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, przewidywanie zdarzeń, modele optymalizacyjne

WROCŁAW, 2024

Streszczenie

Słowa kluczowe: sztuczna inteligencja, przewidywanie zdarzeń, modele optymalizacyjne

Abstract

Keywords: artificial intelligence, 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 rozgrywek w grupie C, do której również należała Polska, Arabii Saudyjskiej nad Argentyną władzę owego kraju ogłosiły dzień wolny od pracy. Ów mecz według wielu ekspertów określany jest jako największa sensacja w całej, niemalże stuletniej, historii mistrzostw świata. To 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 zespołu 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, zawiera on 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 silnie 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 [8]), 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 Pana P.B. Williamsona. W swoim systemie 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 [4]) oczarowało czytelników gazety. Ideę wykorzystania metody *least square* po śmierci Williamsona kontynuował Stefani [33], 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 Pana *M. J. Moroney'a* , który w 1951 roku wykorzystał *rozkładu 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 pana *M.J Maher'a* [26]. Podobna metodologia została użyta przez panów *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 [32]. Jednym z głównych wniosków płynących z ów pracy było przeświadczenie o tym, że takowe sporty są zbyt losowe, aby móc efektywnie przewidywać ich wyniki, co szybko zostało zdementowane w pracy [20]. Faktem jest, iż wystepowanie takich sytuacji, jak czerwona kartka, która znacznie osłabia szansę drużyny, która ją otrzymała, na zwycięstwo, czy nieprawidłowości w sędziowaniu, często prowadziły 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 [20]. Dixon i Coles [13] zmodyfikowali w swojej pracy model Maher'a [26], 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 [6], 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, różnic między nimi oraz liczba meczów brana pod uwagę są niejednoznaczne i silnie zależne od pomysłu autora modelu.

2.2. 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.

Pierwsze prace wykorzystywały znany z szachów ranking **Elo** [16], którego nazwa pochodzi od nazwiska twórcy, amerykańskiego matematyka pochodzenia węgierskiego, *Arpad'a Elo*. W systemie rankingowym Elo każda drużyna zazwyczaj 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.

[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.

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 [16] 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 [1] 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 z takich rankingów, w zasobach internetu dostępnych jest więcej stron internetowych, w których badaczej, jak i entuzjaści piłki nożnej, próbują się prześcignąć ustalając jak najlepsze parametry dla systemu rankingowego ELO, aby jak najlepiej odzwierciedlał realną siłę poszczególnych drużyn.

TO-DO : GAP RATING, BERRAR RATING, PI-RATING, AccuRATE - dopytać, czy to też musi być przedstawiane tak szczegółowo jak ranking ELO?

2.3. 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 podzielić 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*.

2.3.1. 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 [29] . Badania porównawcze wskazują, że omawiana metodyka doskonale radzi sobie z prognozowaniem wyników meczów [24]. 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 [25], 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 [10].

2.3.2. 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 (2.3.6), 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 (2.3.5). 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 [6].

2.3.3. 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 [27], ale również w przewidywaniu meczów z różnych typów rozgrywek [9].

2.3.4. 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 [19].

2.3.5. 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. [31] wykorzystali podejście o nazwie **TabNet** [5], 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 [21]) 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 [30], 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 [23]. Jedna z najnowszych prac, opublikowana w ramach wyzwania 2023 Soccer Prediction Challenge, również wykorzystuje uczenie głębokie w celu rozwiązania problemów przedstawionych przez twórców konkursu [34].

2.3.6. 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* [22] [8] . Różnica między owymi pracami leżała w zaimplemntowanym systemie rankingowym — zwycięzcy całego konkursu skorzystali z piratingu [22], natomiast autorzy drugiej pracy, która mimo lepszego wyniku, nie była brana pod uwagę w rankingu końcowym,[2] 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.

2.4. Analiza wykraczająca poza dane o drużynach

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 wy-

[2] Praca [8] 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

korzystuje 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 [7]. 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 [17].

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 [34]. 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 wzrost rynku zakładów bukmacherskich skłoniło wielu badaczy do eksploracji innych statystyk wchodzących w skład meczów piłkarskich. Z biegiem czasu badacze zauważyli, iż przewidywanie innych, równie ciekawych zdarzeń w meczach piłkarskich 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 [11], powyżej/poniżej 2,5 gola na mecz ([28]), czy ustalonej liczby rzutów rożnych w danym spotkaniu ([35] [18]).

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* [3]. 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*.

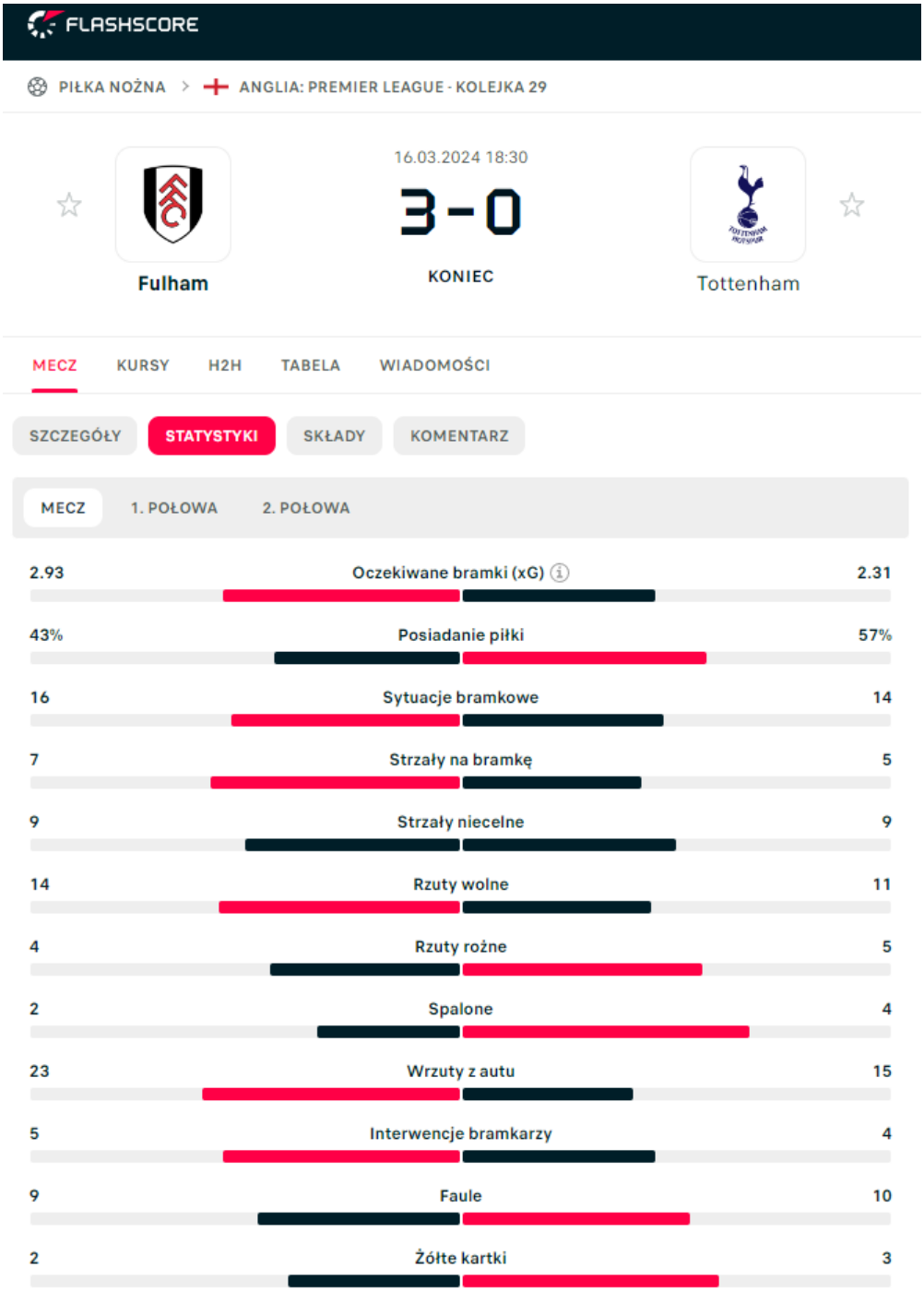
Do analizy zdarzeń w ramach niniejszej pracy wyselekcjonowano dwa najwyższe poziomy rozrgrywkowe dla następujących sześciu krajów:

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.[1]

[1]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 [2]
- country - kolumna przedstawiająca informację z jakiego kraju pochodzi dana drużyna
- name - kolumna przechowująca nazwę danej drużyny[3]

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 315 drużyn z 6 krajów (stan na 17.03.2024), 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.

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

[3] 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 [4]
- 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

[4]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* (stan na 17.03.2024) spotkań rozegranych w ramach ostatnich osmiu sezonów czołowych europejskich lig. [5] [6] [7] [8]

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	2464
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*

3.2.7. Tabela *Odds*

3.2.8. Tabela *Bookmakers*

[5] 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

[6] 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)

[7] 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

[8] 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

Rozdział 4

Budowanie modelu

4.1. Analiza wyselekcjonowanych danych

4.2. System rankingowy

4.2.1. Wpływ czerwonej kartki na wynik meczu

4.3. Metodologie nauczania

4.3.1. Rekurencyjne sieci neuronowe

4.3.2. LSTM: Long short-term memory

4.4. Przykłady działania

Rozdział 5

Szczegóły implementacyjne

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. Porównanie modelu z zakładami bukmacherskimi

6.4.1. Wprowadzenie

6.4.2. Podstawowe pojęcia

Rozdział 7

Podsumowanie

7.1. Omówienie rezultatów badań

7.2. Możliwe rozszerzenia pracy

Literatura

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

[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] https://www.flashscore.pl/ (ostatnia data wizyty: 16.03.2024), strona internetowa wykorzystana do utworzenia zbioru danych.

[4] Urywek wzmianki o systemie williamsona w gaziecie lincoln journal star: https://www.newspapers.com/article/the-knoxville-news-sentinel-williamson-s/115698786/ (ostatnia data wizyty: 02.04.2024).

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

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

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

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

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

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

[11] 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.

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

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

[18] 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.

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

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

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

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

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

[24] 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.

[25] 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.

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

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

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

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

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

[31] 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.

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

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

[34] 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.

[35] 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

Jeśli praca skończyła się wykonaniem jakiegoś oprogramowania, to w dodatku powinna pojawić się instrukcja wdrożeniowa (o tym jak skompilować/zainstalować to oprogramowanie). Przydałoby się również krótkie „*how to*" (jak uruchomić system i coś w nim zrobić – zademonstrowane na jakimś najprostszym przypadku użycia). Można z tego zrobić osobny dodatek.

Dodatek B

Opis załączonej płyty CD/DVD

Tutaj jest miejsce na zamieszczenie opisu zawartości załączonej płyty. Opis ten jest redagowany przed załadowaniem pracy do systemy APD USOS, a więc w chwili, gdy nieznana jest jeszcze nazwa, jaką system ten wygeneruje dla załadowanego pliku. Dlatego też redagując treść tego dodatku dobrze jest stosować ogólniki typu: „Na płycie zamieszczono dokument `pdf` z niniejszej tekstem pracy" – bez wskazywania nazwy tego pliku.

Dawniej obowiązywała reguła, by nazywać dokumenty według wzorca `W04_[nr albumu]_[rok kalendarzowy]_[rodzaj pracy]`, gdzie `rok kalendarzowy` odnosił się do roku realizacji kursu „Praca dyplomowa", a nie roku obrony. Przykładowo wzorzec nazwy dla pracy dyplomowej inżynierskiej w konkretnym przypadku wyglądał tak: `W04_123456_2015_praca inżynierska.pdf`, Takie nazwy utrwalane były w systemie składania prac dyplomowych. Obecnie działa to już inaczej.