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: raz, dwa, trzy

WROCŁAW, 2024

Streszczenie

Słowa kluczowe: raz, dwa, trzy

Abstract

Keywords: one, two, three

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

Wstęp

1.1. Wprowadzenie

1.2. Istniejące podejścia do przewidywania zdarzeń w piłce nożnej

1.3. Znaczenie przewidywania zdarzeń

1.4. Cel i zakres pracy

Rozdział 2

Inżynieria danych

2.1. 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 2.1 przedstawiono rozmaitość statystyk przechowywanych przez opisywaną stronę internetową. Szczegółowy opis statystyk, które zostały wyselekcjonowane jako naistotniejsze znajduje się w sekcji 2.2 *Organizacja danych*.

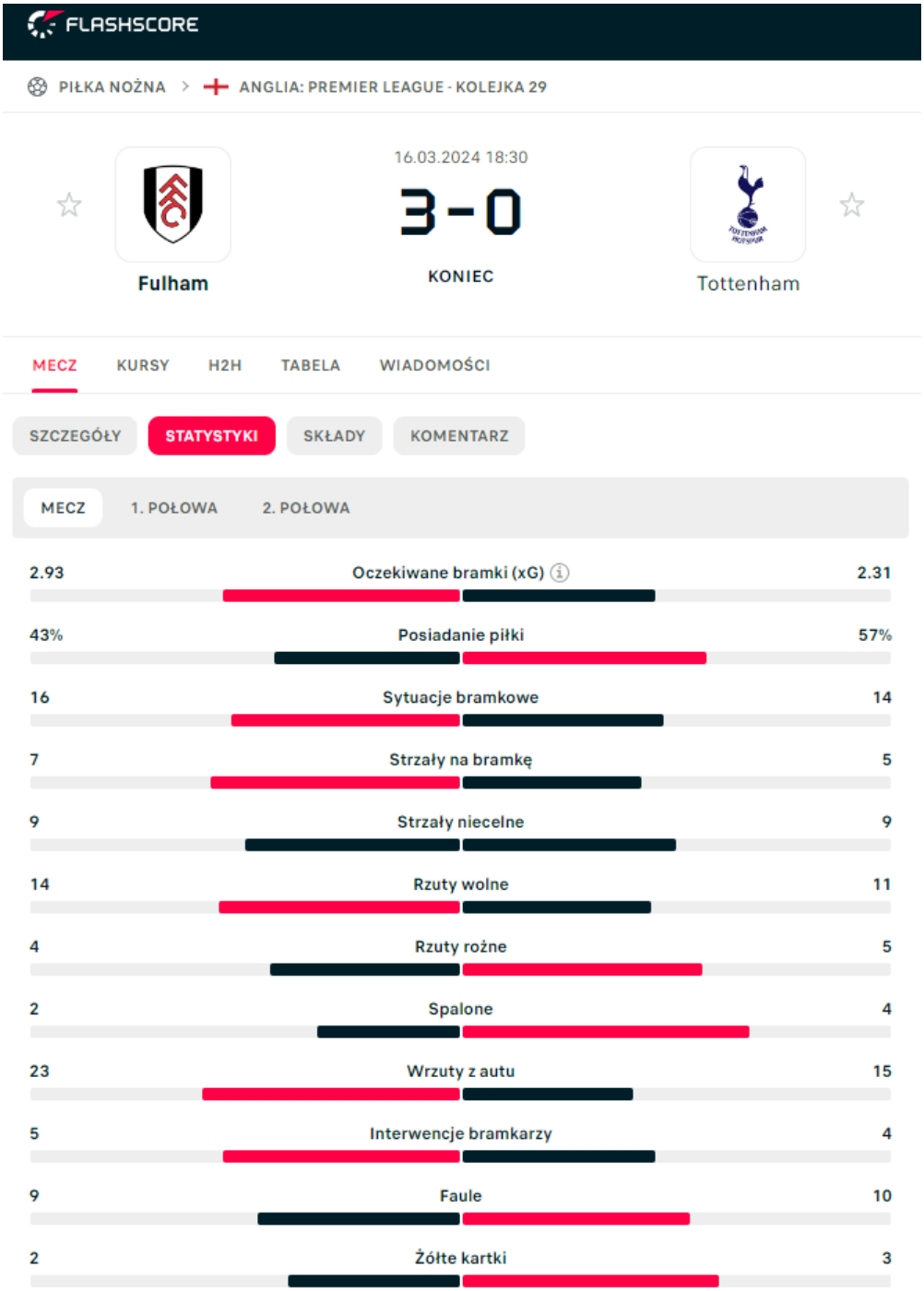
Do analizy zdarzeń w ramach niniejszej pracy wyselekcjonowano dwa najwyższe poziomy rozrgrywkowe dla następujących sześciu krajów:

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. 2.1: 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 2.2



Rys. 2.1: 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

2.2. 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

Poniższe sekcje dokonują szczegółowego opisu wyżej wymienionych encji.

2.2.1. 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. 2.2: 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.

2.2.2. 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. 2.3: Przykład reprezentacji krajów w tabeli Countries

2.2.3. 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 2.4). 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. 2.4: Przykład reprezentacji sezonów w tabeli Seasons

2.2.4. 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. 2.5: Przykład reprezentacji krajowych lig w tabeli Leagues

2.2.5. 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 [2]) 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. 2.6: 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. 2.2: 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

[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 3

Model

3.1. Analiza wyselekcjonowanych danych

3.2. System rankingowy

3.3. Metodologie nauczania

3.4. Przykłady działania

Rozdział 4

Szczegóły implementacyjne

Rozdział 5

Testy

5.1. Porównanie modelu z rzeczywistymi zdarzeniami

5.2. Porównanie modelu względem innych narzędzi

5.3. Porównanie modelu z bukmacherami

itd

Rozdział 6

TMP

Rozdział 7

Podsumowanie

7.1. Omówienie rezultatów badań

7.2. Możliwe rozszerzenia pracy

Literatura

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

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

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.