AKADEMIA GÓRNICZO-HUTNICZA IM. STANISŁAWA STASZICA W KRAKOWIE
Wydział Zarządzania

Samodzielna Pracownia Zastosowań Matematyki w Ekonomii

Praca Dyplomowa

Prognozowanie wyników mistrzostw Europy w piłce nożnej z wykorzystaniem lasów losowych i sieci neuronowych

Forecasting the results of the European football championship using random forests and neural networks

Autor: *Marcin Mika*

Kierunek studiów: *Informatyka i Ekonometria*

Opiekun pracy: *dr hab. Łukasz Lach, prof. uczelni*

Kraków, 2024

Spis treści

Wstęp

Tematem niniejszej pracy licencjackiej jest predykcja wyników mistrzostw Europy w piłce nożnej. Od zawsze byłem fanem tego sportu, a szczególnie spotkań drużyn narodowych. Ponadto, pasjonuje mnie rosnąca rola metod uczenia maszynowego w sporcie, więc ten temat idealnie łączy moje zainteresowania. Celem jest prognoza zwycięzcy turnieju odbywającego się w bieżącym roku za pomocą modelu lasów losowych i sztucznych sieci neuronowych oraz porównanie uzyskanych wyników. Las losowy to model zbudowany z wielu drzew decyzyjnych, na podstawie których obliczana jest ostateczna wartość wyjściowa, a sztuczna sieć neuronów składa się z warstw zawierających węzły, którym trzeba przypisać odpowiednią wagę ustalaną w trakcie procesu uczenia. Podczas pisania niniejszej pracy korzystałem zarówno ze źródeł naukowych jak i popularnonaukowych.

Praca składa się z trzech rozdziałów. W pierwszym skupiłem się na opisie historii prognozowanego zjawiska, przedstawieniu zastosowania analizy danych w sporcie, pokazaniu ogólnej definicji prognozowania oraz przeglądzie badań.

Drugi rozdział zawiera szczegółowe przedstawienie danych, opis sposobu podziału na zbiór treningowy i testowy oraz wytłumaczenie pojęć takich jak drzewo decyzyjne, las losowy i sztuczne sieci neuronowe

W ostatnim rozdziale niniejszej pracy przedstawiłem zasady przyznawania awansu drużynom panujące w badaniu, wyniki uzyskane metodą lasu losowego, prognozy otrzymane za pomocą modelu sztucznych sieci neuronowych oraz zaprezentowałem porównanie tych wyników z przewidywaniami OPTA.

1. Przewidywanie wyników wydarzeń sportowych: historia, znaczenie, techniki.

1.1. Historia turnieju mistrzostw Europy w piłce nożnej

Puchar Europy Narodów, bo tak pierwotnie nazywał się turniej starego kontynentu, po raz pierwszy odbył się w 1960 roku. Uczestnikami zawodów były reprezentacje Francji (gospodarz), Jugosławii, Czechosłowacji oraz ZSRR. W kwalifikacjach wzięło udział tylko 17 zespołów – obecnie więcej drużyn walczy w finałach turnieju o końcowy triumf. Zwycięzcą, a zarazem pierwszym mistrzem Europy, został ZSRR. Od tamtego momentu mistrzostwa Europy niezmiennie są rozgrywane co 4 lata. Wyjątkiem był turniej EURO 2020, a właściwie 2021. Wtedy, z powodu epidemii COVID-19 turniej przełożono na następny rok. Rozgrywki miały format pucharowy – dwa półfinały, mecz o trzecie miejsce oraz finał. Dopiero w 1968 roku turniej nazwano „Mistrzostwami Europy”[1].

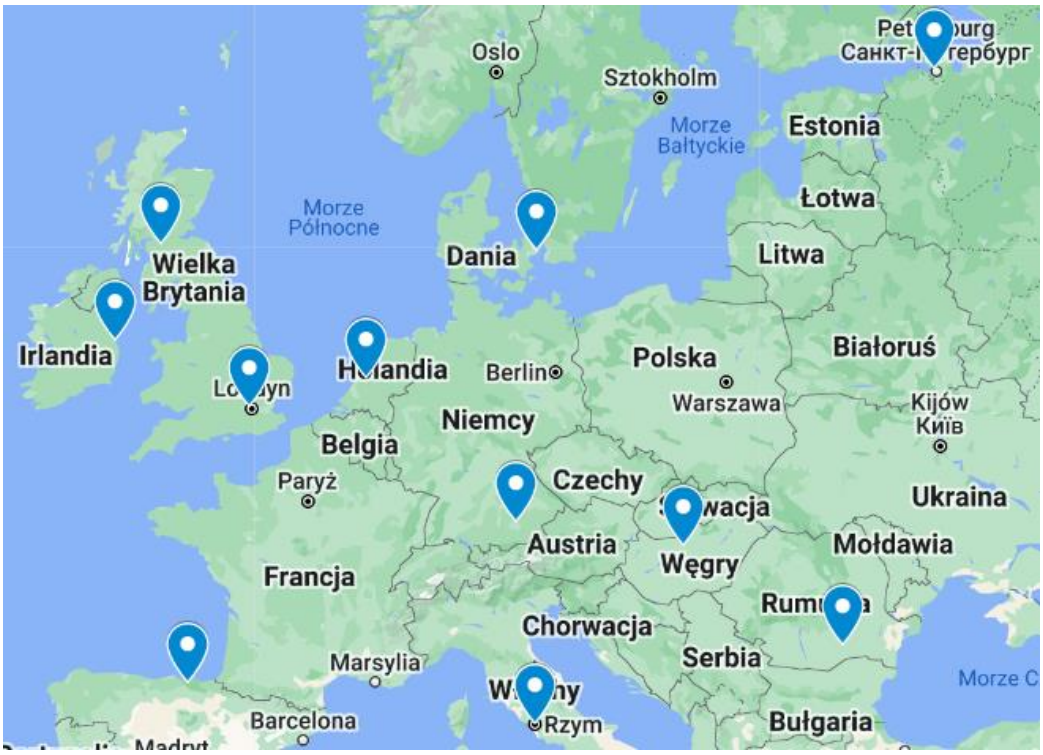
Zmiany w schemacie rozgrywania zawodów wprowadzono w 1980. Od tego czasu gospodarz ma zawsze zagwarantowany udział w turnieju. Wtedy również pierwszy raz w historii rozegrano fazę grupową turnieju, ponieważ brało w nim udział 8 zespołów podzielonych na dwie grupy. Zwycięzcy grup awansowali do finału, gdzie rywalizowali o złoty medal, natomiast przegrani wzięli udział w meczu o trzecie miejsce. Od 1984 roku powrócono do rozgrywania fazy pucharowej turnieju (półfinały i finał) a brązowe medale przyznawano obu przegranym półfinalistom.

Kolejną ważną zmianę wprowadzono w 1996 roku. Po raz kolejny podwojono liczbę uczestników, których umieszczono w czterech grupach zawierających po cztery zespoły. Do ćwierćfinałów awansowały dwa najlepsze zespoły z każdej grupy. Liczba drużyn biorących udział w eliminacjach do turnieju stale rosła, więc wydaje się, że ta reforma była konieczna. Turniej rozgrywany w Anglii okazał się wielkim sukcesem komercyjnym, co uznano za potwierdzenie słuszności wprowadzonych zmian[2].

[1] Adam Pisula, *Historia Mistrzostw Europy w piłce nożnej. Jak zmieniał się turniej na przestrzeni lat?,* https://www.sts.pl/blog/historia-pilkarskich-mistrzostw-europy/ [dostęp: 18.06.2024].
[2] Michał Kołkowski, *Kiedyś elita, dziś pospolite ruszenie. Jak się zmieniały mistrzostwa Europy?,* https://weszlo.com/2019/06/07/kiedys-elita-dzis-pospolite-ruszenie-sie-zmienialy-mistrzostwa-europy/ [dostęp: 18.06.2024].

W 2016 roku ponownie powiększono liczbę reprezentacji biorących udział w finałach mistrzostw starego kontynentu. Do turnieju zakwalifikowały się aż 24 drużyny. Wprowadzone korekty musiały również wpłynąć na przeprowadzenie kwalifikacji do rozgrywek. Po raz pierwszy w historii szanse na udział miały zespoły zajmujące trzecie miejsce w swojej grupie kwalifikacyjnej. Awans do fazy pucharowej finałów również stał się prostszy, ponieważ poza dwiema najlepszymi drużynami w grupie, w turnieju zostawały niektóre zespoły z trzecich miejsc. Dzięki temu mistrzem Europy została reprezentacja Portugalii, która na turnieju we Francji (bo tam rozgrywane były ówczesne finały) wyszła z grupy remisując wszystkie spotkania[3].

Turniej rozgrywany w 2021 roku (chociaż prezentowany pod nazwą EURO 2020) był dość kontrowersyjny. Poprzednie zawody za każdym razem były rozgrywane w jednym lub w dwóch sąsiednich krajach. Tym razem UEFA postanowiła wybrać wielu organizatorów, co wywołało krytykę ze strony fanów, narzekających na brak atmosfery charakterystycznej dla dużej imprezy. Tak wygląda mapa miejsc, w których rozgrywane były mecze w ramach mistrzostw starego kontynentu:



Rysunek 1. Mapa miejsc, w których rozgrywane były mecze w ramach mistrzostw starego kontynentu w roku 2021.
Źródło: https://www.stadiumguide.com/tournaments/uefa-euro-2020/, [dostęp: 18.06.2024].

[3] Ibidem

Zwycięzcy mistrzostw Europy w piłce nożnej prezentują się następująco:

Tabela 1. Klasyfikacja zwycięzców i gospodarzy Mistrzostw Europy

L.p	Rok	Mistrz	Gospodarz
1	1960	ZSRR	Francja
2	1964	Hiszpania	Hiszpania
3	1968	Włochy	Włochy
4	1972	RFN	Belgia
5	1976	Czechosłowacja	Jugosławia
6	1980	RFN	Włochy
7	1984	Francja	Francja
8	1988	Holandia	RFN
9	1992	Dania	Szwecja
10	1996	Niemcy	Anglia
11	2000	Francja	Belgia/Holandia
12	2004	Grecja	Portugalia
13	2008	Hiszpania	Austria/Szwajcaria
14	2012	Hiszpania	Polska/Ukraina
15	2016	Portugalia	Francja
16	2021	Włochy	Europa

Źródło: https://kronika-futbolu.pl/historia/mistrzostwa-europy/ [dostęp: 18.06.2024].

1.2. Analiza danych w sporcie

Z początkiem XXI wieku matematyka zaczęła być coraz częściej wykorzystywana do poprawy wyników zespołów. Szeroko pojęta analiza danych, połączona z prognozowaniem, stała się użytecznym narzędziem w wielu dyscyplinach sportowych – od baseballu po Formułę 1. Oto kilka przykładów, w których statystyka, analiza danych, uczenie maszynowe bądź sztuczna inteligencja poprawiła osiągi drużyn lub rozwinęła rozgrywki:

Baseball: Oakland Athletics

Drużyna Oakland Athetics w 2002 straciła swoje największe gwiazdy, a trener Billy Bean stanął przed wyzwaniem odbudowy zespołu, dysponując przy tym niewielkim budżetem. Wraz ze swoim pomocnikiem ekonomistą (Paul DePodesta) zdecydowali się na ryzykowną taktykę, która polegała na postawieniu na zawodników średnio pożądanych na rynku, ale pasujących do badań opartych na statystycznych analizach, do tej pory ignorowanych przez działaczy drużyn baseballowych. Skutkiem postawienia na jedną kartę było pokonanie przez drużynę historycznego rekordu 19 zwycięstw z rzędu[4]. Na podstawie tej historii nakręcono film o nazwie „Moneyball", który zyskał ogromną popularność.

F1: Mercedes-AMG PETRONAS F1

Obecnie w sportach wyścigowych tysięczne części sekundy decydują o zwycięstwie, więc zespoły walczą na różnych frontach o to, aby wyprzedzić konkurencję. W 2010 roku zespół Mercedes – AMG Petronas powrócił jako konstruktor. Przez następne 10 lat ilość danych dostępnych do analizy znacząco wzrosła i – jak twierdził Geoff Willis, dyrektor ds. inżynierii komercyjnej w Mercedes F1 – skupienie się na danych i traktowanie ich jako podstawowy element kultury firmy przyniosło zrozumienie osiąganych wyników, co skutkowało poprawą osiągów. Optymalizacja samochodu oparta na analizie danych przyczyniła się do znaczącej przewagi nad pozostałymi zespołami, co przyniosło wygraną zespołu.

[4] P. Sawicki, *Bill James, John Henry i excel – jak analityka weszła do świata sportu,* https://newonce.net/artykul/bill-james-john-henry-i-excel-jak-analityka-weszla-do-swiata-sportu, [dostęp: 20.06.2024]

W 2013 roku Formuła 1 wkroczyła w nową erę silników hybrydowych. Zespół mercedesa zaadaptował się najlepiej i najszybciej, czego dowodem są wielokrotne zwycięstwa w Mistrzostwach Konstruktorów[5].

Tenis: IBM AI Draw Analysis

W 2023 firma IBM oraz The All England Lawn Tennis Club ogłosili wprowadzenie nowych narzędzi opartych na działaniu sztucznej inteligencji, mających za zadanie zwiększyć atrakcyjność wielkoszlemowego tenisowego turnieju o nazwie Wimbledon. Jedną z nowości był komentarz generowany przez AI mający na celu wzbogacić ten standardowy, udzielany przez ekspertów. Dodatkowo wprowadzono IBM AI Draw Analysis, która oceniała otrzymaną przez danego zawodnika drogę do finału oraz wybierała jak mogłoby wyglądać najkorzystniejsze losowanie. Analiza ta była oparta na wielu czynnikach takich jak np. dyspozycja zawodnika czy historia starć. Taki system pozwala na odkrycie niespodzianek w turnieju. Do wprowadzonych nowości działających na aplikacji bądź stronie turnieju należały również IBM Power Index Leaderboard (indeks mocy danego zawodnika), IBM Match Insights (analiza przedmeczowa) oraz spersonalizowane Highlights Reels. Nowe narzędzia korzystające z mocy sztucznej inteligencji zdecydowanie rozwinęły rozgrywkę i wprowadziły ją na wyższy poziom[6].

Piłka nożna: Wisła Kraków

Obecny prezes polskiego zespołu piłkarskiego Wisły Kraków, Jarosław Królewski wierzy, że dzięki modelom sztucznej inteligencji można wybrać idealnego trenera, który poprowadzi klub do powrotu do najwyższej klasy rozgrywkowej w Polsce. Na początku bieżącego roku zatrudniono Alberta Rude, który podczas swojej niemal 3 letniej kariery trenerskiej nigdy wcześniej nie pracował w Europie. Początkowo stworzono bazę posiadającą 3 tys. szkoleniowców, którym nadano statystyki opisujące ich styl. Dzięki temu ograniczono wybór najpierw do stu, a następnie do dziesięciu trenerów[7].

[5] A. Philips, *How Data Analytics Emerged as a Competitive Advantage for the Mercedes-AMG Petronas Formula One Team*, https://www.tibco.com/blog/2020/08/27/how-data-analytics-emerged-as-a-competitive-advantage-for-the-mercedes-amg-petronas-formula-one-team/, [dostęp: 20.06.2024].
[6] M. Wójcik, IBM AI Draw Analysis – *IBM wprowadza generatywną sztuczną inteligencję do Wimbledonu,* https://technet-media.pl/artykuly/ibm-ai-draw-analysis-ibm-wprowadza-generatywna-sztuczna-inteligencje-do-wimbledonu [dostęp: 20.06.2024].
[7] J. Kubielas, *Wisła Kraków ma nowego trenera. Wybrała go sztuczna inteligencja,* https://sport.rp.pl/pilka-nozna/art39683621-wisla-krakow-ma-nowego-trenera-wybrala-go-sztuczna-inteligencja [dostęp: 20.06.2024]

Drużyna, którą prowadził, odniosła wielki sukces, wygrywając Puchar Polski, jednak nie udało się jej osiągnąć głównego celu, jakim był awans do Ekstraklasy. Hiszpan odszedł po zakończeniu sezonu, a prezes wybrał kolejnego szkoleniowca za pomocą modelu. Tym razem wybór padł na Kazimierza Moskala. Mimo wielu głosów krytyki, Jarosław Królewski zapowiada dalszy rozwój działu Data Science.

1.3. Prognozowanie

Prognozowanie polega na przewidywaniu przyszłych wydarzeń, zmian. Prowadzone prognozy dotyczą wielu aspektów. W obecnych czasach staramy się przewidzieć wszystko co możliwe, od prognozy meteorologicznej po gospodarczą i wiele innych.[8] Proces prognozowania polega na wnioskowaniu ze znanego o nieznanym. Mając pewien stan wiedzy o jakimś zjawisku chcemy je przewidzieć. Może dotyczyć zarówno wnioskowania na danych czasowych lub czasowo-przekrojowych[9].

„Prognoza jest naukowo uzasadnionym sądem o stanie zjawiska w określonym momencie (okresie) należącym do przyszłości. Użycie w powyższym określeniu słowa „sąd" sygnalizuje niepewność prognozy, a odwołanie się do nauki oznacza, że prognoza musi być racjonalnym wnioskowaniem, prowadzącym od przesłanek do wniosków odnoszących się do przyszłości"[10].

Aby prognozowanie za pomocą modelu ekonometrycznego było możliwe, muszą być spełnione podstawowe założenia ekonometrycznego wnioskowania w przyszłość. Jedno z nich oznajmia, że szacowany model ekonometryczny, opisujący kształtowanie się badanej zmiennej prognozowanej, powinien wykazywać wysoki stopień dopasowania do historycznych danych empirycznych oraz posiadać statystycznie istotne parametry (sprawdzone np. przez test t-studenta dla modeli liniowych). Najlepiej, aby model w przeszłości postawił trafne prognozy[11].

[8] J. Bartman, K. Bajda, *Wykorzystanie sztucznych sieci neuronowych do prognozowania wyników meczów piłkarskich*, Journal of Education, Technology and Computer Science, 2014, 10(2), s. 425–431.
[9] J. B. Gajda, *Prognozowanie i symulacje w ekonomii i zarządzaniu,* C.H.Beck, 2017.
[10] M. Cieślak, P. Dittmann, A. Kania–Gospodarowicz, I. Kuropka, S. Ostasiewicz, B. Radzikowska, *Demografia, metody analizy i prognozowania*, PWN, 1992.
[11] http://www.prognozowanie.info/prognozowanie-ekonometryczne/ [dostęp: 24.08.2024]

Prognozę można podzielić biorąc pod uwagę kilka kryteriów:

- Ze względu na charakter zjawiska rozróżniamy prognozę:
    - Ilościową (gdy stan zjawiska jest wyrażony liczbą)
    - Jakościową (gdy stan zjawiska nie może być zmierzony)
    - Prostą (gdy prognoza dotyczy jednej zmiennej ekonomicznej)
    - Złożoną (gdy podczas prognozowania korzystamy z innych zmiennych prognozowanych)
    - Jednorazową (prognoza stawiana jednokrotnie)
- Ze względu na okres prognozowania rozróżniamy prognozę:
    - Krótkookresową
    - Średniookresową
    - Długookresową
- Ze względu na cel predykcji rozróżniamy prognozę:
    - Badawczą (przyszłość rozpoznawana wszechstronnie)
    - Ostrzegawczą (polega na zwracaniu szczególnej uwagi na sygnały wpływające negatywnie na kształtowanie się zjawisk)
    - Normatywną (prognozy dotyczące norm obowiązujących w przyszłości)
    - Pasywną (zniechęcają do podejmowania określonych działań)
    - Aktywną (pobudzają do podejmowania określonych działań)
- Ze względu na zasięg rozróżniamy prognozę:
    - Całościową (prognoza globalna)
    - Częściową (odcinkowa – dotyczy pewnego aspektu zjawiska)
    - Makroekonomiczną (dotyczy całej gospodarki narodowej lub regionów kraju w skali mikro)
    - Mikroekonomiczną (dotyczy pojedynczej jednostki gospodarczej)
- Ze względu na rodzaj prognozowanego zjawiska rozróżniamy m.in. prognozę:
    - Gospodarczą
    - Społeczną
    - Klimatyczną

Jak wynika z wyżej przedstawionych podziałów, prognozy mogą różnić się miedzy sobą wieloma czynnikami[12].

[12] A. Tatarczak, *Ekonometria – Podręcznik. Studia przypadków,* tom 20, WSEI, 2021 s. 78 – 83.

1.4. Przegląd badań

Prognozowanie wyników sportowych wydarzeń w artykułach naukowych staje się coraz częstszym, ciekawszym i bardziej rozbudowanym tematem. W niniejszym rozdziale przedstawię kilka takich rozwiązań.

Prognozowanie wyników piłkarskiego mundialu – dwa różne podejścia

Sieci Neuronowe

W badaniu A. Hassana, A. R. Akla, I. Hassana oraz C. Sunderland[13] do prognozowania piłkarskich mistrzostw świata w 2018 roku zastosowano sieci neuronowe. W przybliżonym badaniu model prognozuje zwycięstwo lub porażkę w danym spotkaniu. Przewidziano dobrze wyniki meczów w 83,3% w przypadku wygranej i 72,7% w przypadku przegranej badanego zespołu. Dodatkowo model uwzględnił ważność badanych atrybutów. Okazało się, że najbardziej wpływowe atrybuty meczowe to: liczba prób podań o średniej długości, odległość pokonana w strefie 3, liczba podań do strefy ataku, odległość pokonana bez posiadania piłki oraz całkowita odległość pokonana przez drużynę podczas meczu.

Metoda kombinacyjna

Natomiast w swoim badaniu A. Groll, C. Ley, G. Schauberger i H. Van Eetvelde[14] przedstawiają inne podejście, zdecydowanie bliższe niniejszej pracy. Modele regresji Poissona, lasów losowych oraz metod rankingowych przewidują liczbę bramek zdobytych w danym meczu – co kształtuje cały turniej mistrzostw świata w piłce nożnej w 2018 roku. Okazało się, że połączenie lasów losowych z parametrami zdolności zespołów z metod rankingowych poprawia zdolności predykcyjne modelu. Podczas badania przewidziano, że Hiszpania miała największe szanse na wygraną, zaraz przed broniącymi tytułu Niemcami. Prognoza ta nie okazała się trafna, ponieważ turniej zwyciężyli Francuzi wygrywając w finale z Chorwacją, a Hiszpania odpadła w 1/8 z reprezentacją Rosji, będącą gospodarzem turnieju. Niemcy zakończyli udział w rozgrywkach już na fazie grupowej.

[13] A. Hassan, A. R. Akl, I. Hassan, C. Sunderland, *Predicting Wins, Losses and Attributes' Sensitivities in the Soccer World Cup 2018 Using Neural Network Analysis,* Sensors, 2020.
[14] A. Groll, C. Ley, G, Schauberger, H. Van Eetvelde *Prediction of the FIFA World Cup 2018 - A random forest approach with an emphasis on estimated team ability parameters*, Statistical Modelling, 2018.

Prognozowanie wyników spotkań w tenisie ziemnym

W artykule autorstwa P. Sroki oraz J. Trzęsiok[15] celem było opracowanie takiego modelu, który będzie miał większą skuteczność w przewidywaniu zwycięzcy tenisowego meczu od modeli firm bukmacherskich. Skonstruowano dwa ostateczne modele, jeden dla meczy WTA (turnieje damskie) oraz drugi dla meczy ATP (turnieje męskie). Zastosowano metodykę lasów losowych. Zbiór danych zawierał wybrane charakterystyki opisujące mecze z turniejów tenisowych rozegranych w 2015 r. Według autorów cel badania został spełniony, ponieważ dodatnia zdolność predykcyjna wzrosła o 5 punktów procentowych (zarówno dla modelu ATP jak i WTA) względem modeli rynkowych. Określa ona skuteczność predykcji meczów, w których wygrywa faworyzowany zawodnik.

Badanie wyników wybranych konkurencji lekkoatletycznych

W artykule autorstwa P. Cięszczyka i J. Eidera pochylono się nad alternatywnymi sposobami badania wyników sportowych w wybranych konkurencjach atletycznych. Dane pochodziły z igrzysk olimpijskich w 2000 roku. Porównano dwie metody – sztucznych sieci neuronowych oraz regresji liniowej. Okazało się, że sieci neuronowe dały bliższe rezultaty do tych rzeczywistych (w skrajnych przypadkach różnice dochodziły do 500%). Dodatkowo wyniki uzyskiwane za pomocą regresji liniowej w każdej badanej dyscyplinie były bardziej optymistyczne od rzeczywistych rezultatów[16].

[15] P. Sroka, J. Trzęsiok, *Co opowiadają drzewa o tenisie? Predykcja wyników spotkań w tenisie ziemnym z wykorzystaniem drzew klasyfikacyjnych*, Uniwersytet Ekonomiczny we Wrocławiu, 2017.
[16] P. Cięszczyk, J. Eider, *Alternatywne metody badania wyników sportowych w wybranych konkurencjach lekkoatletycznych,* ISSN, 2003.

2. Charakterystyka zbiorów danych i opis wybranych metod.

2.1. Przedstawienie danych

Jak wskazują A. Groll, C. Ley, G. Schauberger i H. Van Eetvelde [17] do prognozowania wyników piłkarskiego turnieju mogą być przydatne dane dotyczące wielu czynników, takich jak ekonomiczny, sportowy, trenerski, drużynowy, lokalizacji. Do każdego z nich autorzy przyporządkowali zmienne, które opisują poszczególne czynniki.

Na tej podstawie do przeprowadzenia własnej analizy wybrałem zmienne, które prezentują się następująco:

Czynnik ekonomiczny:

PKB per capita w cenach stałych (dolary z 2015 r.)[18]

Czynnik sportowy:

Ranking FIFA [19]

Czynnik trenerski:

Informacja, czy trener reprezentacji pochodzi z tego samego kraju[20]

Czynniki drużynowe:

Średnia wartość piłkarza[21]

Średni wiek piłkarza[22]

Liczba zawodników grających w klubach za granicą[23]

Liczba zawodników grających w 10 najlepszych klubach piłkarskich na świecie[24]

[17] A. Groll, C. Ley, G, Schauberger, H. Van Eetvelde, op. cit, 2018, s. 5-6.
[18] World Bank, "World Development Indicators.", https://databank.worldbank.org/source/world-development-indicators [dostęp: 20.04.2024].
[19] "*FIFA World Ranking 1992-2024"*, https://www.kaggle.com/datasets/cashncarry/fifaworldranking [dostęp: 20.04.2024].
[20] https://en.wikipedia.org/wiki/Main_Page [dostęp: 20.04.2024].
[21] https://www.transfermarkt.com [dostęp: 20.04.2024].
[22] Ibidem
[23] https://en.wikipedia.org/wiki/Main_Page [dostęp: 20.04.2024].
[24] Ibidem

Czynnik lokalizacji:

Informacja, czy reprezentacja rozegrała spotkanie w swoim państwie[25]

Tabela 2. Przykładowe wartości zmiennych wykorzystywanych w analizie

Zespół	PKB	Ranking	Trener z kraju	Średnia wartość	Średni Wiek	Gracze za granicą	Top 10 klubów	Gospodarz
Niemcy	43361,18	16	1	29,22	28,2	7	8	1
Szkocja	47923,48	39	1	9,48	27,5	17	1	0

Źródło: opracowanie własne

W badaniu zmienną objaśnianą będzie liczba bramek strzelonych przez daną drużynę w meczu, a zmiennymi objaśniającymi będą różnice zmiennych wymienionych w Tabeli 2 (w przypadku danych liczbowych) i odpowiednie wartości (w przypadku danych faktorowych).

Tabela 3. Wartości zmiennych dotyczące meczu Niemcy – Szkocja.

Zespół	PKB	Ranking	Trener z kraju	Średnia wartość	Średni Wiek	Gracze za granicą	Top 10 klubów	Gospodarz
Niemcy	-4562,31	-23	1	19,740	0,7	-10	7	1
Szkocja	4562,31	23	1	-19,740	-0,7	10	-7	0

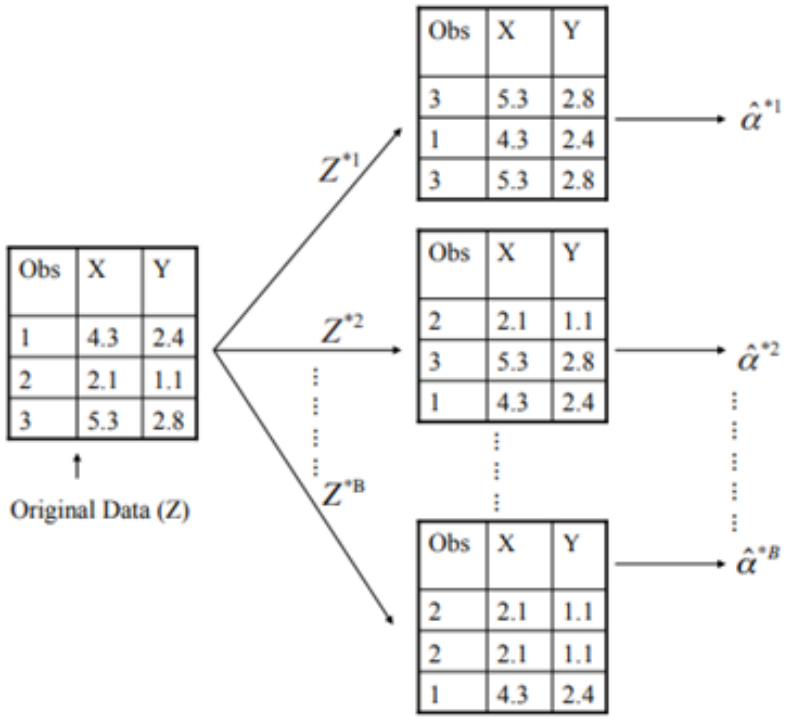
Źródło: opracowanie własne

W Tabeli 3 przedstawiono zmienne objaśniające potrzebne do prognozy liczby zdobytych bramek kolejno dla reprezentacji Niemiec i Szkocji w meczu pomiędzy tymi zespołami. Taki zbiór danych, znajdujący się w Załączniku 2, liczy 462 obserwacje, z czego dla 72 należy prognozować liczbę bramek. Pozostałe 390 obserwacji pochodzi z poprzednich turniejów mistrzostw Europy rozgrywanych w XXI wieku. Dla lasu losowego standaryzacja zmiennych nie jest niezbędna, jednak dla sieci neuronowych i lepszych wyników standaryzuję zmienne liczbowe.

[25] M. Jurisoo, *International football results from 1872 to 2024,* *https*://www.kaggle.com/datasets/martj42/international-football-results-from-1872-to-2017 [dostęp: 20.04.2024].

2.2. Podział na zbiór treningowy i testowy za pomocą metody bootstrap

Zdecydowałem się zastosować metodę bootstrap polegającą na wylosowaniu obserwacji z powtarzaniem, ponieważ zbiór danych jest stosunkowo niewielki. Poniżej przedstawiam schemat takiego losowania.



Rysunek 2. Bootstrap
Źródło: G. James, D. Witten, T. Hastie, R, Tibshirani, *An Introduction to Statistical Learning,* 2021, s. 212.

Na rysunku znajdują się trzy przykładowe zbiory uzyskane metodą losowania z powtarzaniem, które zostały wykorzystane do oszacowania α (estymator obliczony na podstawie danej bootstrapowej próbki danych, których ilość określona jest za pomocą B). Wszystkie niewylosowane obserwacje stanowią zbiór testowy – tak zwane obserwacje out-of-bag (OOB)[26]. Badania wykazują, że po takim losowaniu z powtarzaniem, zbiór OOB liczy około 1/3 początkowych danych[27]. Dzięki tej technice zbiór treningowy składa się z 390 obserwacji (tak jak domyślnie), a zbiór testowy – składający się z obserwacji niedostępnych dla modelu podczas trenowania – liczy 150 obserwacji.

[26] I. Langmore, D. Krasner, "Applied Data Science", 2016, s 102.
[27] B. Efron, R. Tibshirani, *An Introduction to the Bootstrap*, tom. 57, Chapman & Hall, 1993.

2.3. Drzewa decyzyjne

Drzewo decyzyjne to struktura używana do podejmowania decyzji i klasyfikacji. Tworząc graficzny schemat uzyskujemy obraz podobny do korony drzewa. W każdym miejscu podziału drzewa, uzyskujemy odpowiedź „tak" lub „nie" na zadane pytanie. Istnieją dwa rodzaje drzew – klasyfikacyjne i regresyjne. W drzewie pierwszego typu na końcu widnieje odpowiedź, która wskazuje na dopasowanie do odpowiedniej grupy. W badaniu K. Y. Huang i K. J. Chen[28] przewidywane było tylko ostateczne rozstrzygnięcie spotkania (zwycięstwo, remis lub porażka). Ja natomiast chcę przewidzieć liczbę bramek każdej drużyny w jednym meczu. Do tego przyda mi się drzewo regresyjne, które zwraca na końcu wynik liczbowy. Ważną zaletą drzew jest przejrzystość, łatwość interpretacji oraz możliwość korzystania ze zmiennych numerycznych i kategorycznych. Z drugiej strony proces tworzenia drzewa decyzyjnego jest bardzo złożony oraz istnieje duża szansa na zbyt duże dopasowanie modelu do treningowego zbioru danych[29]. Logikę drzewa można przedstawić w następujący sposób (dane po standaryzacji):

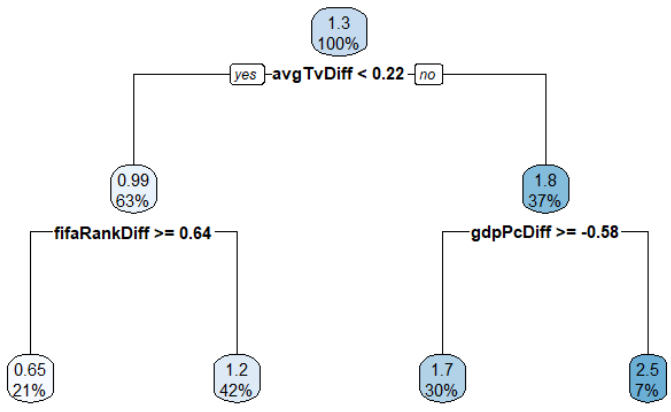
IF (różnica w średniej wartości zawodnika drużyn jest mniejsza niż 0.22)

AND (różnica w rankingu FIFA drużyn jest większa lub równa niż 0.64)

THEN (średnia liczba bramek strzelonych wynosi 0.65)

ELSE IF…

Tak prezentuje się wygląd graficzny drzewa przygotowanego na potrzeby niniejszej pracy:



Rysunek 3. Przykładowe drzewo decyzyjne
Źródło: Opracowanie własne

[28] K. Y. Huang, K. J. Chen, *Multilayer Perceptron for Prediction of 2006 World Cup Football Game,* Journal of Geomatics Science and Technology, 2006.
[29] J. Grus, *Data science od podstaw. Analiza danych w Pythonie. Wydanie II,* Helion, 2018, s.195 – 197.

Przedstawione przeze mnie drzewo ma głębokość wynoszącą tylko 2 poziomy. Poprzez podział danych na podzbiory dla każdej zmiennej objaśniającej i każdej obserwacji wybierany jest najlepszy możliwy schemat drzewa, w którym zmienna prognozowana osiągnie minimalny poziom zróżnicowania. Jest on mierzony za pomocą funkcji straty. W moim badaniu występuje drzewo regresyjne, dlatego zastosowałem funkcję mierzącą średni błąd kwadratowy:

$$Q(R_k) = \frac{1}{N(k)} \sum_{x_i \in R_k} (y_i - \alpha_k)^2 \qquad (1)$$

We wzorze (1):

- R_k to k-ty podzbiór danych,
- $Q(R_k)$ to wartość funkcji kosztu dla podzbioru R_k,
- $N(k)$ to liczba obserwacji w podzbiorze R_k,
- y_i to wartość rzeczywista dla obserwacji x_i
- α_k to przewidywana wartość dla podzbioru R_k.

Gdyby zmienna zależna była zmienną faktorową należałoby zastosować jedną z następujących funkcji:

- Błąd klasyfikacji: sprawdza procent niepoprawnie sklasyfikowanych obserwacji.
- Wskaźnik Giniego: jest używany do oceny jakości podziałów w drzewach decyzyjnych
- Entropia: mierzy różnorodność w zbiorze.

Aby znaleźć podział optymalny lokalnie, stosowany jest algorytm wspinaczki górskiej. Polega on na generowaniu nowych, sąsiednich drzew zmieniając przy tym bieżący model i porównując wynik funkcji kosztu[30].

Podstawowe elementy drzewa to liście, węzły i gałęzie. Węzły reprezentują punkt, w którym dochodzi do podziału. Liście są końcowymi węzłami. Natomiast gałęzie to ścieżki łączące liście z węzłami. Ważne jest dostosowanie parametrów takich jak minimalna liczba obserwacji w węźle (jeżeli danych jest za mało nie dokonujemy następnego podziału i staje się on liściem), minimalna możliwa liczba obserwacji w liściu aby mógł istnieć, głębokość drzewa[31].

[30] M. Walesiak, E. Gatnar, *Statystyczna analiza danych z wykorzystaniem programu R,* PWN, 2012, r.8 (Gatnar) s 238-240.
[31] J. Dean, *Big Data, Data Mining, and Machine Learning,* Wiley, 2014, s. 101-103.

W zaprezentowanym wyżej drzewie występują następujące elementy:

- Węzły:
    - Węzeł korzenny

      Warunek: różnica w średniej wartości rynkowej piłkarza jest mniejsza od 0.22.

      Informacja: średnia liczba bramek zdobytych w całym zbiorze wynosi 1.3.
    - Węzeł lewy

      Warunek: różnica w pozycji w rankingu jest większa lub równa 0.64.

      Informacja: średnia liczba bramek zdobytych w zbiorze spełniającym warunek z węzła korzennego wynosi 0.99. Zbiór ten stanowi 63% wszystkich obserwacji.
    - Węzeł prawy

      Warunek: różnica w wartości współczynnika PKB per capita w cenach stałych (dolary 2015) jest większa lub równa -0.58.

      Informacja: średnia liczba bramek zdobytych w zbiorze nie spełniającym warunku z węzła korzennego wynosi 1.8. Zbiór ten stanowi 37% wszystkich obserwacji.
- Liście (od lewej):
    - Liść nr 1: Średnia liczba bramek zdobytych przez drużynę w zbiorze spełniającym warunki z węzła korzennego i lewego wynosi 0.65. Ten zbiór stanowi 21% wszystkich obserwacji.
    - Liść nr 2: Średnia liczba bramek zdobytych przez drużynę w zbiorze spełniającym warunek spełniający w węźle korzennym i nie spełniającym restrykcji w węźle lewym wynosi 1.2. Ten zbiór stanowi 42% wszystkich obserwacji
    - Liść nr 3: Średnia liczba bramek zdobytych przez drużynę w zbiorze niespełniającym warunku z węzła korzennego i spełniającym restrykcje z węzła prawego wynosi 1.7. Ten zbiór stanowi 30% wszystkich obserwacji.
    - Liść nr 4: Średnia liczba bramek zdobytych przez drużynę w zbiorze nie spełniającym warunków z węzła korzennego i prawego wynosi 2.5. Ten zbiór stanowi 7% wszystkich obserwacji.

- Gałęzie:

  Łączą węzły i liście, w zależności od spełnionego warunku. Lewa gałęź zawsze jest utożsamiana z pozytywną odpowiedzią, a prawa z negatywną.

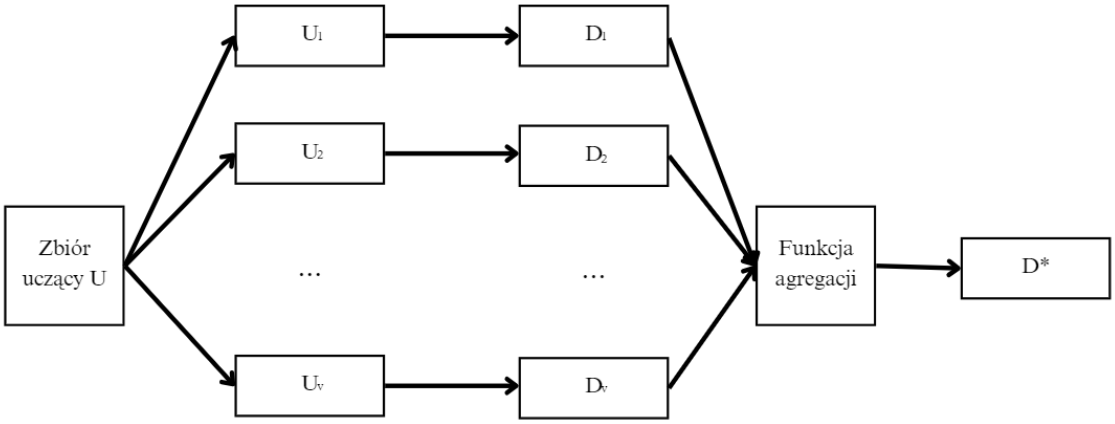
W podanym przykładzie średni błąd kwadratowy wynosi w przybliżeniu 1, co nie jest zadowalającym wynikiem, ponieważ w wielu spotkaniach jedna bramka dla danej drużyny może zmienić rozstrzygnięcie całego meczu. Drzewa decyzyjne są skłonne do wysokiej wariancji co prowadzi do zbyt dużego dopasowania modelu na danych treningowych. Pytania, na które można odpowiedzieć tylko „tak" lub „nie" nie gwarantują najdokładniejszej predykcji[32].

2.4. Las losowy

Las losowy składa się z drzew decyzyjnych i wykorzystuje metodę modyfikowania danych treningowych nazwaną agregacją bootstrapową, której główną zasadę opisałem podczas podziału danych na zbiór treningowy i testowy. Powtarzana jest ona v razy aby uzyskać v zbiorów do treningu a dzięki temu v drzew decyzyjnych wchodzących w skład lasu. Losowane są nie tylko obserwacje, ale również zmienne, które będą brane pod uwagę podczas tworzenia jednego drzewa. Model złożony skonstruowany w ten sposób jest dokładniejszy niż pojedynczy model prosty. Dzieje się tak, ponieważ predykcja otrzymana przy zastosowaniu lasu losowego jest uzyskiwana przez zastosowanie funkcji agregacji. Może być nią policzenie średniej z wyników wszystkich drzew (w przypadku problemu regresji) lub odrzucenie wyników o mniejszym prawdopodobieństwie przez głosowanie (w przypadku problemu klasyfikacji). Dzięki temu zmniejszone jest ryzyko zbyt dużego dopasowania modelu do danych treningowych oraz zwiększona jest jego stabilność. Las, analogicznie jak drzewo decyzyjne, może być zastosowany dla zmiennych liczbowych oraz faktorowych. Wadą lasu z pewnością jest problem w wizualizacji modelu, w przeciwieństwie do drzewa decyzyjnego, którego wygląd zaprezentowałem w poprzednim podrozdziale[33].

[32] O. Theobald, *Machine Learning For Absolute Beginners,* Scatterplot Press, 2017.
[33] M. Szeliga, *Data Science i uczenie maszynowe,* Helion, 2017, s. 126-128.

Tak prezentuje się uproszczony schemat agregacji bootstrapowej:



Rysunek 4. Schemat agregacji bootstrapowej podczas stosowania lasu losowego
Źródło: M. Walesiak, E. Getnar *Statystyczna analiza danych z wykorzystaniem programu R,* r. 2012, rozd. 9 str. 262

Dokładność lasu losowego zależy między innymi od hiperparametrów takich jak:

- Liczby drzew w modelu złożonym (za mała wartość może spowodować, że każda obserwacja powtórzy się kilkukrotnie w budowie modelu)
  Nazwa w funkcji randomForest programu R: ntree
- Liczby zmiennych losowo wybieranych jako zmienne potencjalnie użyte do budowy drzewa (w badaniu problemu klasyfikacji parametr ten powinien mieć wartość równą pierwiastkowi liczby zmiennych a w badaniu problemu regresji równą liczbie zmiennych podzieloną przez 3)
  Nazwa w funkcji randomForest programu R: mtry[34]
- Minimalna liczba obserwacji w liściu (większa liczba powoduje, że będą rosły mniejsze drzewa. Spowoduje to szybsze działanie algorytmu, ale zmniejszy dokładność)
  Nazwa w funkcji randomForest programu R: nodesize

Na podstawie eksperymentu przyjąłem następujące wartości dla wyżej wymienionych parametrów aby wybrać model z najmniejszą wartością średniego błędu kwadratowego:

- ntree: 200, 500, 1000
- mtry: 2, 3, 6
- nodesize: 5, 20, 100

[34] M. Walesiak, E. Getnar *Statystyczna analiza danych z wykorzystaniem programu R*, PWN, 2012, r. 9 str. 270.

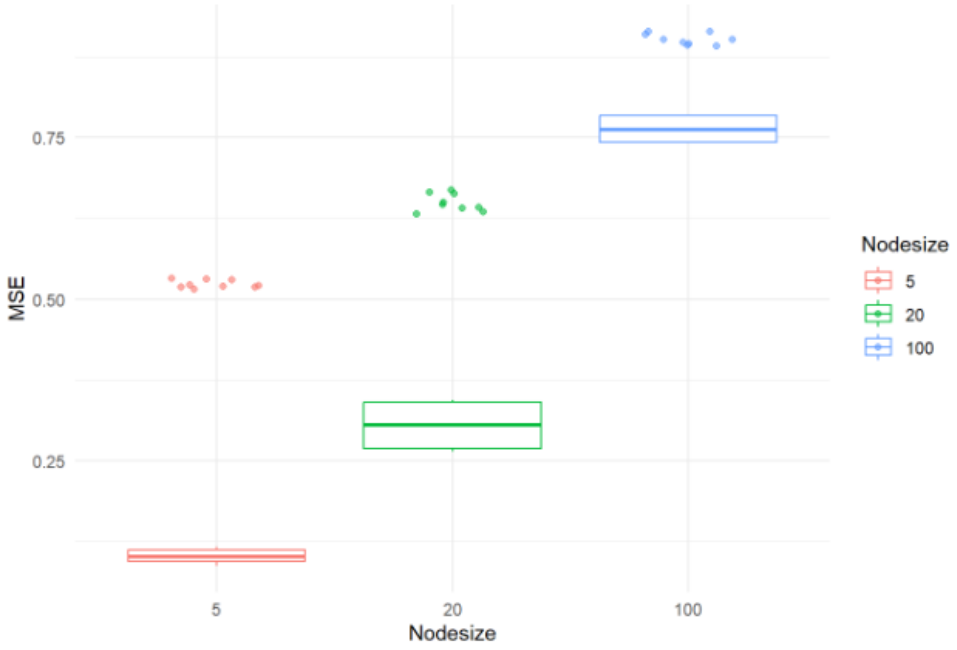
Tabela 4. Modele, które uzyskały MSE na zbiorze testowym mniejsze niż 0.1

ntree	mtry	nodesize	trening_mse	test_mse
200	6	5	0.5324103	0.09321446
500	6	5	0.5195602	0.093646
1000	3	5	0.5152641	0.09712417
1000	6	5	0.5187023	0.08795338

Źródło: Opracowanie własne

Z przedstawionej tabeli wynika, że największy wpływ na dokładność modelu miał parametr dotyczący minimalnej liczby obserwacji w liściu, ponieważ tylko dla wartości równej 5 tego parametru uzyskano wyniki o błędzie mniejszym niż 0.1.

Przedstawiam wykres pudełkowy porównujący wyniki dla różnych wartości:



Rysunek 5. Wykres pudełkowy porównujący uzyskane wyniki dla różnej minimalnej liczby obserwacji w liściu.
Źródło: Opracowanie własne

Na powyższym rysunku zarówno dla danych treningowych (kolorowe punkty) jak i dla danych testowych (pudełka wykresu) wyniki są najlepsze dla modelów z najmniejszą badaną wartością minimalnej liczby obserwacji w liściu. Co ciekawe, w każdym przypadku wyniki na zbiorze testowym są zdecydowanie lepsze niż na zbiorze treningowym. Możliwe, że doszło do tego przez wielokrotny podział zbioru na podzbiory treningowe. Dodatkowo można zastosować analizę ANOVA, aby sprawdzić, czy średnie wartości MSE dla różnych wartości minimalnej liczby obserwacji w liściu są statystycznie różne.

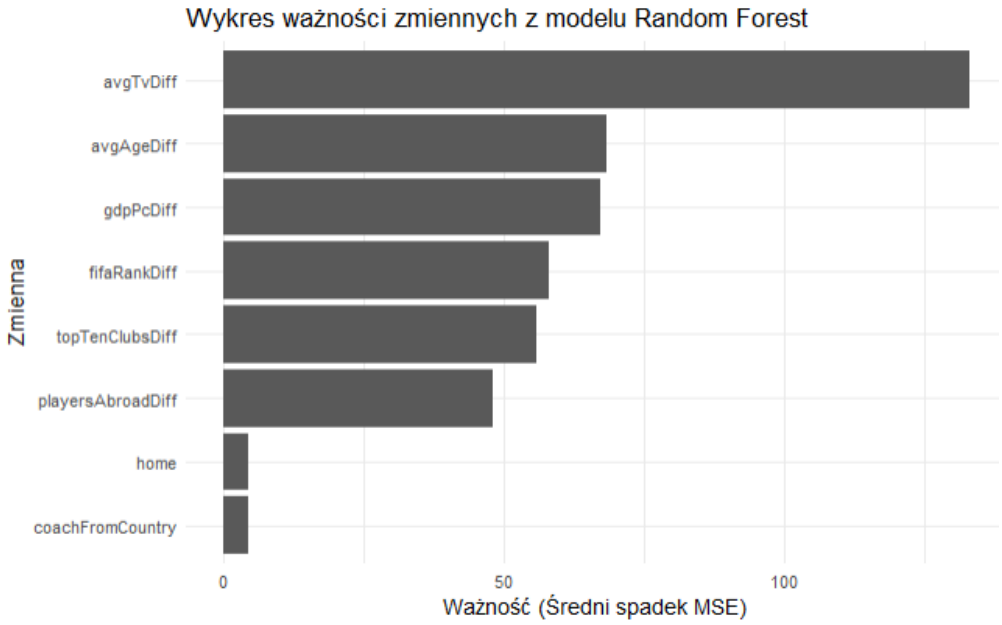
Do dalszej prognozy zastosowałem model o najmniejszym MSE na zbiorze testowym o następujących parametrach:

Tabela 5. Model lasu losowego zastosowany w badaniu

ntree	mtry	nodesize	trening_mse	test_mse
1000	6	5	0.5187023	0.08795338

Źródło: Opracowanie własne

Dodatkowo, używając lasów losowych z łatwością można zmierzyć które zmienne mają największy wpływ na ustalenie ostatecznej wersji modelu. To zjawisko polega na analizie, która zmienna najbardziej zmniejsza wartość średniego błędu kwadratowego. Tak wygląda wykres ważności zmiennych:[35]



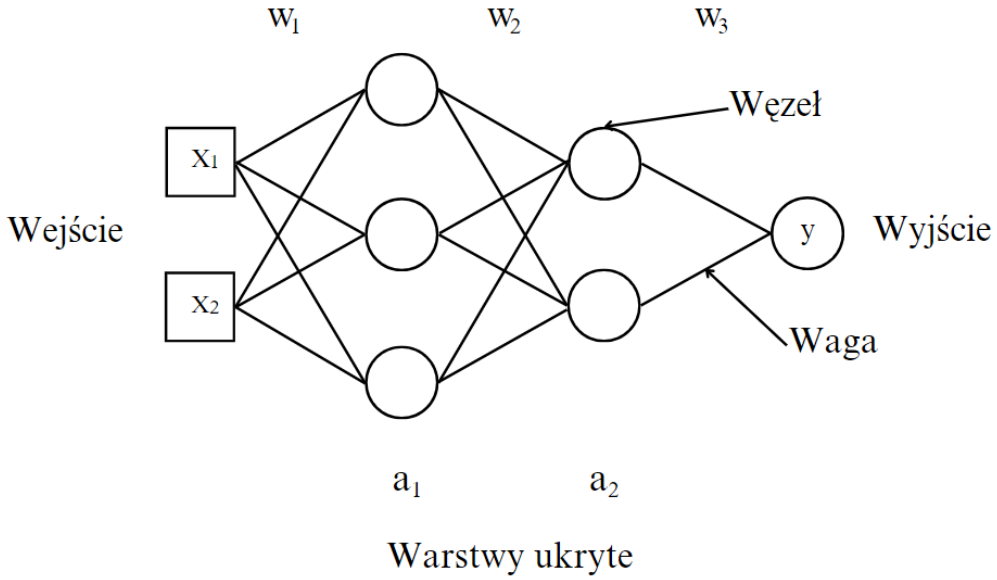
Rysunek 6. Ważność zmiennych.
Źródło: Opracowanie własne

Z wykresu wynika, że zmienna, której usunięcie powodowałoby największy wzrost średniego błędu kwadratowego to różnica w średniej wartości piłkarza reprezentacji. Natomiast zmienne, które mają najmniejszy wpływ na zmianę MSE to zmienne faktorowe (informacja, czy drużyna rozgrywa mecz w swoim państwie oraz czy trener pochodzi z państwa, którego reprezentacje trenuje).

[35] M. Mamczur, *Jak działa las losowy (random forest)?,* 2024, https://miroslawmamczur.pl/jak-dziala-las-losowy-random-forest/ [dostęp: 15.06.2024]

2.5. Sztuczne sieci neuronowe

Sztuczna sieć neuronowa w pewnych sferach przypomina działanie tej biologicznej: rozbudowane struktury zbierają sygnały z neuronów a pojedynczy element przekazuje sygnał wyjściowy. Według R. T. Kneusella[36] to porównanie może wprowadzić w błąd wielu niezaznajomionych z tematem, ponieważ kojarzy się z stworzeniem myślącego mózgu, co nie oddaje do końca natury sztucznej sieci neuronowej. Składa się ona z sygnałów wejściowych, warstw posiadających węzły, wag, funkcji aktywacji, sygnałów wyjściowych, co dobrze odwzorowuje schemat:



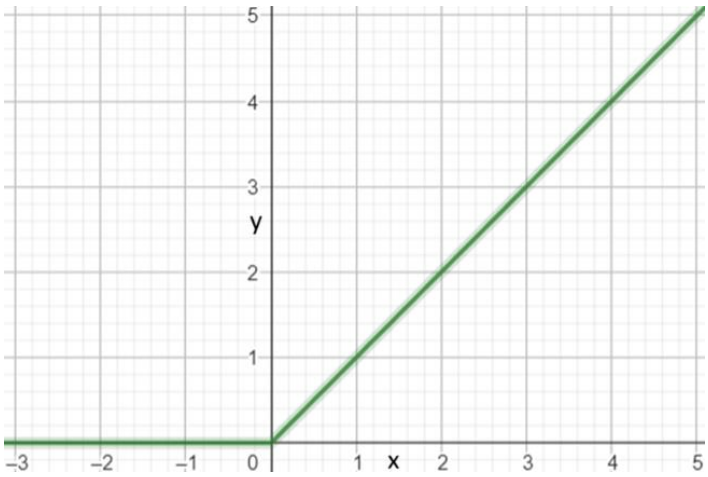
Rysunek 7. Schemat sztucznej sieci neuronowej.

Źródło: R. T. Kneusella, Deep learning – praktyczne wprowadzenie z zastosowaniem środowiska Pythona

Na powyższym rysunku widzimy od lewej sygnały wejściowe x_0 i x_1, które pomnożone przez wagi (symbolizowane przez odpowiednie odcinki) oraz dodane do wartości obciążenia są przekazywane do funkcji aktywacji. Dostarczona wartość jest przekształcona w odpowiedni sposób i przekazywana dalej. Gdy sygnał opuści ostatnią warstwę ukrytą (w tym przykładzie są takie dwie) trafia do warstwy wyjściowej. W przypadku problemu regresji często węzeł w warstwie wyjściowej nie ma funkcji aktywacji, jednak prognozując mistrzostwa Europy w piłce nożnej, a dokładniej liczbę bramek zdobytych przez drużynę zależy mi, aby wartość ostateczna była dodatnia i nie miała ograniczenia z góry[37].

[36] R. T. Kneusella, Deep learning – praktyczne wprowadzenie z zastosowaniem środowiska Pythona, Helion, 2022.
[37] Ibidem.

Idealną funkcją spełniającą te założenia jest f. aktywacji ReLu, której wykres wygląda następująco:



Rysunek 8. Wykres funkcji ReLu.
Źródło: Opracowanie własne

Na powyższym rysunku oś x oznacza sumę sygnałów (wartość wejściowa), a oś y oznacza wartość wyjściową funkcji ReLu dla danej wartości wejściowej. Warto wspomnieć, że podczas badania klasyfikacji w warstwie wyjściowej może znajdować się tyle węzłów ile istnieje grup dopasowania.

Uczenie sieci neuronowej odbywa się na podstawie zbioru treningowego. Jedną z najpopularniejszych metod jej trenowania jest algorytm propagacji wstecznej. Proces składa się z następujących kroków:

1. Ustalenie struktury sieci: liczby warstw, liczby neuronów w warstwach ukrytych, połączeń między węzłami.
2. Losowanie wag.
3. Obliczanie wyniku dla pierwszego przypadku treningowego
4. Obliczanie błędu węzła z warstwy wyjściowej między wartością rzeczywistą a prognozowaną
5. Propagowanie obliczonych błędów do wszystkich poprzednich warstw
6. Powtórzenie procedury
7. Po analizie całego zbioru przeznaczonego do treningu modelu zmiana wszystkich wag
8. Losowanie kolejności obserwacji w zbiorze treningowym i rozpoczęcie kolejnej iteracji

Uczenie modelu zakończy się z chwilą przekroczenia ustawionej liczby iteracji[38].

[38] M. Szeliga, op. cit., s. 204.

Aby wybrać najlepszy model zdecydowałem się na zmianę następujących parametrów:

- Liczba iteracji: Określa jak długo ma trwać trenowanie modelu, a dokładniej ile razy należy przejść przez cały zbiór treningowy

Testowane wartości: 200, 500, 1000

- Liczba węzłów w warstwie ukrytej modelu

Testowane wartości: 4, 8, 16

- Liczba warstw ukrytych

Testowane wartości: 1, 2, 3

Modele z najmniejszym MSE (wartość poniżej 0.2) na zbiorze testowym uzyskane w ten sposób to:

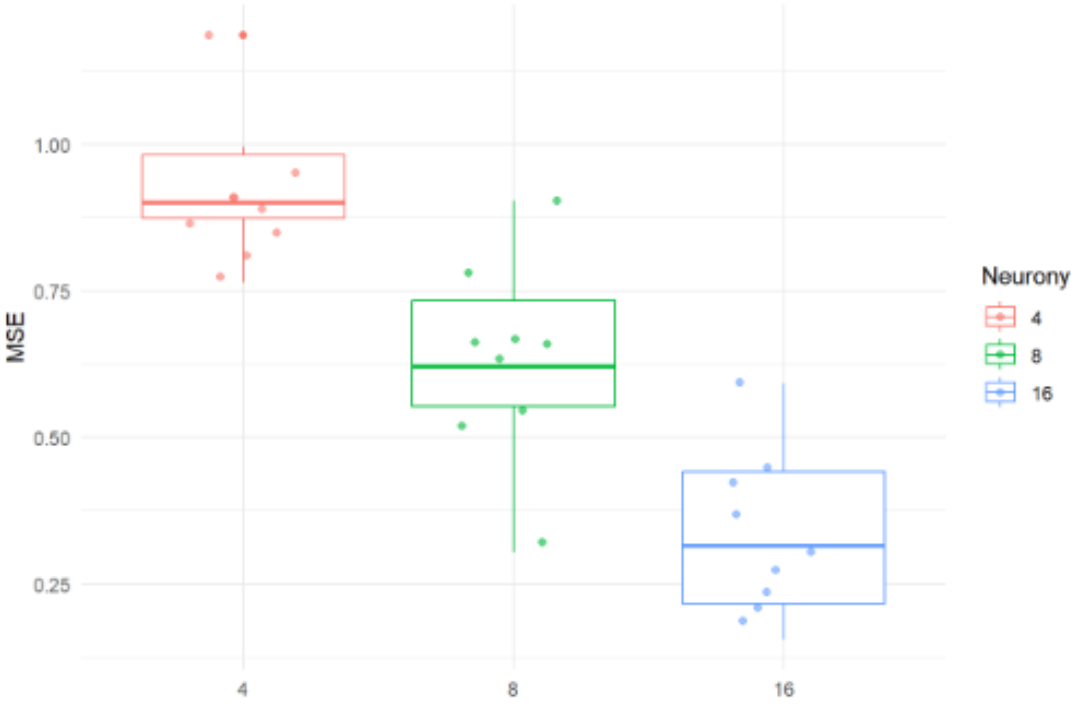
Tabela 6. Najlepsze modele sieci neuronowych

warstwy	epoki	neurony	trening_mse	test_mse
2.0	500.0	16.0	0.2360237	0.1602181
3.0	500.0	16.0	0.2092723	0.1557684

Źródło: Opracowanie własne

W wyżej zaprezentowanej tabeli kolumny oznaczają kolejno liczbę warstw ukrytych, liczbę epok, liczbę neuronów w warstwie ukrytej, wartość MSE uzyskaną na zbiorze treningowym oraz wartość MSE uzyskaną na zbiorze testowym. Można zauważyć, że najlepiej wytrenowane modele powstały gdy liczba epok wynosiła 500 (środkowa testowana wartość) i liczba węzłów w warstwie ukrytej wynosiła 16 (największa testowana wartość). Z tego wynika, że zbyt duża liczba iteracji podczas uczenia modelu może spowodować jego przetrenowanie. Warto zauważyć, że nie ma zbyt dużej różnicy między wynikami zaprezentowanych w tabeli modelów, więc dodanie kolejnej warstwy ukrytej nie spowodowało znacznej poprawy (spadek MSE o około 0.004). Można przypuszczać że na ogół modele z 16 neuronami w warstwie ukrytej były dokładniejsze, ponieważ tylko takie uzyskały MSE mniejsze od 0.2.

Tak prezentuje się wykres słupkowy z podziałem na liczbę węzłów:



Rysunek 9. Wykres pudełkowy porównujący uzyskane wyniki dla różnej liczby węzłów w warstwie ukrytej.
Źródło: Opracowanie własne

Na powyższym wykresie, zarówno dla danych treningowych (kolorowe punkty) jak i dla danych testowych (pudełka wykresu) wyniki są najlepsze dla modelów z największą badaną wartością liczby neuronów. Można byłoby dodatkowo zastosować analizę ANOVA, aby sprawdzić, czy średnie MSE dla różnych wartości liczby neuronów w warstwie są statystycznie różne.

Model, który został wykorzystany w badaniu ma następujące parametry i znajduje się w Załączniku 6:

Tabela 7. Model sieci neuronowej wykorzystany do prognozy turnieju.

warstwy	epoki	neurony	trening_mse	test_mse
3.0	500.0	16.0	0.2092723	0.1557684

Źródło: Opracowanie własne

3. Wyniki

3.1. Zasady awansu drużyn do fazy pucharowej turnieju

W celu zbliżenia badania do rzeczywistego turnieju, starałem się jak najbardziej odwzorować zasady awansu z fazy grupowej do tych rzeczywistych. Jak informuje Regulamin Mistrzostw Europy w Piłce Nożnej UEFA[39] do 1/8 awansują najlepsze dwa zespoły z każdej grupy plus cztery najlepsze z trzecich miejsc.

O kolejności drużyn w grupie decyduje liczba zdobytych punktów we wszystkich spotkaniach fazy grupowej. Za zwycięstwo są przyznawane 3 pkt, za remis 1pkt a za porażkę 0. Aby rozstrzygnąć wynik spotkania, obliczoną przez model liczbę bramek zaokrąglam zgodnie z zasadami matematyki, od 0.5 w górę. Jeżeli zespoły mają taką sama liczbę punktów o układzie tabeli decydują (w wymienionej kolejności):

1. Liczba zdobytych w starciach bezpośrednich między zainteresowanymi zespołami.
2. Lepsza różnica bramek (gole zdobyte - gole stracone) w meczach bezpośrednich.
3. Większa liczba strzelonych bramek w meczach bezpośrednich.
4. Lepsza różnica bramek we wszystkich spotkaniach
5. Większa liczba strzelonych bramek we wszystkich spotkaniach
6. Pozycja w rankingu Ligi Narodów 2022/23

Przyjęta technika różni się trochę od rzeczywistych zasad w turnieju. Między punktem 5 a 6 brana pod uwagę jest niższa suma punktów dyscyplinarnych. Obliczane są one wg wzoru:

- czerwona kartka = 3 punkty
- żółta kartka = 1 punkt
- wykluczenie za dwie żółte kartki w jednym meczu = 3 punkty

Niestety nie miałem możliwości zastosowania tej metody, ponieważ modele nie przewidują zdobytych kartek. Dodatkowo dla ułatwienia zdecydowałem się przyjąć jako punkt 6 ranking Ligi Narodów znajdujący się w Załączniku 4 (stosowany przy rozstrzyganiu pozycji podczas kwalifikacji do turnieju).

[39]Regulations of the UEFA European Football Championship, https://documents.uefa.com/r/Regulations-of-the-UEFA-European-Football-Championship-2022-24/Article-19-Match-system-final-tournament-group-stage-Online [dostęp: 16.06.2024]

Tak wygląda jedna z ostatecznych tabel:

Tabela 8. Wyniki grupy B uzyskane za pomocą lasu losowego

Zespół	Punkty	Różnica bramek	Bramki strzelone	Ranking	Grupa i pozycja
Włochy	7	2	5	3	B1
Hiszpania	5	1	4	1	B2
Chorwacja	2	-1	3	2	B3
Albania	1	-2	3	27	B4

Źródło: Opracowanie własne

W wyżej zaprezentowanej tabeli kolumny oznaczają kolejno drużynę, uzyskane punkty w fazie grupowej, różnicę bramek strzelonych i straconych w fazie grupowej, strzelone bramki w fazie grupowej, miejsce w rankingu ligi narodów oraz grupę i zajętą w niej pozycje. W następnych fazach turnieju, jeżeli zaokrąglona liczba bramek jest równa dla obydwóch drużyn, mnożę obliczone dzięki modelowi wartości przez 1.33, ponieważ czas dogrywki stanowi 1/3 czasu regulaminowego, zaokrąglam i obliczam różnicę. Jeżeli takie działanie dalej nie przyniesie rozstrzygnięcia, o awansie decyduje rzut monetą, który symbolizuje rzuty karne. Tak wygląda przykładowe spotkanie:

Tabela 9. Wynik 1/2 między Anglią a Francją

Zespół	Rywal	Bramki	Różnica	Bramki 2	Różnica po dogrywce
Anglia	Francja	0.6735	0	0.895755	-1
Francja	Anglia	1.4249917	0	1.8952389	1

Źródło: Opracowanie własne

W Tabeli 9 kolumny oznaczają kolejno drużynę, rywala w meczu, bramki zdobyte, różnice bramek między drużyną a rywalem, bramki strzelone po dogrywce, różnicę bramek po dogrywce. W podanym przykładzie Francja przeszła do finału, ponieważ po dogrywce różnica bramek w meczu zmieniła się na ich korzyść. Dla pewności uchwycenia wszystkich możliwych wyników – rzuty karne mogły wpłynąć na ich zmianę – powtórzyłem predykcje turnieju 100 razy dla każdej z metod.

Układ rozgrywanych meczów zależy od tego, które drużyny awansują z trzeciego miejsca. W Załączniku 3 znajdują się dane potrzebne do przygotowania zmiennych objaśnianych dla spotkań w dalszych fazach turnieju. Tak wygląda oficjalny plan rozgrywania meczy, który odwzorowałem w badaniu:

Tabela 10. Zaplanowane spotkania 1/8

Mecz	Drużyny
Mecz 1	WB v 3A/D/E/F
Mecz 2	WA v RC
Mecz 3	WF v 3A/B/C
Mecz 4	RD v RE
Mecz 5	WE v 3A/B/C/D
Mecz 6	WD v RF
Mecz 7	WC v 3D/E/F
Mecz 8	RA v RB

Źródło: https://documents.uefa.com/r/Regulations-of-the-UEFA-European-Football-Championship-2022-24/Article-21-Match-system-final-tournament-knockout-stage-Online [dostęp od: 17.06.2024]

W przedstawionej tabeli W oznacza zwycięzcę danej grupy, R awans z drugiego miejsca a 3 awans jako jedna z najlepszych drużyn z trzeciej lokaty.

Jest wiele możliwości ułożenia spotkań, ponieważ nie wiadomo które zespoły awansują. Poniższa tabela przedstawia wszystkie kombinacje, które zostały odwzorowane w badaniu:

Tabela 11. Wszystkie możliwe przypadki 1/8:

Drużyny z grup, które awansowały z 3 miejsca:	**WB**	**WC**	**WE**	**WF**
A B C D	3A	3D	3B	3C
A B C E	3A	3E	3B	3C
A B C F	3A	3F	3B	3C
A B D E	3D	3E	3A	3B
A B D F	3D	3F	3A	3B
A B E F	3E	3F	3B	3A
A C D E	3E	3D	3C	3A
A C D F	3F	3D	3C	3A
A C E F	3E	3F	3C	3A
A D E F	3E	3F	3D	3A
B C D E	3E	3D	3B	3C
B C D F	3F	3D	3C	3B
B C E F	3F	3E	3C	3B
B D E F	3F	3E	3D	3B
C D E F	3F	3E	3D	3C

Źródło: op. Cit

Wyżej zaprezentowana tabela znajduje się w Załączniku 5. W przypadku, gdy z trzeciego miejsca awansują zespoły z grup A, B, C, D, Zwycięzcy grup B, C, E, F zagrają kolejno z drużynami z grupy A, D, B, C.

3.2. Wyniki uzyskane za pomocą lasu losowego

Tabele grup po symulacji wszystkich spotkań:

Tabela 12. Grupa A, las losowy

Zespół	Punkty
Węgry	7
Niemcy	7
Szwajcaria	1
Szkocja	1

Źródło: Opracowanie własne

Tabela 13. Grupa B, las losowy

Zespół	Punkty
Włochy	7
Hiszpania	5
Chorwacja	2
Albania	1

Źródło: Opracowanie własne

Tabela 14. Grupa C, las losowy

Zespół	Punkty
Anglia	9
Dania	4
Serbia	2
Słowenia	1

Źródło: Opracowanie własne

Tabela 15. Grupa D, las losowy

Zespół	Punkty
Francja	7
Holandia	4
Austria	4
Polska	1

Źródło: Opracowanie własne

Tabela 16. Grupa E, las losowy

Zespół	Punkty
Belgia	9
Ukraina	6
Rumunia	1
Słowacja	1

Źródło: Opracowanie własne

Tabela 17. Grupa F, las losowy

Zespół	Punkty
Portugalia	9
Turcja	6
Czechy	1
Gruzja	1

Źródło: Opracowanie własne

Do dalszej fazy turnieju z pierwszych miejsc w grupie awansowały reprezentacje Węgier, Włoch, Anglii, Francji, Belgii i Portugalii. Drugie miejsca dające miejsce w 1/8 turnieju zajęły drużyny Niemiec, Hiszpanii, Danii, Holandii, Ukrainy oraz Turcji.

Tabela zespołów z trzecich miejsc w grupach, z których awansowały dalej 4 najlepsze zespoły wygląda następująco:

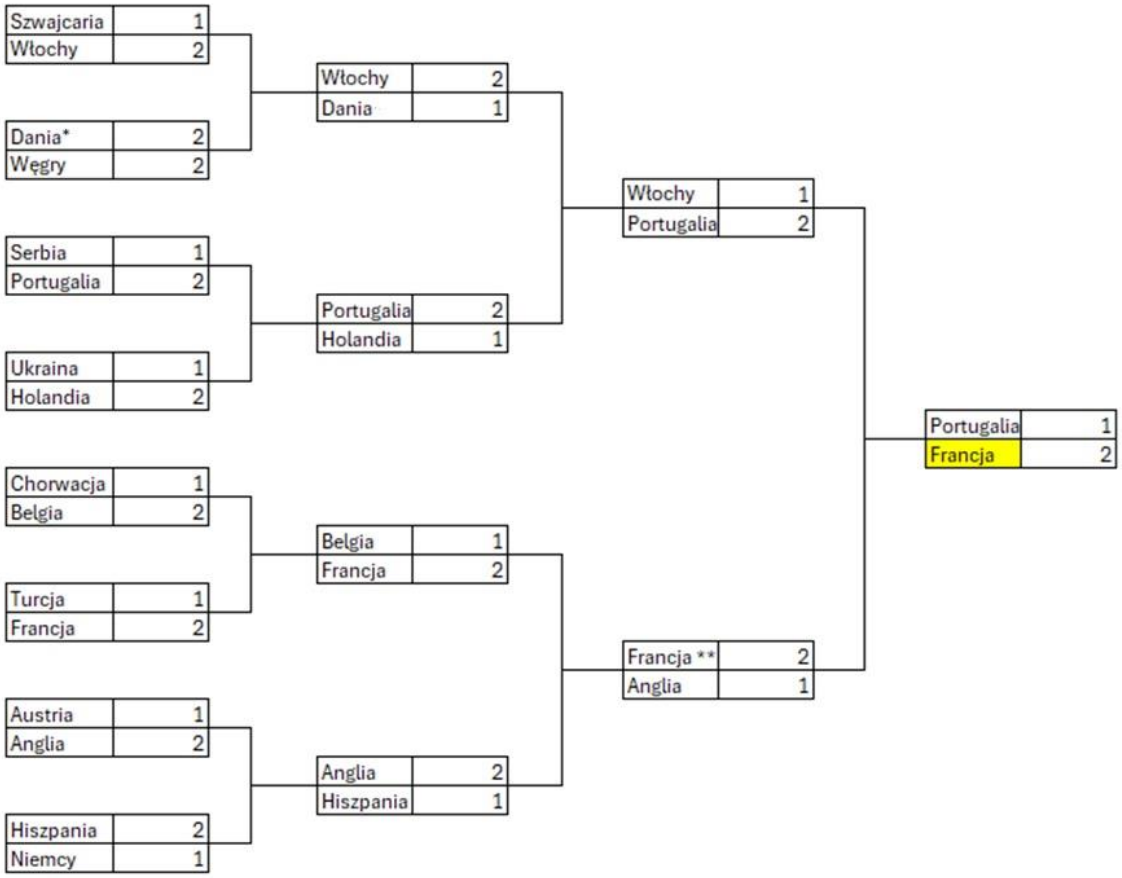
Tabela 18. Drużyny, które zajęły trzecie miejsce w grupach

Zespół	Punkty	Różnica bramek	Bramki strzelone	Ranking
Austria	4	0	4	13
Serbia	2	-1	4	19
Chorwacja	2	-1	3	2
Szwajcaria	1	-2	3	9
Rumunia	1	-2	3	29
Czechy	1	-3	2	14

Źródło: Opracowanie własne

Do następnej fazy turnieju z trzecich miejsc awansowały reprezentacje Austrii, Serbii, Chorwacji i Szwajcarii. Warto zauważyć, że jedna z drużyn potrzebowała tylko punktu, aby pojawić się w 1/8.

Wyniki jednej ze 100 symulacji mistrzostw Europy w piłce nożnej uzyskane za pomocą lasu losowego prezentują się następująco:



Rysunek 10. Turniej uzyskany metodą lasu losowego
Źródło: Opracowanie własne
* Awans w rzutach karnych
** Awans po dogrywce

Wyniki uzyskane za pomocą lasu losowego przyniosły jedno rozstrzygnięcie w rzutach karnych, w 1/8 między Danią a Węgrami. Wynik tego starcia nie wpłynął na ostatecznego zwycięzcę, ponieważ obie te drużyny w przypadku awansu odpadły w następnej fazie w meczu z Włochami. Dodatkowo w półfinale między Anglią a Francją doszło do dogrywki (1:1 w regulaminowym czasie gry), po której zwyciężyli Francuzi.

3.3. Wyniki uzyskane za pomocą sieci neuronowych

Tabele grup po symulacji wszystkich spotkań:

Tabela 19. Grupa A sztuczne sieci

Zespół	Punkty
Szwajcaria	7
Niemcy	5
Węgry	3
Szkocja	1

Źródło: Opracowanie własne

Tabela 20. Grupa B sztuczne sieci

Zespół	Punkty
Hiszpania	9
Włochy	6
Chorwacja	1
Albania	1

Źródło: Opracowanie własne

Tabela 21. Grupa C sztuczne sieci

Zespół	Punkty
Dania	9
Anglia	6
Słowenia	3
Serbia	0

Źródło: Opracowanie własne

Tabela 22. Grupa D sztuczne sieci

Zespół	Punkty
Holandia	5
Polska	4
Francja	4
Austria	1

Źródło: Opracowanie własne

Tabela 23. Grupa E sztuczne sieci

Zespół	Punkty
Belgia	9
Słowacja	4
Rumunia	3
Ukraina	1

Źródło: Opracowanie własne

Tabela 24. Grupa F sztuczne sieci

Zespół	Punkty
Portugalia	7
Czechy	5
Turcja	4
Gruzja	0

Źródło: Opracowanie własne

W wynikach meczów grupowych uzyskanych za pomocą sieci neuronowych było trochę niespodzianek. Patrząc na potencjał kadrowy i przewidywania bukmacherów. Za taką można uznać 2 miejsce Polski czy Słowacji.

Podczas badania testowałem kilka modeli o tych samych parametrach i za każdym razem trafiały się jakieś szokujące rozwiązania. Moim zdaniem może być to powiązane z tym, że sieci neuronowe nie ustalają ważności zmiennych, w przeciwieństwie do lasu losowego.

Tabela zespołów z trzecich miejsc w grupach, z których awansowały dalej 4 najlepsze zespoły wygląda następująco:

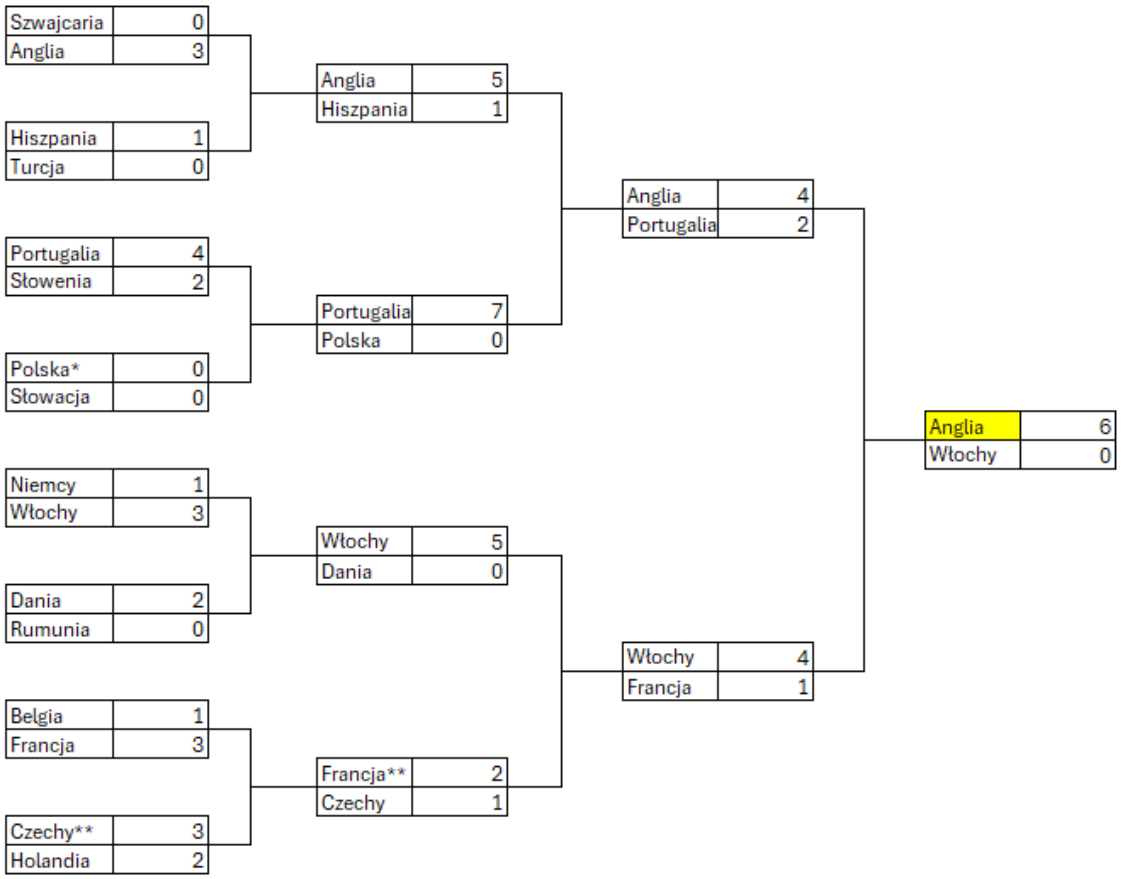
Tabela 25. Drużyny, które zajęły trzecie miejsce w grupach w prognozowaniu sieciami neuronowymi

Zespół	Punkty	Różnica bramek	Bramki strzelone	Ranking
Turcja	4	1	6	35
Francja	4	0	3	12
Słowenia	3	-3	4	25
Rumunia	3	-3	2	29
Węgry	3	-6	1	8
Chorwacja	1	-2	0	2

Źródło: Opracowanie własne

Do następnej fazy turnieju z trzecich miejsc awansowały reprezentacje Turcjii, Francji, Słowenii i Rumunii. Warto zauważyć, że 3 pkt reprezentacji Węgier nie wystarczyły do awansu, która w symulacji za pomocą lasów losowych wyszła z grupy z pierwszego miejsca.

Wyniki jednej ze 100 symulacji mistrzostw Europy w piłce nożnej uzyskane za pomocą sieci neuronowych prezentują się następująco:



Rysunek 11. Turniej uzyskany metodą sieci neuronowej.
Źródło: Opracowanie własne
* Awans w rzutach karnych
** Awans po dogrywce

Na pierwszy rzut oka wyniki uzyskane za pomocą metody sieci neuronowych są bardziej różnorodne, ponieważ przy zastosowaniu lasu losowego najczęściej padał rezultat 2:1 dla jednego z zespołów, a tutaj rozstrzygnięcia spotkań wahają się od 0:0 do 7:0. Rzuty karne w 1/8 między Polską a Słowacją nie spowodowały, wielu zmian, ponieważ obie drużyny odpadały w następnej fazie turnieju w spotkaniu z Portugalią. Podczas wyboru finalnego modelu, napotkałem się na wiele stosunkowo dziwnych wyników, gdzie ostatecznym zwycięzcą turnieju mogły zostać drużyny o obiektywnie niskich szansach na triumf. Mogło być to spowodowane stosunkowo niewielkim zbiorem danych.

3.4. Porównanie z prognozami OPTA

OPTA dokonała predykcji mistrzostw Europy w piłce nożnej, symulując turniej 10 tys. razy i obliczając przy tym procentowe szanse na awans poszczególnych zespołów[40]. Porównanie uzyskanych przez nich wyników z dnia 4 czerwca bieżącego roku i moich predykcji wygląda następująco:

Tabela 26. Zespoły z największymi szansami na ćwierćfinał

OPTA	LAS LOSOWY	SIECI NEURONOWE
Anglia (69.7%)	Anglia	Anglia
Francja (67.2%)	Francja	Francja
Niemcy (66.8%)	Hiszpania	Hiszpania
Hiszpania (66.4%)	Portugalia	Portugalia
Portugalia (64.5%)	Holandia	Włochy
Holandia (54.8%)	Włochy	Dania
Włochy (52.9%)	Belgia	Czechy
Belgia (44.6%)	Dania/Węgry	Polska/Słowacja

Źródło: Opracowanie własne

W prognozie uzyskanej za pomocą lasu losowego drużyny, które uzyskały awans do 1/8 prawie idealnie pokrywają się z tymi, które mają największe szanse wg OPTY. Jedynie Niemcy (gospodarze turnieju) nie zakwalifikowali się do tej fazy na rzecz Danii lub Węgrów. Może to być powiązane z ułatwieniem, które wprowadziłem dotyczące pozycji w grupie – Węgrowie wyprzedzili Niemców przez zastosowanie rankingu Ligi Narodów. Z zespołów premiowanych awansem podczas badania metodą sieci neuronowych prawie połowa nie pokrywa się z wynikami „superkomputera" OPTY.

[40] C. Myson, *Who Will Win Euro 2024? The Opta Predictions*, https://theanalyst.com/eu/2024/06/who-will-win-euro-2024-predictions-opta/, [dostęp: 17.06.2024].

Tabela 27. Zespoły z największymi szansami na półfinał

OPTA (%)	LAS LOSWY	SIECI NEURONOWE
Anglia (45.6%)	Anglia	Anglia
Francja (43.0%)	Francja	Portugalia
Niemcy (39.6%)	Włochy	Włochy
Hiszpania (37.8%)	Portugalia	Francja

Źródło: Opracowanie własne

Finałowa czwórka turnieju ma taki sam skład w prognozie uzyskanej za pomocą lasu losowego jak i sieci neuronowych, natomiast cały turniej miał inny przebieg, co spowodowało różne pary półfinałowe (Anglia – Francja oraz Włochy – Portugalia w metodzie lasu losowego i Anglia – Portugalia oraz Włochy – Francja w metodzie sieci neuronowych).

Tabela 28. Zespoły z największymi szansami na finał i ostateczni zwycięzcy

OPTA (%)	LAS LOSOWY	SIECI NEURONOWE
Anglia (27.1%)	**Francja**	**Anglia**
Francja (25.4%)	Portugalia	Włochy

Źródło: Opracowanie własne

Skład prognozowanych finałów jest różny, natomiast według badania OPTY najbliżej zdobycia trofeum są reprezentacje Francji i Anglii co w pewnym sensie potwierdziło przeprowadzone przeze mnie badanie (Francja zwyciężyła w finale z Portugalią a Anglia w ostatnim spotkaniu turnieju z Włochami).

Podsumowanie

Prognozy turnieju przeprowadzone dwoma technikami, mimo różnych rozstrzygnięć w trakcie turnieju, przewidziały, że do półfinałów awansują te same zespoły. Są to reprezentacje Anglii, Francji, Włoch i Portugalii. Prognoza za pomocą lasów losowych wskazała Francję jako głównego faworyta, natomiast model sztucznych sieci neuronowych docenił reprezentacje Anglii. Dużą zaletą modelu lasu losowego jest stałość wyników. Sieć neuronów, mimo tych samych parametrów podczas treningu, za każdym razem zwraca inny model, którego ostateczne wyniki różnią się. Może to być spowodowane stosunkowo niewielką liczbą obserwacji, bo różnice między otrzymanymi rezultatami są znaczące. Model sztucznych sieci neuronowych na ogół potrafi lepiej uchwycić zależności nieliniowe, natomiast las losowy jest mniej podatny na przetrenowanie, co potwierdza mniejszy średni błąd kwadratowy na zbiorze testowym.

Dodatkowo, prognozując wynik turnieju za pomocą metody lasów losowych uzyskano informację o ważności poszczególnych zmiennych. Największy wpływ na liczbę strzelonych bramek w jednym spotkaniu miała średnia różnica w wartości zespołu, a najmniejszy teren rozgrywanego spotkania i pochodzenie trenera. Pozostałe zmienne takie jak różnica zespołów w średnim wieku zawodnika, różnica państw w wartości PKB per capita, różnica drużyn w ilości zawodników z top 10 klubów piłkarskich w Europie i różnica w ilości zawodników grających w klubach za granicą mają podobny wpływ na ostateczny wynik.

Literatura

J. Bartman, K. Bajda, *Wykorzystanie sztucznych sieci neuronowych do prognozowania wyników meczów piłkarskich*, Journal of Education, Technology and Computer Science, 2014.

J. Dean, *Big Data, Data Mining, and Machine Learning,* Wiley, 2014.

P. Cięszczyk, J. Eider, *Alternatywne metody badania wyników sportowych w wybranych konkurencjach lekkoatletycznych,* ISSN, 2003.

M. Cieślak, P. Dittmann, A. Kania–Gospodarowicz, I. Kuropka, S. Ostasiewicz, B. Radzikowska, *Demografia, metody analizy i prognozowania*, PWN, 1992.

B. Efron, R. Tibshirani, *An Introduction to the Bootstrap*, tom. 57, Chapman & Hall, 1993.

J. B. Gajda, *Prognozowanie i symulacje w ekonomii i zarządzaniu,* C.H.Beck, 2017.

A. Groll, C. Ley, G, Schauberger, H. Van Eetvelde *Prediction of the FIFA World Cup 2018 - A random forest approach with an emphasis on estimated team ability parameters*, Statistical Modelling, 2018.

J. Grus, *Data science od podstaw. Analiza danych w Pythonie. Wydanie II,* Helion, 2018.

A. Hassan, A. R. Akl, I. Hassan, C. Sunderland, *Predicting Wins, Losses and Attributes' Sensitivities in the Soccer World Cup 2018 Using Neural Network Analysis,* Sensors, 2020.

K. Y. Huang, K. J. Chen, *Multilayer Perceptron for Prediction of 2006 World Cup Football Game,* Journal of Geomatics Science and Technology*,* 2006.

R. T. Kneusella, *Deep learning – praktyczne wprowadzenie z zastosowaniem środowiska Pythona*, Helion, 2022.

I. Langmore, D. Krasner, *Applied Data Science*, 2016.

P. Sroka, J. Trzęsiok, *Co opowiadają drzewa o tenisie? Predykcja wyników spotkań w tenisie ziemnym z wykorzystaniem drzew klasyfikacyjnych*, Uniwersytet Ekonomiczny we Wrocławiu, 2017.

A. Tatarczak, *Ekonometria – Podręcznik. Studia przypadków,* tom 20, WSEI, 2021.

M. Walesiak, E. Gatnar, *Statystyczna analiza danych z wykorzystaniem programu R,* PWN, 2012.

Wykaz innych źródeł

https://www.sts.pl/blog/historia-pilkarskich-mistrzostw-europy/ [dostęp: 18.06.2024].

https://weszlo.com/2019/06/07/kiedys-elita-dzis-pospolite-ruszenie-sie-zmienialy-mistrzostwa-europy/ [dostęp: 18.06.2024].

https://newonce.net/artykul/bill-james-john-henry-i-excel-jak-analityka-weszla-do-swiata-sportu, [dostęp: 20.06.2024]

https://www.tibco.com/blog/2020/08/27/how-data-analytics-emerged-as-a-competitive-advantage-for-the-mercedes-amg-petronas-formula-one-team/, [dostęp: 20.06.2024].

https://technet-media.pl/artykuly/ibm-ai-draw-analysis-ibm-wprowadza-generatywna-sztuczna-inteligencje-do-wimbledonu [dostęp: 20.06.2024].

https://sport.rp.pl/pilka-nozna/art39683621-wisla-krakow-ma-nowego-trenera-wybrala-go-sztuczna-inteligencja [dostęp: 20.06.2024]

http://www.prognozowanie.info/prognozowanie-ekonometryczne/ [dostęp: 24.08.2024]

https://databank.worldbank.org/source/world-development-indicators [dostęp: 20.04.2024].

https://www.kaggle.com/datasets/cashncarry/fifaworldranking [dostęp: 20.04.2024]

https://en.wikipedia.org/wiki/Main_Page [dostęp: 20.04.2024]

https://www.transfermarkt.com [dostęp: 20.04.2024]

https://www.kaggle.com/datasets/martj42/international-football-results-from-1872-to-2017 [dostęp: 20.04.2024].

https://miroslawmamczur.pl/jak-dziala-las-losowy-random-forest/ [dostęp: 15.06.2024]

https://documents.uefa.com/r/Regulations-of-the-UEFA-European-Football-Championship-2022-24/Article-19-Match-system-final-tournament-group-stage-Online [dostęp: 16.06.2024]

https://theanalyst.com/eu/2024/06/who-will-win-euro-2024-predictions-opta/, [dostęp: 17.06.2024].

https://www.stadiumguide.com/tournaments/uefa-euro-2020/ [dostęp: 15.06.2024].

https://kronika-futbolu.pl/historia/mistrzostwa-europy/ [dostęp: 15.06.2024]

https://documents.uefa.com/r/Regulations-of-the-UEFA-European-Football-Championship-2022-24/Article-21-Match-system-final-tournament-knockout-stage-Online [dostęp od: 17.06.2024]

Spis tabel

Spis rysunków

Spis równań

Załączniki

Załącznik 1. Kod źródłowy

Załącznik 2. Wyniki wszystkich meczy z turniejów mistrzostw Europy w XXI wieku oraz mecze grupowe prognozowane przez modele.

Załącznik 3. Dane potrzebne do obliczenia zmiennych dla meczów w 1/8, 1/4, 1/2 i finału.

Załącznik 4. Ranking ligi narodów.

Załącznik 5. Zasady układania spotkań po fazie grupowej.

Załącznik 6. Model lasu losowego użyty w niniejszej pracy.

Dostęp: Praca Licencjacka