Uniwersytet Ekonomiczny w Krakowie

Kolegium Nauk o Zarządzaniu i Jakości

Instytut Informatyki, Rachunkowości i Controllingu

Kierunek: Informatyka Stosowana

Specjalność: Inżynieria Oprogramowania



Karol Bułdak

Nr albumu: 221451

Ocena wpływu selekcji cech na dokładność predykcji wyników meczów NBA za pomocą algorytmów uczenia maszynowego

Praca licencjacka

Promotor  
prof. UEK Dariusz Dymek

Kraków 2024

Spis treści

[Wstęp 3](#_Toc176205931)

[Rozdział 1 Wprowadzenie do problemu 4](#_Toc176205932)

[1.1. Po co przewidywać wyniki meczów ligi NBA 4](#_Toc176205933)

[1.2. Przewidywanie wyników meczowych jako problem klasyfikacji 5](#_Toc176205934)

[1.3. Czym jest uczenie maszynowe 7](#_Toc176205935)

[1.4. Statystyki i dane w lidze NBA 11](#_Toc176205936)

[1.5. Charakterystyka zbiorów danych zawierających statystyki ligi NBA 15](#_Toc176205937)

[Rozdział 2 Opis projektu 17](#_Toc176205938)

[2.1. Algorytmy wyboru cech 17](#_Toc176205939)

[2.2. Wybór sposobów predykcji 21](#_Toc176205940)

[2.3. Sposób przeprowadzenia eksperymentów 22](#_Toc176205941)

[2.4. Opis środowiska programistycznego 23](#_Toc176205942)

[Rozdział 3 Opis implementacji 25](#_Toc176205943)

[3.1. Zbieranie danych 25](#_Toc176205944)

[3.2. Proces selekcji danych treningowych i weryfikacji modelu w czasie 26](#_Toc176205945)

[3.3. Przygotowanie modelu bazowego 28](#_Toc176205946)

[3.4. Opis eksperymentów 28](#_Toc176205947)

[3.5. Wyniki eksperymentów 30](#_Toc176205948)

[3.6. Wnioski 32](#_Toc176205949)

[Zakończenie 34](#_Toc176205950)

[Bibliografia 35](#_Toc176205951)

[Spis rysunków 37](#_Toc176205952)

[Spis tabel 38](#_Toc176205953)

# Wstęp

W czasie dynamicznego wzrostu popularności i wartości branży rozrywki sportowej, w tym koszykówki, a w szczególności ligi NBA, coraz częściej wykorzystuje się przewidywanie wyników sportowych rozgrywek za pomocą algorytmów uczenia maszynowego przez specjalistów zatrudnianych przez drużyny. Przyczyną tego zjawiska jest chęć odniesienia drużynowego sukcesu, co skutkuje większym wkładem finansowym od sponsorów oraz rosnącą wartością drużyny. Jakość modelu uczenia maszynowego może być jednym z czynników wpływających na wynik finansowy organizacji.

Celem niniejszej pracy jest zbadanie wpływu wyboru cech (ang. feature selection) na dokładność modeli predykcyjnych w przewidywaniu wyników meczów ligi NBA. Przewidywanie wyników meczów sportowych, zwłaszcza w tak dynamicznej i pełnej zmiennych lidze jak NBA, stanowi istotne wyzwanie badawcze. Właściwy dobór cech, które najlepiej opisują dynamikę gry, jest kluczowym elementem wpływającym na skuteczność algorytmów uczenia maszynowego.

Praca ta koncentruje się na analizie statystyk z poprzednich sezonów NBA i ocenie, które z nich mają największy wpływ na dokładność prognoz. Wykorzystanie odpowiednich cech jest kluczowe dla budowy modelu, który nie tylko poprawnie klasyfikuje wyniki, ale również ma zdolność generalizacji na nowe, nieznane dane. W ramach pracy zastosowane zostaną różne techniki selekcji cech, a także różne metody uczenia maszynowego, aby ocenić ich wpływ na końcowy wynik predykcji. Motywem podjęcia tematu jest jego aktualność oraz zainteresowanie autora pracy jego problematyką.

Praca składa się z trzech rozdziałów. W rozdziale pierwszym zaprezentowano kontekst ligi NBA, opisano czym jest uczenie maszynowe oraz przedstawiono typy oraz najpopularniejsze modele uczenia maszynowego. Drugi rozdział opisuje algorytmy wykorzystane w projekcie, wykorzystane modele uczenia maszynowego oraz opisane zostało środowisko programistyczne. W trzecim rozdziale opisano przebieg eksperymentu, zaprezentowano jego wyniki, oraz wyciągnięto wnioski.

# Wprowadzenie do problemu

## Po co przewidywać wyniki meczów ligi NBA

Liga National Basketball Association to amerykańska liga koszykówki uważana za najlepszą lige świata. NBA jest trzecią najlepiej zarabiającą ligą sportową na świecie, jej dochody to zaskakujące 10,5 miliarda dolarów na sezon (Ozanian, 2023). W rozgrywkach bierze udział 30 drużyn z Ameryki Północnej, każda z nich rozgrywa 82 mecze w sezonie regularnym trwającym od października do kwietnia oraz do 28 meczy w turnieju play off,   
w sumie w całej lidze rozgrywane jest aż do 1320 meczy na sezon. Wysoki poziom ligi   
i efektownie wyglądające akcje zrzeszają oglądających z całego świata, którzy śledzą rozgrywki i kibicują ulubionym drużynom. Każdy prawdziwy fan chce być na bieżąco   
z informacjami, wynikami oraz nowinkami ze świata NBA. Przy takiej skali popularności ligi, takiej mnogości meczy oraz w obliczu tak wielkiego sukcesu finansowego- prognozowanie wyników staje się kluczowym elementem, nie tylko dla samych kibiców, ale również dla trenerów, analityków sportowych, firm bukmacherskich, inwestorów oraz ekspertów telewizyjnych. W świecie NBA każdy stara się przewidzieć jakie czynniki wpływają na sukces drużyny oraz która drużyna wyjdzie z nadchodzącego meczu zwycięsko. Każda   
z wspomnianych osób ma własne przesłanki i motywacje opisujące dlaczego to robi. Kibice mogą robić to aby lepiej zrozumieć gre i czynniki sukcesu. Niektórzy z nich mogą robić to także z chęci zarobku, gdyż firmy bukmacherskie umożliwiają tworzenie zakładów i kuszą fanów koszykówki perspektywą zysku. Firmy te z kolei zatrudniają analityków i ekspertów którzy znają się na problemie predykcji wyników i innych statystyk i ustalają kursy   
w zależności od prawdopodobieństwa wystąpienia wybranego zdarzenia. Zjawisko obstawiania zakładów bukmacherskich z roku na rok zyskuje na popularności, a firmy bukmacherskie notują rekordowe wzrosty w zyskach, w zeszłym roku zyski tej branży wyniosły ponad 10 miliardów dolarów (Lopez, 2024). Dla drużyn, ważną częścią przygotowania taktyki i składu wyjściowego na następne mecze jest analiza poprzednich historycznych starć zespołów   
i wyciąganie wniosków oraz korekta własnych błędów, ale również próba przewidzenia tego, jakimi taktykami drużyna przeciwna będzie starała się zwyciężyć. Współczesne drużyny to już nie tylko sztab szkoleniowy, zawodnicy i trener, ale również wsparcie licznych ekspertów, którzy zza kulis starają się przewidzieć ruchy innych drużyn. Każdy zespół w lidze, który chce zdobyć przewagę inwestuje środki w zaplecze analityków i ekspertów pracujących na wspólny sukces. Prognozowanie wyników spotkań sportowych, szczególnie w kontekście analizy statystyk, stało się nieodłącznym elementem strategii dla wielu zainteresowanych stron (Horvat i in., 2023). Drużyny sportowe, telewizje, redakcje sportowe oraz entuzjaści zakładów bukmacherskich wykorzystują analizę statystyk w celu osiągnięcia różnych korzyści. Liga NBA, jako najpopularniejsza liga koszykarska na świecie, generuje ogromną ilość danych   
z różnych aspektów gry. Od najbardziej podstawowych statystyk takich jak ilość zdobytych punktów, skuteczność rzutów, liczba zbiórek, asyst, aż po bardziej zaawansowane statystyki, które zostaną opisane w następnych podrozdziałach. Jednakże, mimo dostępności tych informacji, prognozowanie wyników meczów nadal pozostaje wyzwaniem z powodu wielu zmiennych wpływających na przebieg gry. Zmienne te mogą obejmować kontuzje zawodników, zmiany w składzie, formę drużyny, a także czynniki zewnętrzne, takie jak harmonogram meczów czy podróże między miastami. W świecie, w którym wyniki są często nieprzewidywalne, sztuka prognozowania staje się coraz bardziej zaawansowana,   
z wykorzystaniem nowoczesnych technologii, takich jak sztuczna inteligencja i analiza big data, które pomagają w tworzeniu bardziej precyzyjnych przewidywań. Pomimo tych narzędzi, nieprzewidywalność sportu, szczególnie w lidze tak dynamicznej jak NBA, pozostaje integralnym elementem, który fascynuje zarówno ekspertów, jak i kibiców. Każdy   
z przewidujących ma własne sposoby na przewidzenie wyniku, niektórzy kierują się intuicją, inni wczytują się w statystyki, niektórzy zaś tworzą skomplikowane modele matematyczne starając się wyliczyć możliwy rezultat spotkania. Na koniec dnia jednak, każdy z nich mierzy się z tym samym problemem- stwierdzeniem czy dana drużyna wygra, czy też przegra. Problem ten możemy nazwać problemem klasyfikacji, gdzie na podstawie zbioru danych chcemy przewidzieć jakiś rezultat.

## Przewidywanie wyników meczowych jako problem klasyfikacji

### Rozwiązanie problemu metodami uczenia maszynowego

Prognozy sportowe zazwyczaj są traktowane jako problem klasyfikacji, w którym przewiduje się jedną klasę (wygrana/przegrana/remis) (Prasetio & Harlili, 2016), a rzadkie przypadki są przewidywane za pomocą wartości numerycznych, takich jak przewidywanie wyniku lub rozbieżności punktowej. Wynik to dokładna liczba punktów zdobytych przez zespół, podczas gdy rozbieżność punktowa to różnica między zdobytymi punktami. Przewidywanie wyniku lub rozbieżności punktowej jest zdecydowanie zadaniem o wiele trudniejszym niż przewidywanie wyników. Przewidywanie wyników związanych z dwiema możliwymi klasami jest mniej wymagające niż przewidywanie wyników związanych z trzema możliwymi klasami. Podobnie łatwiej jest przewidzieć wynik sportów indywidualnych   
w porównaniu do sportów zespołowych, w których występ większej liczby uczestników wpływa na ostateczny wynik (Horvat & Job, 2020). Bunker & Susnjak (2022) przeprowadzili przegląd prac wykorzystujących uczenie maszynowe do przewidywania wyników w sportach zespołowych. Ostatecznym celem było odpowiedzenie na kluczowe pytania badawcze dotyczące prognozowania wyników sportowych. Autorzy doszli do wniosku, że większość badań (65%) brała pod uwagę sztuczne sieci neuronowe w swoich eksperymentach i że sztuczne sieci neuronowe niekoniecznie radzą sobie lepiej niż inne algorytmy uczenia maszynowego. Stwierdzono również, że wcześniejsze badania często wybierały cechy predykcyjne ręcznie, zwykle opierając się na wiedzy badaczy, i że posiadanie dużej bazy danych niekoniecznie prowadzi do wysokiej dokładności. Doszli do wniosku, że sporty   
o niskim wyniku oraz sporty z większą liczbą możliwych wyników zazwyczaj charakteryzują się niższą dokładnością. Autorzy stwierdzili również, że w literaturze dominuje kilka algorytmów uczenia maszynowego, a mianowicie algorytm Regresji Grzbietowej, Metoda Wektorów Nośnych (SVM), K-najbliższych sąsiadów (k-NN) oraz Regresji Logistycznej. Stosowane są zarówno w problemach klasyfikacji, jak i problemach regresji.

### Rozwiązanie problemu metodami statystycznymi

W literaturze omawiane są również metody statystyczne, które mogą być użyte do rozwiązywania problemów klasyfikacyjnych. Do najczęściej wykorzystywanych należą test Chi-kwadrat oraz Liniowa Dyskryminanta Fishera (Clancey, 1984). Te metody są szczególnie użyteczne w analizie zależności między zmiennymi kategorycznymi i mogą być stosowane jako alternatywa dla bardziej złożonych algorytmów uczenia maszynowego. W przypadku prognozowania wyników sportowych, test Chi-kwadrat może być używany do analizy danych kategorycznych, takich jak wyniki meczów (wygrana, przegrana, remis) w odniesieniu do różnych cech drużyn lub zawodników, np. formy zespołu, liczby kontuzji, historii bezpośrednich spotkań itp. Dzięki testowi Chi-kwadrat można na przykład sprawdzić, czy wynik meczu zależy od tego, czy drużyna gra u siebie czy na wyjeździe. Jeśli wartość uzyskana w wyniku testu jest niższa niż przyjęty poziom istotności, można stwierdzić, że istnieje statystycznie istotna zależność między tymi zmiennymi, co może być cenną informacją przy prognozowaniu przyszłych wyników (Sampaio i in., 2015). Liniowa Dyskryminanta Fishera (LDA) jest metodą statystyczną stosowaną głównie w problemach klasyfikacyjnych. Głównym celem LDA jest znalezienie linii (lub płaszczyzny w przypadku większej liczby wymiarów), która najlepiej oddziela klasy (np. wygrana, przegrana, remis) w przestrzeni cech. LDA maksymalizuje stosunek wariancji międzyklasowej do wariancji wewnątrzklasowej, co oznacza, że szuka takich projekcji danych, które najlepiej oddzielają różne klasy (McCabe & Trevathan, 2008). W kontekście sportowym, LDA może być używana do klasyfikacji wyników meczu na podstawie wielu cech, takich jak statystyki dotyczące zawodników, warunki atmosferyczne, poprzednie wyniki, i inne zmienne.

## Czym jest uczenie maszynowe

Gdy pierwszy raz zderzymy się z pojęciem uczenie maszynowe, może wydawać się ono dziwne i niezrozumiałe. W jaki sposób można nauczyć maszynę czegokolwiek? Słownik języka polskiego definiuje uczenie się jako „Przyswajanie sobie pewnego zasobu wiedzy, zdobywanie jakiejś umiejętności” (*uczenie się - definicja, synonimy, przykłady użycia*). Zatem aby nauczyć maszynę czegoś, musi ona posiąść jakąś umiejętność. Podczas zwykłego programowania podajemy szczegółowo kroki potrzebne do rozwiązania problemu, które maszyna następnie wykonuje i tak właśnie rozwiązuje dany problem. Nie możemy jednak nazwać tego procesu uczeniem, ponieważ przecież podaliśmy maszynie na tacy kroki które ma podjąć, by osiągnąć cel. Proces ten nie spełnia definicji uczenia, gdyż maszyna sama nie zdobyła umiejętności, a jedynie wykonała kroki które zostały zadane. Tu z pomocą przychodzą nam algorytmy nauczania maszynowego, są to zestawy matematycznych reguł i technik, które umożliwiają komputerom uczenie się ze zbiorów danych i podejmowanie decyzji bez wyraźnego programowania przez człowieka. Algorytmy uczenia maszynowego budują model matematyczny na podstawie przykładowych danych, zwanych zbiorem uczącym, w celu prognozowania lub podejmowania decyzji bez bycia zaprogramowanym przez człowieka do tego celu. Model ten po otrzymaniu danych na wejściu korzysta z poprzednich doświadczeń które zebrał na zbioru uczącym i może na wyjściu podać rozwiązanie problemu (Burkov, 2019). Od tego jaki problem model może rozwiązać zależy to jaki typ algorytmu został wykorzystany. Typy uczenia maszynowego dzielimy na:

* Uczenie nadzorowane (Supervised Learning): Model jest uczony na danych, które zawierają zarówno wejście, jak i oczekiwane wyjście, czyli etykietę. Celem jest nauczenie modelu mapowania z wejścia na wyjście.
* Uczenie nienadzorowane (Unsupervised Learning): W tym przypadku model jest uczony na danych, które nie posiadają etykiet. Model musi samodzielnie znajdować struktury w danych, szukając wzorców i zależności.
* Uczenie ze wzmocnieniem (Reinforcement Learning): Model uczy się na podstawie interakcji z otoczeniem, podejmując działania i otrzymując nagrody lub kary w zależności od podejmowanych decyzji.

Najpopularniejsze algorytmy wykorzystywane w tych podejściach do uczenia maszynowego to:

* Regresja liniowa (Linear Regression): Umożliwia badanie związku pomiędzy zmiennymi, takimi jak statystyki drużyn lub graczy a wynikiem meczu. Polega ona na dopasowaniu linii do danych treningowych w taki sposób, aby jak najlepiej opisywała zależność między czynnikami objaśniającymi (np. liczba zdobytych punktów, skuteczność rzutów) a czynnikiem objaśnianym, którym może być wynik meczu. Regresja liniowa pozwala na zrozumienie, w jaki sposób różne statystyki wpływają na końcowy rezultat spotkania. Dzięki temu modelowi można określić, które zmienne mają największy wpływ na przewidywanie zwycięzcy, co ułatwia interpretację wyników   
  i podejmowanie decyzji strategicznych. Mimo swojej prostoty, regresja liniowa często stanowi solidną bazę dla bardziej zaawansowanych modeli uczenia maszynowego. Jej przejrzystość i łatwość implementacji sprawiają, że jest chętnie wykorzystywana jako punkt wyjścia w analizach danych sportowych

Obraz zawierający diagram, linia, tekst, Wykres

Opis wygenerowany automatycznie

Rysunek 1 Wykres regresja logistyczna

* Regresja logistyczna (Logistic Regression): Algorytm uczenia maszynowego, który znajduje szerokie zastosowanie w prognozowaniu wyników sportowych. Jego głównym celem jest oszacowanie prawdopodobieństwa, że dany mecz zakończy się zwycięstwem lub porażką, na podstawie zestawu czynników objaśniających, takich jak statystyki drużyn czy graczy. W przeciwieństwie do regresji liniowej, która przewiduje wartości ciągłe, regresja logistyczna koncentruje się na klasyfikacji binarnej, przypisując obserwacje do jednej z dwóch kategorii: zwycięstwo lub porażka. Kluczową zaletą tej metody jest jej zdolność do modelowania nieliniowych zależności między predyktorami a zmienną objaśnianą. Dzięki temu regresja logistyczna może skutecznie radzić sobie   
  z interakcjami i złożonymi wzorcami w danych, co jest szczególnie istotne w przypadku prognozowania wyników sportowych, gdzie wiele czynników wpływa na końcowy rezultat. Ponadto, regresja logistyczna dostarcza nie tylko przewidywań, ale również prawdopodobieństwa przynależności do poszczególnych klas. Ta informacja może być niezwykle cenna dla analityków i trenerów, umożliwiając im ocenę ryzyka   
  i podejmowanie strategicznych decyzji opartych na solidnych podstawach statystycznych.

Obraz zawierający diagram, linia, tekst, Wykres

Opis wygenerowany automatycznie

Rysunek 2 Wykres regresja logistyczna

* K-najbliższych sąsiadów (K-nearest neighbor): Algorytm który znajduje zastosowanie zarówno w zadaniach klasyfikacji, jak i regresji. Jego podstawowa idea opiera się na założeniu, że nowa obserwacja powinna zostać przypisana do tej samej klasy lub otrzymać podobną wartość przewidywaną, co inne obserwacje znajdujące się w jej najbliższym otoczeniu w przestrzeni cech. Algorytm k-NN działa w następujący sposób: dla nowej obserwacji, której klasę lub wartość chcemy przewidzieć, identyfikuje on   
  k najbliższych sąsiadów spośród danych treningowych, biorąc pod uwagę odległość   
  w przestrzeni cech. Następnie, w przypadku klasyfikacji, nowa obserwacja jest przypisywana do klasy, która występuje najczęściej wśród tych k sąsiadów. W przypadku regresji, przewidywana wartość jest średnią wartości docelowych k najbliższych sąsiadów.

Obraz zawierający zrzut ekranu, diagram, linia, Wykres

Opis wygenerowany automatycznie

Rysunek 3 Wykres K-najbliższych sąsiadów

## Statystyki i dane w lidze NBA

Liga NBA jest ligą koszykarską z niezwykle wysokim poziomem rywalizacji. Trzydzieści drużyn rywalizuje ze sobą w sezonie regularnym trwającym 82 mecze rozgrywanym w okresie 24 tygodni, aby znaleźć się w grupie szesnastu najlepszych drużyn i mieć szansę na zagranie w turnieju eliminacyjnym mającym na celu wyłonienie mistrza ligi. Zawodnicy, trenerzy oraz cały sztab drużyny wchodzi na wyżyny swoich umiejętności w celu zmaksymalizowania przewagi którą mogą uzyskać nad innymi (Manley, 1989). Liga NBA jest nie tylko areną intensywnej rywalizacji, ale także ogromnym źródłem danych, które odgrywają kluczową rolę w analizie gry i strategii drużyn. Proces pozyskiwania danych o graczach   
i meczach jest bardzo rozbudowany, obejmując różne źródła informacji. Kluczowe dane dotyczące zawodników są gromadzone zarówno podczas meczów, jak i podczas treningów. Podczas gier, kamery śledzą każdy ruch zawodników, rejestrując ich punkty, asysty, skuteczność rzutów czy liczby zbiórek. Dane te są następnie przetwarzane i archiwizowane, aby stanowić bogate źródło informacji na temat indywidualnych umiejętności i taktyk każdego gracza. Ponadto, liga NBA korzysta z zaawansowanych systemów monitoringu, takich jak kamery umieszczone na koszach, analizujące trajektorie rzutów i ich skuteczność. Oprócz tego, dane są gromadzone również podczas treningów, gdzie nowoczesne technologie pomagają monitorować postępy zawodników, ich obciążenia treningowe, a także ewentualne kontuzje (Georgievski Bojan & Vrtagic Sabahundin, 2021). Obecnie wydajność zawodników i drużyn jest obiektywnie oceniana przy użyciu różnych indeksów efektywności. Prostym i powszechnie stosowanym wskaźnikiem wydajności zawodnika w lidze NBA jest NBA Player Efficiency Index (Manley, 1989). Jego pierwotna forma odnosi się do efektywności pojedynczego zawodnika w jednym meczu, ale może również być uśredniona na przestrzeni kilku spotkań lub obliczana dla całej drużyny. Indeks efektywności zawodnika NBA uwzględnia trzynaście podstawowych elementów gry w koszykówkę, które są uwiecznione w standardowych tabelach statystyk koszykarskich NBA, zwanych box scores (Horvat i in., 2023). Elementy mogą być pozytywne (+), negatywne (-) bądź neutralne (0). Elementy zostały wymienione w Tabela 1.

Tabela 1 Trzynaście podstawowych elementów meczu koszykarskiego.

|  |  |  |
| --- | --- | --- |
| Element (typ) | Skrót | Opis |
| Podstawowy (+) | gmd2F | Rzuty trafione za 2 pkt |
| gmd3F | Rzuty trafione za 3 pkt |
| gmdFT | Rzuty osobiste trafione |
| asts | Asysty |
| blcks | Rzuty zablokowane |
| rbd | Zbiórki defensywne |
| rbo | Zbiórki ofensywne |
| stls | Piłki przechwycone |
| Podstawowy (0) | gat2F  gat3F  gatFT | Próby rzutowe za 2 pkt  Próby rzutowe za 3 pkt  Próby rzutowe osobiste |
| Podstawowy (1) | fls | Faule |
| tos | Straty |

Standardowe elementy gry prowadzą do kilku dodatkowych elementów gry, które nazywamy elementami pochodnymi są one wymienione w Tabela 2. Pierwsze dwa z nich, czyli elementy reprezentujące łączną liczbę zdobytych punktów i łączną liczbę zbiórek, są pozytywnymi wynikami gry (oznaczonymi "+" w pierwszej kolumnie), a pozostałe cztery elementy mają charakter negatywny, oznaczony symbolem "-" w pierwszej kolumnie.

Tabela 2 Osiem pochodnych elementów meczu koszykarskiego.

|  |  |  |
| --- | --- | --- |
| Element (Typ) | Skrót | Opis |
| Pochodny (+) | pts | Punkty łącznie |
| gmdFld | Rzuty trafione z pola za 2 i za 3 punkty łącznie |
| rbs | Zbiórki łącznie |
| Pochodny (0) | gatFld | Próby rzutowe z pola za 2 i za 3 punkty łącznie |
| Pochodny (-) | ms2F | Próby rzutowe z pola za 2 punkty nie trafione |
| ms3F | Próby rzutowe z pola za 3 punkty nie trafione |
| msFld | Rzuty nie trafione z pola za 2 i za 3 punkty łącznie |
| msFT | Rzuty osobiste nie trafione |

Wartości elementów pochodnych wynikają z ich opisu:

Statystyki te to cenne dane opisujące mecze, które możemy użyć do uczenia maszynowego.   
W zależności od tego jakie dane zostaną podane jako wejście do algorytmu dokładność przewidywania wyniku meczu może się różnić.

## Charakterystyka zbiorów danych zawierających statystyki ligi NBA

Charakterystyka zbioru danych stworzonego na podstawie historycznych danych   
z rozegranych meczów obejmuje szczegółową analizę struktury, jakości oraz właściwości tych danych. Taki zbiór danych zawiera informacje dotyczące kluczowych zmiennych i statystyk związanych z przebiegiem poszczególnych spotkań, umożliwiając głęboką analizę zarówno na poziomie drużynowym, jak i indywidualnym. W tym kontekście, istotne jest zrozumienie rozkładów danych, identyfikacja potencjalnych wartości odstających, analiza braków danych oraz jakości danych, co pozwala na efektywne wykorzystanie tych danych w modelowaniu predykcyjnym i innych zaawansowanych technikach analitycznych. Charakterystykę zbioru danych można opisać następująco:

* Dane zawarte w zbiorze to dane numeryczne, które przyjmują wartości całkowite lub wartości zmiennoprzecinkowe,
* Liczba obserwacji zależy od ilości sezonów które bierzemy pod uwagę podczas rozwiązywania problemu klasyfikacji, zwykle aby otrzymać godną reprezentacje tego jak radzą sobie poszczególne drużyny na przestrzeni lat należy zgromadzić dane liczone   
  w tysiącach obserwacji (Loeffelholz i in., 2009),
* Liczba cech które bierzemy pod uwagę powinna być wystarczająca by opisać dane zdarzenie, które chcemy przewidzieć. Najczęściej do modelu przekazujemy komplet statystyk opisujący wydajność drużyn w ofensywie oraz obronie (Leung & Joseph, 2014),
* Braki danych w zbiorach nie występują, wszystkie statystyki w lidze zapisywane są skrupulatnie rejestrowane co eliminuje problem występowania wartości pustych, nie musimy dzięki temu również stosować metod uzupełniania danych, co poprawia skuteczność metod klasyfikacji zastosowanych na podstawie zbioru,
* W zbiorach danych statystyk NBA wartości odstające występują rzadko, ponieważ przypadki, w których drużyny osiągają skrajnie niską lub skrajnie wysoką wydajność, prowadzącą do wyjątkowo dużych różnic punktowych, są stosunkowo sporadyczne,
* Jakość danych statystycznych w NBA jest na najwyższym poziomie, co wynika   
  z zatrudnienia licznego zespołu specjalistów odpowiedzialnych za zbieranie, analizowanie i weryfikację danych. Procesy te minimalizują ryzyko błędów, dzięki czemu dane statystyczne wiernie odzwierciedlają rzeczywisty przebieg wydarzeń na boisku.
* Przy stosowaniu tych danych statystycznych w celu rozwiązywania problemu klasyfikacji występuje konieczność skalowania danych, jeśli chcemy uzyskać jak najwyższy wynik dokładności. Zmienne mają różne jednostki i zakresy wartości, co może wpływać na wyniki modeli klasyfikacyjnych. Algorytmy takie jak SVM, K-najbliższych sąsiadów (k-NN) czy regresja logistyczna są wrażliwe na różnice w skali zmiennych, więc standaryzacja lub normalizacja danych zapewnia, że każda cecha ma równy wpływ na proces uczenia, co poprawia dokładność i efektywność modeli predykcyjnych.

# Opis projektu

## Algorytmy wyboru cech

Celem projektu będzie ocena wpływu selekcji cech na dokładność predykcji wyników meczów koszykówki. Aby to sprawdzić użyte zostaną metody predykcji, których zadaniem będzie przewidzenie, która drużyna wygra mecz na podstawie dostępnych statystyk meczowych z poprzednich sezonów. Parametrami wejściowymi które model będzie analizował i na których podstawie będzie się uczył będą statystyki z historycznych sezonów ligi. Parametry te są podstawowymi danymi wejściowymi używanymi do treningu modelu. Należy wybrać te, które pozwolą na największą dokładność modelu, w tym celu wykorzystano algorytmy selekcji cech. Te algorytmy mają za zadanie wybór najbardziej istotnych cech (zmiennych) w zbiorze danych tak, aby poprawić wydajność modelu predykcyjnego. Jest to szczególnie użyteczne   
w przypadku gdy mamy do czynienia z dużą liczbą cech i chcemy zredukować ich liczbę, eliminując te mniej ważne lub redundantne. W tym projekcie badano to jak wybór cech wpłynie na dokładność modelu, dlatego zdecydowano się na porównanie dokładności trzech modeli wytrenowanych na cechach wybranych przez trzy różne algorytmy selekcji cech.

### Algorytm Sekwencyjnej Selekcji Cech

Pierwszym wybranym algorytmem jest Algorytm Sekwencyjnej Selekcji Cech (ang. Sequential Feature Selection). Algorytm ten jest niezwykle elastyczny, łatwy w implementacji oraz relatywnie niekosztowny obliczeniowo, co czyni go popularnym wyborem w różnych zastosowaniach analitycznych i w nauce o danych. Jego główną zaletą jest zdolność do iteracyjnego redukowania liczby cech wejściowych, przy jednoczesnym zachowaniu lub nawet poprawie wydajności modelu. Proces selekcji cech za pomocą tego algorytmu przebiega   
w kilku krokach:

1. Inicjalizacja: W pierwszej iteracji, algorytm rozpoczyna od przetestowania każdej cechy indywidualnie. Budujemy osobny model dla każdej cechy i oceniamy jego wydajność (np. za pomocą miary dokładności, precyzji, f-score, lub innej miary odpowiedniej dla danego zadania). Spośród wszystkich cech wybieramy tę, która generuje najlepszy wynik.
2. Dodawanie kolejnych cech: W kolejnych iteracjach algorytm dodaje jedną nową cechę spośród jeszcze nie wybranych, do zestawu cech już wybranych. Na każdym kroku oceniamy wydajność modelu z nową kombinacją cech, ponownie wybierając tę cechę, która przynosi największą poprawę.
3. Zatrzymanie procesu: Proces dodawania nowych cech trwa tak długo, aż żadna   
   z pozostałych cech nie poprawi wyniku modelu w sposób znaczący. W momencie, gdy dodanie kolejnej cechy nie przynosi poprawy, algorytm zatrzymuje się, uznając obecny zestaw cech za optymalny.

Algorytm Sekwencyjnej Selekcji Cech jest zatem bardzo przydatnym narzędziem, zwłaszcza w sytuacjach, gdy dysponujemy dużą liczbą cech, a chcemy zredukować ich liczbę, jednocześnie zachowując lub poprawiając wydajność modelu. Działa on efektywnie zarówno w przypadku cech silnie skorelowanych, jak i wtedy, gdy w zestawie danych znajdują się cechy mało informatywne, które mogą pogarszać jakość modelu. Jednym   
z kluczowych atutów tego algorytmu jest jego zdolność do adaptacji w zależności od modelu i danych, co sprawia, że jest szeroko stosowany w różnych dziedzinach, a także w zadaniach takich jak klasyfikacja, regresja, czy prognozowanie. Dzięki swojej prostocie   
i efektywności, Algorytm Sekwencyjnej Selekcji Cech jest często wykorzystywany   
w praktyce, stanowiąc solidną podstawę do dalszych, bardziej zaawansowanych metod selekcji cech.

### Algorytm Select K Best

Następnym użytym algorytmem jest Algorytm Select K Best, który jest jedną   
z najprostszych i jednocześnie efektywnych metod selekcji cech. Jego głównym celem jest wybranie określonej liczby cech (K), które mają największy wpływ na efektywność modelu, opierając się na wybranej miarze oceny ich istotności. Algorytm ten jest szczególnie przydatny w sytuacjach, gdy dysponujemy dużą liczbą cech i chcemy szybko zredukować ich liczbę, jednocześnie skupiając się na tych, które są najbardziej wartościowe dla danego zadania analitycznego. Proces selekcji cech przy użyciu algorytmu Select K Best przebiega w kilku krokach:

1. Ocena cech: Na początku algorytm ocenia każdą cechę indywidualnie, stosując wybraną miarę oceny. Miara ta może być dostosowana do charakteru danych oraz problemu, który próbujemy rozwiązać. Przykładowymi miarami są:

* Statystyka chi-kwadrat: Często stosowana w przypadku danych kategorycznych, pozwala na ocenę istotności poszczególnych cech w kontekście rozkładu wyników,
* Wzajemna informacja (Mutual Information): Jest to miara wzajemnej zależności między dwiema zmiennymi, co pozwala na identyfikację cech, które dostarczają najwięcej informacji o zmiennej docelowej,
* Wartość p dla testu t: Wykorzystywana głównie dla danych numerycznych, umożliwia ocenę, czy różnice w wartościach cech są statystycznie istotne w kontekście zmiennej docelowej.

1. Sortowanie cech: Po ocenie każdej cechy, algorytm sortuje je w malejącej kolejności, bazując na uzyskanych wynikach miary oceny. Cechy, które uzyskały najwyższe wartości, są uznawane za najbardziej istotne.
2. Wybór najlepszych cech: Spośród wszystkich ocenionych i posortowanych cech, algorytm wybiera dokładnie K cech, które uzyskały najwyższe wyniki według wybranej miary oceny. To właśnie te cechy będą następnie użyte w procesie budowy modelu, podczas gdy reszta zostanie odrzucona.

Algorytm Select K Best jest niezwykle użytecznym narzędziem, gdy zależy nam na szybkim   
i prostym zredukowaniu liczby cech. Jego zaletą jest prostota implementacji oraz elastyczność, ponieważ możemy dostosować miarę oceny do specyfiki naszego problemu. Dzięki temu, algorytm ten znajduje szerokie zastosowanie w różnych dziedzinach, takich jak klasyfikacja, regresja, czy analiza danych tekstowych. W praktyce Select K Best jest często używany jako pierwszy krok w procesie selekcji cech, szczególnie gdy mamy do czynienia z danymi wysokowymiarowymi, czyli takimi, gdzie liczba cech jest znacznie większa od liczby próbek. Dzięki jego zastosowaniu możemy szybko ograniczyć zbiór cech do tych najbardziej informatywnych, co nie tylko przyspiesza trening modelu, ale także może prowadzić do poprawy jego dokładności i ogólnej wydajności. Algorytm ten pozwala również na uniknięcie problemu przetrenowania, które może wystąpić, gdy model jest budowany na zbyt dużej liczbie cech, z których nie wszystkie są istotne.

### Algorytm Rekurencyjnej Eliminacji Cech

Ostatnim z użytych algorytmów jest Algorytm Rekurencyjnej Eliminacji Cech (RFE, ang. Recursive Feature Elimination). RFE jest zaawansowaną, iteracyjną metodą selekcji cech, której celem jest systematyczne usuwanie najmniej istotnych cech z zestawu danych, aż do osiągnięcia pożądanej liczby cech. Dzięki temu algorytmowi możemy skoncentrować się wyłącznie na tych cechach, które mają największy wpływ na efektywność modelu, eliminując te, które wnoszą niewiele informacji lub wręcz mogą być szkodliwe dla jego jakości. Proces selekcji cech przy użyciu algorytmu RFE przebiega w kilku kluczowych etapach:

1. Tworzenie modelu bazowego: Na początku procesu budowany jest model bazowy   
   z wykorzystaniem wszystkich dostępnych cech. Może to być dowolny model predykcyjny, np. regresja liniowa, maszyna wektorów nośnych (SVM) lub drzewo decyzyjne. Model ten jest używany jako narzędzie do oceny istotności każdej cechy.
2. Ocena istotności cech: Po zbudowaniu modelu, algorytm ocenia istotność każdej cechy. W zależności od rodzaju modelu bazowego, istotność ta może być mierzona na różne sposoby:

* W przypadku regresji liniowej, istotność cechy może być oceniana na podstawie współczynników wag przypisanych każdej zmiennej w równaniu regresji.
* W przypadku drzew decyzyjnych lub modeli opartych na drzewach, istotność cechy może być oceniana na podstawie znaczenia, czyli tego, jak często i z jakim efektem dana cecha jest wykorzystywana do dzielenia węzłów w drzewie.
* W przypadku maszyn wektorów nośnych, można ocenić istotność cech na podstawie współczynników wektora podporowego.

1. Eliminacja najmniej istotnych cech: Po ocenie istotności cech, te o najniższych wartościach współczynników lub najmniejszym znaczeniu są usuwane z zestawu danych. Algorytm eliminuje jedną lub więcej cech w każdej iteracji, w zależności od wcześniej zdefiniowanej strategii.
2. Powtórzenie procesu: Proces oceny istotności i eliminacji cech jest powtarzany iteracyjnie. Za każdym razem tworzony jest nowy model z pozostałymi cechami, ponownie ocenia się ich istotność, a następnie eliminuje najmniej istotne. Ten cykl powtarza się aż do momentu, gdy w zestawie danych pozostanie określona liczba cech, która jest uznana za optymalną dla modelu.
3. Końcowy wybór cech: Na końcu procesu RFE pozostają tylko te cechy, które   
   w największym stopniu przyczyniają się do wydajności modelu. To właśnie te cechy są wykorzystywane do ostatecznego trenowania modelu, co zazwyczaj prowadzi do lepszej generalizacji i wyższej dokładności na danych testowych.

Algorytm RFE jest szczególnie przydatny, gdy zależy nam na zidentyfikowaniu i zachowaniu tylko tych cech, które są kluczowe dla predykcji, jednocześnie eliminując cechy mniej istotne, które mogą dodawać niepotrzebną złożoność do modelu i zwiększać ryzyko przetrenowania. Jedną z głównych zalet RFE jest jego zdolność do adaptacyjnego dostosowywania się do charakterystyki danych i modelu. W porównaniu do prostszych metod selekcji cech, takich jak Select K Best, RFE uwzględnia interakcje między cechami, co może prowadzić do lepszego zrozumienia struktury danych i lepszych wyników modelu. Ponadto, RFE jest elastyczny   
i może być stosowany z różnymi typami modeli bazowych, co czyni go uniwersalnym narzędziem w zestawie narzędzi każdego analityka danych. Dzięki swojej iteracyjnej naturze, algorytm ten pozwala na głębszą eksplorację i optymalizację zestawu cech, co jest niezwykle cenne w praktyce, zwłaszcza w zadaniach o wysokiej złożoności i dużej liczbie cech.

## Wybór sposobów predykcji

Aby przedstawić szeroki wachlarz dostępnych rozwiązań problemu zdecydowano się na użycie trzech różnych sposobów predykcji. Pierwszym z nich jest Regresja grzbietowa, będąca formą regresji liniowej. Zastosowana została z powodu jej zdolności do radzenia sobie z problemami związanymi z wielokolinearnością danych oraz z przewagą, jaką oferuje   
w przypadku dużej liczby cech. Model ten dodaje do funkcji kosztu współczynnik kary proporcjonalny do kwadratu wielkości współczynników modelu. Ta forma regularyzacji pomaga w stabilizowaniu oszacowań współczynników, co z kolei zmniejsza ryzyko nadmiernego dopasowania modelu (overfitting). W kontekście przewidywania wyników meczów NBA, dane wejściowe składają się z wielu różnych statystyk, z których niektóre mogą być silnie skorelowane. Wielokolinearność między zmiennymi może prowadzić do niestabilności modeli, w wyniku czego ich predykcje stają się mniej wiarygodne. Regresja grzbietowa skutecznie łagodzi ten problem, zmniejszając wariancję oszacowań i zwiększając odporność modelu na szum w danych. Kolejnym wybranym sposobem predykcji jest Regresja Logistyczna. Została wybrana, ponieważ doskonale sprawdza się w problemach klasyfikacyjnych, które można traktować jako zdarzenia binarne (np. wygrana lub przegrana). Umożliwia oszacowanie prawdopodobieństwa przypisania wyniku meczu do jednej z dwóch klas, co jest szczególnie przydatne w tym kontekście. Dodatkowo, regresja logistyczna zapewnia interpretowalność wyników i jest odporna na problem nadmiernego dopasowania, co sprawia, że jest skutecznym narzędziem w analizie danych sportowych. Trzecim sposobem predykcji jest Algorytm K najbliższych sąsiadów. KNN należy do algorytmów nienadzorowanych, które nie zakładają żadnej struktury funkcji predykcyjnej ani nie wymagają wcześniejszego trenowania modelu. Zamiast tego, decyzja o przynależności do danej klasy opiera się na głosowaniu większościowym sąsiadów w przestrzeni cech. KNN został wybrany do przewidywania wyników meczów NBA ze względu na swoją prostotę i efektywność   
w sytuacjach, gdy struktura danych jest skomplikowana lub trudna do modelowania przy użyciu bardziej złożonych metod. Algorytm ten jest nieparametryczny, co oznacza, że nie wprowadza żadnych dodatkowych założeń dotyczących rozkładu danych, następną jego zaletą jest jego zdolność do adaptacji do lokalnych właściwości danych, co sprawia, że jest szczególnie efektywny w sytuacjach, gdy dane są nieregularnie rozłożone lub występują złożone zależności między cechami.

## Sposób przeprowadzenia eksperymentów

Aby rzetelnie zbadać wpływ selekcji cech na dokładność modeli, eksperyment zostanie przeprowadzony kilkukrotnie, z zastosowaniem różnych wartości parametrów selektorów cech odpowiedzialnych za liczbę wybranych cech. Przewidujemy, że różne ilości cech wybranych przez algorytmy przyniosą zróżnicowane rezultaty, a dokładna analiza tych wyników może znacząco ułatwić proces trenowania modeli. W wielu algorytmach stosowanych do selekcji cech złożoność obliczeniowa rośnie wraz z liczbą cech przekazanych do wyboru. W związku   
z tym, korzystne jest zminimalizowanie liczby cech poddawanych selekcji, aby zmniejszyć obciążenie obliczeniowe i skrócić czas przetwarzania. Podczas badania optymalnych parametrów przyjęto trzy poziomy procentowe dla wszystkich cech, wynoszące 5%, 20%   
i 35%. Każdy z tych poziomów reprezentuje różne zakresy liczby cech, co pozwala na bardziej szczegółową analizę i ocenę wpływu poszczególnych parametrów na wyniki badania. Zastosowanie tych trzech poziomów umożliwi ocenę tego jak zmienia się dokładność modeli w zależności od ilości wybranych cech oraz jak różne zestawy cech wpływają na ich efektywność. Niższy poziom procentowy, jak 5%, pozwala na zbadanie, jak dobrze modele radzą sobie z minimalnym zestawem cech, co może być kluczowe w sytuacjach, gdzie priorytetem jest szybkość i efektywność obliczeniowa. Z kolei wyższy poziom, jak 35%, pozwala na ocenę, jak większa liczba cech może wpłynąć na zwiększenie dokładności modeli, choć przy zwiększonym koszcie obliczeniowym. Pośredni poziom 20% stanowi kompromis, pozwalający na znalezienie balansu między złożonością obliczeniową a dokładnością modelu. Dzięki przyjęciu tych trzech poziomów procentowych, uzyskamy pełniejszy obraz wpływu selekcji cech na trenowanie modeli, co w dłuższej perspektywie pozwoli na optymalizację procesu i lepsze dostosowanie parametrów selekcji do specyficznych wymagań badania.

## Opis środowiska programistycznego

Python jest jednym z najczęściej wykorzystywanych języków programowania do implementacji algorytmów uczenia maszynowego, co wynika z jego przejrzystości, prostoty oraz rozbudowanego ekosystemu narzędzi wspierających analizę danych i budowanie modeli. Stworzony przez Guido van Rossum w 1991 roku, Python zyskał uznanie jako wszechstronny język wysokiego poziomu, którego składnia jest zbliżona do języka naturalnego. To sprawia, że Python jest intuicyjny i łatwy do nauki, co znacząco ułatwia proces tworzenia oprogramowania, szczególnie w obszarze uczenia maszynowego. W kontekście uczenia maszynowego, Python oferuje kilka kluczowych narzędzi, które umożliwiają efektywną pracę z danymi oraz tworzenie modeli predykcyjnych. Do najważniejszych z nich należą biblioteki Pandas oraz Scikit-learn. Pandas jest fundamentalną biblioteką w ekosystemie Pythona, służącą do manipulacji i analizy danych. Jej główną strukturą danych jest DataFrame, dwuwymiarowa tabela, która umożliwia przechowywanie danych w formie wierszy i kolumn. Każda kolumna w DataFrame może zawierać różne typy danych, takie jak liczby całkowite, liczby zmiennoprzecinkowe, ciągi znaków, co czyni ją niezwykle elastycznym narzędziem do pracy z danymi. Dzięki Pandas można łatwo wczytywać, przetwarzać, filtrować i analizować dane, co stanowi kluczowy krok w przygotowywaniu danych do modelowania. Gdy dane są już odpowiednio przetworzone i przygotowane za pomocą Pandas, można je przekazać do Scikit-learn, jednej z najpopularniejszych bibliotek w Pythonie, dedykowanej uczeniu maszynowemu. Scikit-learn oferuje szeroki wachlarz algorytmów do klasyfikacji, regresji, klasteryzacji, a także narzędzia do przetwarzania danych, selekcji cech oraz walidacji modelu. Biblioteka ta jest szczególnie ceniona za swoją prostotę, bogatą dokumentację oraz możliwość szybkiego prototypowania modeli. Scikit-learn, w połączeniu z Pandas, umożliwia pełny cykl pracy   
z danymi od ich wstępnego przetwarzania, przez tworzenie i trenowanie modeli, aż po ocenę ich jakości. Dzięki temu Python jest idealnym narzędziem zarówno dla badaczy, jak   
i praktyków. Warto również zwrócić uwagę na Jupyter Notebook, który stanowi doskonałe środowisko do eksploracyjnego programowania, analizy danych oraz wizualizacji wyników. Jupyter Notebook umożliwia tworzenie interaktywnych dokumentów, łączących kod, wyniki obliczeń, wykresy i komentarze w jednym miejscu. To narzędzie jest niezwykle przydatne podczas pracy z Pandas i Scikit-learn, gdyż pozwala na wykonywanie kodu krok po kroku, natychmiastowe przeglądanie wyników oraz dokumentowanie procesu uczenia maszynowego. Podsumowując, Python, z kluczowymi narzędziami takimi jak Pandas, Scikit-learn oraz Jupyter Notebook, oferuje kompleksowe wsparcie w realizacji projektów uczenia maszynowego. Te narzędzia pozwalają na efektywne zarządzanie danymi, szybkie prototypowanie modeli oraz intuicyjną prezentację wyników, co czyni jezyk Python jednym z najważniejszych języków   
w dziedzinie analizy danych i uczenia maszynowego.

# Opis implementacji

## Zbieranie danych

Dane użyte do trenowania modelu pobrano ze strony Basketball Reference za pomocą metody web scrapingu. Plik w formacie CSV który został użyty zawiera tablice wyników   
(z ang. Box score) z każdego meczu sezonu regularnego od sezonu 2015/2016 do sezonu 2022/2023. Każdy rekord to osobny mecz, który został opisany statystykami. Pobrano wszystkie dostępne statystyki, od tych najbardziej podstawowych i oczywistych takich jak która drużyna brała udział, wynik i data meczu aż po statystyki bardziej zaawansowane. Obejmują one szeroki zakres statystyk dla obu drużyn biorących udział w meczu, możemy przyporządkować je do kategorii takich jak:

* Minuty Gry (mp, mp\_max): Całkowita liczba minut, jaką zawodnicy spędzili na boisku.
* Rzuty z Gry (fg, fga, fg%, fg\_max, fga\_max, fg%\_max): Liczba celnych rzutów z gry, liczba prób rzutów z gry i procent celności.
* Rzuty za 3 Punkty (3p, 3pa, 3p%, 3p\_max, 3pa\_max, 3p%\_max): Liczba celnych rzutów za 3 punkty, liczba prób rzutów za 3 punkty i procent celności.
* Rzuty Wolne (ft, fta, ft%, ft\_max, fta\_max, ft%\_max): Liczba celnych rzutów wolnych, liczba prób rzutów wolnych i procent celności.
* Zbiórki (orb, drb, trb, orb\_max, drb\_max, trb\_max, orb%, drb%, trb%): Zbiórki ofensywne, defensywne i całkowite, a także ich procenty.
* Asysty (ast, ast\_max, ast%): Liczba asyst i procent asyst.
* Przechwyty (stl, stl\_max, stl%): Liczba przechwytów i ich procent.
* Bloki (blk, blk\_max, blk%): Liczba bloków i ich procent.
* Straty (tov, tov\_max, tov%): Liczba strat piłki i ich procent.
* Faule (pf, pf\_max): Liczba fauli.
* Punkty (pts, pts\_max): Liczba zdobytych punktów.
* Wskaźniki Efektywności (ts%, efg%, ortg, drtg, ortg\_max, drtg\_max): *True Shooting Percentage* (ts%), *Effective Field Goal Percentage* (efg%), *Offensive Rating* (ortg), *Defensive Rating* (drtg).
* Inne Wskaźniki (3par, ftr, 3par\_max, ftr\_max): *Three-Point Attempt Rate* (3par), *Free Throw Rate* (ftr).
* Czynniki Związane z Użyciem (usg%, usg%\_max): *Usage Rate* (usg%).

## Proces selekcji danych treningowych i weryfikacji modelu w czasie

W projekcie zastosowano metodę backtestingu, która symuluje proces trenowania   
i testowania modelu na danych historycznych w sposób iteracyjny, uwzględniając kolejność czasową danych. Jest to kluczowy aspekt przy analizie szeregów czasowych, gdzie zachowanie chronologii ma bezpośredni wpływ na jakość modelu i jego zdolność do generalizacji.

### Metoda Backtestingu

Funkcja odpowiadająca za backtesting odgrywa centralną rolę w ocenie modelu, iterując przez sezony w zbiorze danych. Proces ten przebiega w następujących krokach:

1. Podział danych na sezony: Dane są uporządkowane według sezonów sportowych, co pozwala na naturalny podział na zbiory treningowe i testowe.
2. Iteracyjny wybór sezonów: Funkcja backtest zaczyna swoją pracę od trzeciego sezonu (ustawienie start=2), co oznacza, że dwa pierwsze sezony są wykorzystywane jako początkowy zbiór treningowy. W każdej iteracji, model jest trenowany na wszystkich dostępnych danych historycznych do danego momentu (czyli na danych ze wszystkich wcześniejszych sezonów), a następnie testowany na bieżącym sezonie. Dzięki temu możliwe jest symulowanie rzeczywistego scenariusza, w którym model predykcyjny korzysta wyłącznie z przeszłych informacji, aby przewidzieć przyszłe wyniki.
3. Gromadzenie wyników predykcji: Po przetestowaniu modelu na danym sezonie, wyniki predykcji są zapisywane. Predykcje te są następnie łączone w jedną, zbiorczą ramkę danych, co umożliwia późniejszą ocenę dokładności modelu na przestrzeni wszystkich sezonów.
4. Ocena modelu: Zebrane wyniki predykcji dla wszystkich sezonów są wykorzystywane do analizy skuteczności modelu, umożliwiając ocenę jego zdolności do generalizacji oraz dokładności na danych testowych.

### Zastosowanie TimeSeriesSplit

Dodatkowo, w projekcie zastosowano metodę TimeSeriesSplit z parametrem n\_splits=3, co oznacza podział danych na cztery części (trzy podziały). W tym podejściu, dane treningowe obejmują wszystkie sezony sprzed bieżącego, natomiast dane testowe zawierają wyłącznie informacje z bieżącego sezonu. Jest to krytyczne podejście w analizie danych czasowych, ponieważ zapewnia, że model jest trenowany wyłącznie na danych historycznych i testowany na przyszłych, co zapobiega tzw. "wyciekowi danych" (data leakage). Wyciek danych może wystąpić, gdy model nieświadomie korzysta z informacji z przyszłości w procesie treningowym, co prowadzi do nienaturalnie wysokiej dokładności, która nie jest osiągalna  
w rzeczywistych warunkach.

### Znaczenie podejścia

Takie podejście jest ważne w kontekście modelowania szeregów czasowych, gdzie predykcje muszą opierać się wyłącznie na dostępnych w danym momencie danych historycznych. Chronologiczny podział danych w TimeSeriesSplit i iteracyjne trenowanie modelu na wcześniejszych sezonach zapewniają, że proces modelowania odzwierciedla rzeczywiste warunki predykcji, co zwiększa wiarygodność wyników i poprawia ich zastosowalność w praktycznych scenariuszach. Metody zastosowane w tym projekcie mają na celu zbudowanie modelu, który nie tylko dokładnie odzwierciedla historyczne trendy, ale również skutecznie prognozuje przyszłe wyniki, co jest kluczowe dla sukcesu w analizach predykcyjnych opartych na danych czasowych.

## Przygotowanie modelu bazowego

Aby ocenić wpływ wybranych cech na dokładność modelu, najpierw opracowano model bazowy, polegający na Regresji Grzbietowej, który posłuży jako punkt odniesienia do porównania wyników wykorzystujących selekcję cech. Model bazowy został wytrenowany na pełnym zestawie dostępnych cech, bez jakiejkolwiek selekcji. Wynikiem tego procesu był model osiągający dokładność na poziomie 53,3301% (zaokrągloną do czterech miejsc po przecinku). Taki wynik jest niezadowalający, zwłaszcza w kontekście analizy zbioru danych, który wskazuje, że drużyna gospodarzy wygrywała mecze w 57,1686% przypadków. Oznacza to, że nawet bez zastosowania metod uczenia maszynowego, jedynie na podstawie informacji o tym, która drużyna jest gospodarzem, można dokładniej przewidzieć wynik meczu.   
W związku z tym, poprawę poziomu dokładności można osiągnąć poprzez zastosowanie selektorów cech, które pomogą zidentyfikować najbardziej istotne zmienne.

## Opis eksperymentów

W eksperymencie zastosowano trzy algorytmy uczenia maszynowego: Ridge Regression, Regresję Logistyczną oraz K-Nearest Neighbors. Dla każdego z tych algorytmów stworzono trzy wersje modeli, różniące się zastosowaną metodą selekcji cech: Sequential Feature Selector, Select K Best oraz Recursive Feature Elimination. W sumie powstało dziewięć różnych modeli. W przypadku regresji logistycznej, aby uzyskać optymalne wyniki, zastosowano dodatkowo Grid Search, który pozwolił na dobranie najlepszych hiperparametrów dla tego modelu. W eksperymencie wyróżnić możemy kroki opisane poniżej.

### Przygotowanie Danych

Dane do eksperymentu pochodziły z wyników meczów NBA z poprzednich sezonów. Kluczowe kroki jakie można wyróżnić podczas przygotowania danych to:

* Sortowanie i czyszczenie danych podczas którego usunięto zbędne kolumny oraz wypełniono braki danych.
* Tworzenie zmiennej celu- dla każdego meczu przypisano wartość binarną do kolumny "target", reprezentującą wygraną lub przegraną zespołu.
* Skalowanie cech- wszystkie cechy numeryczne znormalizowano za pomocą skalera MinMax.

### Tworzenie modeli i selekcja cech

Każdy z modeli był trenowany na danych z wykorzystaniem jednego z trzech selektorów cech:

* Sequential Feature Selector: Selekcja cech oparta na iteracyjnym dodawaniu lub usuwaniu cech w celu minimalizacji błędu.
* Select K Best: Selekcja cech oparta na statystykach, takich jak ANOVA, gdzie wybierane jest określone K najlepszych cech.
* Recursive Feature Elimination: Selekcja cech polegająca na trenowaniu modelu   
  i iteracyjnym usuwaniu najmniej istotnych cech.

### Weryfikacja Modeli

Weryfikacja modeli w ramach eksperymentu opierała się na precyzyjnej ocenie ich zdolności do przewidywania wyników meczów NBA na podstawie danych historycznych. Kluczowym elementem tej oceny była metryka dokładności klasyfikacji, która jest standardową miarą używaną do oceny skuteczności modeli predykcyjnych w zadaniach klasyfikacyjnych. Każdy z testowanych modeli został przeszkolony na historycznych danych dotyczących wyników meczów, a następnie użyty do przewidywania wyników dla zestawu testowego, który składał się z danych, na których model nie był wcześniej trenowany. Weryfikacja polegała na porównaniu wyników przewidzianych przez model z rzeczywistymi wynikami, co pozwoliło ocenić, jak dobrze model radzi sobie z rzeczywistymi danymi. Dokładność klasyfikacji, znana również jako wskaźnik dokładności, została zastosowana jako główna miara oceny wydajności modeli. Jest ona obliczana jako stosunek liczby poprawnych predykcji do całkowitej liczby predykcji, co oznacza, że określa, jaki procent wszystkich przewidywań modelu był zgodny   
z rzeczywistymi wynikami. Dokładność klasyfikacji jest szczególnie istotna w kontekście tego eksperymentu, ponieważ pozwala na bezpośrednie i łatwe zrozumienie, jak skuteczny jest model w kontekście zadania predykcyjnego. Wysoka wartość wskaźnika dokładności wskazuje na to, że model skutecznie uczy się na danych historycznych i jest w stanie dobrze przewidywać przyszłe wyniki. Sposób wyliczenia wskaźnika dokładności można opisać następującym wzorem:

## Wyniki eksperymentów

Wyniki przeprowadzonych eksperymentów zostały zestawione w tabelach, które ilustrują wskaźnik dokładności osiągnięty przez każdy z zastosowanych modeli przy użyciu różnych metod selekcji cech. Analizując tabele można ocenić, jak różne podejścia do selekcji cech wpływają na efektywność predykcyjną poszczególnych modeli.

### Modele Regresji Grzbietowej

Tabela 3 Wyniki modelów Regresji Grzbietowej

|  |  |  |  |
| --- | --- | --- | --- |
|  | 50 | 30 | 7 |
| Algorytm SFS | 63,1702% | 63,1612% | 63,3775% |
| Algorytm Select K Best | 63,6659% | 63,4045% | 62,4313% |
| Algorytm RFE | 63,0621% | 62,9179% | 61,4761% |

### Modele Regresji Logistycznej

Tabela 4 Wyniki modelów Regresji Logistycznej

|  |  |  |  |
| --- | --- | --- | --- |
|  | 50 | 30 | 7 |
| Algorytm SFS | 61,9447% | 62,0077% | 61,8816% |
| Algorytm Select K Best | 63,6478% | 63,4316% | 62,3592% |
| Algorytm RFE | 63,7199% | 63,1162% | 61,6293% |

### Modele algorytmu K-Najbliższych Sąsiadów

Tabela 5 Wyniki modelów K-Najbliższych Sąsiadów

|  |  |  |  |
| --- | --- | --- | --- |
|  | 50 | 30 | 7 |
| Algorytm SFS | 62,6926% | 62,5211% | 61,5431% |
| Algorytm Select K Best | 62,5214% | 62,7647% | 62,3141% |
| Algorytm RFE | 56,3125% | 56,5648% | 56,6009% |

## Wnioski

Na podstawie wyników modeli wytrenowanych na różnych selektorach cech z różną ilością cech do wyboru (50, 30, 7), można wyciągnąć kilka wniosków:

### Regresja Grzbietowa

W przypadku algorytmu SFS, najwyższą dokładność osiągnięto przy 7 cechach (63,3775%), co sugeruje, że zmniejszenie liczby cech może poprawić wydajność modelu   
w przypadku Regresji Grzbietowej. Natomiast w przypadku selektora Select K Best najwyższą dokładność uzyskano przy 50 cechach (63,6659%), co sugeruje, że ten selektor lepiej działa przy większej liczbie cech. W przypadku algorytmu RFE najlepszy wynik uzyskano przy 50 cechach (63,0621%), a dokładność systematycznie spadała wraz z redukcją liczby cech.

### Regresja Logistyczna

Algorytm SFS najwyższą dokładność osiągnął przy 30 cechach (62,0077%), jednak wyniki są zbliżone do tych przy 7 i 50 cechach, co może sugerować, że liczba cech nie ma dużego wpływu na dokładność w tym przypadku. W przypadku Select K Best, najwyższą dokładność uzyskano przy 50 cechach (63,6478%), podobnie jak w przypadku Regresji Grzbietowej. Natomiast algorytm RFE najlepszy wynik uzyskał przy 50 cechach (63,7199%), co również wskazuje na lepszą wydajność przy większej liczbie cech.

### K-Nearest Neighbours

Wykorzystując algorytm SFS, najwyższą dokładność osiągnięto przy 50 cechach (62,6926%), ale z malejącą liczbą cech wyniki systematycznie spadają. W przypadku algorytmu Select   
K Best najlepszy wynik uzyskano przy 30 cechach (62,7647%), ale różnice między różnymi liczbami cech są stosunkowo niewielkie. W przypadku algorytmu RFE dokładność jest znacznie niższa w porównaniu do innych metod, a wyniki są bardzo zbliżone niezależnie od liczby cech, co sugeruje, że ten algorytm nie jest efektywny przy tej metodzie selekcji cech.

### Podsumowanie

Algorytm Select K Best wydaje się być najskuteczniejszym selektorem cech, szczególnie przy 50 cechach, dla modeli Regresji Grzbietowej i Logistycznej. Selektor RFE działa lepiej przy większej liczbie cech, ale jego skuteczność jest niższa w przypadku   
K-Najbliższych Sąsiadów. Natomiast algorytm SFS może działać dobrze przy mniejszej liczbie cech, zwłaszcza w przypadku Regresji Grzbietowej, ale ogólnie daje stabilne wyniki niezależnie od liczby cech. Wybór odpowiedniego selektora cech oraz liczby cech do modelu jest kluczowy dla osiągnięcia optymalnych wyników. Jak dowiódł eksperyment, warto testować różne kombinacje w zależności od rodzaju modelu. Nie można również zapominać   
o zwiększonej złożoności obliczeniowej rosnącej wraz z liczbą cech zadanych do wybrania. Gdy zależy nam na szybkiej analizie zbioru danych warto skorzystać z modelu wykorzystującego mniejszą ilość selektorów, gdyż w większości przypadków zachowuje to dość przyzwoitą dokładność klasyfikacji.

# Zakończenie

W niniejszej pracy podjęta została tematyka związaną z wpływem selekcji cech na dokładność predykcji wyników meczów NBA za pomocą algorytmów uczenia maszynowego. W pracy skupiono się na omówieniu, jak różne metody selekcji cech mogą wpłynąć na efektywność predykcji i jak wybór odpowiednich cech może poprawić jakość modelu. Celem było zbadanie, które cechy spośród dostępnych danych mają największy wpływ na dokładność prognoz, a także weryfikacja, czy zastosowanie algorytmów selekcji cech może znacząco poprawić wyniki predykcji.

Założono, że odpowiednia selekcja cech może prowadzić do znacznego wzrostu dokładności modeli predykcyjnych. W wyniku przeprowadzonych eksperymentów okazało się, że algorytmy selekcji cech rzeczywiście przyczyniły się do poprawy dokładności predykcji. W szczególności, algorytmy takie jak Sequential Feature Selection oraz Select K Best przyniosły świetne rezultaty, osiągając wyższą dokładność w porównaniu do modelu bazowego, który nie korzystał z selekcji cech. Ostatecznie, najbardziej efektywne okazało się zastosowanie algorytmu RFE wraz z metodą Regresji Logistycznej, który umożliwił osiągnięcie dokładności na poziomie 63,7199%.

Wyniki badań potwierdziły zatem postawioną hipotezę, co dowodzi, że odpowiednia selekcja cech jest ważnym elementem w procesie tworzenia skutecznych modeli predykcyjnych. Pozytywnym aspektem wyników jest to, że za pomocą relatywnie prostych metod można znacząco zwiększyć efektywność predykcji, co ma istotne znaczenie w kontekście zastosowań praktycznych, takich jak analiza wyników sportowych czy prognozowanie wyników meczów w czasie rzeczywistym.

Własne przemyślenia na temat uzyskanych wyników prowadzą do wniosku, że przyszłe badania mogłyby skupić się na analizie bardziej zaawansowanych metod selekcji cech, które mogłyby uwzględniać również interakcje między cechami oraz bardziej dynamiczne zmiany w czasie. Dodatkowo, rozważenie integracji z innymi technikami, takimi jak uczenie głębokie czy sieci neuronowe, mogłoby przynieść jeszcze lepsze wyniki predykcyjne.

Na zakończenie należy wskazać, że pomimo uzyskanych pozytywnych rezultatów, istnieją pewne deficyty, które mogą stanowić perspektywy do prowadzenia dalszych badań naukowych. Przede wszystkim, analiza obejmowała dane historyczne, co nie zawsze odzwierciedla dynamiczny charakter rozgrywek sportowych. Kolejnym krokiem mogłoby być zbadanie wpływu innych, bardziej złożonych zmiennych, takich jak kontuzje zawodników czy zmiany w składach drużyn, na dokładność predykcji. Wreszcie, istotnym kierunkiem dalszych badań mogłoby być opracowanie bardziej zaawansowanych modeli hybrydowych, które mogłyby integrować różne podejścia do predykcji w celu uzyskania jeszcze bardziej precyzyjnych prognoz.

# Bibliografia

Bunker, R., & Susnjak, T. (2022). The Application of Machine Learning Techniques for Predicting Match Results in Team Sport: A Review. *Journal of Artificial Intelligence Research*, *73*, 1285–1322. https://doi.org/10.1613/jair.1.13509

Burkov, A. (2019). *The hundred-page machine learning book*. Andriy Burkov.

Clancey, W. J. (1984). *Classification Problem Solving*.

Georgievski Bojan & Vrtagic Sabahundin. (2021). Machine learning and the NBA Game. *Journal of Physical Education and Sport*, *21*(06).

Horvat, T., & Job, J. (2020). The use of machine learning in sport outcome prediction: A review. *WIREs Data Mining and Knowledge Discovery*, *10*(5), e1380. https://doi.org/10.1002/widm.1380

Horvat, T., Job, J., Logozar, R., & Livada, Č. (2023). A Data-Driven Machine Learning Algorithm for Predicting the Outcomes of NBA Games. *Symmetry*, *15*(4), 798. https://doi.org/10.3390/sym15040798

Leung, C. K., & Joseph, K. W. (2014). Sports Data Mining: Predicting Results for the College Football Games. *Procedia Computer Science*, *35*, 710–719. https://doi.org/10.1016/j.procs.2014.08.153

Loeffelholz, B., Bednar, E., & Bauer, K. W. (2009). Predicting NBA Games Using Neural Networks. *Journal of Quantitative Analysis in Sports*, *5*(1). https://doi.org/10.2202/1559-0410.1156

Lopez, G. (2024, kwiecień 5). The Rise of Sports Betting. *The New York Times*. https://www.nytimes.com/2024/04/05/briefing/the-rise-of-sports-betting.html

Manley, M. (1989). *Martin Manley’s Basketball Heaven*. Doubleday Books: New York.

McCabe, A., & Trevathan, J. (2008). Artificial Intelligence in Sports Prediction. *Fifth International Conference on Information Technology: New Generations (itng 2008)*, 1194–1197. https://doi.org/10.1109/ITNG.2008.203

Ozanian, M. (2023, listopad 26). *NBA Valuations 2023*. Forbes. https://www.forbes.com/lists/nba-valuations/

Prasetio, D., & Harlili, Dra. (2016). Predicting football match results with logistic regression. *2016 International Conference On Advanced Informatics: Concepts, Theory And Application (ICAICTA)*, 1–5. https://doi.org/10.1109/ICAICTA.2016.7803111

Sampaio, J., McGarry, T., Calleja-González, J., Jiménez Sáiz, S., Schelling I Del Alcázar, X., & Balciunas, M. (2015). Exploring Game Performance in the National Basketball Association Using Player Tracking Data. *PLOS ONE*, *10*(7), e0132894. https://doi.org/10.1371/journal.pone.0132894

*uczenie się—Definicja, synonimy, przykłady użycia*. (b.d.). Słownik języka polskiego. Pobrano 12 kwiecień 2024, z https://sjp.pwn.pl/slowniki/uczenie%20si%C4%99.html

# Spis rysunków

[Rysunek 1 Wykres regresja logistyczna 9](#_Toc176205646)

[Rysunek 2 Wykres regresja logistyczna 10](#_Toc176205647)

[Rysunek 3 Wykres K-najbliższych sąsiadów 11](#_Toc176205648)

# Spis tabel

[Tabela 1 Trzynaście podstawowych elementów meczu koszykarskiego. 13](#_Toc176205174)

[Tabela 2 Osiem pochodnych elementów meczu koszykarskiego. 14](#_Toc176205175)

[Tabela 3 Wyniki modelów Regresji Grzbietowej 30](#_Toc176205176)

[Tabela 4 Wyniki modelów Regresji Logistycznej 31](#_Toc176205177)

[Tabela 5 Wyniki modelów K-Najbliższych Sąsiadów 32](#_Toc176205178)