A close up of a logo

Description automatically generated

**Kolegium Analiz Ekonomicznych**

**Podyplomowe Studia**

**Inżynieria danych – Big Data**

**Edycja 18**

**Praca końcowa**

**„** **Wykorzystanie modeli uczenia maszynowego do przewidywania końcowego wyniku zespołu w pięciu najlepszych ligach w Europie.”**

Autor: Michał Chruszczewski nr albumu: 83539

Promotor: Mariusz Rafałło

Warszawa 2024

Spis Treści

[Wstęp 3](#_Toc175126047)

[1. Rozdział 1: Dane w sporcie. 5](#_Toc175126048)

[1.1. Wstęp do rozdziału 5](#_Toc175126049)

[1.2. Rys historyczny 5](#_Toc175126050)

[1.3. Gra Liczb 8](#_Toc175126051)

[1.4. Problemy w analizie danych piłkarskich 12](#_Toc175126052)

[1.5. Podsumowanie rozdziału 14](#_Toc175126053)

[2. Rozdział 2: Uczenie maszynowe, regresja logistyczna, drzewa decyzyjne i lasy losowe wykorzystane do badań. 16](#_Toc175126054)

[2.1. Wstęp do rozdziału 16](#_Toc175126055)

[2.2. Budowanie modeli od A do Z 16](#_Toc175126056)

[2.3. Eksploracyjna Analiza Danych 18](#_Toc175126057)

[2.4. Regresja logistyczna 23](#_Toc175126058)

[2.5. Drzewa decyzyjne 25](#_Toc175126059)

[2.6. Las losowy 28](#_Toc175126060)

[2.7. Ocena modeli klasyfikacyjnych 29](#_Toc175126061)

[2.8. Podsumowanie rozdziału 33](#_Toc175126062)

[3. Rozdział 3: Analiza statystyk piłkarskich w 5 najlepszych ligach europejskich. 35](#_Toc175126063)

[3.1. Wstęp do rozdziału 35](#_Toc175126064)

[3.2. Eksploracyjna analiza danych 36](#_Toc175126065)

[3.3. Modele uczenia maszynowego 40](#_Toc175126066)

[3.4. Wnioski płynące z analizy 47](#_Toc175126067)

[3.5. Podsumowanie rozdziału 49](#_Toc175126068)

[Zakończenie 50](#_Toc175126069)

[Bibliografia 51](#_Toc175126070)

[Spis tabel 52](#_Toc175126071)

[Spis rysunków 52](#_Toc175126072)

## Wstęp

Uczenie maszynowe w kontekście piłki nożnej to obszar wciąż niedostatecznie zgłębiony, o czym świadczy niewielka liczba publikacji poświęconych tej tematyce. Jest to zaskakujące, biorąc pod uwagę, że coraz częściej w dyskusjach o futbolu pojawiają się pojęcia takie jak oczekiwane gole czy szanse na dośrodkowanie, a podczas transmisji meczów firmy takie jak Oracle prezentują prognozy wyliczane na podstawie modeli matematycznych, pokazując procentowe szanse na zwycięstwo poszczególnych drużyn. Statystyki stały się nieodłącznym elementem każdej rozmowy o piłce nożnej – zarówno w domowym zaciszu, jak i w telewizyjnych studiach.

Mając na uwadze powyższe, a także będąc od lat fanem futbolu i gier strategicznych, takich jak Football Manager, postanowiłem podjąć wyzwanie i napisać pracę poświęconą analizie piłkarskiej. Temat ten fascynuje mnie także w życiu prywatnym, dlatego mam nadzieję, że realizacja tego projektu pozwoli mi lepiej zrozumieć mechanizmy rządzące sportem. Ponadto, dzięki doświadczeniu zawodowemu w tworzeniu modeli ryzyka w bankowości, chciałem zbudować solidne podstawy teoretyczne w dziedzinie uczenia maszynowego oraz zrealizować własny projekt, aby sprawdzić swoje umiejętności.

Temat mojej pracy, jak sugeruje tytuł, dotyczy przewidywania pozycji zespołów w tabeli na koniec sezonu z wykorzystaniem algorytmów uczenia maszynowego oraz różnych statystyk zgromadzonych w bazie danych. Celem badania jest sprawdzenie, jak dokładnie można przewidzieć ostateczny sukces klubu, opierając się wyłącznie na danych z boiska.

Praca składa się z trzech rozdziałów, z których każdy dotyczy innego aspektu poruszanego tematu. Pierwszy rozdział koncentruje się na analizie danych w sporcie, prezentując krótki rys historyczny oraz najważniejsze trendy w tej dziedzinie. W ostatnim podrozdziale omawiam problemy, które mogą pojawić się przy tworzeniu modeli predykcyjnych w piłce nożnej, stanowiąc jednocześnie wstęp do dalszych rozważań.

Drugi rozdział to podbudowa teoretyczna dla eksploracyjnej analizy danych oraz tworzenia modeli uczenia maszynowego. Na podstawie własnego doświadczenia zawodowego i literatury przedstawiam proces budowy modelu – od jego kodowania, przez testy, aż po wdrożenie. Podkreślam również kluczową rolę eksploracyjnej analizy danych, która stanowi fundament pracy nad modelem.

Ostatni rozdział pracy ma charakter praktyczny. Na podstawie danych ze strony fbref.com analizuję szanse poszczególnych drużyn na znalezienie się w strefie spadkowej, środku tabeli, europejskich pucharach lub grupie walczącej o mistrzostwo. Wykorzystując modele regresji logistycznej, drzewa decyzyjnego oraz lasu losowego, klasyfikuję kluby piłkarskie do odpowiednich grup. Dane użyte w analizie pochodzą z lat 2016-2024 i obejmują pięć najlepszych lig europejskich: angielską, niemiecką, francuską, hiszpańską i włoską.

Projekt ten wymagał nie tylko pracy opisowej, ale również znacznego nakładu pracy związanej z programowaniem i zbieraniem danych. Korzystając ze strony fbref.com, musiałem pobrać i przetworzyć dane za pomocą skryptów w języku Python. Duża część kodu obejmuje również tworzenie modeli uczenia maszynowego. Wszystkie pliki związane z analizą danych stanowią załącznik do niniejszej pracy i są jej istotną częścią.

Mam nadzieję, że lektura mojej pracy dostarczy czytelnikowi podstawowych informacji na temat analizy piłkarskiej i wzbudzi zainteresowanie tym tematem. Uważam, że dziedzina ta jest wciąż niewystarczająco zbadana i ma wiele do zaoferowania.

## Rozdział 1: Dane w sporcie.

### Wstęp do rozdziału

Rozdział pierwszy pracy został poświęcony zagadnieniu danych w sporcie. Zanim zbudowane zostaną modele oraz przeanalizowane dane w celu zbadania problemu postawionego we wstępie należy przyjrzeć się czym w ogóle jest ta dziedzina, jaka jest jej historia i jakie trudności napotyka. Poszczególne podrozdziały opisują najważniejsze zagadnienia związane z analizą danych w sporcie, koncentrując się przede wszystkim na piłce nożnej, ponieważ to ona jest przedmiotem niniejszej pracy.

W pierwszej kolejności opisany został rys historyczny, który pokazuje rozwój tej dziedziny. Następnie w podrozdziale „Gra Liczb” wprowadzono pojęcia związane z analizą danych w sporcie oraz pokazano jej proste przykłady na podstawie różnych danych i grafik pochodzących od dostawcy Opta. Na samym końcu przedstawiono problemy, które wiążą się z omawianą dziedziną. Pozwoli to na lepsze zrozumienie tematu oraz da podbudowę pod analizę dokonaną w ostatnim rozdziale pracy.

### Rys historyczny

Dane w sporcie były wykorzystywane od zarania dziejów (np. do liczenia punktów lub bramek), w pewnym sensie podstawą rywalizacji jest ocenienie, która drużyna była lepsza na podstawie policzalnych faktów, które miały miejsce w czasie zdarzenia sportowego. Jednakże do początków dwudziestego pierwszego wieku, większość ludzi nie zastanawiało się głębiej nad statystykami i ich roli w ulubionej grze. Liczyła się dla nich najważniejsza liczba, która widniała na tablicy wyników. Po części dzięki popularności gier piłkarskich, a w szczególności Footbal Manager (wcześniej Championship Manager) część kibiców zaczęła patrzeć na piłkę nożną przez pryzmat liczb, poszczególne atrybuty zawodników stały się policzalne.[[1]](#footnote-1) Nareszcie można było porównać drybling zawodników, którzy występowali na tej samej pozycji lub skuteczność wślizgów obrońców, a każdy kibic mógł sprawdzić, jak poradziłby sobie, jako menadżer swojego zespołu.[[2]](#footnote-2)

W sporcie do statystycznej rewolucji doszło w latach 90. dwudziestego wieku, kiedy to Michael Lewis zaczął stosować metodę analizy statystycznej zwaną Moneyball.[[3]](#footnote-3) W trakcie trwania wspomnianego dziesięciolecia szybko przeniknęła ona do innych sportów takich jak koszykówka, futbol amerykański, ale i również do piłki nożnej, która jest przedmiotem niniejszej pracy.[[4]](#footnote-4) Dużą rolę w przełomie odegrała organizowana przez Sloan School of Management konferencja Sports Analytics. Po raz pierwszy odbyła się ona w 2006 roku i już w trakcie pierwszej jej edycji poświęcono jeden z paneli futbolowi (albo jak mówią Amerykanie- soccerowi). Nie przyciągnęła ona jednak wielu zainteresowanych, a jedynym dużym klubem, który wysłał na konferencję swoich przedstawicieli była FC Barcelona.[[5]](#footnote-5) Jeden z jej uczestników, Ted Knutson, wykazał się jednak profetyzmem i powiedział: „Kluby piłkarskie wykazują się niezwykłym zaangażowaniem, gdy chodzi o podpisywanie kontraktów z zawodnikami i szkoleniowcami, lecz nie dotyczy to analityków. Ale to się zmieni”.[[6]](#footnote-6) Jak trafnie zauważyli Anderson i Sally w swojej książce „Gra liczb”: „Seven words have long dominated football: That’s the way it’s always been done.” (pol. „Siedem słów od dawna dominowało w futbolu: Tak zawsze to było robione.”)[[7]](#footnote-7) W piłce nożnej panuje niezwykła niechęć do wszelkich zmian, a trenerzy często podejmują decyzje ufając swojej intuicji. Od pierwszej konferencji organizowanej przez MIT minęło niemal 20 lat i każdy poważny klub piłkarski zatrudnia armię pracowników, która analizuje różne dane płynące wprost z boiska, nie ograniczając się tylko i wyłącznie do strzelonych goli, czy asyst, ale skupiając się również na detalach, które na pierwszy rzut oka nie mają wielkiego znaczenia dla przebiegu spotkania.[[8]](#footnote-8) Ma to na celu znalezienie przewag, które pozwolą osiągnąć końcowy sukces, który definiowany jest różnie, czasami jako końcowy triumf, a czasami, jako uniknięcie relegacji. Pokazuje to, że analiza danych mimo tradycjonalizmu piłkarskiego środowiska stała się nieodzownym elementem trenerskiego warsztatu, a jej skuteczność przełamała początkową niechęć.

Jednakże „romans” piłki nożnej i statystyk zaczął się o wiele wcześniej, a chyba jego najbardziej znanym przykładem jest Charle Reep, były pilot RAF, który z kartką w ręce, żmudnie zapisywał poszczególne boiskowe zdarzenia, wyciągając na ich podstawie wnioski.[[9]](#footnote-9) Niestety nie używał on żadnych naukowych metod i jego konkluzje były często zbytnim uproszczeniem, które zatrzymało rozwój angielskiej piłki nożnej na wiele lat.[[10]](#footnote-10) Chyba najbardziej szkodliwą z jego tez było przekonanie o tym, że akcje składające się z małej ilości podań prowadzą do największej liczby bramek, ponieważ piłka transportowana jest od razu w strefę zagrożenia dla przeciwnika[[11]](#footnote-11).

Rys. 1 Dystrybucja podań i przejęć piłki w latach 1953- 1967, autor Charles Reep (źródło: Anderson Ch., Sally D., op. cit., s. 18)

Jak widać na powyższym wykresie (Rys. 1) według Reepa duża liczba podań skutkowała zwiększeniem szans na stratę piłki. Z dzisiejszego punktu widzenia twierdzenie takie jest niezwykle kontrowersyjne, szczególnie w zderzeniu z piłką nożną w wydaniu Pepa Guardioli i innych hiszpańskich szkoleniowców odnoszących sukcesy w licznych turniejach. Wydaje się, że angielski analityk obserwując brytyjski futbol omyłkowo założył, że każde podanie musi być długie, bo zwiększy to szansę drużyny na zdobycie bramki. Zapomniał o takich parametrach, jak dokładność, szybkość operowania piłką itp. Miało to niebagatelny wpływ na angielski futbol, który zaczął słynąć z stylu gry „Kick n’ rush” (pol. „Kopnij i biegnij”).[[12]](#footnote-12) Prace Reepa i innych podobnych mu pasjonatów stanowiły podwaliny dla pracy dzisiejszych statystyków, analityków zajmujących się futbolem zarówno amatorsko, jak i profesjonalnie.

### Gra Liczb

Tytuł tego podrozdziału to zarazem tytuł książki wydanej przez Chrisa Andersona i Davida Sally’ego w 2013 roku, świetnie oddający czym staje się piłka nożna w XXI wieku.[[13]](#footnote-13) Coraz więcej firm konsultingowych zaczyna skupiać się tylko i wyłącznie na piłce nożnej dostarczając danych, paliwa do wszelkich analiz i wykorzystania ich w zdobyciu przewagi nad przeciwnikiem (najbardziej znaną z nich jest Opta).[[14]](#footnote-14) Stało się to niemal pewnym rytuałem, że w trakcie transmisji sportowych, czy też w pomeczowym studio, eksperci powołują się na różne, mniej i bardziej wyszukane statystyki, próbując wytłumaczyć, dlaczego jeden zespół osiągnął przewagę nad drugim lub próbując przewidzieć wynik spotkania, czy też końcową pozycję danej drużyny w lidze.

Zanim zaczniemy głębiej wchodzić w świat liczb związanych z piłką nożną, w pierwszej kolejności należy stworzyć pewną siatkę pojęciową, która bezpośrednio odnosi się do omawianych zjawisk. Najbardziej podstawowym terminem jest „analiza”, której znakomitą definicję za angielską Wikipedią cytują Anderson i Sally w swojej książce: “Analysis is the process of observing and breaking down a complex topic or substance into smaller parts to gain a better understanding of it.” (pol. „Analiza to proces obserwowania i rozkładania złożonego tematu lub substancji na mniejsze części w celu lepszego zrozumienia go.”)[[15]](#footnote-15) Oddaje to również istotę analizy w kontekście piłki nożnej, gdzie obserwatorzy starają się rozłożyć na czynniki dość skomplikowane wydarzenie jakim jest mecz piłki nożnej w celu zbudowania przewagi lub wyjaśnienia boiskowych zdarzeń. Jak dopowiadają Anderson i Sally: „(…) is much more than just spreadsheets and statistics: it is an openness to data and information of all kinds- formal, informal, categorized, disorganized, observed, recorded, remembered etc. – and it is a determination to find whatever truth, patterns and correspondences they may contain” (pol. "(…) to znacznie więcej niż tylko arkusze kalkulacyjne i statystyki: to otwartość na dane i informacje wszelkiego rodzaju - formalne, nieformalne, skategoryzowane, niezorganizowane, obserwowane, rejestrowane, zapamiętywane itd. - oraz determinacja, by znaleźć jakąkolwiek prawdę, wzorce i korelacje, które mogą zawierać.")[[16]](#footnote-16) W ramach pojęcia analizy w sporcie należy wyróżnić jej dwa typy:

Analiza sportowa

Analiza pozaboiskowa

Analiza boiskowa

Rys. 2. Rodzaje analizy w sporcie (oprac. własne na podst. https://en.wikipedia.org/wiki/Sports\_analytics)

W niniejszej pracy skupimy się przede wszystkim na analizie boiskowej (tzn. statystyk płynących z meczów), pozostawiając to, co dzieję się poza meczami na marginesie. Niemniej jednak w tym miejscu należy wspomnieć, że często wyjaśnieniem wydarzeń boiskowych są wydarzenia dziejące się w gabinetach prezesów lub na trybunach.

Na podstawie przywołanej wcześniej definicji analizy, kluczem jest zbadanie danych i informacji pochodzących z wydarzeń sportowych. W związku z tym warto poznać znaczenie obu tych pojęć. Pierwsze z nich to „zbiory wartości, które przekazują informacje, opisując ilość, jakość, fakt, statystyki, inne znaczenia lub sekwencje symboli, które mogą być dalej interpretowane i przetwarzane.”[[17]](#footnote-17) W kontekście piłki nożnej będą to wszelkie statystyki, które dotyczą meczu, zawodników, trenerów itd. Do danych zaliczymy na przykład liczbę strzałów, liczbę podań, wartość zawodnika czy liczbę wygranych meczów przez trenera. Jeśli chodzi o informację to definiuje się jako „(…) właściwość pewnych obiektów, relacja między elementami zbiorów pewnych obiektów, której istotą jest zmniejszanie niepewności (nieokreśloności)”[[18]](#footnote-18). Będzie to np. wyciągnięty wniosek na podstawie jakiejś statystyki. Dobrym przykładem podstawowej informacji płynącej z futbolowych statystyk jest wyłonienie zwycięzcy meczu na podstawie strzelonych bramek. Stanowi podstawę jakiekolwiek analizy piłkarskiej i stał się podstawą warsztatu trenerskiego menadżerów na całym świecie. W Niemczech, nieco ironicznie, takich szkoleniowców nazywa się „laptopowymi trenerami”[[19]](#footnote-19) ze względu na to, że nie rozstają się oni ze swoimi komputerami i tabletami. Jednakże mimo nieco prześmiewczego tonu większość komentatorów i dziennikarzy zdaje sobie sprawę z tego, że dane i analiza stały się nieodzownym elementem piłki nożnej i stanowią niezwykle ważny składnik sukcesu wielu drużyn na całym świecie. Nawet tak relatywnie niewielkie kluby, jak FC Koln zatrudniają kilku pełnoetatowych i kilkudziesięciu niepełno etatowych analityków[[20]](#footnote-20) do rozpracowywania rywali i wgryzania się w szczegóły własnych statystyk.

Dostarczycielem paliwa dla wszelkich analiz, czyli danych, stały się wyspecjalizowane firmy (np. Opta) które w ciągu jednego meczu potrafią dostarczyć ponad dwa tysiące różnych danych (podań, strzałów, itp.), rejestrowanych co jedną sekundę w trakcie trwania spotkania.[[21]](#footnote-21) Stało się to możliwe dzięki rozwojowi Big Data, który to pozwolił rejestrować niespotykane dotąd ilości danych, co stało się podstawą do ich istnienia Opta.[[22]](#footnote-22) Wzrost liczby danych na przestrzeni lat obrazuje poniższy wykres, dostarczony przez driblab. Pokazuje on jak zmieniała się liczba zawodników w różnych bazach danych na przestrzeni lat. Widać niemal dziesięciokrotny wzrost w liczbie gromadzonych danych pomiędzy rokiem 2016 i 23.

Rys. 3. Liczba zawodników w bazach danych (źródło: <https://www.driblab.com/driblab-en/football-data-the-largest-coverage-in-the-market/>)

Wystarczy rzut oka na statystyki, które dostarczane są przez np. Opta, żeby zobaczyć, jak zaawansowanych analiz można dokonać. Dane takie, jak podania w ostatnią tercję boiska, czy też liczba podań ofensywnych potrafią rzucić nowe światło na mecz. Poniżej przykład zaawansowanych statystyk dostarczanych przez Opta dla potrzeb analiz pomeczowych w angielskiej telewizji Sky Sports.

A diagram of a match

Description automatically generated

Rys.4. Analiza podań Liverpoolu, Man City i Chelsea (źródło: https://m.allfootballapp.com/news/EPL/EPL-club-styles-revealed-in-advanced-Opta-stats-City-dominate-attack-matters/1743637)

Nawet z tak prostej grafiki, jak powyższa możemy wyciągnąć kilka wniosków odnoszących się do stylu gry poszczególnych drużyn. Widać, że Man City starannie konstruował akcje i grał futbol oparty na posiadaniu piłki. Natomiast Liverpool, który w tamtym sezonie został mistrzem Anglii[[23]](#footnote-23), preferował bezpośrednie ataki, wykonywał mniej podań w trakcie trwania akcji. Co ciekawe mimo tego, że Man City stworzył w tamtym sezonie najwięcej sytuacji dogodnych do strzelenia bramki to drużyna z miasta Beatles’ów została koronowana, jako najlepsza w tamtym sezonie.[[24]](#footnote-24) Jest to dowód na to, że ważnym elementem piłki nożnej nadal jest szczęściem, a używanie statystyk i dogłębna analiza danych zwiększa szanse na sukces, ale nie jest jego gwarancją.

Jak zostało wspomniane w niniejszej pracy, nie tylko profesjonaliści zaczęli większą uwagę przykładać do danych. Również zwykli kibice coraz częściej w swoich dyskusjach odnoszą się do różnych statystyk, które udostępniane są nieodpłatnie przez takie serwisy, jak sofascore.com, czy fbref.com. Mogą oni znaleźć tam wiele szczegółów dotyczących ich ulubionych drużyn odnoszących się zarówno do gry defensywnej, jak i ofensywnej poszczególnych zespołów.[[25]](#footnote-25) Podnosi to jakość dyskusji na tematy piłkarskie, a także zmusza ekspertów do posługiwania się liczbami w swoich analizach, aby zadowolić najbardziej wybrednych widzów.

### Problemy w analizie danych piłkarskich

Mimo licznych korzyści płynących z analizy danych w piłce nożnej, należy pamiętać o pewnych problemach, które się z nią wiążą. Doskonale opisał je Christoph Biermann w swojej książce „Piłkarscy Hakerzy. O rewolucji w futbolu i sztuce zbierania danych”. Pokazał on na konkretnych przykładach jakie pułapki czekają na analityków piłkarskich, przypominając o tym, że oprócz twardych danych, futbolem w równej mierze rządzą niepoliczalne czynniki takie jak wola walki, czy też łut szczęścia. Chyba najlepszym tego przykładem jest mecz pomiędzy FC Barcelona, a Celticiem Glasgow, który odbył się 7 listopada 2012 roku, a jego statystyki wyglądały następująco:

|  |  |  |
| --- | --- | --- |
| Celtic F.C. - Wikipedia |  | FC Barcelona - Wikipedia |
| 25 % | Posiadanie piłki | 75% |
| 5 | Sytuacje bramkowe | 23 |
| 5 | Strzały na bramkę | 14 |
| 303 | Podania | 999 |
| 44% | Procent celnych podań | 87% |

Tabela 1. Statystyki meczu pomiędzy Celticiem Glasgow a FC Barcelona z dn. 7 listopada 2012 roku (oprac. własne na podst. https://www.flashscore.pl/mecz/rLLV8Mnm/#/szczegoly-meczu/statystyki-meczu/0)

Na podstawie suchej analizy danych meczowych i nie znając wcześniej końcowego wyniku bez wątpienia wskazalibyśmy FC Barcelonę, jako zwycięzcę tego pojedynku. Mimo ogromnej przewagi Blaugrana przegrała ten mecz 2-1, pokazuje to, że pojedynczy wynik meczu często zależy od szczęścia, a najlepsza drużyna nie zawsze wygrywa. W ostatecznym rozrachunku to FC Barcelona wygrała grupę Ligi Mistrzów i doszła, aż do półfinału ulegając późniejszemu zwycięzcy całych rozgrywek- FC Bayernowi.[[26]](#footnote-26) Ostateczna tabela Grupy G, w której doszło do starcia Dumy Katalonii oraz Celticu wyglądała następująco:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Msc. | Drużyna | Mecze | Zwycięstwa | Remisy | Porażki | Różnica bramek | Punkty |
| 1. | FC Barcelona | 6 | 4 | 1 | 1 | +6 | 13 |
| 2. | Celtic Glasgow | 6 | 3 | 1 | 2 | +1 | 10 |
| 3. | SL Benfica | 6 | 2 | 2 | 2 | 0 | 8 |
| 4. | Spartak Moskwa | 6 | 1 | 0 | 5 | -7 | 3 |

Tabela 2. Grupa G Ligi Mistrzów 2012-13

(źródło https://pl.wikipedia.org/wiki/Liga\_Mistrzów\_UEFA\_(2012/2013))

Niemniej jednak, zdarza się, że również na dystansie całych rozgrywek zespół, który gra statystycznie gorzej od rywali zdobywa tytuł. Dowodem na to jest chociażby przywoływany w niniejszej pracy przykład sezonu 2019/2020 w angielskiej Premier League, kiedy to statystycznie najlepszym zespołem był Manchester City, a mistrzem został Liverpool FC. Jest to kolejny dowód na to, że statystyki nie zawsze oddają pewny nieuchwytny element w futbolu, który czasami nazywany jest szczęściem, czasami wolą walki i zwycięstwa. Jeden z analityków reprezentacji Niemiec, po zakończonych sukcesem Mistrzostwach Świata w 2014 roku, pokusił się nawet o stwierdzenie, że: „Jako analitycy przestaliśmy zwracać uwagę na niemal wszystkie parametry, którym przyglądaliśmy się w poprzednich latach, poczynając od czasu spędzonego przy piłce po liczbę podań do przodu. Jesteśmy coraz bardziej przekonani, że brakuje nam danych, które dostarczają prawdziwych informacji na temat rzeczy faktycznie decydujących o sukcesie w piłce nożnej”.[[27]](#footnote-27) Jest to hipoteza, która nie przetrwała próby czasu, ponieważ kluby zaczęły przeznaczać coraz większe sumy na analizę danych, algorytmy uczenia maszynowego itp., aby zwiększyć swoje szanse na końcowy sukces, ale obrazuje dobrze, że czasami sukces wynika z pewnego składnika, którego w żaden sposób nie oddają skrupulatnie odnotowywane statystyki.

Oprócz przytoczonego wcześniej problemu z niepoliczalnym czynnikiem w piłce nożnej, istnieje szereg bardziej przyziemnych faktorów, które wpływają negatywnie na analizę danych. Jedną z głównych przeszkód jest niejednolitość statystyk zbieranych przez automatyczne systemy. Przez to często dane muszą być weryfikowane przez człowieka, który ręcznie zlicza np. wykonane podania, czy oddane strzały. Uwidoczniło się to chociażby Mistrzostw Świata w Rosji, kiedy to różni dostawcy danych dostarczali rozbieżne informacje o przebiegniętych przez zespoły kilometrach.[[28]](#footnote-28) Stanowi to niemały problem w analizie, ponieważ uniemożliwia rzetelnej przeprowadzenie ze względu na to, że nieprecyzyjne dane mogą prowadzić do błędnych wniosków.

### Podsumowanie rozdziału

Analiza danych sportowych, a w szczególności tych futbolowych jest niezwykle wielowątkowym i skomplikowanym zagadnieniem, które od wielu dziesięcioleci zajmuje ludzi na całym świecie. Od czasu pionierów takich, jak Charles Reep minęło wiele lat, a technologia niezwykle poszła do przodu pozwalając na zbieranie różnorodnych statystyk, które tworzą dokładny obraz sportowych zmagań. Dzięki firmom takim, jak Opta futbol z roku na rok staje się coraz bardziej „Grą Liczb”, czyli policzalnym zdarzeniem, które składa się z wielu czynników takich, jak strzały, podania i akcje defensywne. Niemniej jednak, mimo coraz większej policzalności piłki nożnej w dalszym ciągu szczęście odgrywa istotną rolę w ostatecznym wyniku, a drużyny mogą wygrać tytuł swoją wolą walki tak, jak Liverpool w 2020 roku.

Na podstawie danych piłkarskich z fbref.com zostanie dokonana analiza statystyk oraz zostaną stworzone modele uczenia maszynowego, wyjaśniające, jakie wydarzenia meczowe są najbardziej istotne dla powodzenia zespołu i czy w ogóle da się na ich podstawie dokonać takich prognoz. Wcześniej jednak należy zapoznać się z metodami i algorytmami, które zostaną do tego wykorzystane, co opisane zostało w rozdziale drugim niniejszej pracy.

## Rozdział 2: Uczenie maszynowe, regresja logistyczna, drzewa decyzyjne i lasy losowe wykorzystane do badań.

### Wstęp do rozdziału

Po poznaniu pojęć związanych z analizą danych w sporcie, należy podjąć próbę odpowiedzenia na pytania zawarte we wstępie do niniejszej pracy. Nie jest to jednak możliwe bez zagłębienia się w koncepcje związane z eksploracyjną analizą danych oraz uczeniem maszynowym, ponieważ to one pozwolą nam na zbadanie statystyk pochodzących z kilku ostatnich sezonów pięciu najlepszych lig europejskich.

Ze względu na to, że problem badawczy dotyczy klasyfikacji wieloklasowej, tzn. które miejsce zajmie drużyna na koniec danego sezonu mając określone statystyki, to do przeprowadzenia analiz zostały użyte modele oparte na metodach klasyfikacji: regresja logistyczna, drzewa decyzyjne oraz lasy losowe. Zanim jednak przejdziemy do ich opisu należy skupić się na ścieżce, którą należy przejść, aby zbudować model uczenia maszynowego.

### Budowanie modeli od A do Z

W pierwszej kolejności, zanim przejdziemy do szczegółów tworzenia modeli, należy nakreślić drogę, po której każdy programista, *data scientist* lub inna osoba zajmująca się uczeniem maszynowym, musi przejść, aby skutecznie zbudować las losowy, czy też sztuczną sieć neuronową. Zasadniczo wyróżnia się pięć etapów tworzących ten proces[[29]](#footnote-29):

1. Przygotowanie danych
2. Eksploracyjna Analiza Danych
3. Tworzenie modeli
4. Testowanie i walidacja modeli
5. Iteracyjne ulepszanie modeli

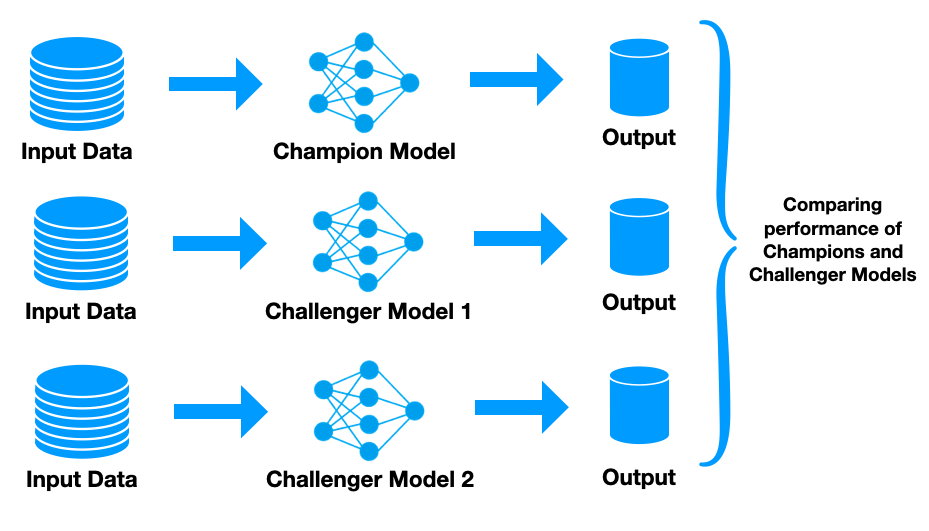
Kroki te są niezbędne do stworzenia nowego modelu i implementowane są w niemal każdej organizacji, która specjalizuje się taką działalnością. W tym rozdziale omówione zostaną pierwszy i ostatni z nich, ponieważ osobne podrozdziały poświęcone zostaną Eksploracyjnej Analizie Danych oraz tworzeniu i walidacji modeli.

Przygotowanie danych jest niezbędnym, wstępnym etapem jakiejkolwiek analizy[[30]](#footnote-30), polega ono przede wszystkim na ich przekształceniu w taki sposób, aby możliwe było zbudowanie algorytmu uczenia maszynowego.[[31]](#footnote-31) Nie jest to katalog enumeratywny, ale może ono się składać z:

1. Wybrania sposobu radzenia sobie z brakującymi rekordami
2. Kodowanie zmiennych kategorialnych
3. Zmiana typów zmiennych
4. Transformacja zmiennych, aby bardziej odpowiadały rozkładowi normalnemu
5. Skalowanie zmiennych

Wszystkie te działania nakierowane są na takie uporządkowanie zbioru danych, aby łatwo można było załadować go do modelu i przeprowadzić wyczerpującą eksploracyjną analizę danych.

Ostatnim krokiem w tworzeniu modelu jest jego iteracyjne ulepszanie. Jednym z podejść jest stosowanie schematu *Champion- Challenger[[32]](#footnote-32).* Polega on na tym, że w fazie utrzymania modelu (tzn. już po jego testowaniu i oddaniu na produkcję), tworzy się nowe algorytmy zmieniające niektóre parametry (np. ilość warstw ukrytych w sieci neuronowej), aby polepszyć jakość prognoz. Gdy nowy model (*challenger*) zostanie utworzony porównuje się jego wyniki z dotychczas używanym (*champion*). W momencie, kiedy *challenger* uzyska np. lepszą czułość i jego skuteczność zostanie oceniona wyżej, staje się nowym *championem.* W praktyce wygląda to tak, że modele przechodzą okresową walidację (np. raz na pół roku) podczas, której inżynierowie zmieniają ich parametry lub tworzą je od zera (stosując inne algorytmy). Po ewaluacji nowy model jest wdrażany na produkcję. Poniższy schemat dobrze oddaje istotę tegoż procesu (Rys. 5.):



Rys. 5 Schemat *Champion-Challenger* w modelowaniu (źródło: https://deeplearninguniversity.com/mlops/mlops-model-testing/)

### Eksploracyjna Analiza Danych

Jak zostało wspomniane w pierwszym podrozdziale, tuż po przygotowaniu danych do tworzenia modelu należy je dogłębnie zbadać w celu dostrzeżenia podstawowych zależności i sprawdzenia, co tak naprawdę jesteśmy w stanie z nich wyczytać. Służy temu eksploracyjna analiza danych.[[33]](#footnote-33) W kolejnych akapitach tego podrozdziału przejdę przez konkretne metody, którymi posługuje się ta dziedzina. Zostały one w większości zastosowane również do przeprowadzenia analiz opisanych w rozdziale 3 pracy.

#### Miary położenia

Pierwszym krokiem eksploracyjnej analizy danych jest zazwyczaj określenie wokół jakich wartości skupione są dane numeryczne. W tym celu wykorzystuje się miary położenia, a najbardziej podstawową z nich jest średnia arytmetyczna. Definiuje się ją, jako iloraz sumy zaobserwowanych wartości zmiennej mierzalnej przez liczbę obserwacji.[[34]](#footnote-34) Jej wzór wygląda następująco:

gdzie:  
𝑥̅ – symbol średniej arytmetycznej, 𝑥i – warianty cechy mierzalnej,  
n – liczebność badanej zbiorowości.

Jednak ze względu na to, że zwykła średnia arytmetyczna jest nieodporna na wartości skrajne, stosuje się jej dwa warianty: średnią ważoną oraz średnią ucinaną.[[35]](#footnote-35) Pierwsza z nich jest ilorazem sumy iloczynów określonych wartości oraz przypisanych im wag i sumy tych wag. Wyraża się ją wzorem:

gdzie:  
𝑥̅ – symbol średniej ważonej, 𝑥i – warianty cechy mierzalnej,  
*n* – liczebność badanej zbiorowości, *w*- przypisane wagi

W ten sposób niektórym wartościom można przypisać większe znaczenie, niż innym. W przypadku średniej ucinanej, do obliczenia średniej pomijamy pewną liczbę wartości skrajnych. Ma ona wzór:

gdzie:  
𝑥̅ – symbol średniej ucinanej, 𝑥i – warianty cechy mierzalnej,  
*n* – liczebność badanej zbiorowości, *p*- liczba odciętych wartości skrajnych

Oprócz trzech wymienionych wcześniej średnich można zbadać również medianę, która jest estymatorem odpornym.[[36]](#footnote-36) Oznacza to, że nie jest ona podatna na duże odchylenie wartości skrajnych, ponieważ wyznacza wartość środkową.

#### Miary rozproszenia

W następnej kolejności, po ustaleniu wokół jakich wartości skupione są dane należy sprawdzić, czy są one bardzo rozproszone i potrzebują np. normalizacji lub standaryzacji. Najczęściej używanymi miarami rozproszenia są odchylenie standardowe oraz wariancja, które określają średnią odległość wartości od średniej arytmetycznej.[[37]](#footnote-37) Pomagają one zbadać, czy dane są bardzo rozproszone (np. średnia z liczb 1 i -1 oraz 1000 i -1000 jest taka sama, natomiast w drugim przykładzie odległość wartości od średniej jest znacznie większa). Wariancja jest średnią kwadratów odchyleń każdej wartości od średniej. Wyraża ją wzór:

gdzie:  
 – symbol wariancji, *Xi* – pojedyncza wartość,  
*n* – liczebność badanej zbiorowości, - średnia aretmetyczna

Odchylenie standardowe jest natomiast pierwiastkiem kwadratowym z wariancji:

gdzie:  
 *s*– symbol odchylenia standardowego, – wariancja

Częściej używa się odchylenia standardowego, ponieważ jak można zauwazyć ma ono tę samą skalę, co oryginalne dane, ułatwia to intepretację wyników.

#### Rozkład danych

Pracując nad danymi można również zbadać rozkład danych. Najlepiej nadają się do tego różne wykresy. W pracy skupię się na tych, które zostały użyte do badania danych w rozdziale 3. i są to wykresy pudełkowe oraz histogramy. Tworzy się je w celu zbadania próbki i ocenienia w jaki sposób rozkładają się wartości (np. czy jakaś wartość, albo grupa wartości mają przewagę nad innymi).

Wykres pudełkowy przedstawia wartość minimalną, maksymalną, medianę oraz pierwszy i trzeci kwartyl danych[[38]](#footnote-38) . W wariancie złożonym, zamiast wartości minimalnej oraz maksymalnej znajdziemy wartości Q3 + 1,5 \* rozkład międzykwartylowy (max) oraz Q1 – 1,5 \* rozkład międzykwartylowy (min), prezentowane są również wartości odstające. Wykresy pudełkowe w wersji prostej (Rys. 6) i złożonej (Rys. 7) wyglądają następująco:

A diagram of a rectangular object with red and blue text

Description automatically generated

Rys. 6 Wykres pudełkowy prosty (źródło:

https://mfiles.pl/pl/index.php/Wykres\_pudełkowy)

A diagram of a diagram

Description automatically generated

Rys.7 Wykres pudełkowy złożony (źródło:

https://mfiles.pl/pl/index.php/Wykres\_pudełkowy

Na obu wykresach możemy znaleźć mnóstwo informacji na temat rozkładu danych i ocenić, czy należy podjąć działania w celu ich normalizacji lub standaryzacji.

Kolejnym sposobem sprawdzenia rozłożenia zbioru są histogramy, które pokazują liczebność wartości w próbie. Na osi x prezentowane są poszczególne wartości, oś y natomiast pokazuje ich liczebność.[[39]](#footnote-39) Przykładowy histogram wygląda następująco (Rys. 8.):

Rys. 8. Histogram podstawowy (źródło: https://mfiles.pl/pl/index.php/Histogram)

Wykres ten, podobnie jak wykres pudełkowy, pomaga zrozumieć, czy dalsze przekształcenia wartości są konieczne, aby móc ją użyć w tworzeniu modelu. Szczególnie przydatny jest on w sprawdzeniu skośności rozkładu, czyli ocenieniu asymetrii (czy więcej wartości występuje po lewej lub prawej stronie osi x).[[40]](#footnote-40)

#### Korelacja

Ostatnią metodą eksploracyjnej analizy danych opisaną w pracy jest korelacja. Definiuje się ją, jako „współzależność liniową zmiennych losowych”.[[41]](#footnote-41) Obrazuje ona w jaki sposób dane kolumny (parametry) w danych są między sobą powiązane. Korelacja może być dodatnia (gdy jedna zmienna rośnie to druga także) lub ujemna (gdy jedna zmienna maleje to druga rośnie), a jej zakres zamyka się między -1 i 1, gdzie 0 oznacza brak korelacji, a wartości skrajne (-1,1) reprezentują doskonałą korelację ujemną lub dodatnią. Jej wzór wygląda następująco:

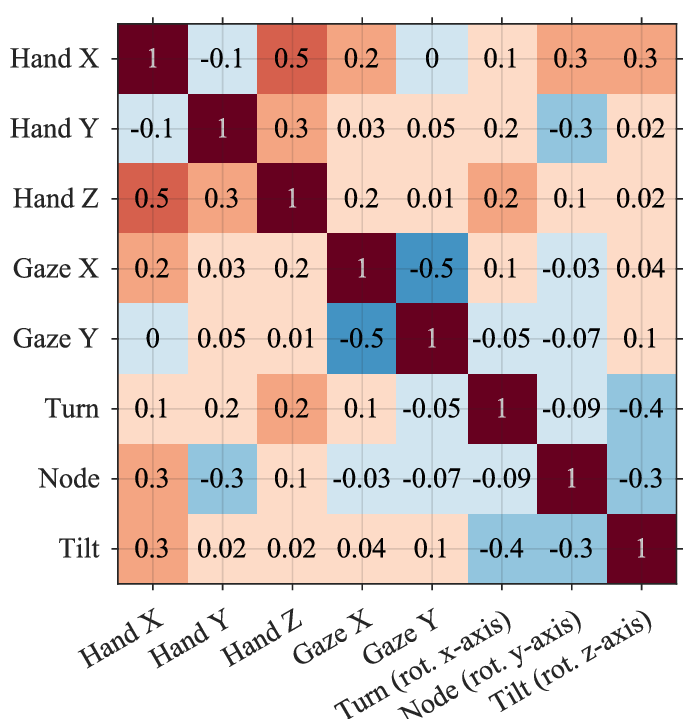
1. Dla zmiennych liczbowych (korelacja Pearsona)

gdzie:  
 – symbol korelacji, *E*– wartość oczekiwana,  
 – odchylenie standardowe z populacji

1. Dla zmiennych kategorialnych (korelacja rang Spearmana)

gdzie:  
 – symbol korelacji, - liczba par danych  
– kwadraty różnic pomiędzy rangami odpowiadających sobie wartości cech i

Na podstawie obliczonych wartości korelacji rysuje się matrycę, która prezentuje zależności pomiędzy wybranymi lub wszystkimi zmiennymi.



Rys. 9. Matryca korelacji (źródło: https://www.researchgate.net/figure/An-example-of-a-correlation-matrix-Each-entry-in-the-correlation-matrix-is-a-Pearsons\_fig1\_366861559)

### Regresja logistyczna

Pierwszy model, który został wykorzystany do analizy danych pochodzących ze strony fbref.com to regresja logistyczna. Zanim przejdziemy do wyjaśnienia tego konkretnego typu, należy wyjaśnić czym jest regresja. Definuje się ją, jako metodę mającą na celu zbadanie siły relacji pomiędzy zmiennymi objaśniającymi oraz objaśnianą.[[42]](#footnote-42) Jej celem jest sprawdzenie, jak historycznie kształtowała się ta zależność i na tej podstawie stworzenie modelu predykcyjnego, który będzie w stanie przewidzieć ją dla nowych danych.

Regresja logistyczna koncentruje się na analizie szans, czyli stosunku liczby sukcesów do liczby porażek. W praktyce oznacza to, że jeśli prawdopodobieństwo, iż dana zmienna przyjmie wartość 1, wynosi 0,5, i jest równe prawdopodobieństwu przyjęcia wartości 0, to szansa na wystąpienie jednego z tych zdarzeń wynosi 1. Co do zasady składa się ona z dwóch części: logitu oraz odpowiedzi funkcji logistycznej. Pierwszy z elementów mapuje dowolną wartość z przedziału ( do prawdopodobieństwa (0,1)[[43]](#footnote-43). Natomiast odpowiedź funkcji logistycznej jest to odwrotność logitu, która pozwala na pozostanie w zakresie (0,1), czyli de facto sprowadza wynik funkcji do wartości binarnej, wyraża się ją równaniem:

gdzie:  
 – prawodpobieństwo,

*e* – stała Eulera  
 – wyraz wolny

– współczynnik regresji

– wartość zmiennych objasniających

Ze wzgledu na swój kształt, funkcja logarytmiczna często nazwana jest sigmoidalną. Poniższy wykres prezentuje jej najbardziej typowy wygląd (Rys. 10).

Rys. 10 Przykład wykresu funkcji logistycznej dla wartości (-10,10) (oprac. własne)

Najbardziej podstawowy typ regresji logistycznej przyjmuje wartości 0 lub 1, jednak ze względu na to, że w niniejszej pracy rozwiązywany jest problem, który dotyczy bardziej skompliowanej klasyfikacji (liczby całkowite od 1 do 20), należy wprowadzić pewną modyfikcaję i omówić również podtyp regresji logistycznej jakim jest wielomianowa regresja logistyczna. W przeciwieństwie do podstawowego typu, za jej pomocą można dokonać predykcji zmiennej zależnej, która jest nominalna, ale nie dychotomiczna (binanrna).[[44]](#footnote-44) W związku z powyższym idealnie nadaje się ona do przewidywania zmiennych, które przyjmują wartości inne, niż (0,1), a jednocześnie w założeniu są one kategorialne. W uproszczeniu wygląda to tak, że tworzony jest układ równań regresji logistycznej, które to równania odnoszą się do poszczególnych możliwych stanów zmiennej objaśnianej. Następnie na podstawie zmiennych niezależnych przewiduje się prawodpodobieństwo na uzyskanie danego stanu i porównuje wyniki poszczególnych równań.[[45]](#footnote-45)

Podsumowując, regresja logistyczna jest prostym sposobem tworzenia modeli predykcyjnych, których celem jest przewidywanie zmiennych zależnych, kategorialnych. W niniejszej pracy, ze względu na charakter badań, został użyty model regresji wielomianowej, pozwalający na przewidywanie zmiennych niebinarnych. W następnej kolejności opisane zostaną modele drzewa decyzyjnego oraz las losowy.

### Drzewa decyzyjne

Kolejnym algorytmem, który został wykorzystany do modelowania w rozdziale trzecim pracy są drzewa decyzyjne. Definiuje się je jako: „graficzną prezentację podziału przestrzeni m wymiarowej na rozłączne fragmenty (segmenty przestrzeni, wielowymiarowe kostki)”[[46]](#footnote-46). Z pojęciem drzew decyzyjnych łączy się szereg definicji oraz procesów, które należy poznać, aby w pełni zrozumieć charakter oraz sposób działania tych algorytmów. Zanim jednak to nastąpi, warto na wstępie dodać, że w kontekście tej pracy używany jest algorytm CART (*Classification and Regression Trees*; pol. Drzewa Klasyfikacyjne oraz Regresyjne), które jest odmianą drzew decyzyjnych używanym powszechnie w statystyce oraz uczeniu maszynowym.[[47]](#footnote-47) Jest to algorytm tworzący drzewa decyzyjne, które służą do klasyfikacji i regresji, polegający na rekurencyjnym dzieleniu danych na mniejsze podzbiory w sposób maksymalizujący jednorodność klas lub minimalizujący błąd w każdej podgrupie.

Kluczowym elementem w tworzeniu drzew decyzyjnych jest podział rekursywny, czyli wielokrotnie powtarzany podział danych, którego celem jest stworzenie maksymalnie jednorodnych grup.[[48]](#footnote-48) Podział odbywa się na podstawie kryterium podziału, tzn. wartości predyktora dzielącej zbiór na dwie części. Takie podziały tworzy się najczęściej na podstawie nieczystości Giniego. Jest ona miarą tego, jak często dany element zestawu danych byłby niepoprawnie zaklasyfikowany, gdyby jego klasyfikacji dokonywano by losowo, zgodnie z rozkładem klas w danym zbiorze[[49]](#footnote-49). Wyraża go równanie:

gdzie:  
– nieczystość Giniego,

– to proporcja elementów należących do klasy i  
 – liczba klas

Inną, często wykorzystywaną miarą jest entropia, czyli miara niepewności związanej z przewidywaniem klasy, wskazująca na to, jak chaotycznie rozłożone są klasy w danym zbiorze danych. Im wyższa entropia, tym większa niepewność i większa różnorodność klas w zbiorze. Wyraża ją równanie:[[50]](#footnote-50)

gdzie:  
– entropia w zbiorze danych S,

– prawdopodobieństwo wystąpienia danej klasy

Na podstawie nieczystości Giniego lub entropii, dokonuje się w kolejnych węzłach oceny tego, do której klasy przypisać podzbiór danych. Na samym końcu znajdują się liście drzewa, które są zakończeniem zbioru reguł „jeśli to” mających na celu, jak najlepszą klasyfikację danych.[[51]](#footnote-51) Oczywiście nie jest to koniec tworzenia drzewa klasyfikacyjnego, ponieważ kluczowym elementem jest również przycinanie mające na celu zredukowanie nadmiernego dopasowania algorytmu do danych treningowych.[[52]](#footnote-52) Polega ono na wstecznym redukowaniu węzłów i gałęzi drzewa do określonej głębokości (np. 4 poziomów), aby tak stworzony model można było użyć do przetestowania na danych walidacyjnych, a w następnej kolejności do predykcji wartości zmiennej objaśnianej w zupełnie nowym zbiorze danych. Przykładowa reprezentacja graficzna drzewa decyzyjnego wygląda następująco (Rys. 11.)

A diagram of a graph

Description automatically generated

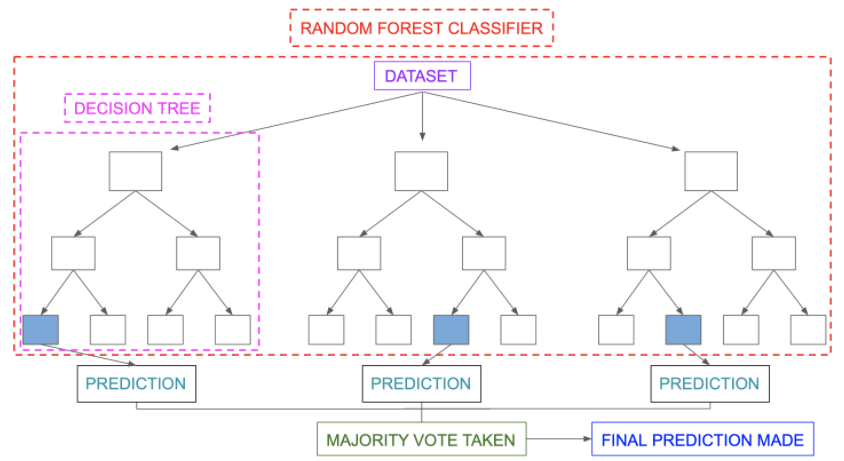
Rys. 11. Drzewo decyzyjne- Wynik meczu na podstawie goli strzelonych i straconych (oprac. własne – kod python)

Powyżej zaprezentowano drzewo składające się tylko z dwóch węzłów, w praktyce może być ich znacznie więcej. To samo tyczy się możliwej liczby stanów zmienne objaśniającej, w tym przykładzie występują tylko dwa (przegrana, wygrana), natomiast w modelu stworzonym w rozdziale 3 jest ich 20 (miejsca w tabeli od 1 do 20).

Podsumowując, drzewa decyzyjne to jeden z modeli uczenia maszynowego umożliwiający przewidywanie np. klas. Jego graficzną reprezentacją jest diagram składający się z węzłów, gałęzi oraz liści, które dzielą zbiór danych na coraz mniejsze części, używając do tego różnych miar dopasowania (np. entropii lub nieczystości Giniego). W następnym rozdziale opiszę las losowy, który jest rozwinięciem drzew decyzyjnych i teoretycznie pozwala na stworzenie bardziej dokładnych modeli predykcyjnych.

### Las losowy

Ostatnim algorytmem uczenia maszynowego, który został użyty do analizy danych w rozdziale trzecim jest las losowy. Jak zostało wspomniane w poprzednim rozdziale jest to poniekąd rozwinięcie drzew decyzyjnych, ponieważ do jego budowy wykorzystuje się wiele takich modeli, aby jak najlepiej przewidzieć zmienną objaśnianą. Kluczowym pojęciem związanym z tym algorytmem jest *bagging* (skrót od *bootstrap aggregating*). Jest to model opracowany przez L. Breimana w 1996 r.[[53]](#footnote-53), opierający się na dwóch ideach *bootstraping-*u, czyli tworzeniu wielu zestawów danych poprzez losowe próbkowanie z zastępowaniem z oryginalnego zestawu danych oraz agregacji, która łączy powstałe w wyniku *bootstraping*-u modele.[[54]](#footnote-54) W lasach losowych w każdej iteracji losowana jest określona liczba zmiennych oraz rekordów na podstawie, których budowany jest model drzewa decyzyjnego. Następnie próba ta zwracana jest do zbioru.[[55]](#footnote-55) Te sam zestaw danych, w różnych konfiguracjach może być losowany wielokrotnie. Co do zasady przyjęło się, że każde drzewo powinno zawierać √P zmiennych użytych do jego tworzenia, gdzie P to ogólna liczba zmiennych objaśniających. Na samym końcu wyniki są agregowane w taki sposób, aby model zwracał jak najdokładniejszą wartość zmiennej zależnej. Jest to jego największa przewaga nad algorytmem drzewa decyzyjnego, ponieważ ostateczny wynik modelu jest wypadkową wielu różnych konfiguracji zmiennych i rekordów. Schematycznie las losowy wygląda tak, jak został przedstawiony poniżej (Rys. 12.):



Rys. 12. Schemat lasu losowego (źródło: https://www.analyticsvidhya.com/blog/2021/10/an-introduction-to-random-forest-algorithm-for-beginners/)

Podsumowując, lasy losowe wykorzystują dwa algorytmy do przewidywania zmiennej objaśnianej. Pierwszym z nich są drzewa decyzyjne, które są tworzone na podstawie danych wylosowanych przez drugi z algorytmów, czyli bagging. Na tej podstawie tworzy się model, który świetnie radzi sobie zarówno z problemami klasyfikacyjnym, jak i regresją. W następnym, ostatnim podrozdziale opisane zostaną metody oceny modeli klasyfikacyjnych.

### Ocena modeli klasyfikacyjnych

Kluczowym elementem tworzenia modeli uczenia maszynowego jest ich ewaluacja, która polega na sprawdzeniu, jak model radzi sobie z danymi testowymi. Obejmuje to zbadanie jego precyzji, czułości, zagregowanej miary F1 i macierzy pomyłek.[[56]](#footnote-56) Jednak czasami miary te są niewystarczające do zbadania jakości modelu. W związku z tym, warto poznać krzywą ROC i AUC, które lepiej, niż podstawowe miary radzą sobie z oceną modeli klasyfikacji. Ponadto w celu, jak najdokładniejszego sprawdzenia danych należy przeprowadzić walidację krzyżową, która pomoże ocenić rzeczywistą skuteczność stworzonych modeli. Należy również zwrócić uwagę na to, że tworzone modele są wieloklasowe i wygląd np. macierzy pomyłek różni się od modelu, którego zmienna objaśniana ma charakter binarny.

#### Precyzja, czułość, miara F1 i macierz pomyłek

Podstawowymi metrykami używanymi do oceniana modeli klasyfikacyjnych są precyzja, czułość oraz miara F1. Stanowią one podstawę badania tego, jak model radzi sobie z danymi testowymi. Jak wspomniane zostało we wstępie, kluczowym krokiem przed rozpoczęciem jakiejkolwiek oceny jest podział zbioru danych na uczący i walidacyjny, a następnie przeprowadzenie ewaluacji na tym drugim.[[57]](#footnote-57)

Precyzja jest to miara, która pokazuje ile rekordów wskazanych przez algorytm, jako należących do danej klasy, w rzeczywistości do niej należy.[[58]](#footnote-58) Oblicza się ją stosując poniższy wzór:

Czułość jest kolejną miarą, która pozwala nam zrozumieć skuteczność modelu. Definiuje się ją, jako iloraz prawdziwie pozytywnych rekordów wskazanych przez algorytm oraz tych oznaczonych, jako prawdziwie pozytywne i fałszywie negatywne. Oznacza to, że bada on stosunek rekordów, które zostały wytypowane przez model, jako należące do danej klasy liczby rekordów, jaka powinna zostać wskazana. Zależność opisuje poniższy wzór:[[59]](#footnote-59)

Miara F1 stara się uśrednić wyniki precyzji i czułości stosując średnią harmoniczną, w związku z tym jej wzór wygląda następująco:[[60]](#footnote-60)

Oprócz wymienionych wcześniej miar warto również stworzyć macierz pomyłek, która obrazuje w jaki sposób rekordy zostały przypisane do danych klas, porównując to jak powinny być przypisane prawidłowo. Wygląda ona następująco (Tabela 3.):[[61]](#footnote-61)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **1** | **2** | **3** | **4** |
| **1** | 91 | 3 | 4 | 2 |
| **2** | 4 | 90 | 3 | 3 |
| **3** | 4 | 4 | 88 | 4 |
| **4** | 6 | 8 | 6 | 80 |

Tabela 3. Przykładowa macierz pomyłek dla klas 1 do 4 (oprac. własne)

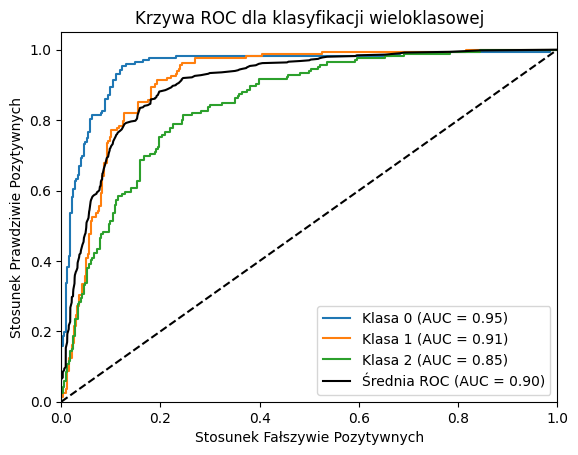
Jak widać w zamieszczonej powyżej tabeli, która pokazuje hipotetyczną macierz pomyłek dla klas 1 do 4 (pierwsza kolumna i rząd), kolejne komórki wskazują liczbę zaklasyfikowanych poprawnie (zielone) i niepoprawnie (czerwone) przypadków.

#### Krzywa ROC i AUC

W celu lepszego zbadania modeli klasyfikacyjnych, oprócz wymienionych powyżej miar, można ocenić za pomocą krzywej ROC (*ang. Receiver Operating Characteristic*), która posługując się czułością oraz miarą fałszywie pozytywnych klasyfikacji (FPK= ) w układzie współrzędnych obrazuje jakość modelu.[[62]](#footnote-62) Istotnym jest zauważenie, że wykres ten należy czytać od prawej do lewej, gdzie w prawym górnym rogu wykresu prawdopodobieństwo równe jest 1, natomiast w lewym dolnym rogu równa się ono 0.[[63]](#footnote-63) Krzywa ROC pomaga nam znaleźć punkt odcięcia prawdopodobieństwa dla naszego modelu, tzn. takiej jego wartości, aby model klasyfikował dane rekordy zgodnie z naszymi oczekiwaniami. Domyślnie dla macierzy pomyłek punkt odcięcia ustawiony jest na 0,5, czyli wszystkie rekordy powyżej będą klasyfikowane, jako 1, a wszystkie poniżej, jako 0.

Z pojęciem krzywej ROC wiąże się również AUC (*ang. Area Under Curve;* pol. Powierzchnia pod wykresem), która wprost wskazuje zdolność predykcyjną dla modelu. Jeśli AUC= 0,5 oznacza to, że klasyfikator jest losowy. Im wyższa wartość AUC, tym jakość przewidywań jest lepsza.

Jak wskazano wielokrotnie w niniejszej pracy, rozważany problem klasyfikacyjny ma charakter wieloklasowy. W związku tym należy przyjąć pewną taktykę posługiwania się krzywą ROC-AUC, która doskonale radzi sobie z klasyfikatorami binarnymi, ale wymaga pewnych zmian w przypadku kilku klas.[[64]](#footnote-64) Jednym ze sposobów jest podejście jeden przeciwko wszystkim, który polega na tym, że dzielimy zbiór na jedną z klas i resztę, a następnie sprawdzamy, jak model poradził sobie z klasyfikacją dla niej. Powtarzamy procedurę dla wszystkich możliwych klas i obliczamy średnią AUC, która będzie w tym wypadku miarą skuteczności predykcyjnej modelu. Na wykresie będzie wyglądać to następująco (Rys. 13):[[65]](#footnote-65)



Rys. 13. Krzywa ROC dla klasyfikacji wieloklasowej (oprac. własne – kod python)

Jak można zauważyć naniesienie kilku krzywych ROC na wykres znajdzie również zastosowanie przy porównaniu kilku modeli, pozwalając na łatwe ich porównanie i szybki znalezienie odpowiednich punktów odcięcia dla każdego z nich.

#### Walidacja krzyżowa

W poprzednim punkcie opisane zostały poszczególne miary używane do oceny jakości modeli. Jednak aby zwiększyć pewność tego badania należy zastosować walidację krzyżową, która polega na podziale zbioru uczącego na podzbiory. Liczbę powstałych w ten sposób części danych oznacza się literką *k*, a sama metoda nazywana jest przez to *k-fold crossvalidation* (pol. K-złożona walidacja krzyżowa)[[66]](#footnote-66). Po podziale, należy jeden z podzbiorów oznaczyć, jako walidacyjny, a resztę wykorzystać do trenowania modelu. Następnie procedurę należy powtórzyć do momentu, aż każdy rekord będzie wykorzystany w próbie walidacyjnej.[[67]](#footnote-67) Za każdym razem, gdy trenujemy i walidujemy model, sprawdzane są również miary jakości (tzn. precyzja, czułość i metryka F1). W ostatnim kroku, sprawdzamy zagregowane wyniki modelu i dokonujemy ostatecznej oceny, czy dany model ma zdolność predykcyjną na podstawie dostarczonych danych i stworzonego algorytmu.

### Podsumowanie rozdziału

Tworzenie modeli klasyfikacyjnych to nie tylko budowa samych algorytmów, ale bardziej złożony proces składający się zazwyczaj z kilku kroków, których nie wolno omijać. W pierwszej kolejności należy przygotować dane w sposób, który pozwoli na załadowanie ich do algorytmu. Kluczowe jest tutaj zastosowanie odpowiedniego podejścia do pustych wartości oraz uzgodnienie typów danych, które będą użyte przy tworzeniu modelu. Następnie należy przeprowadzić dogłębną analizę eksploracyjną, dającą podstawy do poznania zbioru służącego do tworzenia algorytmu. W pracy zaproponowano kilka metod przeprowadzenia jej: sprawdzenie miar położenia, rozproszenia, rozkładu danych oraz korelacji. Po analizie danych możemy przejść do tworzenia modeli. Ze względu na to, że problematyka pracy związana jest z klasyfikacją wieloklasową, opisane zostały trzy modele służące do jej przeprowadzenia: regresja logistyczna, drzewa decyzyjne oraz lasy losowe. Algorytmy te powinny pozwolić na predykcję zmiennej objaśnianej, czyli ostatecznej pozycji zespołu w lidze. Na samym końcu należy ocenić jakość stworzonego modelu korzystając z miar takich, jak czułość, precyzja oraz F1, a także tworząc macierz pomyłek. W celu lepszej ewaluacji jakości można posłużyć się również krzywą ROC i AUC. Warto przeprowadzić również walidację krzyżową, aby upewnić się, że model daje podobne wyniki z użyciem różnych danych walidacyjnych.

W następnym rozdziale kroki, które zostały opisane powyżej posłużą do przeanalizowania danych piłkarskich oraz stworzenia na ich podstawie modeli uczenia maszynowego.

## Rozdział 3: Analiza statystyk piłkarskich w 5 najlepszych ligach europejskich.

### Wstęp do rozdziału

Ostatni rozdział pracy poświęcony został praktycznej analizie oraz tworzeniu modeli uczenia maszynowego na podstawie danych piłkarskich pochodzących ze strony fbref.com, która jest bazą różnorakich statystyk. W rozdziale tym postaram się odpowiedzieć na pytanie, czy na podstawie pewnych danych da się przewidzieć, które miejsce drużyna zajmie na końcu rozgrywek. Na tej podstawie spróbuję wytypować statystyki, które mają kluczowe znaczenie dla przewidywania końcowej pozycji w lidze. W związku z tym, że niektóre statystyki mają oczywisty przeważający wpływ na końcowy wynik, bazując na moim doświadczeniu i powszechnej wiedzy, zdecydowałem się wykluczyć takie parametry, jak: zwycięstwa(W), przegrane(L), remisy(D), a także liczbę punktów (Pts). Usunięte zostały również pochodne tych zmiennych. W sposób zbyt oczywisty da się je powiązać z ostatecznym miejsce drużyny w tabeli, dlatego parametry te mogłyby nazbyt ważyć w ostatecznym modelu i w związku z tym umniejszać innym, które są może mniej rzucającymi się w oczy, ale również mają ogromny wpływ na zmienną objaśnianą (pozycję na koniec sezonu). Pokazuje to poniższy wykres (Rys. 13.):

A graph of different colored rectangles

Description automatically generated

Rys. 14. Korelacja zmiennej objaśnianej z usuniętymi statystykami (oprac. własne na podst. fbref.com)

Jak widać niemal wszystkie wymienione zmienne wykazują silną dodatnią lub ujemną korelację z parametrem Rk.

Zanim przejdziemy do tworzenia modeli, w następnym podrozdziale dokonana zostanie eksploracyjna analiza danych, która ma na celu zaznajomienie czytelnika ze zbiorem, który będzie omawiany oraz pokaże podstawowe zależności pomiędzy różnymi parametrami a zmienną zależną.

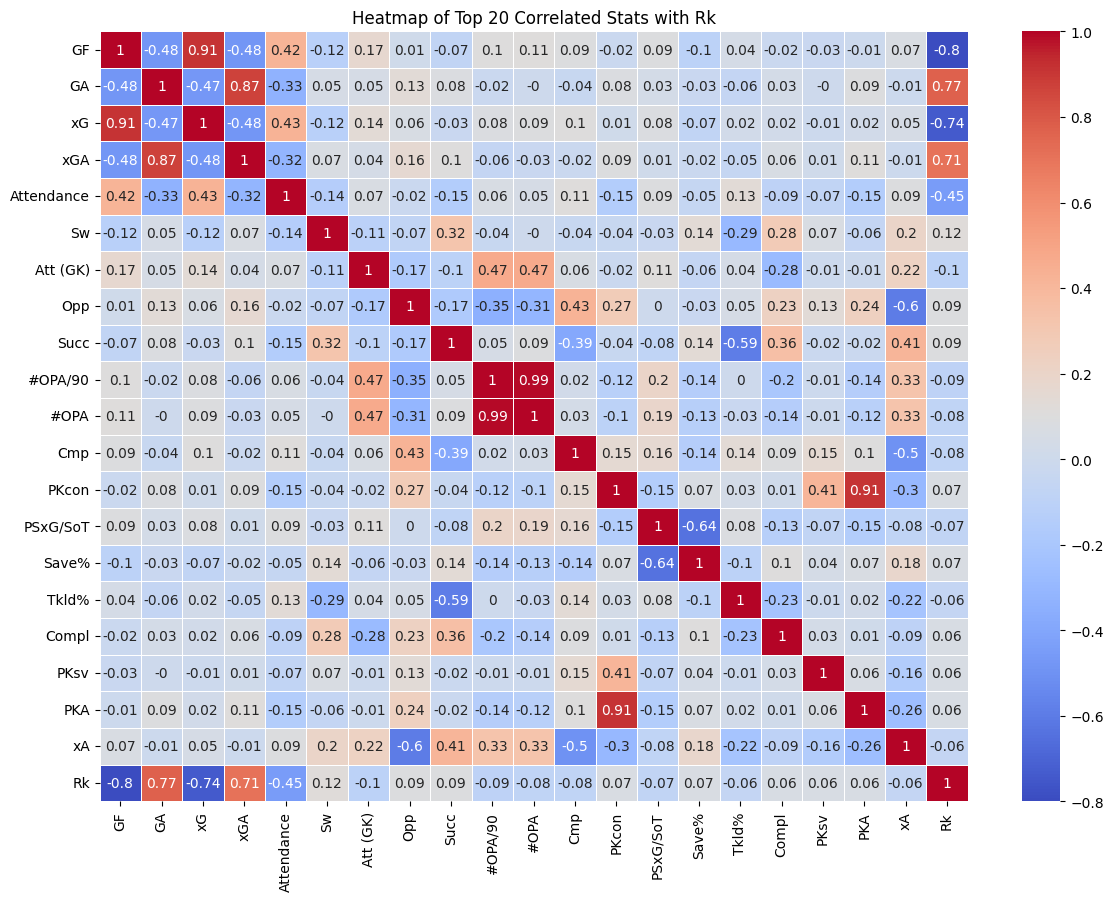
### Eksploracyjna analiza danych

Na samym wstępie warto poznać strukturę omawianego zbioru i zbadać jego kształt. Podstawowe statystyki prezentują się następująco (Tabela 4.):

|  |  |
| --- | --- |
| Liczba wierszy | Liczba kolumn |
| 684 | 182 |

Tabela 4. Podstawowe kształty zbioru danych (oprac. własne na podst. fbref.com)

Już na pierwszy rzut oka można zauważyć, że tabela jest bardzo rozbudowana horyzontalnie, co będzie miało wpływ w dalszej analizie, ponieważ konsekwencją tego będzie selekcja zmiennych, która zostanie użyta w modelach. Dla potrzeb eksploracyjnej analizy danych, aby pokazać pewne trendy posłużę się selekcją zmiennych na podstawie korelacji ze zmienną zależną i wybiorę 10 parametrów o najwyższym dodatnim i 10 o najniższym ujemnym związku z „Rk” Rys. 14.). Jak widać na poniższej matrycy korelacji, występuje bardzo słabe połączenie pomiędzy zmienną zależną a innymi parametrami. Może to utrudnić budowę modeli (szczególnie regresji logistycznej) i sugeruje słaby związek pomiędzy pozycją zespołu w lidze, a jakimiś konkretnymi statystykami. Należy nadmienić, że dość silna korelacja występuje pomiędzy pozycją w lidze, a liczbą widowni na trybunach.



Rys. 15. Korelacja pomiędzy zmienną objaśnianą a innymi parametrami (oprac. własne na podst. fbref.com)

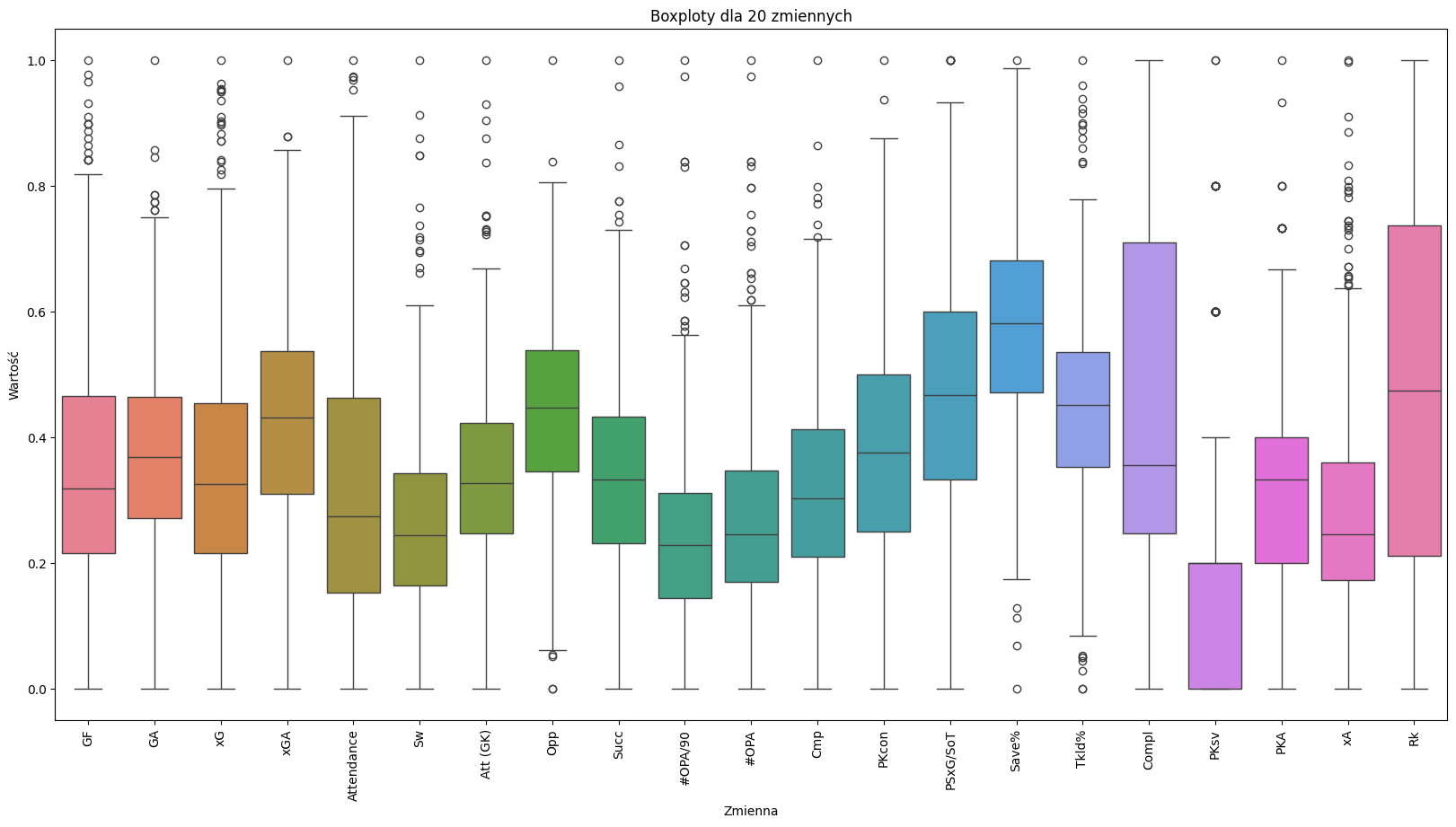
W następnej kolejności zbadamy miary położenia i rozproszenia wytypowanych w poprzednim kroku parametrów. Przed prezentacją dokonano standaryzacji, ze względu na to, że różne parametry posługiwały się rozbieżnymi skalami. Wyniki zaprezentowano w tabeli poniżej (Tabela 5.):

|  |  |  |  |
| --- | --- | --- | --- |
| **Zmienna** | **Średnia arytmetyczna** | **Odchylenie standardowe** | **Mediana** |
| **GF** | 0.356459 | 0.193250 | 0.318182 |
| **GA** | 0.373434 | 0.158521 | 0.369048 |
| **xG** | 0.359093 | 0.190476 | 0.324962 |
| **xGA** | 0.428718 | 0.162499 | 0.431200 |
| **Attendance** | 0.317734 | 0.222536 | 0.273840 |
| **Sw** | 0.261114 | 0.139153 | 0.243781 |
| **Att (GK)** | 0.343610 | 0.139506 | 0.327430 |
| **Opp** | 0.438484 | 0.142671 | 0.446809 |
| **Succ** | 0.341401 | 0.157844 | 0.332613 |
| **#OPA/90** | 0.247834 | 0.139354 | 0.227666 |
| **#OPA** | 0.273169 | 0.149396 | 0.245763 |
| **Cmp** | 0.321533 | 0.153564 | 0.302326 |
| **PKcon** | 0.375914 | 0.170080 | 0.375000 |
| **PSxG/SoT** | 0.449220 | 0.189700 | 0.466667 |
| **Save%** | 0.575986 | 0.154137 | 0.580645 |
| **Tkld%** | 0.445588 | 0.145011 | 0.451713 |
| **Compl** | 0.427598 | 0.235753 | 0.355072 |
| **PKsv** | 0.188012 | 0.196386 | 0.200000 |
| **PKA** | 0.315205 | 0.160474 | 0.333333 |
| **xA** | 0.282356 | 0.158516 | 0.245023 |
| **Rk** | 0.488920 | 0.298333 | 0.473684 |

Tabela 5. Miary położenia i rozproszenia dla analizowanych danych piłkarskich po standaryzacji (oprac. własne na podst. fbref.com)

Jak widać w tabeli powyżej, mediany i średnie nie różnią się zbytnio od siebie, co sugeruje, że dane nie zawierają wielu wartości odstających i nie mają one dużego wpływu na zbiór. Niemniej jednak wartości odchylenia standardowego mogą sugerować duże rozproszenie rekordów wokół średniej.

W celu lepszego zobrazowania zmiennych i potwierdzenia przypuszczeń płynących z analizy miał położenia oraz rozproszenia, warto zwizualizować sobie je za pomocą opisanych w rozdziale drugim wykresów. W pierwszej kolejności, aby sprawdzić wartości odstające oraz kwartyle należy stworzyć wykresy pudełkowe. (Rys. 16)



Rys. 16. Wykresy pudełkowe dla 20 najbardziej skorelowanych ze zmienną objaśnianą parametrów (oprac. własne na podst. fbref.com)

Jak widać na powyższym wykresie, pomimo tego, że dane są w większości symetryczne, występuje pewna ilość wartości odstających. Jednakże ze względu na to, że zbiór ma ograniczoną liczbę rekordów zdecydowano się je pozostawić.

Dowodem na symetryczność większości zmiennych są również histogramy (Rys. 17.), które poza nielicznymi wyjątkami pokazują, że dane mają rozkład normalny, gdzie wartość mediany zbliżona jest do średniej arytmetycznej. W związku z tym można ocenić, iż zbiór jest gotowy do przygotowania na nim modeli uczenia maszynowego.

A screenshot of a graph

Description automatically generated

Rys. 17. Histogramy dla 20 najbardziej skorelowanych ze zmienną objaśnianą parametrów (oprac. własne na podst. fbref.com)

### Modele uczenia maszynowego

W kolejnych punktach niniejszego podrozdziału zostaną zaprezentowane modele uczenia maszynowego, które zostały stworzone na podstawie zbadanych danych. Każdy z nich oceniony został stosując miary i metody opisane w rozdziale drugim pracy. W ostatnim punkcie niniejszego podrozdziału zostaną one zestawione ze sobą i porównane. Ze względu na wysoką granularność danych, zmienna objaśniana „Rk” została przekształcona i pogrupowana tworząc nowy parametr „label”, która zawiera cztery kategorie: *relegation* („Rk” 16-20)*, league\_average* („Rk”: 10-15)*, european\_cups*(„Rk”: 5-9)*, champions\_group* („Rk”: 1-4) odpowiadające aspiracjom poszczególnych zespołów.

Nie zastosowano klasycznego podziału danych na treningowe i walidacyjne, ponieważ użyto walidacji krzyżowej, uwzględniając małą liczbę rekordów wydaje się być to strategią optymalną w celu, jak najpełniejszego wykorzystania zbioru.

Jak wiadomo, kluczowym w modelach klasyfikacyjnych jest optymalizacja modeli na podstawie punktu odcięcia, który określa stopień prawdopodobieństwa, przy którym rekord ma być zakwalifikowany jako 0 lub 1. Dla potrzeb tej pracy, ze względu na to, że mamy do czynienia z klasyfikacją wieloklasową, punkt odcięcia został wybrany na podstawie parametru AUC, czyli dla prawdopodobieństwa, które wyznacza najwyższą wartość tego parametru.

#### Regresja logistyczna

Pierwszym utworzonym modelem jest regresja logistyczna wieloklasowa. Macierz pomyłek, która powstała na podstawie wykorzystania modelu i walidacji krzyżowej wygląda następująco (Rys. 18.):

A blue squares with numbers

Description automatically generated

Rys.18. Regresja logistyczna- macierz pomyłek (oprac. własne na podst. fbref.com)

Jak widać model poradził sobie dobrze z oceną najniższej klasy, czyli *relegation*, z dużą dokładnością estymując, czy na podstawie dostarczonych statystyk zespół spadnie z ligi, a także *champions\_group*, która pokazuje szansę na walkę o mistrzostwo. Potwierdzają to również miary czułości, precyzji i F1 (Tabela 6).

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **prezycja** | **czułość** | **F1** | **Liczność kategorii** |
| **champions\_group** | 0.80 | 0.84 | 0.82 | 140 |
| **european\_cups** | 0.68 | 0.51 | 0.58 | 175 |
| **league\_average** | 0.64 | 0.54 | 0.59 | 210 |
| **relegation** | 0.63 | 0.91 | 0.74 | 159 |
| **średnia ważona metryk** | 0.68 | 0.68 | 0.67 | 684 |

Tabela 6. Regresja logistyczna- miary oceny modelu (oprac. własne na podst. fbref.com)

Model uzyskał również wysoką średnią wartość AUC dla czterech kategorii, co świadczy o jego wysokiej możliwości do rozróżnienia pomiędzy klasami, zaprezentowano to na poniższym wykresie (Rys. 19.):

A graph of a number of different colored lines

Description automatically generated

Rys. 19. Regresja logistyczna- krzywe ROC dla poszczególnych klas i uśredniona krzywa ROC (oprac. własne na podst. fbref.com)

Mimo ograniczonej zdolności przewidywania dla środkowych klas, model ten ma pewną wartość predykcyjną i może być używany do przewidywania, czy dany zespół odniesie sukces w lidze.

#### Drzewo decyzyjne

Kolejnym modelem stworzonym w celu przewidywania pozycji zespołu na koniec ligowych zmagań jest drzewo decyzyjne. Tak, jak w przypadku regresji logistycznej w pierwszej kolejności zbadamy macierz pomyłek utworzoną na podstawie walidacji krzyżowej (Rys. 20.)

A blue squares with numbers

Description automatically generated

Rys. 20. Drzewo decyzyjne- macierz pomyłek (oprac. własne na podst. fbref.com)

Jak widać powyżej, model poradził sobie dość przeciętnie z przewidywaniem klas. Potwierdzają to również metryki jego jakości (Tabela 7.)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **prezycja** | **czułość** | **F1** | **liczność kategorii** |
| **champions\_group** | 0.78 | 0.74 | 0.76 | 140 |
| **european\_cups** | 0.52 | 0.50 | 0.51 | 175 |
| **league\_average** | 0.56 | 0.64 | 0.60 | 210 |
| **relegation** | 0.74 | 0.67 | 0.70 | 159 |
| **średnia ważona metryk** | 0.64 | 0.63 | 0.63 | 684 |

Tabela 7. Drzewo decyzyjne- miary oceny modelu (oprac. własne na podst. fbref.com)

Stosunkowo niski jest również AUC, co sugeruje, że model nie najlepiej poradził sobie z rozpoznaniem poszczególnych klas (Rys. 21.). Powoduje to jego mniejszą wartość predykcyjną niż wcześniej omawianej regresji logistycznej wieloklasowej.

A graph with different colored lines

Description automatically generated

Rys. 21. Drzewo decyzyjne- krzywe ROC dla poszczególnych klas i uśredniona krzywa ROC (oprac. własne na podst. fbref.com)

Mimo ograniczonej przydatności drzewa decyzyjnego do oceny przewidywania pozycji zespołu w lidze może stanowić punkt odniesienia dla innych modeli, ponadto w następnym punkcie zostanie opisany las losowy, który jest złożony z wielu drzew decyzyjnych i przez to dużo lepiej poradził sobie ze stawianym przed nim zadaniem.

#### Las losowy

Ostatnim modelem stworzonym do zbadania danych z fbref.com jest las losowy (złożony z 500 drzew decyzyjnych). Jak pokazuje zamieszczona poniżej macierz pomyłek, najlepiej poradził sobie on ze stawianym przed nim zadaniem (Rys. 22.). Model ten jest w stanie przewidywać zarówno niższe, jak i wyższe klasy. Problematyczne dla niego jest dokładna predykcyjne grupie mistrzowskiej (*champions\_group*), gdzie dużo rekordów należących w rzeczywistości do tej klasy zostało przypisane do niższych.

A blue squares with numbers and a number

Description automatically generated

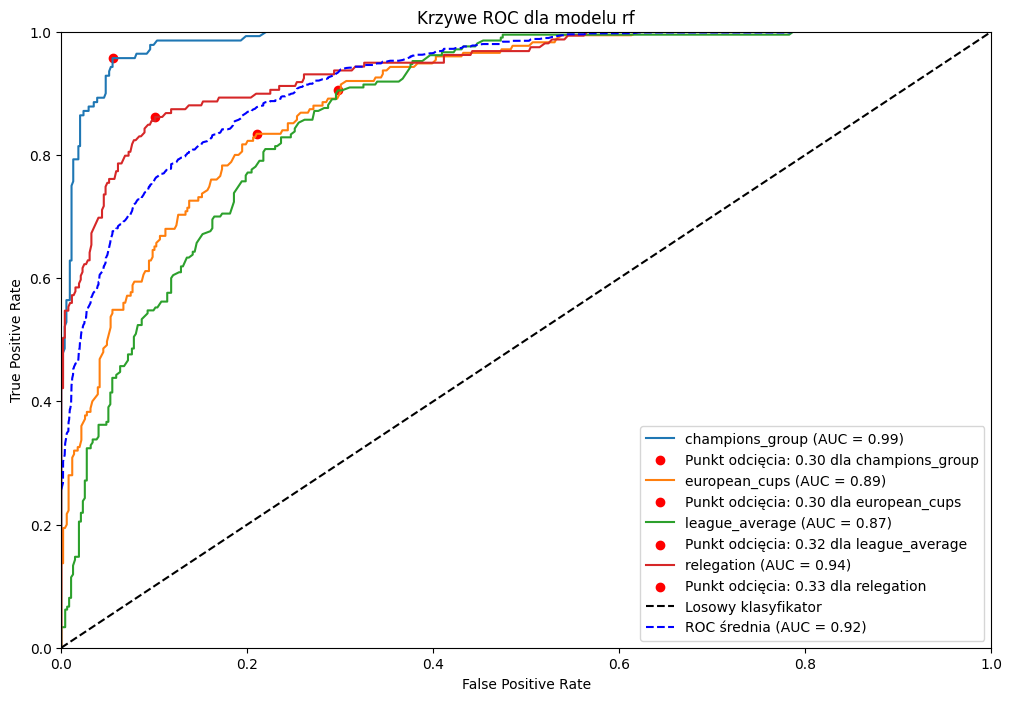
Rys. 22. Las losowy- macierz pomyłek (oprac. własne na podst. fbref.com)

Również metryki jakości prezentują się najlepiej w przypadku lasu losowego (Tabela 8.) i świadczą o wysokiej jego zdolności do przewidywania klas na różnych zestawach danych.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **prezycja** | **czułość** | **F1** | **liczność kategorii** |
| **champions\_group** | 0.91 | 0.61 | 0.74 | 140 |
| **european\_cups** | 0.62 | 0.63 | 0.62 | 175 |
| **league\_average** | 0.65 | 0.69 | 0.67 | 210 |
| **relegation** | 0.72 | 0.86 | 0.79 | 159 |
| **średnia ważona metryk** | 0.71 | 0.70 | 0.70 | 684 |

Tabela 8. Lasy losowe- miary oceny modelu (oprac. własne na podst. fbref.com)

Las losowy stworzony na podstawie danych charakteryzuje się także wysokim średnim AUC wynoszącym 0.92. Pokazuje to jego wysoką zdolność do przewidywania klas, co jest kluczową metryką w przypadku rozważanego przypadku użycia. Krzywa ROC i AUC zostały zaprezentowane na wykresie poniżej. (Rys. 23.)



Rys. 23. Las losowy- krzywe ROC dla poszczególnych klas i uśredniona krzywa ROC (oprac. własne na podst. fbref.com)

Biorąc pod uwagę powyższe rozwiązania można dojść do wniosku, iż omawiany model najlepiej poradził sobie ze stawianym przed nim zadaniem i może być bardzo pomocny do przewidywania ostatecznej pozycji zespołu w lidze. Średnia zdolność poszczególnych modeli do przewidywania klas została zaprezentowana łącznie na poniższym wykresie, co stanowi niezbity dowód przeważającej jakości lasu losowego (Rys. 22.)

A graph of a positive rate

Description automatically generated with medium confidence

Rys. 24. Uśrednione krzywe ROC dla wszystkich omawianych modeli (oprac. własne na podst. fbref.com)

### Wnioski płynące z analizy

Znając już wyniki predykcji i dokonując analizy zbioru należy zastanowić się na praktycznymi wnioskami, które można wyciągnąć na ich podstawie. W innym wypadku stworzone modele, wykresy itd. są jedynie danymi, które nie niosą żadnych użytecznych dla analityków piłkarskich informacji.

W pierwszej kolejności należy zauważyć, że analiza statystyk rzeczywiście ma sens i na jej podstawie można sprawdzić, jakie miejsce zespół zajmie w lidze. Zachęca to do starannego zbierania danych i przetwarzania ich. Można na podstawie statystyk z pierwszej połowy sezonu sprawdzić, czy nasz zespół zajmuje miejsce należne mu na podstawie jego gry, ponieważ to czasami nieuchwytny czynnik szczęścia gra rolę i mimo wspaniałej postawy na boisku drużyna przegrywa mecze.

Ponadto biorąc pod uwagę, że najlepiej sprawdził się algorytm lasu losowego, można sprawdzić, które czynniki odgrywały najważniejszą rolę przy tworzeniu modeli. Wynik zaprezentowano w tabeli poniżej (Tabela 9.)

|  |  |
| --- | --- |
| **Zmienna** | **Ważność** |
| Gole strzelone | 0.093472 |
| Gole stracone | 0.075657 |
| Oczekiwane gole strzelone | 0.069860 |
| Oczekiwane gole stracone | 0.054807 |
| Widownia | 0.023220 |
| Dośrodkowania przeciwko zespołowi | 0.009006 |
| Kontakty w swoim polu karnym | 0.007651 |
| Akcje prowadzące do gola na 90 minut | 0.007003 |
| Okazje bramkowe stworzone ze stałych fragmentów gry | 0.006957 |
| Przegrane próby odzyskania piłki | 0.006883 |
| Strzały na 90 minut | 0.006863 |
| Liczba odzyskanych piłek | 0.006759 |
| Liczba podań wykonanych przez bramkarza | 0.006711 |
| Procent skutecznych podań wykonanych przez bramkarza | 0.006694 |

Tabela 9. Ważność cech użytych przez model lasu losowego (oprac. własne na podst. fbref.com)

Analizując tabelę powyżej można dojść do wielu wniosków, widać przede wszystkim większe znaczenie goli strzelonych nad straconymi (tak samo w przypadku oczekiwanych statystyk), sugeruje to, że większe znaczenie ma ofensywne nastawienie zespołu, niż szczelna defensywa. Ponadto rzuca się w oczy istotność statystyk generowanych przez bramkarzy, szczególnie dotyczące podań, kontaktów we własnym polu karnym, co może sugerować, że trenerzy powinni przykładać dużą uwagę do tego elementu gry.

Zmienną, która jest najbardziej niespodziewana w całym zestawieniu okazuje się być liczba widowni, która pokazuje, że kibice na trybunach w rzeczywistości są dwunastym zawodnikiem i mają duży wpływ na powodzenie lub niepowodzenie zespołu. Stanowi to ogromne wyzwanie przed działami marketingu w klubach, które powinny dbać o to, aby jak najwięcej kibiców zasiadało w każdy weekend na widowni.

Wyciągnięte powyżej wnioski są jedynie fragmentem tego, co można dowiedzieć się analizując wyniki modelu. Pokazuje to jednak, że nawet proste badanie może pomóc w odkryciu pewnych cech, które powinno się poprawić, aby podjąć walkę o mistrzostwo, albo w przypadku słabszych klubów znaleźć się na bezpiecznym miejscu w środku tabeli.

### Podsumowanie rozdziału

W ostatnim rozdziale pracy opisano praktyczny proces analizy danych oraz tworzenia modeli uczenia maszynowego na ich podstawie. Na samym początku przeprowadzono eksploracyjną analizę danych, która stanowi fundament do dalszych rozważań, posługując się najbardziej skorelowanymi ze zmienną objaśnianą cechami zwizualizowano i opisano reprezentatywną próbkę.

W dalszej kolejności stworzono trzy modele uczenia maszynowego: regresję logistyczną, drzewo decyzyjne oraz las losowy. Na podstawie takich metryk, jak czułość, precyzja, F1, ROC i AUC okazało się, że ostatni z wymienionych modeli najlepiej nadaje się do predykcji zmiennej objaśnianej na podstawie zbioru.

Ostatnim krokiem całego procesu było wyciągnięcie praktycznych wniosków płynących z analizy i modeli. Oczywiście ze względu na rozmiar zbioru i ilość różnych zmiennych w piłce nożnej jest to tylko wstęp, który udowadnia jednak, że uczenie maszynowe w futbolu ma sens i powinno być szerzej używane.

## Zakończenie

Używanie modeli uczenia maszynowego w piłce nożnej wciąż jest uważane za nowatorskie podejście. Nie wszyscy trenerzy są w stanie przekonać się do nowych metod. Jak starałem udowodnić się w tej pracy, metody uczenia maszynowego naprawdę działają i mogą pomóc w osiąganiu określonych celów. Przeprowadzone przeze mnie badanie danych z fbref.com jest dobrym przykładem tego, że można stworzyć skuteczny algorytm, działający na określonym zbiorze statystyk, który jest w stanie zbadać jakie elementy wymagają poprawy i jakie są składniki sukcesu każdej drużyny.

Temat danych piłkarskich jest niezwykle rozbudowany, przede wszystkim ze względu na ilość zagrań i zdarzeń, które rejestruje się w trakcie spotkania. Stanowi to nie lada wyzwanie dla analityka, który musi na podstawie wiedzy i algorytmów skupić się wyłącznie na tych, które niosą rzeczywiście jakąś wartość. W pracy skupiłem się na tym, aby odkryć, czy w ogóle istnieją takie zależności, tzn. czy z dostarczonych danych da się wyczytać cokolwiek ponad losowy szum. Na podstawie przeprowadzonej analizy z całą pewnością można stwierdzić, że statystyki piłkarskie niosą w sobie dużo informacji, które okazują się być przydatne i tylko od analityków zależy, czy je należycie wykorzystają.

W dalszych rozważaniach nad tematem, chciałbym zejść na niższy poziom granularności i sprawdzić na przykład, czy strzelone gole zależą od innych zmiennych i da się poprawić swoją ofensywę skupiając się na kilku składowych każdego ataków (np. na podaniach w ostatnią tercję boiska). Uważam, że dopiero na takim poziomie szczegółowości, analiza danych piłkarskich przynosi realną wartość dla klubów piłkarskich, ponieważ mogą one się skupić na poprawie konkretnych aspektów swojej gry. Niemniej jednak, takie badania, jak przeprowadzone w pracy mogą pomóc zrozumieć składowe sukcesu każdego zespołu. Jak już wspomniałem w rozdziale trzecim, niezwykle zaskoczyła mnie na przykład istotność widowni w kontekście sukcesu danej drużyny. Takie odkrycia pozwalają spojrzeć na piłkę nożną z innej strony i były dla mnie główną motywacją do napisania tej pracy.

## Bibliografia

1. https://en.wikipedia.org/wiki/Championship\_Manager\_(video\_game), [dostęp: 29.07.2024]
2. Anderson Ch., Sally D., *The Numbers Game*, Milton Keynes (UK), 2013, s. 256
3. Kuper S., Szymański S., *Futbonomia*, Kraków 2017, wyd. 1, s. 20
4. Biermann Ch., *Piłkarscy hakerzy*. O rewolucji w futbolu i sztuce zbierania danych, Kraków, 2022, s. 146
5. https://www.bbc.com/sport/football/61648608 [dostęp: 29.07.2024]
6. https://en.wikipedia.org/wiki/Long\_ball [dostęp: 2.08.2024]
7. https://www.statsperform.com/about/ [dostęp: 29.07.2024]
8. https://en.wikipedia.org/wiki/Analysis\_(disambiguation) [dostęp: 29.07.2024]
9. https://pl.wikipedia.org/wiki/Dane [dostęp: 29.07.2024]
10. https://pl.wikipedia.org/wiki/Informacja [dostęp: 01.08.2024]
11. https://www.nytimes.com/athletic/3502290/2022/08/14/domenico-tedesco-red-bull-leipzig/ [dostęp: 01.08.2024]
12. https://fbref.com/en/comps/9/2019-2020/2019-2020-Premier-League-Stats [dostęp: 2.08.2024]
13. Fbref.com[dostęp: 2.08.2024]
14. https://pl.wikipedia.org/wiki/Liga\_Mistrzów\_UEFA\_(2012/2013) [dostęp: 5.08.2024]
15. Dean J. , *Big data, Data mining, Machine Learning,* Wiley, 2014
16. https://ekordo.pl/uczenie-maszynowe-przygotowanie-danych-wprowadzenie/ [dostęp: 7.08.2024]
17. https://www.datarobot.com/blog/introducing-mlops-champion-challenger-models/ [dostęp: 7.08.2024]
18. Bruce P., Bruce A., Gedeck P., *Statystyka praktyczna w data science*, O’Reilly, 2020
19. https://stat.gov.pl/metainformacje/slownik-pojec/pojecia-stosowane-w-statystyce-publicznej/2937,pojecie.html [dostęp:
20. ttps://www.ue.katowice.pl/fileadmin/\_migrated/content\_uploads/13\_G.Trzpiot\_Wybrane\_statystyki\_odporne.pdf [dostęp:8.0
21. ttps://pogotowiestatystyczne.pl/slowniki/wariancja/ [dostęp: 8.08.2024]
22. https://mfiles.pl/pl/index.php/Wykres\_pudełkowy [dostęp: 8.08.2024]
23. https://mfiles.pl/pl/index.php/Histogram [dostęp: 8.08.2024]
24. https://predictivesolutions.pl/skosnosc-i-kurtoza [dostęp: 8.08.2024]
25. https://encyklopedia.pwn.pl/haslo/korelacja;4008870.html [dostęp: 8.08.2024]
26. https://www.investopedia.com/terms/r/regression.asp [dostęp: 9.08.2024]
27. Kwak, Chanyeong, and Alan Clayton-Matthews. "Multinomial logistic regression." *Nursing research*51.6 (2002): 404-410.
28. Trzpiot, Grażyna, and Alicja Ganczarek. "Drzewa decyzyjne w statystycznej analizie decyzji na przykładzie wirtualnych łańcuchów dostaw." *Acta Universitatis Lodziensis. Folia Oeconomica* 271 (2012).
29. https://www.learndatasci.com/glossary/gini-impurity/ [dostęp: 12.08.2024]
30. https://towardsdatascience.com/decision-trees-explained-entropy-information-gain-gini-index-ccp-pruning-4d78070db36c
31. Breiman, Leo. "Bagging predictors." *Machine learning* 24 (1996): 123-140.
32. Schonlau, Matthias, and Rosie Yuyan Zou. "The random forest algorithm for statistical learning." *The Stata Journal* 20.1 (2020)
33. https://towardsdatascience.com/accuracy-precision-recall-or-f1-331fb37c5cb9 [dostęp: 13.08.2024]
34. https://ksopyla.com/data-science/precision-recall-f1-miary-oceny-klasyfikatora/ [dostęp: 13.08.2024]
35. https://www.analyticsvidhya.com/blog/2022/02/k-fold-cross-validation-technique-and-its-essentials/ [dostęp 13.08.2024
36. Kjell J., Kuhn M., *Applied Predictive Modelling*, Springer 2013

## Spis tabel

|  |  |
| --- | --- |
| **Tabela** | **Źródło** |
| Tabela 1. Statystyki meczu pomiędzy Celticiem Glasgow a FC Barcelona z dn. 7 listopada 2012 roku | oprac. własne na podst. https://www.flashscore.pl/mecz/rLLV8Mnm/#/szczegoly-meczu/statystyki-meczu/0) |
| Tabela 2. Grupa G Ligi Mistrzów 2012-13 | https://pl.wikipedia.org/wiki/Liga\_Mistrzów\_UEFA\_(2012/2013) |
| Tabela 3. Przykładowa macierz pomyłek dla klas 1 do 4 | oprac. własne |
| Tabela 4. Podstawowe kształty zbioru danych | oprac. własne na podst. fbref.com |
| Tabela 5. Miary położenia i rozproszenia dla analizowanych danych piłkarskich po standaryzacji | oprac. własne na podst. fbref.com |
| Tabela 6. Regresja logistyczna- miary oceny modelu | oprac. własne na podst. fbref.com |
| Tabela 7. Drzewo decyzyjne- miary oceny modelu | oprac. własne na podst. fbref.com |
| Tabela 8. Lasy losowe- miary oceny modelu | oprac. własne na podst. fbref.com |
| Tabela 9. Ważność cech użytych przez model lasu losowego | oprac. własne na podst. fbref.com |

## Spis rysunków

|  |  |
| --- | --- |
| **Rysunek** | **Źródło** |
| Rys. 1 Dystrybucja podań i przejęć piłki w latach 1953- 1967, autor Charles Reep | Anderson Ch., Sally D., op. cit., s. 18 |
| Rys. 2. Rodzaje analizy w sporcie | oprac. własne na podst. https://en.wikipedia.org/wiki/Sports\_analytics |
| Rys. 3. Liczba zawodników w bazach danych | https://www.driblab.com/driblab-en/football-data-the-largest-coverage-in-the-market/) |
| Rys.4. Analiza podań Liverpoolu, Man City i Chelsea | https://m.allfootballapp.com/news/EPL/EPL-club-styles-revealed-in-advanced-Opta-stats-City-dominate-attack-matters/1743637 |
| Rys. 5 Schemat *Champion-Challenger* w modelowaniu | https://deeplearninguniversity.com/mlops/mlops-model-testing/ |
| Rys. 6 Wykres pudełkowy prosty | https://mfiles.pl/pl/index.php/Wykres\_pudełkow |
| Rys. 7 Wykres pudełkowy złożony | https://mfiles.pl/pl/index.php/Wykres\_pudełkow |
| Rys. 8. Histogram podstawowy | https://mfiles.pl/pl/index.php/Histogram |
| Rys. 9. Matryca korelacji | https://www.researchgate.net/figure/An-example-of-a-correlation-matrix-Each-entry-in-the-correlation-matrix-is-a-Pearsons\_fig1\_366861559 |
| Rys. 10 Przykład wykresu funkcji logistycznej dla wartości (-10,10) | oprac. własne |
| Rys. 11. Drzewo decyzyjne- Wynik meczu na podstawie goli strzelonych i straconych | oprac. własne – kod python |
| Rys. 12. Schemat lasu losowego | https://www.analyticsvidhya.com/blog/2021/10/an-introduction-to-random-forest-algorithm-for-beginners/) |
| Rys. 13. Krzywa ROC dla klasyfikacji wieloklasowej | oprac. własne – kod python |
| Rys. 14. Korelacja zmiennej objaśnianej z usuniętymi statystykami | oprac. własne na podst. fbref.com |
| Rys. 15. Korelacja pomiędzy zmienną objaśnianą a innymi parametrami | oprac. własne na podst. fbref.com |
| Rys. 16. Wykresy pudełkowe dla 20 najbardziej skorelowanych ze zmienną objaśnianą parametrów | oprac. własne na podst. fbref.com |
| Rys. 17. Histogramy dla 20 najbardziej skorelowanych ze zmienną objaśnianą parametrów | oprac. własne na podst. fbref.com |
| Rys. 18. Regresja logistyczna- macierz pomyłek | oprac. własne na podst. fbref.com |
| Rys. 19. Regresja logistyczna- krzywe ROC dla poszczególnych klas i uśredniona krzywa ROC | oprac. własne na podst. fbref.com |
| Rys. 20. Drzewo decyzyjne- macierz pomyłek | oprac. własne na podst. fbref.com |
| Rys. 21. Drzewo decyzyjne- krzywe ROC dla poszczególnych klas i uśredniona krzywa ROC | oprac. własne na podst. fbref.com |
| Rys. 22. Las losowy- macierz pomyłek | oprac. własne na podst. fbref.com |
| Rys. 23. Las losowy- krzywe ROC dla poszczególnych klas i uśredniona krzywa ROC | oprac. własne na podst. fbref.com |
| Rys. 24. Uśrednione krzywe ROC dla wszystkich omawianych modeli | oprac. własne na podst. fbref.com |

1. <https://en.wikipedia.org/wiki/Championship_Manager_(video_game)>, [dostęp: 29.07.2024] [↑](#footnote-ref-1)
2. Anderson Ch., Sally D., *The Numbers Game,* Milton Keynes (UK), 2013, s. 256 [↑](#footnote-ref-2)
3. Kuper S., Szymański S., *Futbonomia,* Kraków 2017, wyd. 1, s. 20 [↑](#footnote-ref-3)
4. *Ibidem* [↑](#footnote-ref-4)
5. Biermann Ch., *Piłkarscy hakerzy. O rewolucji w futbolu i sztuce zbierania danych,* Kraków, 2022, s. 146 [↑](#footnote-ref-5)
6. *Ibidem,* s. 146 [↑](#footnote-ref-6)
7. Anderson Ch., Sally D, *op. cit.,* s.1 [↑](#footnote-ref-7)
8. *Ibidem* [↑](#footnote-ref-8)
9. <https://www.bbc.com/sport/football/61648608> [dostęp: 29.07.2024] [↑](#footnote-ref-9)
10. *Ibidem* [↑](#footnote-ref-10)
11. *Ibidem* [↑](#footnote-ref-11)
12. <https://en.wikipedia.org/wiki/Long_ball> [dostęp: 2.08.2024] [↑](#footnote-ref-12)
13. Anderson Ch., Sally D., *op. cit.* [↑](#footnote-ref-13)
14. <https://www.statsperform.com/about/> [dostęp: 29.07.2024] [↑](#footnote-ref-14)
15. <https://en.wikipedia.org/wiki/Analysis_(disambiguation)> [dostęp: 29.07.2024] [↑](#footnote-ref-15)
16. Anderson Ch., Sally D., *op. cit.*, s. 6 [↑](#footnote-ref-16)
17. <https://pl.wikipedia.org/wiki/Dane> [dostęp: 29.07.2024] [↑](#footnote-ref-17)
18. <https://pl.wikipedia.org/wiki/Informacja> [dostęp: 01.08.2024] [↑](#footnote-ref-18)
19. <https://www.nytimes.com/athletic/3502290/2022/08/14/domenico-tedesco-red-bull-leipzig/> [dostęp: 01.08.2024] [↑](#footnote-ref-19)
20. Anderson Ch., Sally D., *op. cit.*, s. 14 [↑](#footnote-ref-20)
21. *Ibidem*, s.10 [↑](#footnote-ref-21)
22. *Ibidem,* s. 22 [↑](#footnote-ref-22)
23. https://fbref.com/en/comps/9/2019-2020/2019-2020-Premier-League-Stats [dostęp: 2.08.2024] [↑](#footnote-ref-23)
24. *Ibidem* [↑](#footnote-ref-24)
25. Fbref.com[dostęp: 2.08.2024] [↑](#footnote-ref-25)
26. https://pl.wikipedia.org/wiki/Liga\_Mistrzów\_UEFA\_(2012/2013) [dostęp: 5.08.2024] [↑](#footnote-ref-26)
27. Bierman Ch. *Op. cit.,* s.154 [↑](#footnote-ref-27)
28. *Ibidem,* s. 150-151 [↑](#footnote-ref-28)
29. Dean J. , *Big data, Data mining, Machine Learning,* Wiley, 2014 [↑](#footnote-ref-29)
30. *Ibidem* [↑](#footnote-ref-30)
31. https://ekordo.pl/uczenie-maszynowe-przygotowanie-danych-wprowadzenie/ [dostęp: 7.08.2024] [↑](#footnote-ref-31)
32. https://www.datarobot.com/blog/introducing-mlops-champion-challenger-models/ [dostęp: 7.08.2024] [↑](#footnote-ref-32)
33. Bruce P., Bruce A., Gedeck P., *Statystyka praktyczna w data science,* O’Reilly, 2020 [↑](#footnote-ref-33)
34. https://stat.gov.pl/metainformacje/slownik-pojec/pojecia-stosowane-w-statystyce-publicznej/2937,pojecie.html [dostęp: 8.08.2024] [↑](#footnote-ref-34)
35. Bruce P., Bruce A., Gedeck P., *op. cit.,* s.25 [↑](#footnote-ref-35)
36. https://www.ue.katowice.pl/fileadmin/\_migrated/content\_uploads/13\_G.Trzpiot\_Wybrane\_statystyki\_odporne.pdf [dostęp:8.08.2024] [↑](#footnote-ref-36)
37. https://pogotowiestatystyczne.pl/slowniki/wariancja/ [dostęp: 8.08.2024] [↑](#footnote-ref-37)
38. https://mfiles.pl/pl/index.php/Wykres\_pudełkowy [dostęp: 8.08.2024] [↑](#footnote-ref-38)
39. https://mfiles.pl/pl/index.php/Histogram [dostęp: 8.08.2024] [↑](#footnote-ref-39)
40. https://predictivesolutions.pl/skosnosc-i-kurtoza [dostęp: 8.08.2024] [↑](#footnote-ref-40)
41. https://encyklopedia.pwn.pl/haslo/korelacja;4008870.html [dostęp: 8.08.2024] [↑](#footnote-ref-41)
42. https://www.investopedia.com/terms/r/regression.asp [dostęp: 9.08.2024] [↑](#footnote-ref-42)
43. Bruce P., Bruce A., Gedeck P., *op. cit.,* s.193 [↑](#footnote-ref-43)
44. Kwak, Chanyeong, and Alan Clayton-Matthews. "Multinomial logistic regression." *Nursing research* 51.6 (2002): 404-410. [↑](#footnote-ref-44)
45. Kjell J., Kuhn M., *Applied Predictive Modelling*, Springer 2013 [↑](#footnote-ref-45)
46. Trzpiot, Grażyna, and Alicja Ganczarek. "Drzewa decyzyjne w statystycznej analizie decyzji na przykładzie wirtualnych łańcuchów dostaw." *Acta Universitatis Lodziensis. Folia Oeconomica* 271 (2012). [↑](#footnote-ref-46)
47. Bruce P., Bruce A., Gedeck P., *op. cit.,* s.228 [↑](#footnote-ref-47)
48. *Ibidem,* s. 229 [↑](#footnote-ref-48)
49. https://www.learndatasci.com/glossary/gini-impurity/ [dostęp: 12.08.2024] [↑](#footnote-ref-49)
50. https://towardsdatascience.com/decision-trees-explained-entropy-information-gain-gini-index-ccp-pruning-4d78070db36c [dostęp: 12.08.2024] [↑](#footnote-ref-50)
51. Bruce P., Bruce A., Gedeck P., *op. cit.,* s.229 [↑](#footnote-ref-51)
52. *Ibidem* [↑](#footnote-ref-52)
53. Breiman, Leo. "Bagging predictors." *Machine learning* 24 (1996): 123-140. [↑](#footnote-ref-53)
54. Johnson K., Kuhn M., *op. cit.*, s. 198- 200 [↑](#footnote-ref-54)
55. Schonlau, Matthias, and Rosie Yuyan Zou. "The random forest algorithm for statistical learning." *The Stata Journal* 20.1 (2020): 3-29. [↑](#footnote-ref-55)
56. Bruce P., Bruce A., Gedeck P., *op. cit.,* s.205. [↑](#footnote-ref-56)
57. https://towardsdatascience.com/accuracy-precision-recall-or-f1-331fb37c5cb9 [dostęp: 13.08.2024] [↑](#footnote-ref-57)
58. https://ksopyla.com/data-science/precision-recall-f1-miary-oceny-klasyfikatora/ [dostęp: 13.08.2024] [↑](#footnote-ref-58)
59. *Ibidem* [↑](#footnote-ref-59)
60. *Ibidem* [↑](#footnote-ref-60)
61. *Ibidem* [↑](#footnote-ref-61)
62. https://media.statsoft.pl/\_old\_dnn/downloads/krzywe\_roc\_czyli\_ocena\_jakosci.pdf [dostęp: 16.08.2024] [↑](#footnote-ref-62)
63. *Ibidem* [↑](#footnote-ref-63)
64. https://towardsdatascience.com/multiclass-classification-evaluation-with-roc-curves-and-roc-auc-294fd4617e3a [dostęp: 16.08.2024) [↑](#footnote-ref-64)
65. *Ibidem* [↑](#footnote-ref-65)
66. https://www.analyticsvidhya.com/blog/2022/02/k-fold-cross-validation-technique-and-its-essentials/ [dostęp 13.08.2024] [↑](#footnote-ref-66)
67. *Ibidem* [↑](#footnote-ref-67)