, a $K = 20$ dla meczów towarzyskich) oraz liczby zdobytych bramek w meczu (parametr K jest zwiększany o połowę, jeśli mecz zostanie wygrany dwoma bramkami, o $\frac{3}{4}$, jeśli mecz zostanie wygrany trzema bramkami i o $\frac{3}{4} + \frac{(N-3)}{8}$, jeśli mecz zostanie wygrany czterema lub więcej bramkami, gdzie N to różnica bramek). Powyżej przytoczono tylko jeden rodzaj takich rankingów Elo, w zasobach internetu dostępnych jest więcej stron internetowych, w których badacze, jak i entuzjaści piłki nożnej, próbują się prześcignąć ustalając jak najlepsze parametry dla systemu rankingowego Elo, aby jak najlepiej odzwierciedlał realną siłę poszczególnych drużyn.

Ranking PI

Ranking PI autorstwa Anthonego Constantinou i Norman'a Fenton'a zadebiutował w 2013 roku [8]. Charakteryzuje się on tym, iż każda drużyna ma przypisane dwa rankingi — domowy, jak i wyjazdowy. Wynika to z faktu, iż drużyny częściej wygrywają grając na swoich obiektach [7] [40]. Ostateczny ranking obliczany jest jako średnia arytmetyczna z rankingu domowego i wyjazdowego danej drużyny:

$$R_i = \frac{R_{i,H} + R_{i,A}}{2} \tag{2.5}$$

Gdzie R_i oznacza aktualny ranking i-tej drużyny, $R_{i,H}$ przedstawia ranking ów drużyny w meczach domowych, a $R_{i,A}$ - w meczach wyjazdowych. Użycie liter **H** i **A** jednoznacznie wskazuje odpowiednio na odwołanie się do drużyny gospodarzy (home team) i drużyny gości (away team). Należy zauważyć, iż ranking *PI* jest pionierem jeśli chodzi o wykorzystanie zjawiska domowej przewagi w wydarzeniach sportowych. W celu aktualizacji rankingu obliczana jest oczekiwana różnica bramek między drużynami. Dokonywane jest to na podstawie rankingów gospodarzy jak i gości. Jeśli rzeczywisty wynik meczu jest lepszy z perspektywy danej drużyny niż oczekiwano, jej ocena jest zwiększana na podstawie różnicy między faktyczną liczbą bramek, a ustaloną przez system rankingowy. Warto zauważyć, iż jeśli druzyna rozgrywa mecz domowy, to jedynie jej rating **H** ulega zmianie, **A** pozostaje bez zmian. Analogiczna sytuacja występuje dla drużyny przyjezdnej. Dokładny sposób aktualizacji rankingów domowych jak i wyjazdowych został opisany w oryginalnej pracy autorów [8].

Ranking Berrar'a

Ranking Berrar'a powstał jako rozszerzenie podejścia zaproponowanego przez Constantinou oraz Fenton'a. Do wcześniej zaprezentowanego podziału rankingu na domowy i wyjazdowy, Berrar wraz z Lopezem i Dubitzkym postanowili wprowadzić pojęcie *siły ofensywnej* oraz *słabości defensywnej* danej drużyny. Aplikując zapropnowane rozwiązania, zarówno do meczów

[6] `https://eloratings.net/` (ostatnia data wizyty: 02.04.2024), strona internetowa prezentująca ranking ELO piłkarskich drużyn narodowych

[7] Zjawisko to w żargonie piłkarskim określane jest jako **przewaga własnego boiska**

domowych jak i wyjazdowych, otrzymano cztery cechy, według których dokonywane są późniejsze analizy. Cały ranking opiera się na przewidywaniu oczekiwanej liczby bramek na podstawie wyżej wspomnianych ocen drużyn biorących udział w spotkaniu. Sama predykcja liczby bramek dokonywana jest z wykorzystaniem funkcji logistycznej przypominającej sigmoid [29]. Autorzy pracy motywują użycie takiego podejścia małą liczbą bramek występujących w meczach piłkarskich: zauważają, iż rzadkością jest wystąpienie większej liczby bramek niż 5 [8] [6]. Aby przedstawić mechanizm ustalania siły ofensywnej jak i defensywne drużyny należy wprowadzić pomocnicze zmienne. Niech:

- att_H oznacza ofensywną ocene drużyny gospodarza
- def_H oznacza defensywną ocenę drużyny gospodarza
- att_A i def_A — analogicznie jak powyżej z tą różnicą, że wartości odnosza się do drużyny przyjezdnej

wtedy rankingi aktualizują się zgodnie z poniższym wzorcem:

$$att_H = att_H + \omega_{att_H} * (goals_H - e\_goals_H) \tag{2.6}$$
$$def_H = def_H + \omega_{def_H} * (goals_H - e\_goals_H) \tag{2.7}$$
$$att_A = att_A + \omega_{att_A} * (goals_A - e\_goals_A) \tag{2.8}$$
$$def_A = def_A + \omega_{def_A} * (goals_A - e\_goals_A) \tag{2.9}$$

Gdzie zmienna oznaczona słowem $goals$ oznacza liczbę bramek zdobytych w danym meczu przez daną drużynę, a e_goals - liczbę bramek przewidywaną przez model. W oparciu o powyższe parametry ustalana jest oczekiwana liczba bramek w następujący sposób:

$$e\_goals_H(att_H, def_A) = \frac{\alpha_H}{1 + exp(-\beta_H(att_H + def_A) - \gamma_h)} \tag{2.10}$$

Definicję wszystkich parametrów wykorzystanych w ramach konstrukcji rankingu oraz sposoby ich ustalania dostępne są dla czytelnika w oficjalnej pracy autorów [6]

Raking GAP

Ranking GAP (ang. Generalized Attacking Performance) [42] powstał jako kolejne rozszerzenie filozofii stosowanej w rankingu PI. Dla danego meczu piłkarskiego niech S_H i S_A będą miernikami wydajności drużyny gospodarza oraz drużyny gości, gdzie definicja wydajności jest ścisle określana przez użytkownika. Podejście zastosowane w rankingu GAP jest na tyle uniwersalne, iż przy pomocy tego samego zestawu formuł, zmieniając jedynie mierniki, czyli parametry wejściowe, można przewidywać wyniki tych samych meczów. Przez parametry wejściowie rozumie się statystyki meczowe użyte do predykcji. W swojej pracy jako najpopularniejsze używane meczowe obserwacje autor wymienia liczbę bramek, strzałów na bramkę, ale również i tak nieoczywiste statystyki jak liczba rzutów rożnych. Nic nie stoi na przeszkodzie, aby dowolnie mieszać ze sobą ów aspekty danych pochodzących z meczu piłkarskiego, aby uzyskać jak najdokładniejszą skuteczność predykcji rezultatów meczowych.

W procesie tworzenia rankingu GAP każda drużyna otrzymuje oddzielne oceny wydajności pfensywnej i defensywnej dla meczów u siebie i na wyjeździe. W oryginalnej pracy oznaczono to w następujący sposób:

- H_i^a — ofensywna ocena wydajności i-tej drużyny w meczach domowych
- H_i^d — defensywna ocena wydajności i-tej drużyny w meczach domowych
- A_i^a — ofensywna ocena wydajności i-tej drużyny w meczach wyjazdowych

[8] W zbiorze danym utworzonym w ramach pracy jedynie GAP1 meczów zakończyło się sumą bramek większą niż 5. Stanowi to około 5 procent wszystkich wyników

- A_i^d — defensywna ocena wydajności i-tej drużyny w meczach wyjazdowych

Oceny otrzymane w ramach zastosowania rankingu GAP odnoszą się do oczekiwanej skuteczności drużyny w ataku bądź obronie podczas starcia z przeciętną drużyna w danej lidze. Najlepsze drużyny w ramach danych rozgrywek będą miały bardzo wysokie mierniki ofensywne (skorelowane jest to z większą liczbą zdobytych bramek czołowych drużyn), a bardzo niskie mierniki defesywne (związane jest to z faktem, iż najlepsze drużyny, to z reguły te, które tracą najmniej bramek).

Zestaw poniższych reguł przedstawia mechanizm aktualizacji ocen po każdym z meczów:

$$H_i^a = max(H_i^a + \lambda\phi_1(S_h - \frac{H_i^a + A_j^d}{2}), 0) \tag{2.11}$$

$$A_i^a = max(A_i^a + \lambda(1-\phi_1)(S_h - \frac{H_i^a + A_j^d}{2}), 0) \tag{2.12}$$

$$H_i^d = max(H_i^d + \lambda\phi_1(S_a - \frac{A_j^a + H_i^d}{2}), 0) \tag{2.13}$$

$$A_i^d = max(A_i^d + \lambda(1-\phi_1)(S_a - \frac{A_j^a + H_i^d}{2}), 0) \tag{2.14}$$

$$\tag{2.15}$$

Analogiczny zestaw reguł został zdefiniowany dla drugiej drużyny biorącej udział w starciu:

$$A_j^a = max(A_j^a + \lambda\phi_2(S_a - \frac{A_j^a + H_i^d}{2}), 0) \tag{2.16}$$

$$H_j^a = max(H_j^a + \lambda(1-\phi_2)(S_a - \frac{A_j^a + H_i^d}{2}), 0) \tag{2.17}$$

$$A_j^d = max(A_j^d + \lambda\phi_2(S_h - \frac{H_i^a + A_j^d}{2}), 0) \tag{2.18}$$

$$H_j^d = max(H_j^d + \lambda(1-\phi_2)(S_h - \frac{H_i^a + A_j^d}{2}), 0) \tag{2.19}$$

$$\tag{2.20}$$

Dla danego meczu mówi się, iż drużyna grająca u siebie przewyższyła oczekiwania, jeśli jej pomeczowa ocena jest lepsza niż średnia z dotychczasowej oceny ofensywnej rozpatrywanej drużyny i oceny defensywnej rywala. Dochodzi wtedy do aktualizacji danych o wartości wyznaczone na podstawie powyższych wzorów. Naturalnym jest, iż analogiczny mechanizm funkcjonuje dla defensywnych aspektów rankingu. Parametr λ determinuje wpływ ostatniego meczu na ranking (związane jest to z faktem, iż mecze bliższe teraźniejszości z perpsektywy analizy są istotniejsze [8]),parametr ϕ_1 odpowiada za wpływ rezultatu meczu domowego na mecz wyjazdowy, a ϕ_2 — meczu wyjazdowego na domowy. Wszystkie dodatkowe informacje na temat rankingu, parametrów, oraz metody ich wyznaczania dostepne są w oryginalnej pracy autora [42]

2.4. Predykcja oparta o uczenie maszynowe

Same systemy rankingowe mogą być wykorzystywane jako modele predykcyjne. Dzięki takim systemom można wydedukować zwycięzce na podstawie tego, która drużyna w aktualnym momencie ma wyższy ranking, jednak wielu badaczy zauważyło, iż użycie takich konstrukcji jako cechy modelu może skutkować otrzymaniem o wiele lepszych rezultatów. Rozwój technologii oraz zwiększenie dostępności do coraz to większych zbiorów danych umożliwiły prężny wzrost

zainteresowania zaawansowanymi metodami analizy danych w sporcie. Uczenie maszynowe oferuje nowe możliwości w prognozowaniu wyników meczów piłkarskich poprzez wykorzystanie zaawansowanych algorytmów, które potrafią wyodrębnić z danych istotne zależności i wzorce. Prace naukowe wykorzystujące uczenie maszynowe do osiągnięcia jak najdokładniejszych predykcji można podzelić na grupy ze względu na użyte metdologie. Ponadto, z uwagi na istnienie konstrukcji opartych wyłącznie na systemach rankingowych i statystyce oraz wyłącznie na uczeniu maszynowym, modele, które wykorzystują oba podejścia nazwano modelami *hybrydowymi*.

2.4.1. Konstrukcje Bayes'a

Sieć Bayes'a (ang. Bayes' network) — W drugiej dekadzie XXI wieku popularną metodologią uczenia maszynowego wykorzystywaną w pracach związanych z przewidywaniem wyników były *sieci Bayes'a*. Takowa sieć charakterzyje się tym, iż jest to probabilistyczny model graficzny, który reprezentuje zbiór zmiennych i ich warunkowych zależności za pomocą skierowanego grafu acyklicznego [35] . Badania porównawcze wskazują, że omawiana metodyka doskonale radzi sobie z prognozowaniem wyników meczów [27]. Na podstawie powyższych wniosków, do predykcji wyników meczów angielskiej Premier League greccy badaczy wykorzystali autorski system rankingowy o nazwie *AccuRATE* w połączeniu z kilkoma metodami uczenia maszynowego, w tym między innymi z naiwnym klasyfikatorem Bayes'a [30], który zdaniem autorów, może być postrzegany jako specjalny typ sieci bayesowskiej, opierający się na założeniu, że atrybuty są niezależne i żadne inne atrybuty nie wpływają na wyniki. Praca, która uplasowała się w czołowej trócje konkursu 2017 Soccer Predicition Challenge opierała się na systemie rankingowym *pi-rating*, który następnie wykorzystywany był przez hybrydową sieć Bayes'a (ang. Hybrid Bayesian Network) w celu wyłonienia zwycięzcy starcia [9].

2.4.2. Struktury drzewiaste

Drzewa decyzyjne (ang. decision trees) — Nadzorowane podejście do uczenia się stosowane w statystyce i eksploracji danych. Drzewa decyzyjne są wykorzystywane jako model predykcyjny do wyciągania wniosków na temat zbioru obserwacji, co idealnie dopasowuje się do prezentowanego problemu. Wiele prac naukowych wskazuje jakoby struktury drzewiaste, wraz z użyciem wzmocnień gradientowych (2.4.6), były jednym z lepszych podejść do przewidywania wyników. Na przestrzeni ostatnich pięciu lat głównym rywalem w tej dziedzinie okazują się być modele oparte o uczenie głębokie (2.4.5). Ponadto, istnieją prace wskazujące na to, iż inne struktury drzewiaste, a dokładniej **drzewa losowe** (ang. random forrest) mimo swojej niepozornej nazwy mogą konkurować na równym poziomie z modelami decyzyjnymi [4].

2.4.3. Wektory nośne

Wektory nośne (ang. **SVM**, support vector machine) — forma klasyfikatora, którego nauka polega na wyznaczeniu hiperpłaszczyzny rozdzielającej z maksymalnym marginesem przykłady należące do dwóch klas. Ów metdologia została użyta w wielu pracach skupiających się na przewidywaniach wyników konkretnej ligii [32], ale również w przewidywaniu meczów z różnych typów rozgrywek [7].

2.4.4. K-najbliższych sąsiadów

K-najbliższych sąsiadów (ang. **KNN**, K-Nearest Neighbor) — nadzorowany algorytm leniwego uczenia się stosowany w uczeniu maszynowym. Oznacza to, że przechowuje on dane szkoleniowe prezentowane przez nadzorców (ang. supervisors) i porównuje je z innymi danymi w celu

prognozowania. Wykorzystanie takiego podejścią z dużą liczbą dostepnych meczów pozwala na wykrywanie wzorców w wcześniej rozgrywanych spotkaniach, na podstawie których mogą być generowane predykcje nadchodzących spotkań [21].

2.4.5. Uczenie głębokie

Uczenie głębokie (ang. deep learning) — typ uczenia maszynowego, które używa sztucznych sieci neuronowych, aby umożliwić systemom cyfrowym uczenie się i podejmowanie decyzji na podstawie danych bez struktury i etykiet. W ciągu ostatnich kilku lat głębokie uczenie zyskało na popularności w przewidywaniu wyników meczów piłki nożnej, ze względu na swoje sukcesy w wielu innych dziedzinach życia takich jak wizja komputerowa, czy przetwarzanie języka naturalnego, czego konsekwencją jest duża liczba prac, w której modele wykorzystują wspomnianą metdologię uczenia. W swojej pracy Razali i in. [38] wykorzystali podejście o nazwie **TabNet** [3], które można interpretować jako głęboką sieć neuronową zaprojektowaną specjalnie dla danych przechowywanych w formie tabel. Panowie Danisik, Lacko oraz Farkas porównali wydajność modelu *długiej pamięci krótkotrwałej* (ang. **LSTM** long-short term memory [23]) z klasyfikacją, przewidywaniem numerycznym, losowym zgadywaniem wyników jak i z przewidywaniami bukmacherskimi [12]. Rahman wykorzystał głębokie i sztuczne sieci neuronowej **ANN** (ang. artificial neural network) do przewidywania wyników meczów Mistrzostw Świata rozegranych w 2018 roku [37], natomiast Panowie Jain, Tiwari oraz Sardar wykorzystali rekurencyjne sieci neuronowe (ang. **RNN** - recurrent neural network) wraz w wcześniej wspomnianym LSTM'em do przewidywania wyników angielskiej pierwszej klasy rozgrywkowej [26]. Jedna z najnowszych prac, opublikowana w ramach wyzwania 2023 Soccer Prediction Challenge, również wykorzystuje uczenie głębokie w celu rozwiązania problemów przedstawionych przez twórców konkursu [44].

2.4.6. Wzmocnienie gradientowe

Wzmocnienie gradientowe (ang. gradient boosting) — podejście do nauczania głębokiego zakładające budowanie drzew decyzyjnych, z których każde kolejne w iteracyjnym procesie staje się doskonalsze od poprzedniego. Ostateczny model agreguje całą serię drzew. Jedne z dwóch najlepszych prac przesłanych w ramach wyzwania 2017 Soccer Prediction Challenge opierały się właśnie na drzewach decyzyjnych wzmocnionych gradientowo. Obie prace oparły implementację swojego modelu o strukturę *XGBoost* [25] [6] . Różnica między owymi pracami leżała w zaimplemntowanym systemie rankingowym — zwycięzcy całego konkursu skorzystali z piratingu [25], natomiast autorzy drugiej pracy, która mimo lepszego wyniku, nie była brana pod uwagę w rankingu końcowym,[9] zaimplementowali własny system, nazwany później od nazwiska jednego z twórców, systemem *Berrar'a*. Ponadto w pierwszej z prac wykorzystano RDN-Boost, jednak ze względu na lepsze osiągnięcia modelu opartego na XGBoost stanowił on jedynie punkt odniesienia dla testów.

[9]Praca [6] mimo faktu, iż osiągnęła najlepszy wynik ze wszystkich przesłanych prac nie została ogłoszona jako zwycięzca ze względu na fakt, iż została zredagowana przez osoby, które były również zaangażowane w organizację konkursu

2.5. Wykorzystanie kursów i zakładów bukmacherskich

2.6. Konkurs 2017 Soccer Prediction Challenge

Ze względu na rosnące zainteresowanie tematem związanym z przewidywaniem wyników meczów piłkarskich wydawnictwo *Springer* zorganizowało konkurs, który okazał się być kamieniem milowym dla rozpatrywanego obszaru. Od tego momentu większość prac naukowych wykorzystuje opublikowany w ramach wyzwania zbiór danych, bądź wyniki otrzymane przez modele w ramach konkursu, w celu porównania działania swojego dzieła z innymi rozwiązaniami znanymi naukowemu światu. W specjalnej edycji czasopisma *Machine Learning* autorzy podsumowali całe wydarzenie prezentując najlepsze prace wraz z opisem ich działania oraz elementami, które odróżniały je od reszty stawki [5]. W ramach wyzwania każdy z uczestników otrzymał dostęp do bazy danych, która po zakończeniu współzawodnictwa została upubliczniona do wglądu dla każdego użytkownika internetu wraz z dokładnym opisem zawartości [14]). Baza ta zawiera ponad 216 000 meczów z 52 lig piłkarskich i 35 krajów z okresu między 2000 a 2017 rokiem. W ramach wyzwania każdy z modeli po intensywny trenowaniu na otrzymanej bazie testowej miał za zadanie przewidzieć wyniki 206 spotkań rozgrywanych między 31 marcem, a 30 kwietniem sezonu 2016/17. Do zmierzenia poprawności przewidywań modelów wykorzystano **RPS** (ang. Ranked Probability Score), który idealnie sprawuje się w sytuacji, gdy prognozy wyrażone są jako rozkłady prawdopodobieństwa wyników [18].

2.7. Konkurs 2023 Soccer Prediction Challenge

Po sześciu latach od poprzedniej edycji w minionym roku odbyła się kolejna, w której badacze z różnych zakątków świata mięli za zadanie sprostać lekko zmodyfikowanemu wyzwaniu. W tej wersji, w odróżnieniu od poprzedniej z 2017 roku, badacze mieli do rozwiązania 2 problemy - problem wyznaczenia poprawnego zwycięzcy starcia bądź remisu oraz problem wyznaczenia dokładnego wyniku (poza ustaleniem zwycięstwa gospodarzy musieli również podać dokładny wynik jakim zakończy się spotkanie, np. 1:0). Mimo faktu, iż modele miały zostać oceniane osobno pod względem umiejętności przewidywania dokładnych wyników oraz pod względem umiejętności wyboru zwycięzcy, aby owoc pracy badaczy został przyjęty do wyzwania musiał potrafić przewidywać oba zdarzenia. Prawidłowość oceny otrzymanych wyników w teście wyznaczania dokładnego wyniku miała zostać oceniona na podstawie średniego błędu kwadratowego (**RMSE** ang. root mean square error) między rzeczywistymi wynikami a przewidywanymi wynikami, natomiast w teście wyznaczania zwycięzcy (bądź remisu) użyto znanej z poprzedniej edycji formuły **RPS**. Niestety, w momencie redagowania pracy tylko nieliczne artykuły opisujące działanie przesłanych modelów ujrzały światło dzienne [44]. Aktualnie jedyne informacje zawarte na stronie organizatora to lista rankingowa przedstawiająca osiągnięcia poszczególnych uczestników [2]. Należy zauważyć, iż uruchomienie kolejnej części konkursu związanej z przewidywaniem wyników w piłce nożnej podkreśla znaczenie tego problemu oraz rosnące zapotrzebowanie na skuteczne rozwiązania w tej dziedzinie.

2.8. Przewidywanie pozostałych zdarzeń

Podstawowym problemem w przewidywaniu wyników meczów piłkarskich jest możliwość występowania remisów, co oznacza, iż są trzy istniejące rezultaty spotkania. Naturalnym jest, iż łatwiej jest przewidzieć zdarzenie mając do wybory tylko i wyłącznie dwie opcje. Powyższe rozumowanie oraz diametralny wzrost rynku związanego z zakładami bukmacherskimi skłoniły wielu badaczy do eksploracji innych statystyk wchodzących w skład meczów piłkarskich. Z

biegiem czasu badacze zauważyli, iż przewidywanie innych, równie ciekawych zdarzeń, aniżeli samych końcowych rezultatów, okazuje się być nie mniej interesującym wyzwaniem. Istnieją prace, które badają występowanie takich zdarzeń jak popularne *BTTS* (ang. both teams to score, innym popularnym skrótoweczem używanym zamiennie jest BTS, z pominięcie T odpowiadającemu wyrazowi *the*), których celem jest ustalenie, czy w wybranym meczu obie drużyny strzelą przynajmniej jedną bramkę [10], powyżej/poniżej 2,5 gola na mecz ([33]), czy ustalonej liczby rzutów rożnych w danym spotkaniu ([46] [19]).

Rozdział 3

Inżynieria danych

3.1. Zbiory danych

Przewidywanie zdarzeń w sporcie, jak i w innych dziedzinach życia, opiera się silnie na odkrywaniu wzorców w historycznych danych. Bazując na tym fakcie, istotnym jest, aby analizowane zbiory były dokładne i obszerne, czyli zawierały dużą liczbę wpisów. W celu przeprowadzenia analizy, stworzono własnoręcznie zbiór danych, opierając się na informacjach ze stron internetowych o charakterze kronikarskim. Jedną z takich stron, która została użyta w celu pozyskania istotnych informacji na temat spotkań, była strona o nazwie *flashscore*[1]. Dzięki wpisom uzyskanym z przytoczonej strony utworzono zbiór zorganizowany w formie relacyjnej bazy danych nazwanej *ekstrabet*, który w dalszej części pracy będzie służył jako podstawa do analizy, a następnie wnioskowania wyników przyszłych zdarzeń. Na ilustracji 3.1 przedstawiono rozmaitość statystyk przechowywanych przez opisywaną stronę internetową. Szczegółowy opis statystyk, które zostały wyselekcjonowane jako naistotniejsze znajduje się w sekcji 3.2 *Organizacja danych*.

Do analizy zdarzeń w ramach niniejszej pracy wyselekcjonowano dwa najwyższe poziomy rozrgrywkowe dla następujących sześciu krajów:

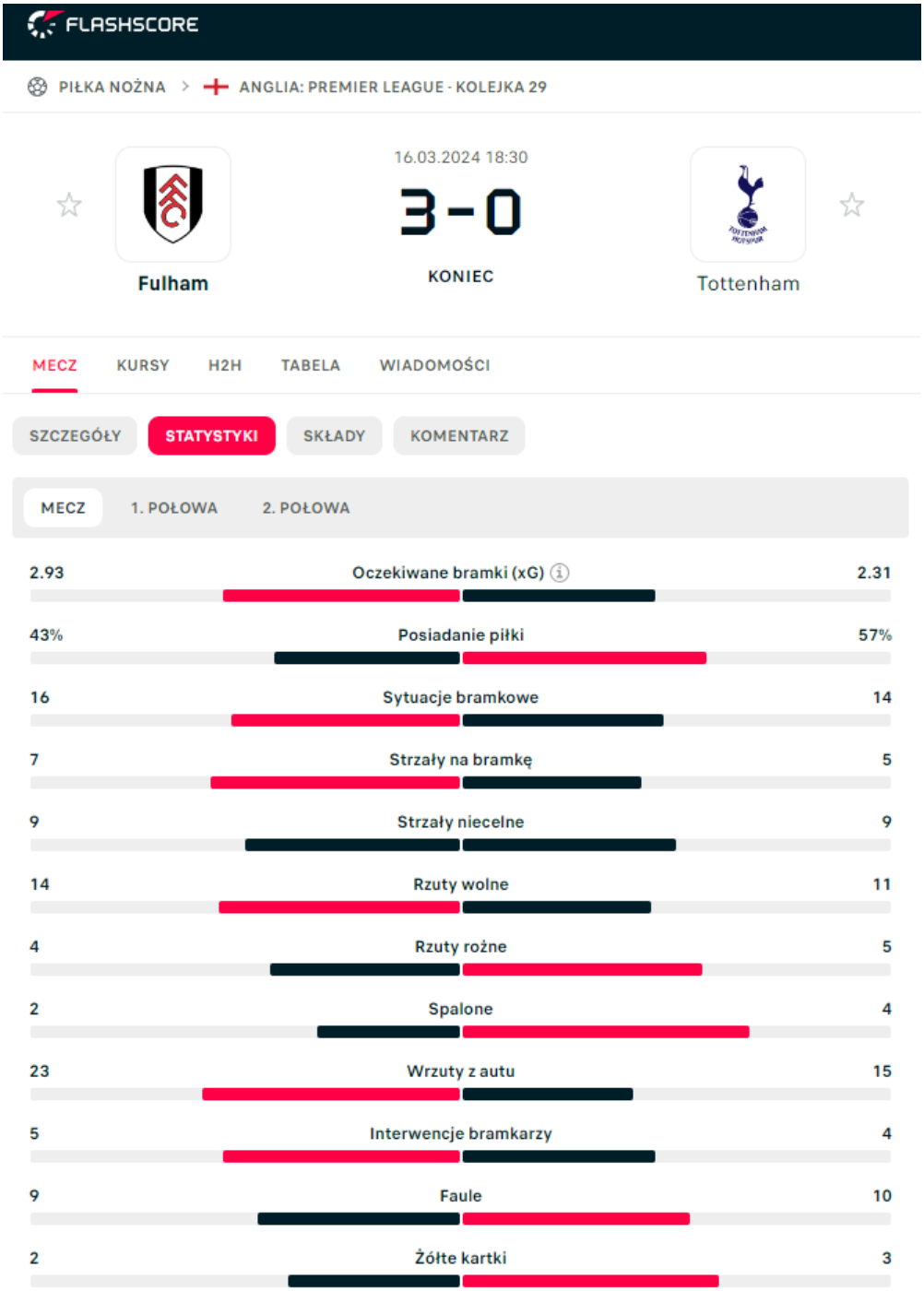
Nazwa kraju	Pierwsza liga	Druga liga
Polska	Ekstraklasa	1. Liga
Anglia	Premier League	Championship
Francja	Ligue 1	Ligue 2
Włochy	Serie A	Serie B
Hiszpania	LaLiga	LaLiga2
Niemcy	Bundesliga	2. Bundesliga

Tab. 3.1: Tabela zawierająca kraje oraz nazwy (stan na 16.03.2024) dwóch najwyższych poziomów rozgrywkowych, których analiza spotkań jest głównym celem badawczym pracy. Należy mieć na uwadze fakt, iż nazwy lig w przyszłości mogą ulec zmianie, ze względu na potencjalną zmianę sponsora tytularnego ligii czy reorganizację rozgrywek

Domyślnie zebrano dane ze spotkań, które odbywały się w ciągu ostatnich osmiu sezonów (włącznie z sezonem 2016/17), jednakże selekcja spotkań z niektórych ligi, takich jak polska **1. Liga** została skrócona ze względu na braki w zapisie kronikarskim wyselekcjonowanych cech.[2]

[1] `https://www.flashscore.pl/` (ostatnia data wizyty: 19.05.2024), strona internetowa wykorzystana do utworzenia zbioru danych

[2] Dokładna liczba sezonów, a co za tym idzie spotkań, dla każdej z lig, zostanie przedstawiona w sekcji odpowiedzialnej za organizację danych 3.2



Rys. 3.1: Przykładowa strona ze statystykami dla pojedynczego spotkania (tu: spotkanie między drużynami Fulham oraz Tottenham rozgrywane dnia 16.03.2024 w ramach 29 kolejki spotkań angielskiego najwyższego poziomu rozgrywkowego zwanego **Premier League** zakończone wynikiem 3:0 dla gospodarzy

3.2. Organizacja danych

Zgodnie z informacją przedstawioną w poprzedniej sekcji, wszystkie dane na temat spotkań zostały przedstawione przy pomocy relacyjnej bazy danych. W skład bazy wchodzą następujące tabele:

1. Teams
2. Countries

3. Matches
4. Seasons
5. Leagues
6. Events
7. Odds
8. Bookmakers

Poniższe sekcje dokonują szczegółowego opisu wyżej wymienionych encji.

3.2.1. Tabela *Teams*

Tabela *Teams* przechowuje identyfikatory oraz informacje o drużynach, które rozegrały przynajmniej jedno spotkanie w ramach analizowanego zbioru danych. W skład ów tabeli wchodzą następujące kolumny

- id - kolumna reprezentująca unikalny identyfikator nadawany każdej drużynie, która rozegrała przynajmniej jeden mecz w ramach dowolnych rozgrywek objętych badaniem [3]
- country - kolumna przedstawiająca informację z jakiego kraju pochodzi dana drużyna
- name - kolumna przechowująca nazwę danej drużyny[4]

id	country	name
1	1	Śląsk Wrocław
2	1	Jagiellonia Białystok
3	1	Lech Poznań
4	1	Raków Częstochowa
5	1	Legia Warszawa

Rys. 3.2: Przykład reprezentacji drużyn w tabeli Teams

W ramach badania wybrano 323 drużyn z 6 krajów (stan na 16.05.2024), których osiągnięcia będą przewidywane przez utworzony model.

3.2.2. Tabela *Countries*

Prezentowana tabela zawiera identyfikatory oraz nazwy krajów, których drużyny uczestniczyły w przynajmniej jednym spotkaniu w ramach analizowanego zbioru danych. Struktura tabeli skupia się wyłącznie na dwóch polach: identyfikatorze i nazwie kraju. Ze względu na ich prostotę, szczegółowy opis w postaci listy został pominięty. Badanie skupia się na pięciu czołowych krajach europejskich pod względem osiągnięć piłkarskich, a także na rodzimym - Polsce. Polska została szczególnie uwzględniona w dalszej analizie, ze względu na znikomą liczbę badań tego typu dla ów kraju.

[3] Jeśli nie zaznaczono inaczej, wszystkie tabele w bazie danych posiadają identyfikatory oznaczone przy pomocy skrótu **id**

[4] Należy mieć na uwadzę fakt, iż podobnie jak z nazwami lig, z biegiem czasu nazwy drużyn również mogą ulegać zmianom

id	name
1	Polska
2	Anglia
3	Francja
4	Niemcy
5	Włochy

Rys. 3.3: Przykład reprezentacji krajów w tabeli Countries

3.2.3. Tabela *Seasons*

Tabela przechowująca unikalne identyfikatory sezonów branych pod uwage w ramach selekcji spotkań. Sezon piłkarski w Europie trwa od lipca lub sierpnia do maja bądź czerwca (w krajach o systemie jesień-wiosna) i od marca lub kwietnia do listopada bądź grudnia (w krajach o systemie wiosna-jesień). Ze względu na fakt, iż wszystkie analizowane ligii grają systemem jesień-wiosna sezon podawany jest jako dwa lata oddzielone symbolem /. Jeśli wynika to z kontekstu, pierwsze dwie cyfry oznaczające tysiąclecie oraz stulecie są pomijane przy podawaniu drugiego roku, co oznacza, iż sezon rozpoczęty jesienią 2016 roku, a zakończony wiosną 2017, oznaczono jako 2016/17 (więcej przykładzie na zrzucie ekranu 3.4). Wyjątkiem byłby sezon 1999/2000, jednakże wykracza on poza ramy czasowe prezentowanego badania. Lata, w których odbywają się rozgrywki przechowywane są w kolumnie *years*

id	years
1	2023/24
2	2022/23
3	2021/22
4	2020/21
5	2019/20

Rys. 3.4: Przykład reprezentacji sezonów w tabeli Seasons

3.2.4. Tabela *Leagues*

Tabela przechowująca informację na temat lig objętych badaniem. Poza standardowym identyfikatorem ligi znajdują się tam informację o kraju, w którym ona funkcjonuje (pole *country*, oraz o nazwie, którą nosi (pole *name*). W nazwach niektórych lig mogą pojawiać się sponsorzy tytularni, którzy moga ulegać częstym zmianom.

id	name	country
1	PKO BP Ekstraklasa	1
2	Premier League	2
3	Ligue 1	3
4	Bundesliga	4
5	Serie A	5

Rys. 3.5: Przykład reprezentacji krajowych lig w tabeli Leagues

3.2.5. Tabela *Matches*

Tabela przechowująca wszystkie niezbędne do wnioskowania dane o spotkaniach rozegranych w rozpatrywanych ligach. Poniżej przedstawiono wszystkie cechy, które zostały wyselekcjonowane jako najistotniejsze do przewidywania wybranych w pracy zdarzeń:

- league - liga, w ramach której rozegrano spotkanie [5]
- season - sezon, w trakcie którego rozegrano dany mecz
- game_date - data rozegrania spotkania w formacie rrrr-mm-dd hh:mm:ss
- home_team - drużyna, która w danej rywalizacji była gospodarzem
- away_team - drużyna, która w danym starciu była gościem
- home_team_goals - liczba bramek zdobyta przez drużynę gospodarzy
- away_team_goals - liczba bramek uzyskana przez drużynę gości
- home_team_xg - wartość wskaźnika xG (Oczekiwane bramki - liczba bramek, które dana drużyna powinna strzelić na podstawie jakości i liczby oddanych strzałów [15]) drużyny gospodarzy.
- away_team_xg - wartość wskaźnika xG drużyny gości
- home_team_bp - czas posiadania piłki (bp - ball possesion) przez gospodarzy wyrażony w procentach
- away_team_bp - czas posiadania piłki przez gości wyrażony w procentach
- home_team_sc - wykreowane szanse strzeleckie (sc - scoring chances) przez drużynę organizującą spotkanie
- away_team_sc - utworzone szanse strzeleckie przez drużynę gości
- home_team_sog - liczba strzałów celnych na bramkę (sog - shots on goal) lokalnego zespołu
- away_team_sog - liczba strzałów celnych na bramkę drużyny przyjezdnej
- home_team_fk - liczba rzutów wolnych (fk - free kicks) wykonanych przez miejscowy zespół
- away_team_fk - liczba rzutów wolnych wykonanych przez gości
- home_team_ck - liczba rzutów rożnych (ck - corner kick) wykonanych przez zespół grający na własnym terenie
- away_team_ck - liczba rzutów rożnych wykonanych przez gości
- home_team_off - liczba spalonych (off - offsides) gospodarzy
- away_team_off - liczba spalonych drużyny przyjezdnej
- home_team_fouls - liczba fauli (fouls - faule) gospodarzy
- away_team_fouls - liczba fauli gości
- home_team_yc - liczba żółtych kartek (yc - yellow card) lokalnego zespołu
- away_team_yc - liczba żółtych kartek przyjezdnego zespołu
- home_team_rc - liczba czerwonych kartek (rc - red card) lokalnej drużyny
- away_team_rc - liczba czerwonych kartek gości

id	league	season	home_team	away_team	game_date	home_team_goals	away_team_goals
1	1	2	13	19	2023-05-27 17:30:00	3	0
2	1	2	3	2	2023-05-27 17:30:00	2	0
3	1	2	5	1	2023-05-27 17:30:00	3	1
4	1	2	21	7	2023-05-27 17:30:00	0	0
5	1	2	9	20	2023-05-27 17:30:00	3	0

Rys. 3.6: Przykład reprezentacji meczów w tabeli Matches. Ze względów objętościowych większa część kolumn w prezentowanym zrzucie ekranu zostala pominięta

[5] Ze względu na przedstawienie bazy danych w postaci normalnej, w tabeli **Matches** pole *league* oznaczone jest przy pomocy unikalnego identyfikatora ligi. Podobny mechanizm dotyczy pól: *seasons*, *home_team* i *away_team*

W ramach badania wybrano *34681* [6] spotkań rozegranych w ramach ostatnich osmiu sezonów czołowych europejskich lig. [7] [8] [9] [10]

Nazwa ligii	Liczba sezonów	Liczba spotkań w sezonie	Suma spotkań
Ekstraklasa	8	3*306 / 1*240 / 4*296	2342
1. Liga	7	5*309 / 1*306 / 1*308	2159
Premier League	8	380	3040
Championship	8	557	4456
Bundesliga	8	308	2462
2. Bundesliga	8	308	2464
LaLiga	8	380	3040
LaLiga2	8	468	3744
Serie A	8	7* 380 / 1*381	3041
Serie B	8	4 * 390 / 1*388 / 1*352 / 1*472 / 1*470	3242
Ligue 1	8	1 * 306 / 4 * 384 / 1 * 279 / 1 * 382	2503
Ligue 2	6	2 * 380 / 3 * 382 / 1 * 280	2186

Tab. 3.2: Tabela przedstawia analizowane ligii wraz z liczbą sezonów objętą analizą, liczbą spotkań przypadającą na sezon oraz sumaryczną pobraną liczbą spotkań. Różnice w liczbie spotkań na sezon dla niektórych lig zostały wyjaśnione w przypisach na ów stronie

3.2.6. Tabela *Events*

Tabela *Events* zawiera wszystkie rozpatrywane zdarzenia wraz z najpopularniejszymi wartościami, jakie mogą osiągnąć. Przez wartość rozumiemy liczbę wystąpień danego zdarzenia. Zgodnie z ustaloną nomenklaturą, wartości podawane są w następującej formie: **<drużyna-/mecz> <powyżej/poniżej> <wartość> <nazwa_zdarzenia>**. Przykładem jednego z wpisów w tabeli Events zgodnie z powyższym wzorcem jest: *Gospodarz powyżej 3.5 gola*. Należy zauważyć, iż często jeśli dane zdarzenie odnosi się do meczu, słowo to jest pomijane (przykład: *Powyżej 1.5 gola*)

[6] W bazie danych znajdują się również przyszłe spotkania, nie posiadające wyników, które zostały wyselekcjonowane w celach treningowych

[7] Polska Ekstraklasa do sezonu 2020/21 składała się z 16 zespołów. W sezonie 2020/21 postanowiono dokonać ekspansji ligi do 18 zespołów, jednakże ze względu na pandemię część spotkań w ramach grup mistrzowskich oraz spadkowych nie została rozegrana, mistrz i spadkowicz zostali wyłonieni bezpośrednio po fazie zasadniczej

[8] 1. Liga w ramach wyłonienia ostatniej drużyny awansującej do najwyższej klasy rozgrywkowej rozgrywa baraż między drużynami z miejsc 3-6. Drużyny z miejsc 3 i 6 oraz 4 i 5 rywalizują ze sobą w bezpośrednim pojedynku, zwycięzcy awansują do finału, którego triumfator awansuje do Ekstraklasy. Taki schemat funkcjonuje od sezonu 2019/20. W sezonie 2018/19 nie było mini-playoff'ów o awans (stąd tylko 306 spotkań), natomiast w sezonie 2017/18 rozegrany został jedynie dwumeczowy baraż o utrzymanie (stąd 308 spotkań)

[9] W Serie A w sezonie 2022/23 twórcy reformy rozgrywek nie przewidzieli sytuacji, w której na ostatnim miejscu spadkowym znajdą się dwie drużyny o tej samej liczbie punktów oraz o tym samym rekordzie meczów bezpośrednich. Ze względu na ów fakt, między drużynami Spezia oraz Verona musiał zostać rozegrany dodatkowy mecz o pozostanie we włoskiej najwyższej klasie rozgrywkowej. W Serie B różnice w liczbie meczów wynikają z licznych zmian w systemie rozgrywek

[10] W Ligue 1 sezon 2020/21 ze względu na pandemię nie został ukończony, stąd znaczący spadek liczby meczów w jednym z sezonów. Inne dodatkowe zmiany we francuskiej piłce odnośnie liczby spotkań na sezon (dotyczy również Ligue 2) wynikają z ich zapału do częstego reformowania rozgrywek

id	name
1	Zwycięstwo gospodarzy
2	Remis
3	Zwycięstwo gości
4	Gospodarz nie przegra
5	Gość nie przegra

Rys. 3.7: Przykład reprezentacji najpopularniejszych zdarzeń w tabeli Events

W ramach badania wyselekcjonowano siedem popularnych zdarzeń występujących w meczach piłkarskich

- Rezultat meczu — Zwycięstwo gospodarza, remis, bądź zwycięstwo gości
- Podwójna szansa — Gospodarz nie przegra bądź gość nie przegra
- BTTS — Czy obie drużyny strzelą bramkę?
- Liczba bramek
- Liczba rzutów rożnych
- Liczba fauli
- Liczba żółtych kartek

Warto zauważyć, iż zdarzenia zawierające w swojej nazwie *liczba* będą rozpatrywane z persepktywy meczu jak i z perpsektywy każdej z drużyn.

3.2.7. Tabela *Odds*

Tabela odds zawiera informację na temat kursów wystawionych przez danego bukmachera na dane zdarzenie w analizowanym meczu. Stąd naturalnie kolumny, które wchodzą w skład ów tabeli to:

- match — pole zawierające id analizowanego spotkania
- bookmaker — kolumna zawierająca nazwę bukmachera
- event — nazwa zdarzenia
- odds — ustalony kurs

3.2.8. Tabela *Bookmakers*

Tabela *Bookmakers* zawiera nazwy bukmacherów, od których pobrane zostały kursy do dalszej analizy. Zawiera ona jedynie identyfikator bukmachera, jak i jego nazwę.

id	name
1	Superbet
2	Betclic
3	Fortuna
4	STS
5	LvBet
6	Betfan

Rys. 3.8: Nazwy bukmacherów branych pod uwage w ramach przeprowadzonego badania

W ramach badania wybrano sześciu najpopularniejszych legalnych bukmacherów operujacych na terenie Rzeczpospolitej Polski.

3.2.9. Zależności między tabelami

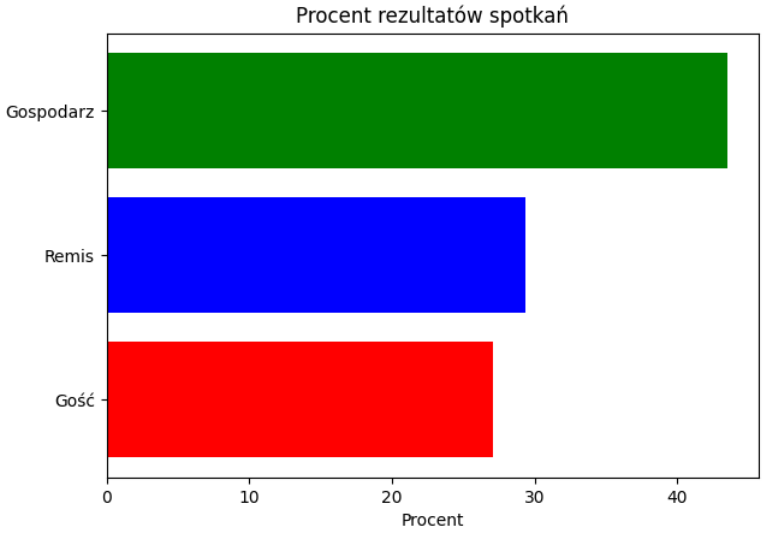
Rozdział 4

Budowanie modelu

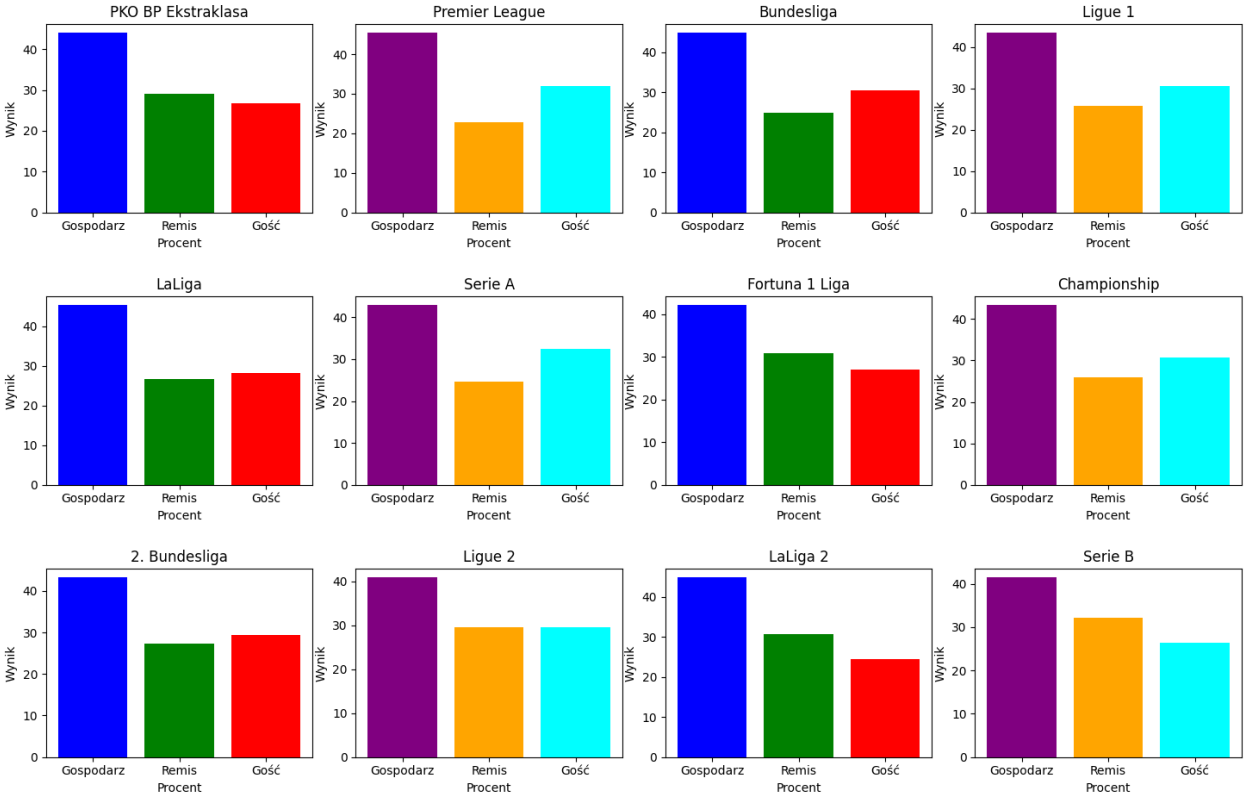
4.1. Analiza wyselekcjonowanych zdarzeń

Wynik meczu

Najbardziej pożądanym aspektem do przewidywania z perspektywy wszystkich grup społecznych jest ten najważniejszy, czyli wynik meczu piłkarskiego. Jak wcześniej wskazano, ze względu na możliwość uzyskania trzech rezultatów jest to jedno z trudniejszych zdarzeń do poprawnego wytypowania. Innym czynnikiem przedstawionym na wykresach 4.1 oraz 4.2 utrudniającym realizację zaplanowanego badania jest istnienie fenomenu przewagi własnego boiska. Zdecydowanie utrudnia to analizę, szczególnie w sytuacji, gdy drużyną gości jest ekipa będącą silniejsza z perspektywy rankingu.



Rys. 4.1: Wykres przedstawiający sumaryczny rozkład rezultatów spotkań we wszystkich analizowanych meczach



Rys. 4.2: Wizualizacja rozkładu rezultatów we wszystkich analizowanych ligach. Należy zauważyć, iż w żadnej lidzie nie zachodzi sytuacja, w której gospodarz wygrywałby rzadziej niż w 40 procentach spotkań, co potwierdza, iż ta niespodziewana własność spotkań ligowych jest zjawiskiem uniwersalnym

Przy przewidywaniu wyników meczów w ramach rozgrywek ligowych należy mieć na uwadzę szerszy kontekst otaczający dane spotkanie. Należy zauważyć, iż istnieją sytuacje, w których jedna z drużyn, a czasem nawet obie, nie wychodzi na spotkanie zmotywowana do walki o punkty w 100 procentach. Może to wynikać z wielu czynników:

- Napięty terminarz — Dotyczy najlepszych drużyn, które ze względu na uczestnicwo w meczach pucharowych rozgrywają dwa spotkania tygodniowo. Ze względu na duża eksploatację zawodników, w teoretycznie łatwiejszych spotkaniach ligowych takowe drużyny często sięgają po rezerwowy skład w celu odciążenia podstawowych zawodników
- Priorytet dla pucharów — Dotyczy drużyn z niższych pozycji w tabeli, które ze względu na uczestnictwo w pucharach mogą zaistnieć na arenie krajowej bądź międzynarodowej. Podobnie jak w przypadku napiętego terminarza, drużyny te mogą całe swoje przygotowanie taktyczne poświęcić przyszłemu meczowi pucharowemu zaniedbując lekko najbliższe spotkanie ligowe
- Początek i koniec sezonu — Z reguły dotyczy drużyn okupujących, dla których rozegranie najbliższego, bądź najbliższych spotkań, jest nieistotne z perspektywy pozycji w tabeli

Liczba goli

Czy obie strzelą?

Ciągły wzrost dostępnych danych piłkarskich stwarza bezprecedensowe możliwości badawcze dla lepszego zrozumienia dynamiki piłki nożnej. Podczas gdy wiele badań koncentruje się tylko i wyłącznie na przewidywaniu, która drużyna wygra mecz, bądź ile bramek padnie w meczu, inne interesujące pytania, takie jak to, czy obie drużyny w danym meczu zdobędą bramkę, są często

zaniedbywane [10]. **BTTS** (ang. Both Teams To Score), gdyż takowym skrótowcem określą się ów problem, ...

4.2. Rozwiązywanie popularnych problemów

4.2.1. Problem początku i końca sezonu

Piłkarskie ligi zasadniczo rozgrywane są w systemie kołowym. Jest to sposób prowadzenia rozgrywek, nie tylko piłkarskich, w którym rywalizacja polega na bezpośrednich pojedynkach między wszystkimi drużynami. W systemie kołowym każda z drużyn musi rozegrać jedno spotkanie ze wszystkimi przeciwnikami biorącymi udział w lidze, przez co zwyczajowo ów sposób prowadzenia rywalizacji nazywany jest systemem *każdy z każdym* (ang. *round-robin*). W piłce nożnej aby zniwelować wcześniej wspomniany atut własnego boiska rozgrywany jest podwójny system kołowy - jedno spotkanie między drużynami A i B odbywa się na stadionie należącym do zespołu A, a drugie, na obiekcie ekipy B. Dodanie dodatkowej kopii zbioru spotkań z zamienionymi gospodarzami w piłkarskim żargonie nazywane jest *meczem i wyjazdem* (ang. *double round-robin*).

Pojedynczy sezon ligowy determinowany jest przez liczbę meczów niezbędnych do rozegrania, aby zrealizować pełny format systemu podwójnego kołowego. Zakładając, iż dane rozgrywki posiadają X drużyn, należy rozegrać $\frac{X*(X-1)}{2}$ spotkań, aby wyłonić zwycięzce. Z perspektywy pojedynczej ekipy należy przeprowadzić więc $2*(X-1)$ meczów. Biorąc pod uwagę zasady mające na celu ochronę zdrowia piłkarzy między dwoma meczami danej ekipy musi być odstęp minimum 72 godzin. Zauważając, iż poza rozgrywkami ligowymi drużyny prowadzą również rywalizację w przeróżnych pucharach (krajowych bądź europejskich), rywalizacje ligowe rozgrywane są cyklicznie co tydzień [1].

Problem początku sezonu, zwany również okresem przejściowym, stanowi wyzwanie dla przewidywania wyników meczów ze względu na znaczne zmiany w składach drużyn i restrukturyzacje klubów, które często mają miejsce w tym okresie, ze względu na obowiązuje *okienko transferowe*, czyli moment, w którym zawodnicy mogą legalnie zmianiać swoje barwy klubowe. Ze względu na liczne rewolucje prowadzone w tym okresie, zauważalnym jest fakt, iż pierwsze ligowe mecze są zdecydowanie bardziej nieprzewidywalne niż pozostałe. Dodatkowo na początku sezonu każda drużyna jest zmotywowana do udowodnienia swojej wartości. Nowe zespoły w lidze lub zespoły z niższymi aspiracjami mogą sprawić niespodzianki, pokonując teoretycznie silniejszych rywali.

Problem końca sezonu ...

Mając to na uwadze, badacze w swoich pracach podjęli próbę niwelacji negatywnych efektów powyższych zjawisk. Najpopularniejszym podejściem jest rezygnacia z brania pod uwagę n pierwszych, jak i n ostatnich meczów, gdzie n jest indywidualnie dobierane przez badacza [42] [6]. Wtedy pozostałe mecze przewidywanę są na podstawie poprzednich spotkań, które z perpsektywy rankingów są pozbawione egzystencji, dostarczają jedynie suchych statystyk odnośnie bramek, wyniku i innych wyszczególnionych cech. W ten sposób, przy problemie początku sezonu, dano czas drużynom na rozgrzanie się, a rozgrywkom, na powrót do naturalnego przebiegu, a przy problemie końca sezonu, uniknięto analizy meczów, których wynik mógł być znacznie wypaczony ze względu na chociażby wykorzystanie niestandardowych taktych czy rezerwowych zawodników.

W ramach prezentowanego badania przyjęto strategię, w której pierwsze, jak i ostatnie mecze sezonu traktowane są ze zwiększoną łagodnością. Wprowadzono dwa współczynniki: **współ-**

[1] W szczególnych przypadkach mecze, bądź całe kolejki mogą zostać przesunięte. Dzieje się tak między innymi w czasie świat, czy specjalnych przerw reprezentacyjnych

czynnik ograna, oraz **współczynnik relaksu** których zadaniem jest zmniejszanie korzyściu bądź strat rankingowych płynących z rozgrywania meczów w skrajnych etapach sezonu.

Współczynnik ograna ...

Współczynnik relaksu ...

4.2.2. Problem awansów i spadków

4.2.3. Problem aktualności

4.2.4. Przewaga własnego boiska

Wpływ czerwonej kartki na wynik meczu

TO-DO: Opis recency problem

4.3. System rankingowy

4.3.1. Motywacja

4.3.2. Opis systemu

4.3.3. Przykłady działania

4.4. Metodologie nauczania

4.4.1. Rekurencyjne sieci neuronowe

4.4.2. LSTM: Long short-term memory

4.4.3. Pojęcie okna

4.5. Przykłady działania

Rozdział 5

Szczegóły implementacyjne

5.1. Struktura projektu

5.1.1. Moduł Ratings

Moduł *Ratings* zawiera zaimplementowane systemy rankingowe w ramach realizacji modeli.

5.1.2. Moduł DataPrep

5.1.3. Moduł Model

5.1.4. Moduł Main

5.1.5. Dokumentacja

Rozdział 6

Walidacja modelu

6.1. Porównanie modelu z rzeczywistymi zdarzeniami

6.2. Porównanie modelu względem innych narzędzi

6.3. Wykrywanie anomalii

6.3.1. Śląsk Wrocław 2023/24

6.3.2. Piast Gliwice 2020/21

6.4. Testy

6.4.1. Testy modelu względem rzeczywistych wyników

6.4.2. Testy wyników względem innych rozwiązań

6.4.3. Wnioski

6.5. Porównanie modelu z zakładami bukmacherskimi

6.5.1. Wprowadzenie

6.5.2. Typy zakładów

6.5.3. Podatek

6.5.4. EV Bets

6.5.5. Przykłady rozbieżności kursów

6.5.6. Przykłady wygenerowanych zakładów

6.5.7. Eksperymenty

6.5.8. Wnioski

Rozdział 7

Podsumowanie

7.1. Omówienie rezultatów badań

7.2. Możliwe rozszerzenia pracy

Literatura

[1] Przykład publikacji systemu williamsona w gaziecie lincoln journal star: https://www.newspapers.com/article/the-knoxville-news-sentinel-williamson-s/115698786/ (ostatnia data wizyty: 02.04.2024).

[2] https://sites.google.com/view/2023soccerpredictionchallenge/important-dates (ostatnia wizyta 02.04.2024), strona internetowa wyzwania 2023 soccer prediction challenge organizowanego przez wydawnictwo springer.

[3] S. O. Arik, T. Pfister. Tabnet: Attentive interpretable tabular learning, 2020.

[4] R. Baboota, H. Kaur. Predictive analysis and modelling football results using machine learning approach for english premier league. *International Journal of Forecasting*, 2019.

[5] D. Berrar, P. Lopes, J. Davis, W. Dubitzky. Guest editorial: special issue on machine learning for soccer. *Machine Learning*, 108, 10 2018.

[6] D. Berrar, P. Lopes, W. Dubitzky. Incorporating domain knowledge in machine learning for soccer outcome prediction. *Mach. Learn.*, 108(1):97–126, jan 2019.

[7] H. Chen. Neural network algorithm in predicting football match outcome based on player ability index. *Advances in Physical Education*, 09:215–222, 01 2019.

[8] A. Constantinou, N. Fenton. Determining the level of ability of football teams by dynamic ratings based on the relative discrepancies in scores between adversaries. *Journal of Quantitative Analysis in Sports*, 9:37–50, 01 2013.

[9] A. C. Constantinou. Dolores: a model that predicts football match outcomes from all over the world. *Machine Learning*, 108:49–75, 2018.

[10] I. B. da Costa, L. B. Marinho, C. E. S. Pires. Forecasting football results and exploiting betting markets: The case of "both teams to score". *International Journal of Forecasting*, 38(3):895–909, 2022.

[11] N. Danisik, P. Lacko, M. Farkas. Football match prediction using players attributes. strony 201–206, 08 2018.

[12] N. Danisik, P. Lacko, M. Farkas. Football match prediction using players attributes. strony 201–206, 08 2018.

[13] M. J. Dixon, S. Coles. Modelling association football scores and inefficiencies in the football betting market. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 46, 1997.

[14] W. Dubitzky, P. Lopes, J. Davis, D. Berrar. The open international soccer database for machine learning. *Machine Learning*, 108, 01 2019.

[15] H. H. Eggels. Expected goals in soccer:explaining match results using predictive analytics. 2016.

[16] E.Jones. Artificial neural network approach for football scores prediction. *Journal of Networking and Communication Systems (JNACS)*, 6:13–21, 01 2023.

[17] A. E. Elo. *The Rating of Chessplayers, Past and Present*. Arco Pub., New York, 1978.

[18] E. S. Epstein. A scoring system for probability forecasts of ranked categories. *Journal of Applied Meteorology (1962-1982)*, 8(6):985–987, 1969.

[19] A. D. Fitt, C. J. Howls, M. Kabelka. Valuation of soccer spread bets. *The Journal of the Operational Research Society*, 57(8):975–985, 2006.

[20] A. Groll, C. Ley, G. Schauberger, H. Eetvelde. Prediction of the fifa world cup 2018 - a random forest approach with an emphasis on estimated team ability parameters, 06 2018.

[21] U. Haruna, J. Maitama, M. Mohammed, R. Raj. *Predicting the Outcomes of Football Matches Using Machine Learning Approach*, strony 92–104. 01 2022.

[22] I. D. Hill. Association football and statistical inference. *Journal of The Royal Statistical Society Series C-applied Statistics*, 23:203–208, 1974.

[23] S. Hochreiter, J. Schmidhuber. Long short-term memory. *Neural computation*, 9:1735–80, 12 1997.

[24] B. Holmes, I. G. McHale. Forecasting football match results using a player rating based model. *International Journal of Forecasting*, 40(1):302–312, 2024.

[25] O. Hubáček, G. Šír, F. Železný. Learning to predict soccer results from relational data with gradient boosted trees. *Machine Learning*, 108, 01 2019.

[26] S. Jain, E. Tiwari, P. Sardar. *Soccer Result Prediction Using Deep Learning and Neural Networks*, strony 697–707. 01 2021.

[27] A. Joseph, N. Fenton, M. Neil. Predicting football results using bayesian nets and other machine learning techniques. *Knowledge-Based Systems*, 19(7):544–553, 2006. Creative Systems.

[28] S. Kampakis, A. Adamides. Using twitter to predict football outcomes. 11 2014.

[29] K. Kempf-Leonard. *Encyclopedia of Social Measurement, Pages 641-651*. Number t. 1 serii Encyclopedia of Social Measurement. Elsevier Science, 2005.

[30] G. Kyriakides, K. Talattinis, G. Stephanides. A hybrid approach to predicting sports results and an accurate rating system. *International Journal of Applied and Computational Mathematics*, 3, 03 2017.

[31] M. Maher. Modelling association football scores. *Statistica Neerlandica*, 36:109–118, 1982.

[32] M.-C. Malamatinos, E. Vrochidou, G. Papakostas. On predicting soccer outcomes in the greek league using machine learning. *Computers*, 11:133, 08 2022.

[33] R. Mattera. Forecasting binary outcomes in soccer. *Annals of Operations Research*, 325:1–20, 08 2021.

[34] D. Palinggi, J. F. Ramos, J. Torres-Sospedra, S. Trilles Oliver, J. Huerta. Using weather condition and advanced machine learning methods to predict soccer outcome. 06 2019.

[35] J. Pearl. Chapter 2 - bayesian inference. J. Pearl, redaktor, *Probabilistic Reasoning in Intelligent Systems*, strony 29–75. Morgan Kaufmann, San Francisco (CA), 1988.

[36] C. Pipatchatchawal, S. Phimoltares. Predicting football match result using fusion-based classification models. strony 1–6, 06 2021.

[37] M. Rahman. A deep learning framework for football match prediction. *SN Applied Sciences*, 2, 02 2020.

[38] N. Razali, A. Mustapha, N. Arbaiy, P. Lin. Deep learning for football outcomes prediction based on football rating system. wolumen 2644, strona 040007, 11 2022.

[39] C. Reep, B. Benjamin. Skill and chance in association football. 1968.

[40] M. Ruano, R. Pollard, J.-C. Luis-Pascual. Comparison of the home advantage in nine different professional team sports in spain. *Perceptual and motor skills*, 113:150–6, 08 2011.

[41] R. Stefani. Football and basketball predictions using least squares. *IEEE Transactions on Systems, Man, and Cybernetics*, 7:117–121, 01 1977.

[42] E. Wheatcroft. A profitable model for predicting the over/under market in football. *International Journal of Forecasting*, 36, 01 2020.

[43] F. Wunderlich, D. Memmert. A big data analysis of twitter data during premier league matches: do tweets contain information valuable for in-play forecasting of goals in football? *Social Network Analysis and Mining*, 12, 12 2021.

[44] C. Yeung, R. Bunker, R. Umemoto, K. Fujii. Evaluating soccer match prediction models: A deep learning approach and feature optimization for gradient-boosted trees, 2023.

[45] S. K. M. Yi, M. Steyvers, M. D. Lee, M. J. Dry. The wisdom of the crowd in combinatorial problems. *Cognitive Science*, 36(3):452–470, 2012.

[46] S. Yip, Y. Zou, R. T. H. Hung, K. F. C. Yiu. Forecasting number of corner kicks taken in association football using compound poisson distribution, 2023.

Dodatek A

Instrukcja wdrożeniowa

UWAGA: Poniższy opis przedstawia sposób instalacji odpowiednich pakietów dla komputerów korzystających z systemu operacyjnego **WINDOWS 10**. Użytkownik chcąc zainstalować aplikację wraz z jej komponentami na komputerze z innym systemem operacjnym zobowiązany jest do samodzielnego zapoznania się ze wszystkimi komendami bądź mechanizami umożliwającymi instalację wskazanych pakietów.

Poniższa lista prezentuje podstawowe wymagania jakie musi spełnić komputer czytelnika, aby poprawnie uruchomić utworzone w ramach pracy oprogramowanie:

- Mysql
- Python
- Pakiety: numpy, pandas, matplotlib
- Pakiety: keras i tensorflow

Dodatek B

Opis załączonej płyty CD/DVD

Na załączonej płycie CD znajdują się wszystkie kody źródłowe wymagane w celu poprawnego uruchomienia programu. Poniżej przedstawiono schemat ułożenia katalogów oraz plików znajdujących się na dołączonym do pracy nośniku danych.

TO-DO: Graf z rozmieszczeniem plików na płycie CD