AKADEMIA GÓRNICZO-HUTNICZA IM. STANISŁAWA STASZICA W KRAKOWIE
WYDZIAŁ GEOLOGII, GEOFIZYKI I OCHRONY ŚRODOWISKA

KATEDRA GEOINFORMATYKI I INFORMATYKI STOSOWANEJ

Projekt dyplomowy

Wykorzystanie metod uczenia maszynowego w predykcji wyników zawodów sportowych

Autor: *Damian Gortych*
Kierunek studiów: *Inżynieria i Analiza Danych*
Opiekun pracy: *dr hab. inż. Tomasz Danek*

Kraków, 2022

Spis treści

Wstęp

Uczenie maszynowe jako obszar sztucznej inteligencji jest uznawane za połączenie matematyki, statystyki oraz informatyki. W ostatnich latach przechodzi ono okres intensywnego rozwoju oraz wzrostu zainteresowania, zarówno w celach komercyjnych, jak i pracach naukowych. Jedną z dziedzin jego zastosowania jest przewidywanie rezultatów okołosportowych, a wśród nich osiągnięć przyznawanych w koszykówce.

Celem niniejszej pracy jest predykcja wyniku wyboru zwycięzcy nagrody MVP w lidze NBA z wykorzystaniem technik uczenia maszynowego. Część techniczna przygotowana została przy użyciu języka Python wraz z wybranymi bibliotekami.

Rozdział pierwszy zawiera wprowadzenie do teorii związanej z koszykówką oraz nagrodą MVP. Pomaga w zrozumieniu zagadnień poruszanych w dalszej części pracy. Rozdział drugi poświęcony jest szczegółowemu opisowi zgromadzonych danych oraz wstępnemu ich przygotowaniu. W rozdziale trzecim zaprezentowano analizę deskryptywną zbioru. Szczególną uwagę zwrócono na niezbalansowanie oraz zmienność w czasie. Rozdziały czwarty oraz piąty skupiają się na przedstawieniu zastosowanego podejścia w tworzeniu modeli uczenia maszynowego. Omówiony zostaje schemat oceny oraz wyboru ostatecznego rozwiązania.

1 Wprowadzenie do teorii związanej z tematem pracy

Celem rozdziału jest wyjaśnienie najważniejszych pojęć kluczowych do zrozumienia problematyki pracy. Ma on za zadanie wprowadzić do tematu gry w koszykówkę oraz struktury rozgrywek ligi NBA. Służy to ostatecznie opisaniu i zrozumieniu zasad wręczania nagrody MVP.

1.1 Gra w koszykówkę

Koszykówka jest obecnie trzecim najpopularniejszym sportem na świecie ustępując jedynie piłce nożnej oraz krykietowi [1]. Za datę jej powstania uznaje się 21 grudnia 1891 roku. Wówczas nauczyciel wychowania fizycznego James Naismith wymyślił ją jako sposób na zachowanie sprawności fizycznej i zdrowia przez uczniów podczas zimy. Głównymi jej zasadami były [2]:

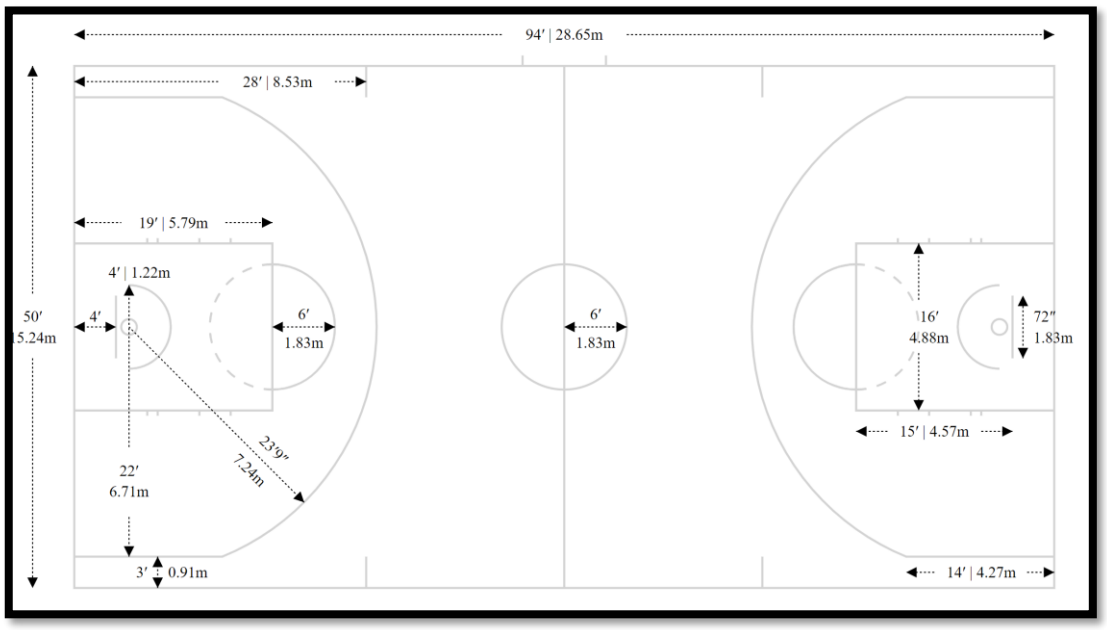
- Mecz jest rozgrywany okrągłą piłką wyłącznie przy użyciu rąk.
- Gracz nie może poruszać się z piłką.
- Gracz może znajdować się w dowolnym miejscu boiska.
- Zabroniony jest brutalny kontakt fizyczny.
- Kosz powinien być umiejscowiony wysoko nad boiskiem.

Bardzo szybko zdobyła ona popularność zarówno w Stanach Zjednoczonych, jak i na całym świecie. Dowodem tego było umieszczenie jej jako pokazowej dyscypliny na Igrzyskach Olimpijskich już w roku 1904. Oznaczało to również dynamicznie wprowadzane zmiany w jej zasadach [3]:

- Dodano możliwość kozłowania piłki.
- Wprowadzono rzuty za jeden i dwa punkty.
- Zamknięte kosze zamieniono na otwarte obręcze z tablicą.
- Liczbę graczy z jednej drużyny znajdujących się jednocześnie na boisku ustalono na pięć.

Przełomowym momentem było dodanie w późniejszych latach linii rzutów za trzy punkty. Miało to na celu poprawienie widowiskowości meczu i poskutkowało zmianą

strategii zdobywania punktów [4]. Rysunek 1 przedstawia obecne wymiary boiska do koszykówki stosowane w lidze NBA.



Rysunek 1 *Wymiary boiska do koszykówki w lidze NBA*
***Źródło*:** [5]

1.2 Liga NBA

Liga NBA (National Basketball Association) jest amerykańsko-kanadyjską zawodową ligą koszykówki uznawaną za największą na świecie. Została założona w 1946 roku jako Basketball Association of America (BAA), a swoją obecną nazwę przyjęła w 1949 po połączeniu się z National Basketball League (NBL) [6]. Najwyższe stanowisko jako jej komisarz pełni obecnie Adam Silver.

1.2.1 System rozgrywek

W skład ligi NBA wchodzi 30 drużyn, które podzielone są równo na dwie konferencje, które z kolei dzielą się na trzy dywizje każda. Raz do roku rozgrywany jest sezon składający się z kilku etapów [7].

Tabela 1 *Orientacyjny kalendarz dla sezonu 2020-2021 ligi NBA*

Data	Etap sezonu
11-19 Grudnia 2020	Mecze przedsezonowe
22 Grudnia 2020 – 4 Marca 2021	Pierwsza połowa sezonu regularnego
5-10 Marca 2021	Weekend gwiazd
11 Marca – 16 Maja 2021	Druga połowa sezonu regularnego
18-21 Maja 2021	Turniej Play-In
22 Maja – 22 Lipca 2021	Faza Play-off

Źródło**:** *Opracowanie własne na podstawie* [8]

Tabela 1 przedstawia orientacyjny kalendarz sezonu 2020-2021. Rozpoczyna się on od gier wstępnych, które nie mają znaczenia dla wyniku rozgrywek, a służą jedynie jako możliwość zaprezentowania się przez nowych zawodników.

Sezon regularny rozpoczyna zmagania o zwycięstwo. W jego trakcie zespoły rozgrywają po 82 mecze: 52 z zespołami ze swojej konferencji oraz 30 z drugiej. Po jego zakończeniu 10 najlepszych drużyn z każdej z nich awansuje do dalszej fazy rozgrywek Play-off lub Play-In i zawalczy o zwycięstwo całej ligi. To właśnie po zakończeniu sezonu regularnego i tylko na jego podstawie wręczana jest nagroda MVP, której wyniku predykcja jest tematem tej pracy.

1.2.2 Nagroda MVP

Nagroda MVP (Most Valuable Player) jest statuetką dla najlepszego gracza ligi w sezonie regularnym. Wprowadzona została w roku 1956 i początkowo aż do roku 1980 wybierana była przez zawodników NBA. Obecnie powinność tę sprawuje panel dziennikarzy sportowych oraz radiowych ze Stanów Zjednoczonych oraz Kanady. Każdy z nich tworzy listę rankingową dla pięciu kandydatów, którzy otrzymują następnie punktację przedstawioną w tabeli 2. W 2010 roku dodano także jeden wspólny głos fanów, do którego przyczynić może się każdy za pośrednictwem głosowania internetowego.

Tabela 2 *Punktacja głosowania NBA MVP w zależności od zajętego miejsca*

Numer miejsca	Liczba przyznawanych punktów
Miejsce pierwsze	10 punktów
Miejsce drugie	7 punktów
Miejsce trzecie	5 punktów
Miejsce czwarte	3 punkty
Miejsce piąte	1 punkt

Źródło: *Opracowanie własne na podstawie* [9]

Zawodnik z największa ilością zdobytych punktów jest ogłaszany najbardziej wartościowym graczem ligi i otrzymuje nagrodę MVP. Tabela 3 przedstawia zwycięzców od roku 2000. Najwięcej statuetek w historii zdobył Kareem Abdul-Jabbar (sześć). Jedynym jednogłośnym zwycięzcą, który otrzymał pierwsze miejsce w każdym ze 131 głosów był Stephen Curry w sezonie 2015-2016 [10].

Tabela 3 *Zwycięzcy nagrody MVP w lidze NBA w latach 2000-2022*

Rok otrzymania nagrody	Imię i nazwisko zwycięzcy
2000	Shaquille O'Neal
2001	Allen Iverson
2002	Tim Duncan
2003	Tim Duncan
2004	Kevin Garnett
2005	Steve Nash
2006	Steve Nash
2007	Dirk Nowitzki
2008	Kobe Bryant

2009	LeBron James
2010	LeBron James
2011	Derrick Rose
2012	LeBron James
2013	LeBron James
2014	Kevin Durant
2015	Stephen Curry
2016	Stephen Curry
2017	Russell Westbrook
2018	James Harden
2019	Giannis Antetokounmpo
2020	Giannis Antetokounmpo
2021	Nikola Jokic
2022	Nikola Jokic

Źródło: *Opracowanie własne na podstawie* [11]

2 Dane

Celem rozdziału jest przedstawienie źródła zdobytych danych oraz opisanie podejścia zastosowanego w celu wstępnego przygotowania ich do późniejszej predykcji z wykorzystaniem uczenia maszynowego. Zawiera on także opis wszystkich zawartych zmiennych numerycznych.

2.1 Źródło danych

Kaggle jest internetową platformą społecznościową dla osób pracujących z danymi oraz miłośników uczenia maszynowego. Zawiera ponad 50,000 materiałów, z których każdy może swobodnie korzystać.

Przygotowanie danych rozpoczęto od pobrania zbioru, który za pośrednictwem wspomnianej platformy udostępnił Omri Goldstein [12]. Zawiera on 53 kolumny z dokładnymi statystykami wszystkich graczy począwszy od roku 1950, aż do roku 2018. Wzbogacony został o dane z lat 2019-2022, a także dodatkowe trzy kolumny z informacjami o wynikach głosowania na nagrodę MVP. Pobrano je ze strony basketball-reference, która gromadzi zaawansowane statystyki w temacie koszykówki.

2.2 Wstępne przygotowanie danych

W pierwszym kroku usunięte zostały dane przed rokiem 1982, ponieważ w tym okresie większość z nich nie było zbieranych. Skutkowało to zmniejszeniem procentu brakujących obserwacji w całym zbiorze z 31 % do 0.42 %. Następnie sprawdzone zostało występowanie duplikatów. Przykładowy znaleziony rezultat przedstawiono na rysunku 2. Widać na nim, że dane o zawodniku pojawiają się trzy razy dla tego samego roku. Ponadto, jeden z wierszy jest sumą pozostałych. Wynika to z faktu, że gracze w trakcie sezonu mogą zmienić drużynę którą reprezentują. W temacie tej pracy istotne są jedynie pełne dane, więc statystyki rozdzielone na poszczególne zespoły należało usunąć.

	Tm	Year	Player	Pos	Age	G	GS	MP	FG	FGA
65	TOT	1982	Charlie Criss	PG	33.0	55	20.0	1392.0	222	498
66	ATL	1982	Charlie Criss	PG	33.0	27	0.0	552.0	84	210
67	SDC	1982	Charlie Criss	PG	33.0	28	20.0	840.0	138	288

Rysunek 2 *Przykład zduplikowanych danych*

Źródło: *Opracowanie własne*

	Tm	Year	Player	Pos	Age	G	GS	MP	FG	FGA	FG%	3P	3PA	3P%
11314	HOU	2006	Josh Davis	PF	25.0	1	0.0	0.0	0	0	NaN	0.0	0.0	NaN
11631	SAS	2006	Alex Scales	SG	27.0	1	0.0	0.0	0	0	NaN	0.0	0.0	NaN
12594	GSW	2008	Stephane Lasme	SF	25.0	1	0.0	0.0	0	0	NaN	0.0	0.0	NaN

Rysunek 3 *Przykład brakujących danych*

Źródło: *Opracowanie własne*

Ostatnim etapem było podjęcie działań w zakresie pozostałych brakujących wartości. Z ukazanego przykładu na rysunku 3 określić można powód ich występowania w zbiorze. Wartości niektórych kolumn obliczane są na podstawie ilorazu innych. Oznacza to, że w przypadku liczby zero w dzielniku jako wynik otrzymywany jest NaN (Not a Number), czyli zapis interpretowany, w przypadku pracy z danymi, jako wartość brakująca. Dla przykładu zmienna FG% (procent rzutów z gry) obliczana jest zgodnie ze wzorem:

$$FG\% = \frac{FG}{FGA} \qquad \textbf{(1)}$$

gdzie:

FG – trafione rzuty z gry

FGA – oddane rzuty z gry

Korzystając z tej obserwacji, trywialny staje się wniosek, iż w celu pozbycia się wartości NaN należy zastąpić je zerem. W ten sposób zakończone zostało wstępne przygotowanie danych.

2.3 Opis danych

Spośród wszystkich zebranych zmiennych opisujących zawodnika oraz jego statystyki, do dalszej pracy wybranych zostało 52. Tabela 4 opisuje wszystkie wartości numeryczne występujące w danych, które użyte zostały do predykcji wyniku nagrody MVP.

Tabela 4. *Opis zmiennych numerycznych zawartych w danych*

Nazwa zmiennej	**Opis znaczenia**
Age	Wiek gracza
G	Liczba rozegranych meczy
GS	Liczba rozegranych meczy jako starter
MP	Liczba rozegranych minut na boisku
FG	Trafione rzuty z gry
FGA	Oddane rzuty z gry
FG%	Procent rzutów z gry (FG/FGA)
3P	Trafione rzuty za trzy punkty
3PA	Oddane rzuty za trzy punkty
3P%	Procent rzutów za trzy punkty ($3P/3PA$)
2P	Trafione rzuty za dwa punkty
2PA	Oddane rzuty za dwa punkty
2P%	Procent rzutów za dwa punkty ($2P/2PA$)
eFG%	Efektywny procent rzutów z gry ($(FG + 0.5 * 3P)/ FGA$)
FT	Trafione rzuty wolne
FTA	Oddane rzuty wolne

FT%	Procent rzutów wolnych (FT/FTA)
ORB	Zbiórki ofensywne
DRB	Zbiórki defensywne
TRB	Wszystkie zbiórki
AST	Asysty
STL	Przechwyty
BLK	Bloki
TOV	Straty
PF	Faule
PTS	Zdobyte punkty
PER	Ocena wydajności gracza [13]
TS%	Prawdziwa skuteczność rzutowa ($PTS\ /\ (2\ *\ TSA)$)
DRB%	Procent zbiórek defensywnych ($100\ *\ (DRB\ *\ (Tm\ MP\ /\ 5))\ /\ (MP\ *\ (Tm\ TRB\ +\ Opp\ TRB))$)
ORB%	Procent zbiórek ofensywnych ($100\ *\ (ORB\ *\ (Tm\ MP\ /\ 5))\ /\ (MP\ *\ (Tm\ ORB\ +\ Opp\ DRB))$)
TRB%	Procent wszystkich zbiórek ($100\ *\ (TRB\ *\ (Tm\ MP\ /\ 5))\ /\ (MP\ *\ (Tm\ TRB\ +\ Opp\ TRB))$)
AST%	Procent asyst ($100\ *\ AST\ /\ (((MP\ /\ (Tm\ MP\ /\ 5))\ *\ Tm\ FG)\ -\ FG)$)
STL%	Procent przechwytów ($100\ *\ (STL\ *\ (Tm\ MP\ /\ 5))\ /\ (MP\ *\ Opp\ Poss)$)
BLK%	Procent bloków ($100\ *\ (BLK\ *\ (Tm\ MP\ /\ 5))\ /\ (MP\ *\ (Tm\ FGA\ -\ Opp\ 3PA))$)
TOV%	Procent strat ($100\ *\ TOV\ /\ (FGA\ +\ 0.44\ *\ FTA\ +\ TOV)$)
USG%	Procent użyteczności ($100\ *\ ((FGA\ +\ 0.44\ *\ FTA\ +\ TOV)\ *\ (Tm\ MP\ /\ 5))\ /\ (MP\ *\ (Tm\ FGA\ +\ 0.44\ *\ Tm\ FTA\ +\ Tm\ TOV))$)
OWS	Ofensywny udział zawodnika w wygranej [14]
DWS	Defensywny udział zawodnika w wygranej [14]

WS	Udział zawodnika w wygranej [14]
WS/48	Udział zawodnika w wygranej na 48 minut [14]
OBPM	Ofensywny wpływ zawodnika na mecz [15]
DBPM	Defensywny wpływ zawodnika na mecz [15]
BPM	Wpływ zawodnika na mecz [15]
VORP	Wartość zawodnika w porównaniu ze zmiennikiem [15]
Votes_first	Liczba głosów z pierwszym miejscem
Points_max	Maksymalna ilość punktów do zdobycia w głosowaniu
Points_won	Zdobyte punkty w głosowaniu
Award_share	Udział w nagrodzie ($Points\_won \ / \ Points\_max$)

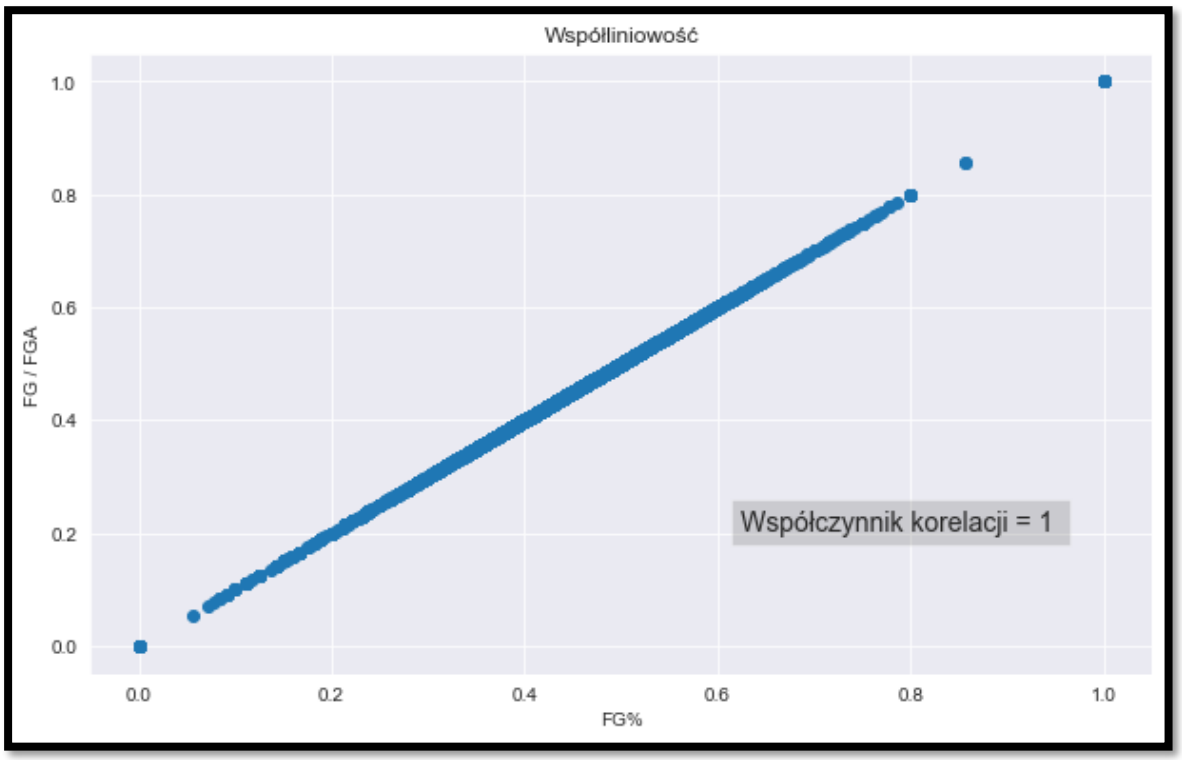
Źródło: *Opracowanie własne na podstawie* [16]

3 Analiza deskryptywna danych

Celem rozdziału jest analiza danych pod względem optymalizacji podejścia do predykcji zwycięzcy nagrody MVP. Skupia się ona na zbadaniu ogólnej charakterystyki zbioru oraz zależności w nim występujących.

3.1 Współliniowość

Na etapie wstępnego przetwarzania danych ustalone zostało, że niektóre statystyki obliczane są na podstawie pozostałych. Większość z nich korzysta ze stosunkowo skąplikowanych wzorów, jednakże wartość kilku otrzymywana jest w oparciu o prosty iloraz.



Rysunek 4 *Wykres przedstawiający współliniowość w danych*

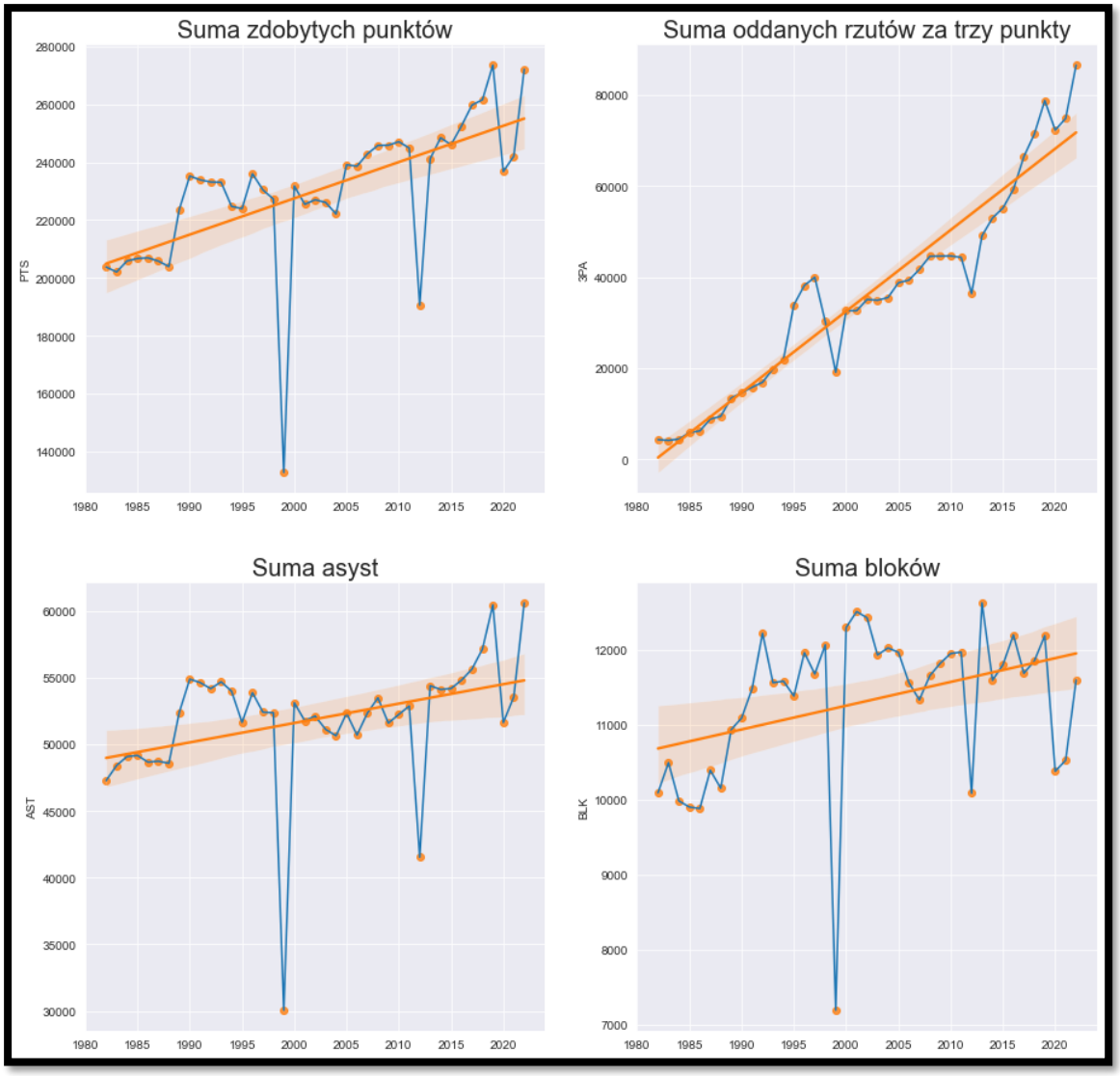
Źródło: *Opracowanie własne*

Rysunek 4 przedstawia wykres zależności zmiennej FG% od ręcznie obliczonej wartości $FG \,/\, FGA$, których współczynnik korelacji wynosi dokładnie jeden. Służy on

jako przykład zjawiska współliniowości, które występuje w przypadku silnego powiązania pomiędzy zmiennymi objaśniającymi. Najczęściej wymaga ono usunięcia ze zbioru nadmiarowych cech, ponieważ prowadzą one do bezcelowego skomplikowania modelu uczenia maszynowego. W związku z tym dane zostały zredukowane o kolumny FG%, 3P%, 2P% oraz FT%.

3.2 Zmienność w czasie

Koszykówka jako dyscyplina sportowa stale się rozwija. Dzieje się to za sprawą postępu graczy, którzy w każdym roku stają się szybsi, zwiększają swoją skuteczność oraz inteligentniej poruszają się po boisku. Rysunek 5 przedstawia wykresy czterech statystyk, które w znaczym stopniu zmieniły się na przestrzeni lat. Największą różnicę widać w linii trendu dla sumy oddanych rzutów za trzy punkty, które w ostatniej dekadzie zdefiniowały styl rozgrywania meczu.



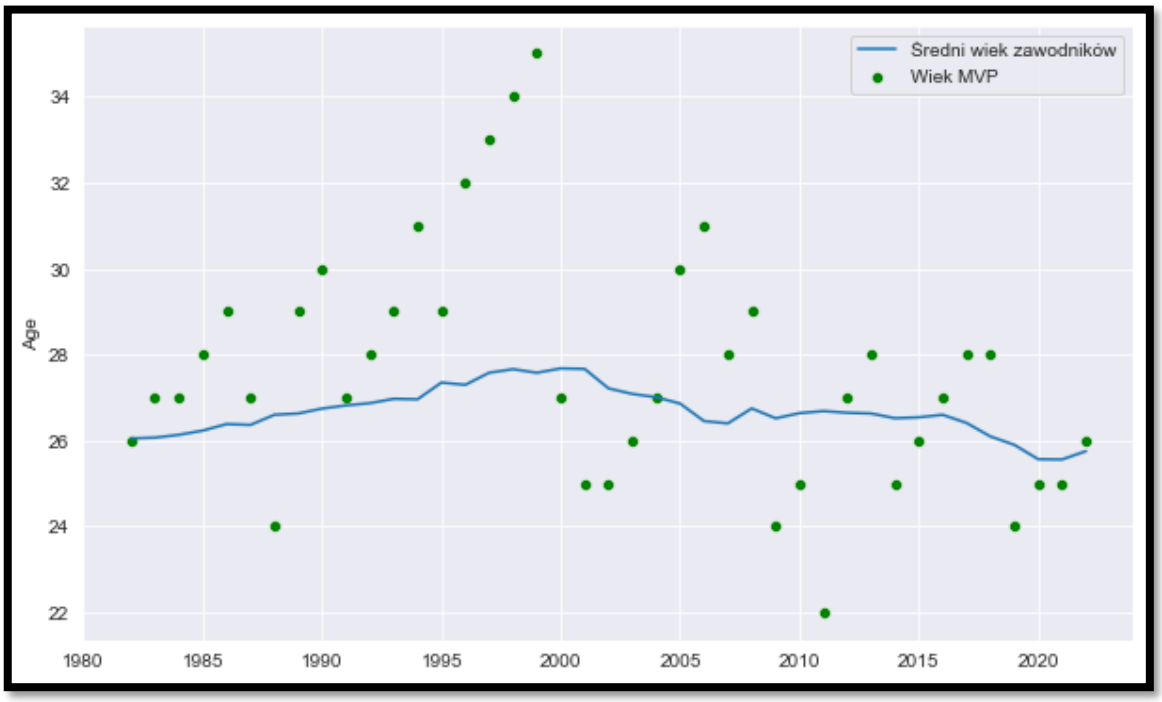
Rysunek 5 *Wykresy sumy wybranych statystyk dla kolejnych lat*

Źródło: *Opracowanie własne*

Na wykresach można również zauważyć cztery spadki wartości pojedynczych punktów kolejno w latach 1999, 2012, 2020 oraz 2021. Dwa pierwsze spowodowane są wprowadzeniem przez ligę lokautów (ang. *NBA Lockaout*), które zmniejszyły liczbę rozgrywanych meczy w trakcie sezou regularnego do 50 w sezonie 1998-99 oraz do 66 w sezonie 2011-2012. Pozostałe dwa przypadają na okres pandemi COVID-19, która również ograniczyła o kilkanaście liczbę odbytych przez dryżyny spotkań.

Wartościowych informacji dostarcza także wykres wieku zawodników otrzymujących nagrodę MVP w kolejnych latach przedstawiony na rysunku 6. Zauważyć

można, że wzrasta liczba graczy ze statuetką, których wiek jest niższy od średniej dla całej ligi, która ponadto sama obniża się w XXI wieku.



Rysunek 6 *Wykres wieku zawodników otrzymujących nagrodę MVP w kolejnych latach*
Źródło: *Opracowanie własne*

Biorąc pod uwagę wykonaną analizę zmienności danych w czasie, sformułowano wniosek, iż zachodzi ona w stopniu wystarczającym do wzięcia jej pod uwagę w tworzeniu modelu uczenia maszynowego. Oznacza to konieczność podejścia do zebranych danych na zasadach podobnych do przypadku szeregu czasowego.

3.3 Korelacja oraz trend

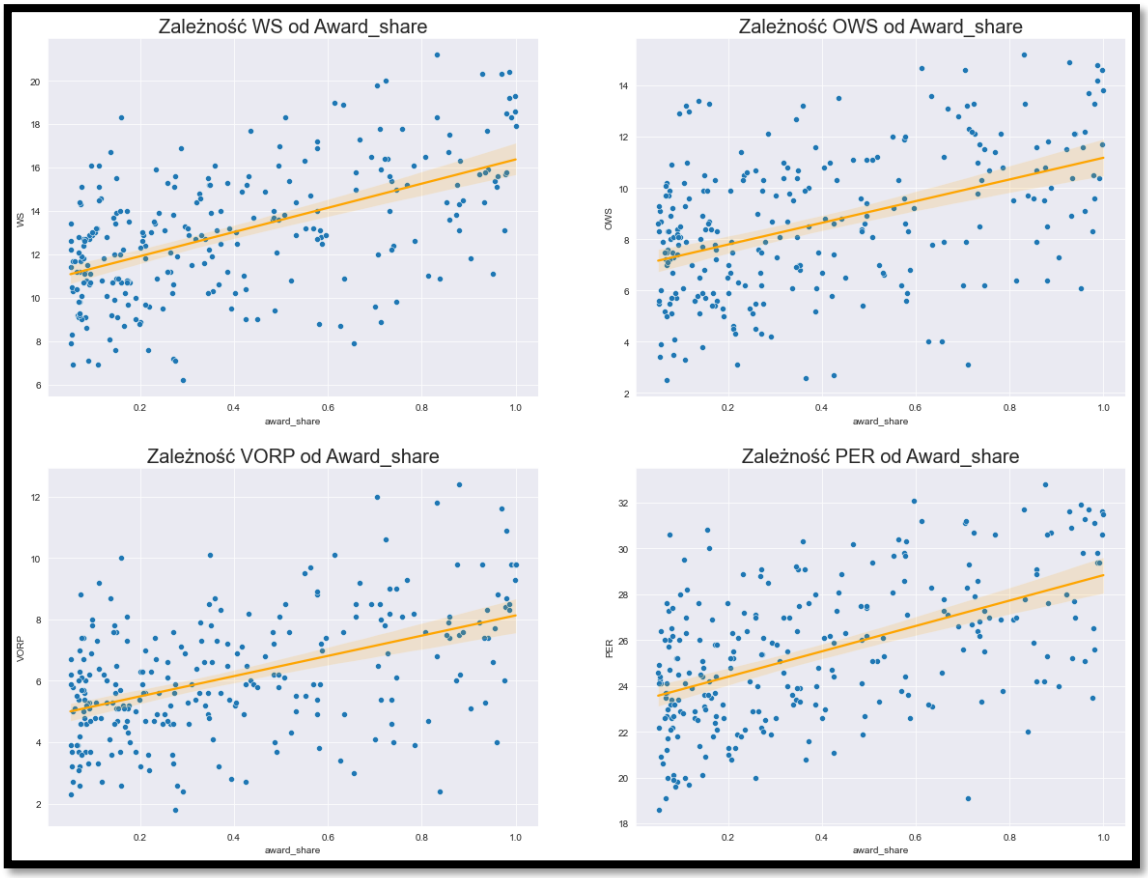
W przygotowaniach do tworzenia modelu uczenia maszynowego, kluczowe jest wykonanie korelacji pomiędzy danymi w zbiorze. Nacisk kładziony jest głównie na zależności w stosunku do zmiennej objaśnianej, której wartość jest celem predykcji. W związku z tym, że jest ona typu kategorycznego i posiada dwie możliwe wartości zdecydowano się na wybranie do obliczeń współczynnika korelacji Spearmana. Rysunek 7 przedstawia dziewięć cech o najwyższych jego wartościach względem kolumny MVP. Ponadto zawiera on również ich średnie miary dla graczy w zależności od otrzymania statuetki służące w celach porównawczych. Wartości współczynnika korelacji wynoszące

w przybliżeniu 0.08 oznaczają, że nie można bezpośrednio wnioskować na temat zmiennej MVP na podstawie badanych cech. Jednakże pozwalają one wstępnie ustalić ranking ich wpływu na wyniki budowanych modeli uczenia maszynowego.

	Korelacja	MVP	Bez MVP
WS	0.082726	15.895122	2.631269
OWS	0.082346	10.880488	1.351693
VORP	0.082285	7.802439	0.619454
PER	0.081786	28.031707	12.740332
WS/48	0.081696	0.264488	0.071896
BPM	0.081173	8.758537	-1.957246
PTS	0.081121	2090.487805	530.050082
OBPM	0.080965	6.765854	-1.554143
FG	0.080441	758.292683	200.660655

Rysunek 7 *Korelacja oraz średnie wartości wybranych zmiennych*
Źródło: *Opracowanie własne*

Innym ważnym elementem charakterystyki zbioru, który warto zbadać jest trend. W celu jego wizualizacji stworzone zostały wykresy czterech najlepszych zmiennych wybranych na podstawie korelacji. Zbadano ich zależność od statystyki award_share, która oznacza ułamek możliwych do otrzymania głosów. Zgodnie z tym gracz z największą jej wartością otrzymuje w danym roku nagrodę MVP. Pod uwagę nie były brane obserwacje z wartością wspomnianej zmiennej mniejszą niż 0.05, co oznacza otrzymanie 5 % wszystkich głosów. Wyniki wykonanej wizualizacji przedstawione zostały na rysunku 8.



Rysunek 8 *Wykresy zależności wybranych zmiennych od award_share*

Źródło: *Opracowanie własne*

3.4 Niezbalansowanie

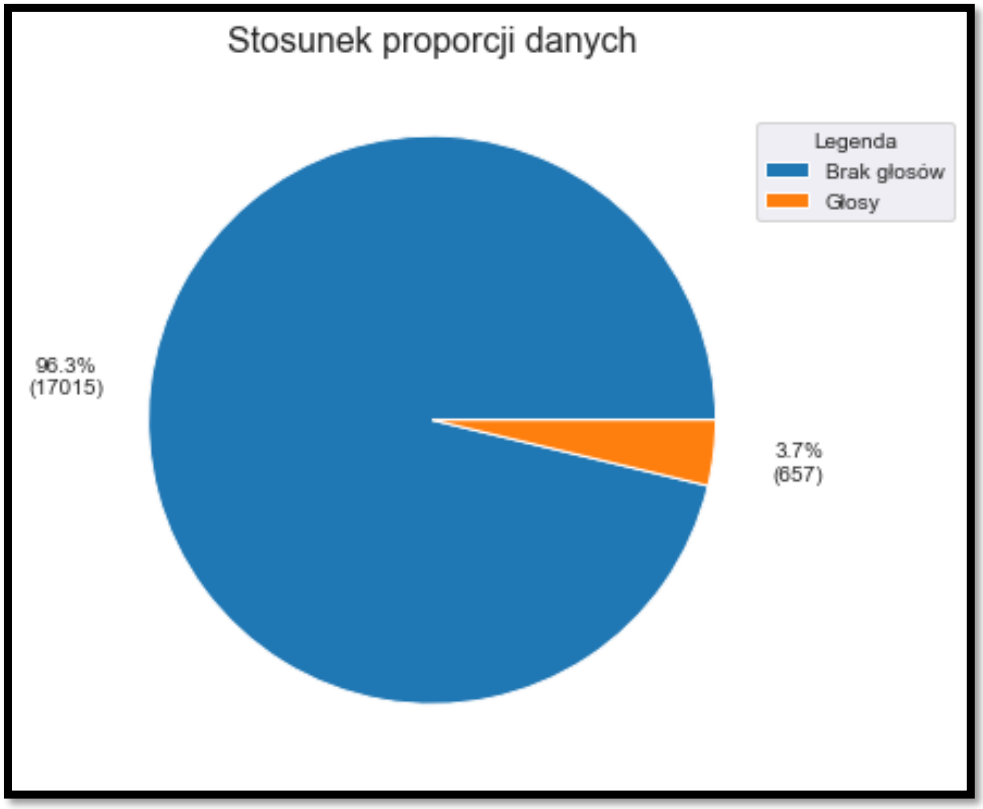
Niezbalansowanie danych jest sytuacją, w której występują znaczące różnice w wielkości klas zmiennej objaśnianej. Określenia tego możemy także użyć dla dużej dysproporcji rozkładu zmiennej numerycznej, na przykład występowanie wartości zero w większości zbioru. Zjawisko to jest nieporządane, ponieważ prowadzi do osłabienia modeli uczenia maszynowego oraz utrudnia ich ewaluacje.

3.4.1 Znaczenie problemu

Jednym z możliwych sposobów predykcji wyniku nagrody jest wykonanie klasyfikacji i poprawne przydzielenie binarnej kategorii zmiennej MVP. W podejściu tym występuje jednak znaczący problem, którym jest niezbalansowanie klas wspomnianej cechy.

Każdego roku tylko jeden zawodnik zostaje najlepszym graczem ligi, co oznacza, że analizowana zmienna zawiera 41 wartości jeden oraz 17631 wartości zero.

Drugim sposobem jest skorzystanie z regresjii i predykcja numerycznej zmiennej award_share. Rysunek 9 przedstawia wykres stosunku proporcji obserwacji w zależności od otrzymania chociaż jednego głosu. Również w tym przypadku z powodu znacznej liczby wartości zero, która wynika z charakterystyki głosowania na MVP, występuje niezbalansowanie utrudniające zbudowanie skutecznych modeli.



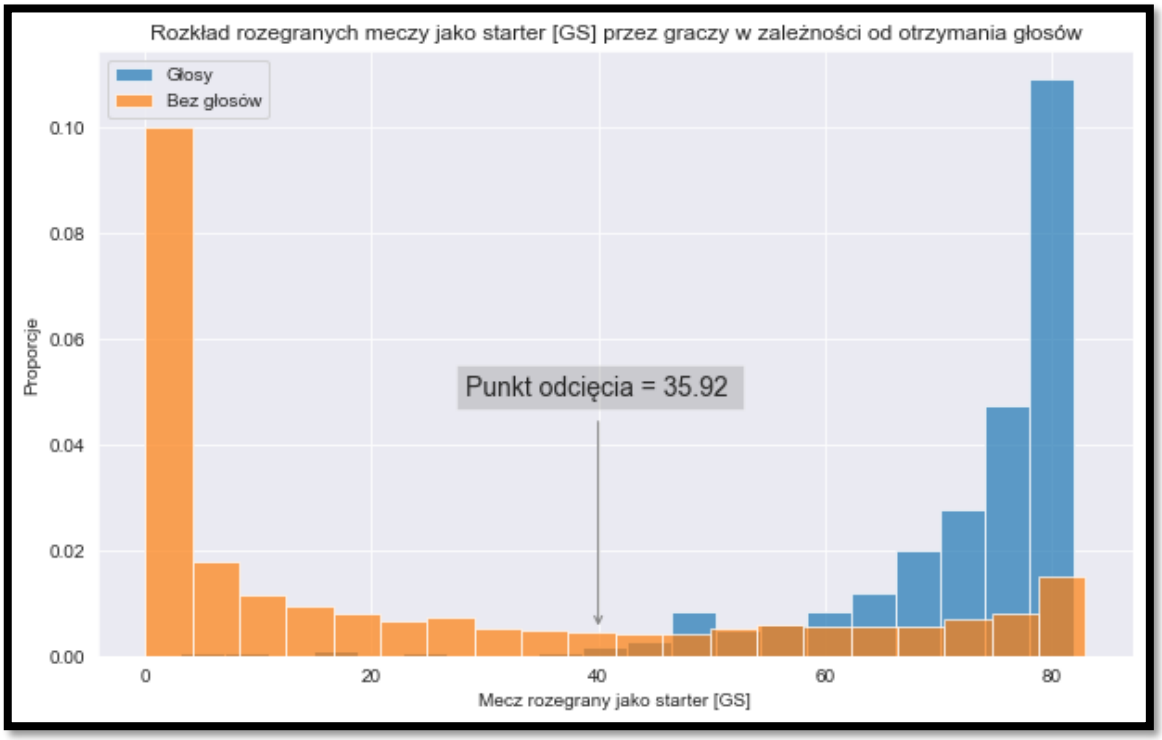
Rysunek 9 *Wykres przedstawiający niezbalansowanie danych*

Źródło: *Opracowanie własne*

3.4.2 Statystyczne balansowanie danych

W celu zbalansowania danych w pierwszej kolejności zastosowano podejście statystyczne. Polegało ono na wyborze zmiennych o największym potencjale posiadania różnic wartości pomiędzy klasami, a następnie obliczeniu punktu odcięcia na podstawie reguły trzy sigma. Rysunek 10 zawiera wykres rozkładu pierwszej badanej zmiennej GS.

Widać na nim, że liczba rozgrywanych meczy jako zawodnik wyjściowego zespołu jest w większości dużo niższa dla graczy, którzy nie otrzymali żadnego głosu.



Rysunek 10 *Wykres rozkładu zmiennej GS dla dwóch klas otrzymanych ze względu na zdobyte głosy*

Źródło: *Opracowanie własne*

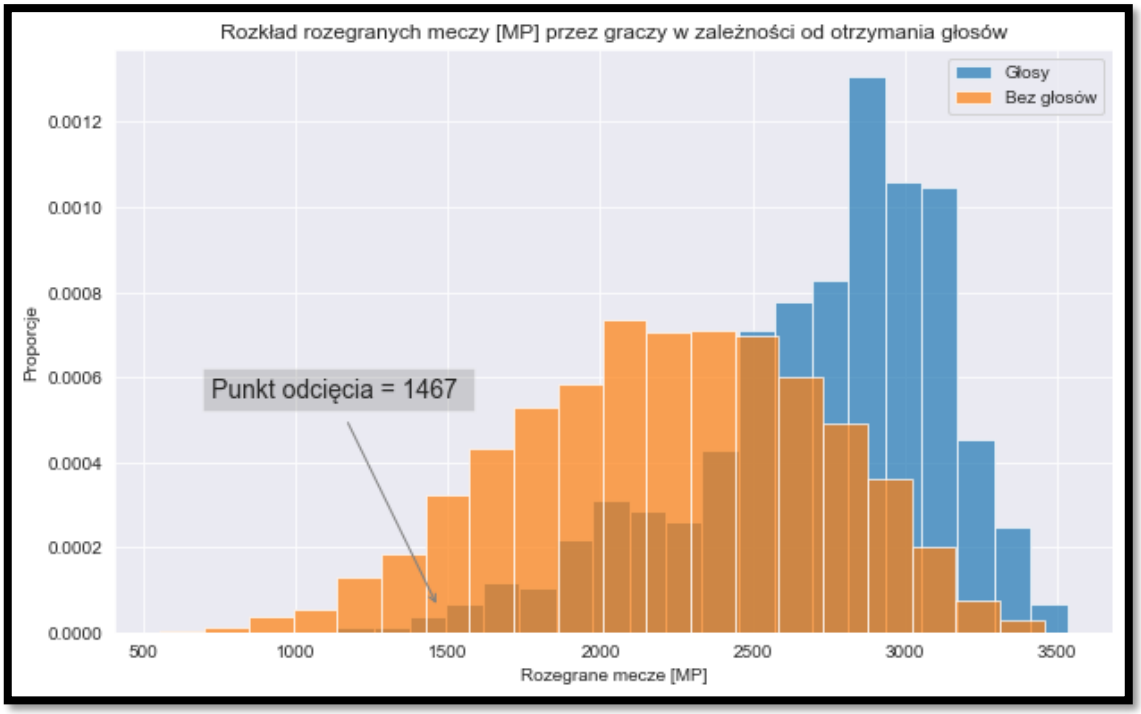
Przed usunięciem wartości odstających ważnym elementem było sprawdzenie, czy obserwacje sklasyfikowane jako takie i posiadające niezerową wartość award_share są istotne w dalszej analizie. Ich przykład dla zmiennej GS ukazany jest na rysunku 11. Jako kryterium istotności zastosowano wartość award_share większą niż 0.02, co oznacza otrzymanie ponad 2 % głosów. Zgodnie z nim wszystkie obserwacje odstające zostały uznane za nieistotne i usunięte ze zbioru.

	Year	Player	G	GS	award_share
0	1982	Michael Cooper	76	14.0	0.004
1	1991	Kevin McHale	68	10.0	0.001
2	1992	Detlef Schrempf	80	4.0	0.001
3	1995	Michael Jordan	17	17.0	0.011
4	1995	Dennis Rodman	49	26.0	0.009
5	1996	Magic Johnson	32	9.0	0.007
6	1999	Darrell Armstrong	50	15.0	0.002
7	1999	Rasheed Wallace	49	18.0	0.001
8	2008	Manu Ginobili	74	23.0	0.007
9	2010	Manu Ginobili	75	21.0	0.002
10	2021	Derrick Rose	50	3.0	0.010

Rysunek 11 *Obserwacje odstające posiadające niezerową wartość award_share dla zmiennej GS*

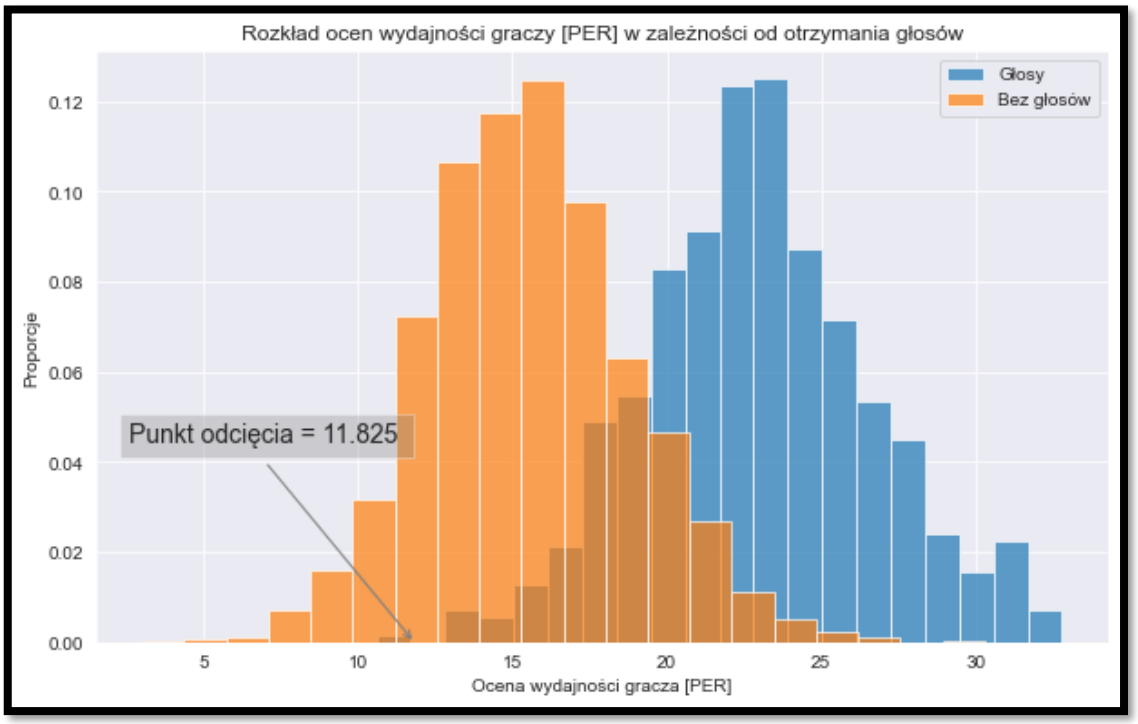
Źródło: *Opracowanie własne*

Schemat ten został następnie powtórzony dla czterech innych cech: MP, PER, VORP oraz BPM, a dalszy proces nie prowadził do otrzymania zauważalnych różnic w zbalansowaniu danych. Rysunki 12 i 13 zawierają wykresy rozkładów dwóch z nich wraz z zaznaczoną wartością punktu odcięcia obliczonego na podstawie reguły trzy sigma dla klasy zawierającej obserwacje o niezerowej liczbie otrzymanych głosów. Należy zaznaczyć, że dane dla kolejnych badanych zmiennych pozbawione są tych usuniętych poprzednio.



Rysunek 12 *Wykres rozkładu zmiennej MP dla dwóch klas otrzymanych ze względu na zdobyte głosy*

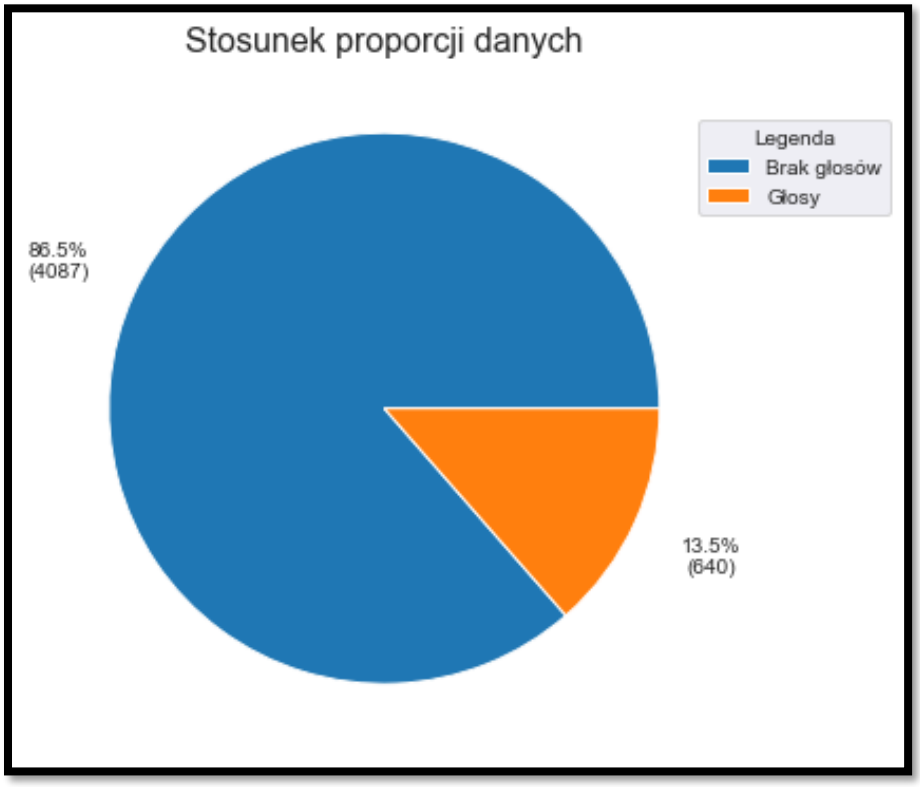
Źródło: *Opracowanie własne*



Rysunek 13 *Wykres rozkładu zmiennej PER dla dwóch klas otrzymanych ze względu na zdobyte głosy*

Źródło: *Opracowanie własne*

W rezultacie podejścia statystyczego do zbalansowania danych otrzymano stosunek proporcji pomiędzy klasami przedstawiony na rysunku 14. Porównując go do rysunku 9, ze zbioru usuniętych zostało 12,945 wierszy, w tym jednie 17 zawierających statystyki graczy, którzy otrzymali chociaż jeden głos.



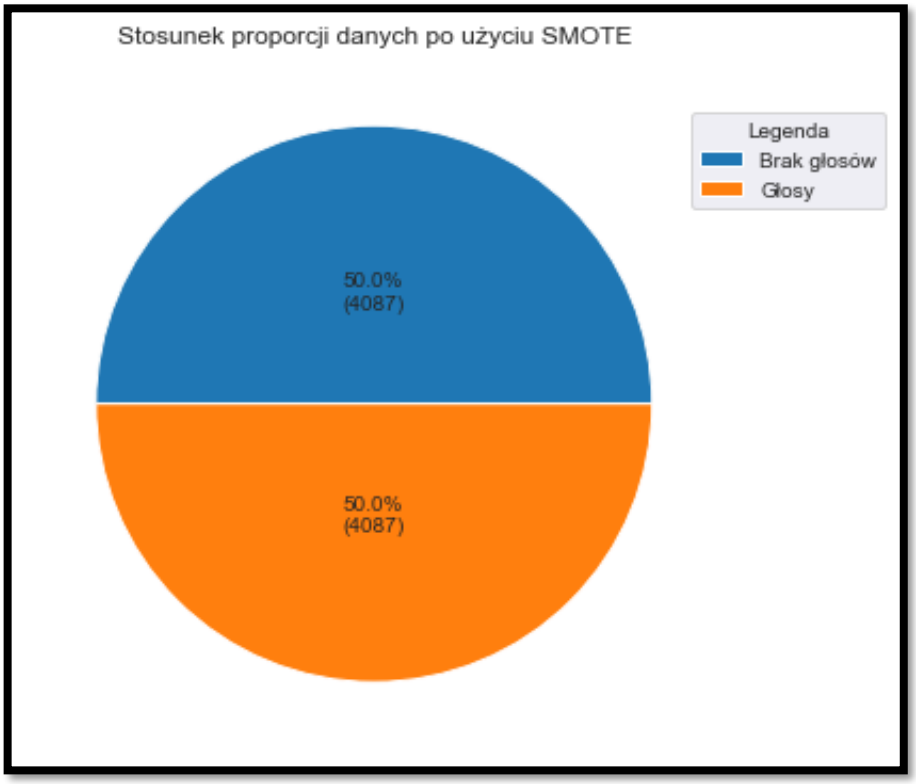
Rysunek 14 *Wykres przedstawiający stosunek proporcji danych po przeprowadzonym balansie statystycznym*
Źródło: *Opracowanie własne*

3.4.3 Wykorzystanie nadpróbkowania

Nadpróbkowanie (ang. *oversampling*) jest techniką pozwalającą na zbalansowanie danych. Polega na powiększeniu mniejszej z klas o sztuczne obserwacje w celu wyrównania jej z większą. Podejście to zostało wybrane, ponieważ liczność zbioru jest stosunkowo niewielka, a dodanie kolejnych obserwacji może pozytywnie wpłynąć na tworzone modele uczenia maszynowego.

Spośród wielu algorytmów nadpróbkowania wybór padł na SMOTE (ang. *Synthetic Minority Oversampling Technique*). Losuje on obserwacje ze zbioru mniejszej klasy, a następnie dodaje sztuczne wartości leżące na przecięciu się linii łączących ich

najbliższych sąsiadów w przestrzeni cech [17]. Proces ten jest powtarzany aż do otrzymania oczekiwanego balansu. Rysunek 15 przedstawia stosunek proporcji danych po zastosowaniu omówionego algorytmu.



Rysunek 15 *Wykres przedstawiający stosunek proporcji danych po zastosowaniu algorytmu SMOTE*
Źródło: *Opracowanie własne*

4 Opis podejścia do predykcji

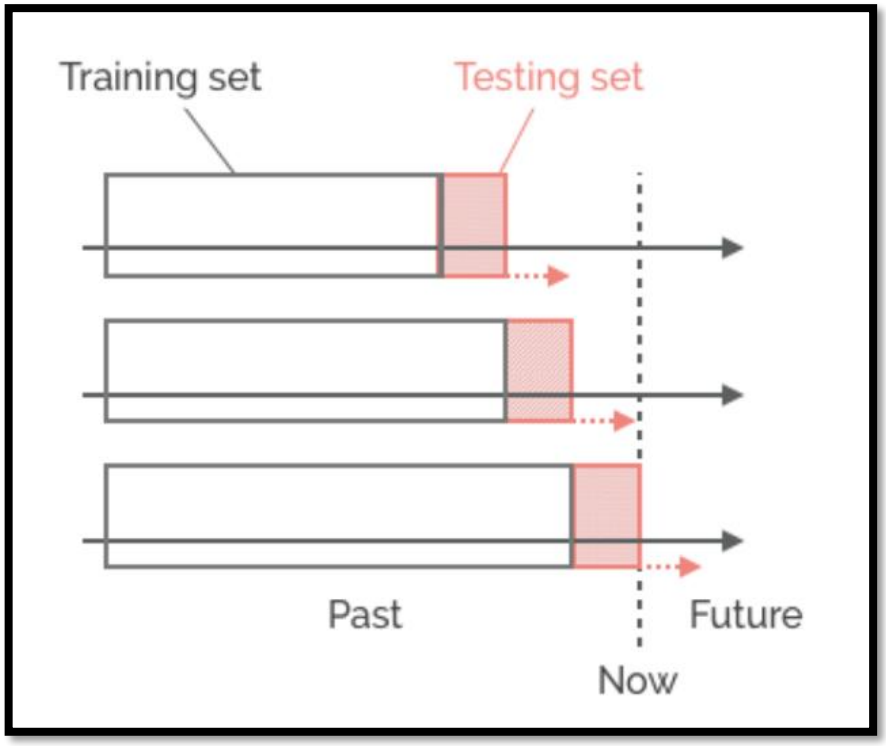
Celem rozdziału jest opisanie podejścia do predykcji wyniku nagrody MVP, która jest głównym tematem omawianej pracy. Skupia się on na strategii budowy modeli uczenia maszynowego oraz przedstawia użyte algorytmy. Zawiera on także schemat oceny ich działania.

4.1 Strategia

Zasadniczą część pracy rozpocząć należy od przedstawienia strategii wykorzystanej do predykcji wyniku nagrody MVP. W tym celu posłużono się dwiema metodami uczenia maszynowego: klasyfikacją oraz regresją. W obu z nich do oceny skuteczności posłużono się trzema zbiorami danych przygotowanymi na etapie balansowania. Są to:

- dane niezbalansowane,
- dane zbalansowane metodą statystyczną,
- dane zbalansowane metodą statystyczną oraz algorytmem SMOTE.

Na podstawie wniosków wyciągniętych z badania zmienności w czasie podjęto decyzje o zastosowaniu testowania wstecznego z powiększającym się oknem (*ang. backtesting with expanding window)*. Rysunek 16 zawiera schemat poglądowy tego podejścia. Polega ono na ewaluacji modelu przy pomocy zbioru treningowego (biały prostokąt) zawierającego wyłącznie dane historyczne względem zbioru testowego (czerwony prostokąt), który dzielony jest na odcinki czasowe. Następnie w kolejności chronologicznej dla każdego z nich wykonywane są trening modelu oraz predykcja [18].



Rysunek 16 *Schemat poglądowy testowania wstecznego z powiększającym się oknem*
Źródło: [18]

W procesie implementacji testowania wstecznego w problemie predykcji MVP jako zbiór treningowy wybrano dane z lat 1982-2009, natomiast jako zbiór testowy dane z lat 2010-2022. Główną uwagę skupiono na najnowszych statystykach, ponieważ z założenia model ma służyć w celu przewidzenia rezultatów przyszłych nieznanych sezonów. Stosunek zbliżony do 70:30 z jakim podzielono obserwacje pozwolił na maksymalizację wielkości danych stosowanych do treningu przy jednoczesnym zachowaniu optymalnej liczby wyników przeznaczonych do oceny.

4.2 Klasyfikacja

Klasyfikacja polega na przydzieleniu etykiety klasy dla obserwacji które przewidujemy. Jednym z wyróżnianych jej typów jest wariant binarny, w którym występują dwie możliwe wartości zmiennej objaśnianej zapisywane jako 0 i 1 oznaczające kolejno fałsz oraz prawdę. Predykcja wyniku wyboru nagrody MVP jest jednym z jego przykładów.

Podczas procesu budowy i ewaluacji modelów klasyfikacyjnych posłużono się możliwością predykcji prawdopodobieństwa przynależności obserwacji do danej klasy.

Umożliwiło to stworzenie rankingu pomagającego w ocenie, który jest dokładniej przedstawiony w kolejnym podrozdziale. W ten sposób zapobiegnięto również możliwości otrzymania kilku zwycięzców w tym samym roku.

Spośród pięciu testowanych algorytmów, w finalnej wersji pracy pozostawiono i udoskonalono dwa, które osiągnęły najlepsze wyniki. Są nimi :

- Las losowy (ang. *Random forrest*),
- Wzmocnienie gradientowe (ang. *Gradient boosting*).

4.3 Regresja

Regresja polega na estymacji wielkości nieznanej zmiennej na podstawie innych cech objaśniających. W problemie predykcji wyniku wyboru nagrody MVP użyto jej w celu wyznaczenia wartości statystyki award_share, która oznacza ułamek możliwych do otrzymania głosów. Zgodnie z tym, zawodnik z największym jej wskaźnikiem otrzymuje w danym roku nagrodę najlepszego gracza ligi.

W pierwszym kroku stworzono prosty model regresji liniowej służący jako punkt odniesienia w ocenie. Tak samo jak w przypadku klasyfikacji, na podstawie zwróconych wartości sporządzono ranking dla poszczególnych lat. Ostatecznie wybrano dwa najlepsze algorytmy:

- Las losowy (ang. *Random forrest*),
- Wzmocnienie gradientowe (ang. *Gradient boosting*).

4.4 Schemat oceny modeli

Rysunek 17 przedstawia wyniki oceny modelu regresji logistycznej dla danych zbalansowanych statystycznie. Służy on jako przykład ewaluacji podejścia klasyfikacyjnego, która stosowana była dla wszystkich modeli a następnie jej wyniki zestawiane ze sobą i porównywane w celu wybrania najlepszego rozwiązania.

```
Brier Score:  0.0013181849685843933
              precision    recall  f1-score   support

         0.0       1.00      1.00      1.00      6504
         1.0       0.64      0.69      0.67        13

    accuracy                           1.00      6517
   macro avg       0.82      0.85      0.83      6517
weighted avg       1.00      1.00      1.00      6517

       Year                 Player  MVP  MVP_pred_proba  Rank
11375  2010           LeBron James  1.0        0.907515   1.0
11951  2011           Derrick Rose  1.0        0.794699   1.0
12280  2012           LeBron James  1.0        0.868584   1.0
12736  2013           LeBron James  1.0        0.941197   1.0
13126  2014           Kevin Durant  1.0        0.798579   1.0
13585  2015          Stephen Curry  1.0        0.121383   2.0
14070  2016          Stephen Curry  1.0        0.885389   1.0
14899  2017      Russell Westbrook  1.0        0.979188   1.0
15122  2018           James Harden  1.0        0.060062   4.0
15485  2019  Giannis Antetokounmpo  1.0        0.038535   3.0
16010  2020  Giannis Antetokounmpo  1.0        0.025849   3.0
16789  2021           Nikola Jokić  1.0        0.925129   1.0
17356  2022           Nikola Jokić  1.0        0.873488   1.0
Skuteczność predykcji pierwszego miejsca:  69.23076923076923 %
Skuteczność predykcji minimum drugiego miejsca:  76.92307692307693 %
Skuteczność predykcji minimum trzeciego miejsca:  92.3076923076923 %
```

Rysunek 17 *Wyniki oceny modelu regresji logistycznej dla danych zbalansowanych statystycznie*

Źródło: *Opracowanie własne*

W pierwszym kroku obliczana jest metryka Brier'a (ang. *Brier score*), która oznacza różnicę średniokwadratową pomiędzy przewidywanym prawdopodobieństwem a rzeczywistymi wartościami. Służy ona jako wyjściowe spojrzenie na działanie modelu. Następnie wyświetlane są statystyki związane z macierzą pomyłek (ang. *confusion matrx*). Są one silnie zależne od badanego zbioru danych, więc używane są do porównań w obrębie jednego z nich. W dalszej kolejności prezentowane są rzeczywiste dane zawodników którzy otrzymali nagrodę MVP w latach dla których wykonywany jest test. Poza predykowaną wartością prawdopodobieństwa najbardziej kluczowy jest ranking. Oznacza on miejsce jakie przydzielone zostało danej obserwacji przez model. Daje on możliwość zdobycia informacji o jakości predykcji w konkretnym roku, a także umożliwia stworzenie wartości skuteczności modelu w zależnosci od miejsca, które

przedstawione są w dolnej części rysunku 17. Jest on najbardziej kluczową i ostateczną częścią procesu wyboru najlepszego rozwiązania.

```
MSE: 0.01182744409821526
MAE: 0.07208034790719611
      Year                 Player  MVP  Rank
3286  2010           LeBron James  1.0   1.0
3461  2011           Derrick Rose  1.0   2.0
3537  2012           LeBron James  1.0   2.0
3640  2013           LeBron James  1.0   1.0
3738  2014           Kevin Durant  1.0   1.0
3851  2015          Stephen Curry  1.0   2.0
3967  2016          Stephen Curry  1.0   1.0
4169  2017      Russell Westbrook  1.0   1.0
4230  2018           James Harden  1.0   2.0
4303  2019  Giannis Antetokounmpo  1.0   2.0
4418  2020  Giannis Antetokounmpo  1.0   1.0
4559  2021           Nikola Jokić  1.0   1.0
4677  2022           Nikola Jokić  1.0   1.0
Skuteczność predykcji pierwszego miejsca:  61.53846153846154 %
Skuteczność predykcji minimum drugiego miejsca:  100.0 %
Skuteczność predykcji minimum trzeciego miejsca:  100.0 %
```

Rysunek 18 *Wyniki oceny modelu regresji liniowej dla danych zbalansowanych statystycznie*
Źródło: *Opracowanie własne*

W przypadku podejścia regresyjnego, metrykę Brier'a oraz macierz pomyłek zastąpiono błędem średniokwadratowym (ang. *mean squared error)* MSE oraz średnim błędem bezwzględnym (ang. *mean absolute error*). Służą one również jako wyjściowa metoda oceny modelu, natomiast pozostałe elementy w tym ranking pozostały bez zmian. Przykład wyników oceny regresji liniowej dla danych zbalansowanych statystycznie przedstawiony jest na rysunku 18.

5 Wyniki

Celem rozdziału jest przedstawienie wyników jakie uzyskały dwa najlepsze algorytmy dla metod klasyfikacji oraz regresji. Zawiera on zbiorcze zestawienia skuteczności oraz szczegółowe oceny dla najefektywniejszych modeli.

5.1 Klasyfikacja

Zbiorcze zestawienie wyników predykcji zwycięzcy nagrody MVP dla dwóch najlepszych algorytmów klasyfikacyjnych przedstawone jest w tabeli 5. Zarówno w przypadku lasu losowego jak i wzmocnienia gradientowego najskuteczniejsze okazały się modele działające na danych zbalansowanych statystycznie z wykorzystaniem SMOTE. Ich szczegółowe oceny znajdujące się na rysunkach 19 oraz 20, służące uzupełniająco, sugerują zdecydowany wybór wzmocnienia gradientowego jako zwycięskiego modelu klasyfikacyjnego. Przewidział on pozytywne etykiety zbioru testowego zaledwie z jednym błędem, umieszczając zwycięzcę z 2015 roku na drugim miejscu w rankingu.

Tabela 5. *Zestawienie skuteczności najlepszych algorytmów klasyfikacyjnych*

Las losowy			
Skuteczność / Zbiór danych	Skuteczność min. pierwszego miejsca	Skuteczność min. drugiego miejsca	Skuteczność min. trzeciego miejsca
Dane niezbalansowane	53.85 %	92.31 %	92.31 %
Dane zbalansowane statystyczne	69.23 %	92.31 %	92.31 %
Dane zbalansowane statystycznie + SMOTE	76.92 %	92.31 %	92.31 %
Wzmocnienie gradientowe			
Skuteczność / Zbiór danych	Skuteczność min. pierwszego miejsca	Skuteczność min. drugiego miejsca	Skuteczność min. trzeciego miejsca

Dane niezbalansowane	76.92 %	76.92 %	100 %
Dane zbalansowane statystyczne	76.92 %	84.62 %	100 %
Dane zbalansowane statystycznie + SMOTE	92.31 %	100 %	100 %

Źródło: *Opracowanie własne*

```
Brier Score:  0.0045542990538020404
              precision    recall  f1-score   support

         0.0       0.99      1.00      1.00      1492
         1.0       1.00      0.31      0.47        13

    accuracy                           0.99      1505
   macro avg       1.00      0.65      0.73      1505
weighted avg       0.99      0.99      0.99      1505

      Year                 Player  MVP  MVP_pred_proba  Rank
3286  2010           LeBron James  1.0        0.328269   1.0
3461  2011           Derrick Rose  1.0        0.015487   7.0
3537  2012           LeBron James  1.0        0.285657   1.0
3640  2013           LeBron James  1.0        0.507622   1.0
3738  2014           Kevin Durant  1.0        0.543956   1.0
3851  2015          Stephen Curry  1.0        0.391995   1.0
3967  2016          Stephen Curry  1.0        0.577209   1.0
4169  2017      Russell Westbrook  1.0        0.303152   2.0
4230  2018           James Harden  1.0        0.355802   1.0
4303  2019  Giannis Antetokounmpo  1.0        0.293995   2.0
4418  2020  Giannis Antetokounmpo  1.0        0.298894   1.0
4559  2021           Nikola Jokić  1.0        0.476370   1.0
4677  2022           Nikola Jokić  1.0        0.528262   1.0
Skuteczność predykcji pierwszego miejsca:  76.92307692307693 %
Skuteczność predykcji minimum drugiego miejsca:  92.3076923076923 %
Skuteczność predykcji minimum trzeciego miejsca:  92.3076923076923 %
```

Rysunek 19 *Wyniki oceny klasyfikatora lasu losowego dla danych zbalansowanych statystycznie + SMOTE*

Źródło: *Opracowanie własne*

```
Brier Score:  0.0090042940250073778
              precision    recall  f1-score   support

         0.0       1.00      1.00      1.00      1492
         1.0       1.00      0.85      0.92        13

    accuracy                           1.00      1505
   macro avg       1.00      0.92      0.96      1505
weighted avg       1.00      1.00      1.00      1505

      Year                 Player  MVP  MVP_pred_proba  Rank
3286  2010           LeBron James  1.0        0.829993   1.0
3461  2011           Derrick Rose  1.0        0.888807   1.0
3537  2012           LeBron James  1.0        0.893523   1.0
3640  2013           LeBron James  1.0        0.994510   1.0
3738  2014           Kevin Durant  1.0        0.994178   1.0
3851  2015          Stephen Curry  1.0        0.271928   2.0
3967  2016          Stephen Curry  1.0        0.979550   1.0
4169  2017      Russell Westbrook  1.0        0.995766   1.0
4230  2018           James Harden  1.0        0.967690   1.0
4303  2019  Giannis Antetokounmpo  1.0        0.309029   1.0
4418  2020  Giannis Antetokounmpo  1.0        0.969991   1.0
4559  2021           Nikola Jokić  1.0        0.996346   1.0
4677  2022           Nikola Jokić  1.0        0.998953   1.0
Skuteczność predykcji pierwszego miejsca:  92.3076923076923 %
Skuteczność predykcji minimum drugiego miejsca:  100.0 %
Skuteczność predykcji minimum trzeciego miejsca:  100.0 %
```

Rysunek 20 *Wyniki oceny klasyfikatora wzmocnienia gradientowego dla danych zbalansowanych statystycznie + SMOTE*
Źródło: *Opracowanie własne*

5.2 Regresja

Dla podejścia regresyjnego zbiorcze zestawienie wyników predykcji przedstawione jest w tabeli 6. Identycznie jak w przypadku klasyfikacji, oba algorytmy osiągnęły najlepszą skuteczność dla danych zbalansowanych statystycznie z wykorzystaniem SMOTE. Korzystając z dokładnych ocen ukazanych na rysunkach 21 oraz 22, najlepszym modelem regresyjnym ponownie wybrany został ten działający na wzmocnieniu gradientowym. Przewidział on bezbłędnie wszystkie pozytywne etykiety zbioru testowego, a także otrzymał satysfakcjonujące wartości metryk MSE oraz MAE.

Tabela 6. *Zestawienie skuteczności najlepszych algorytmów regresyjnych*

Las losowy			
Skuteczność / Zbiór danych	Skuteczność min. pierwszego miejsca	Skuteczność min. drugiego miejsca	Skuteczność min. trzeciego miejsca
Dane niezbalansowane	61.54 %	84.62 %	84.62 %
Dane zbalansowane statystyczne	69.23 %	92.31 %	92.31 %
Dane zbalansowane statystycznie + SMOTE	76.92 %	92.31 %	92.31 %
Wzmocnienie gradientowe			
Skuteczność / Zbiór danych	Skuteczność min. pierwszego miejsca	Skuteczność min. drugiego miejsca	Skuteczność min. trzeciego miejsca
Dane niezbalansowane	61.54 %	92.31 %	92.31 %
Dane zbalansowane statystyczne	53.85 %	84,62 %	92.31 %
Dane zbalansowane statystycznie + SMOTE	100 %	100 %	100 %

Źródło: *Opracowanie własne*

```
MSE: 0.006741246851824197
MAE: 0.02553570728016874
      Year                 Player  MVP  Rank
3286  2010          LeBron James  1.0   1.0
3461  2011          Derrick Rose  1.0   6.0
3537  2012          LeBron James  1.0   1.0
3640  2013          LeBron James  1.0   1.0
3738  2014          Kevin Durant  1.0   1.0
3851  2015         Stephen Curry  1.0   2.0
3967  2016         Stephen Curry  1.0   1.0
4169  2017     Russell Westbrook  1.0   2.0
4230  2018          James Harden  1.0   1.0
4303  2019  Giannis Antetokounmpo  1.0   2.0
4418  2020  Giannis Antetokounmpo  1.0   2.0
4559  2021          Nikola Jokić  1.0   1.0
4677  2022          Nikola Jokić  1.0   1.0
Skuteczność predykcji pierwszego miejsca:  61.53846153846154 %
Skuteczność predykcji minimum drugiego miejsca:  92.3076923076923 %
Skuteczność predykcji minimum trzeciego miejsca:  92.3076923076923 %
```

Rysunek 21 *Wyniki oceny modelu regresyjnego lasu losowego dla danych zbalansowanych statystycznie + SMOTE*
Źródło: *Opracowanie własne*

```
MSE: 0.007790412639506482
MAE: 0.02441919768821072
      Year                 Player  MVP  Rank
3286  2010          LeBron James  1.0   1.0
3461  2011          Derrick Rose  1.0   1.0
3537  2012          LeBron James  1.0   1.0
3640  2013          LeBron James  1.0   1.0
3738  2014          Kevin Durant  1.0   1.0
3851  2015         Stephen Curry  1.0   1.0
3967  2016         Stephen Curry  1.0   1.0
4169  2017     Russell Westbrook  1.0   1.0
4230  2018          James Harden  1.0   1.0
4303  2019  Giannis Antetokounmpo  1.0   1.0
4418  2020  Giannis Antetokounmpo  1.0   1.0
4559  2021          Nikola Jokić  1.0   1.0
4677  2022          Nikola Jokić  1.0   1.0
Skuteczność predykcji pierwszego miejsca:  100.0 %
Skuteczność predykcji minimum drugiego miejsca:  100.0 %
Skuteczność predykcji minimum trzeciego miejsca:  100.0 %
```

Rysunek 22 *Wyniki oceny modelu regresyjnego wzmocnienia gradientowego dla danych zbalansowanych statystycznie + SMOTE*
Źródło: *Opracowanie własne*

6 Wnioski

Przede wszystkim pierwszym spostrzeżeniem po analizie działania modeli jest dominacja ich skuteczności dla danych zbalansowanych statystycznie wraz z zastosowaniem SMOTE. Osiągnęły one najbardziej zadowalające wyniki oraz wartości metryk dla każdego algorytmu oraz zastosowanego podejścia.

Kolejnym wnioskiem jest znacząca przewaga wzmocnienia gradientowego nad lasem losowym, która jest zauważalna w każdym badanym przypadku. Wynikać może ona z faktu, iż algorytm wzmocnienia gradientowego jest lepiej dostosowany do działania na niezbalansowanych danych poprzez wzmacnianie wpływu klasy z etykietami pozytywnymi. Ponadto wykonuje on optymalizacje w przestrzeni funkcji, co ułatwia korzystanie z funkcji straty [19].

Rysunki 23 oraz 24 przedstawiają wyniki dwóch najlepszych modeli na zbiorze testowym powiększonym o lata 2000-2009. Osiągnęły one podobne do siebie rezultaty, ale nieco gorsze niż poprzednio. Jednakże biorąc pod uwage, że w latach tych wystąpiło wiele kontrowersyjnych wyborów nagrody MVP [20], należy uznać je za satysfakcjonujące.

```
MSE: 0.008498323898250714
MAE: 0.0247517143487648
      Year                 Player  MVP  Rank
2066  2000       Shaquille O'Neal  1.0   1.0
2149  2001          Allen Iverson  1.0   4.0
2248  2002             Tim Duncan  1.0   1.0
2368  2003             Tim Duncan  1.0   1.0
2500  2004          Kevin Garnett  1.0   1.0
2676  2005             Steve Nash  1.0   7.0
2797  2006             Steve Nash  1.0   6.0
2926  2007          Dirk Nowitzki  1.0   1.0
2981  2008            Kobe Bryant  1.0   5.0
3154  2009           LeBron James  1.0   1.0
3286  2010           LeBron James  1.0   1.0
3461  2011           Derrick Rose  1.0   1.0
3537  2012           LeBron James  1.0   1.0
3640  2013           LeBron James  1.0   1.0
3738  2014           Kevin Durant  1.0   1.0
3851  2015          Stephen Curry  1.0   1.0
3967  2016          Stephen Curry  1.0   1.0
4169  2017      Russell Westbrook  1.0   1.0
4230  2018           James Harden  1.0   1.0
4303  2019  Giannis Antetokounmpo  1.0   1.0
4418  2020  Giannis Antetokounmpo  1.0   1.0
4559  2021           Nikola Jokić  1.0   1.0
4677  2022           Nikola Jokić  1.0   1.0
Skuteczność predykcji pierwszego miejsca:  82.6086956521739 %
Skuteczność predykcji minimum drugiego miejsca:  82.6086956521739 %
Skuteczność predykcji minimum trzeciego miejsca:  82.6086956521739 %
```

Rysunek 23 *Wyniki oceny najlepszego modelu regresyjnego dla powiększonego zbioru testowego*

Źródło: *Opracowanie własne*

```
Brier Score:  0.0021112434008379445
              precision    recall  f1-score   support

         0.0       1.00      1.00      1.00      2714
         1.0       0.94      0.70      0.80        23

    accuracy                           1.00      2737
   macro avg       0.97      0.85      0.90      2737
weighted avg       1.00      1.00      1.00      2737

      Year                 Player  MVP  MVP_pred_proba  Rank
2066  2000       Shaquille O'Neal  1.0        0.974422   1.0
2149  2001          Allen Iverson  1.0        0.010950   1.0
2248  2002             Tim Duncan  1.0        0.996085   1.0
2368  2003             Tim Duncan  1.0        0.992395   1.0
2500  2004          Kevin Garnett  1.0        0.915364   1.0
2676  2005             Steve Nash  1.0        0.287442   2.0
2797  2006             Steve Nash  1.0        0.003149   5.0
2926  2007          Dirk Nowitzki  1.0        0.249641   1.0
2981  2008            Kobe Bryant  1.0        0.017632   1.0
3154  2009           LeBron James  1.0        0.974253   1.0
3286  2010           LeBron James  1.0        0.829993   1.0
3461  2011           Derrick Rose  1.0        0.888807   1.0
3537  2012           LeBron James  1.0        0.893523   1.0
3640  2013           LeBron James  1.0        0.994510   1.0
3738  2014           Kevin Durant  1.0        0.994178   1.0
3851  2015          Stephen Curry  1.0        0.271928   2.0
3967  2016          Stephen Curry  1.0        0.979550   1.0
4169  2017      Russell Westbrook  1.0        0.995766   1.0
4230  2018            James Harden  1.0        0.967690   1.0
4303  2019  Giannis Antetokounmpo  1.0        0.309029   1.0
4418  2020  Giannis Antetokounmpo  1.0        0.969991   1.0
4559  2021           Nikola Jokić  1.0        0.996346   1.0
4677  2022           Nikola Jokić  1.0        0.998953   1.0
Skuteczność predykcji pierwszego miejsca:  86.95652173913044 %
Skuteczność predykcji minimum drugiego miejsca:  95.65217391304348 %
Skuteczność predykcji minimum trzeciego miejsca:  95.65217391304348 %
```

Rysunek 24 *Wyniki oceny najlepszego modelu klasyfikacyjnego dla powiększonego zbioru testowego*
Źródło: *Opracowanie własne*

W rezultacie jako ostateczne rozwiązanie problemu predykcji wyniku wyboru zwycięzcy nagrody MVP wybrano dwa modele korzystające z danych zbalansowanych statystycznie wraz z wykorzystaniem algorytmu SMOTE:

- model klasyfikacyjny wzmocnienia gradientowego,
- model regresyjny wzmocnienia gradientowego.

Dają one podobne rezultaty i oba powinny być używane w przewidywaniu przyszłych zwycięzców statuetki.

7 Podsumowanie

Celem pracy była predykcja wyniku wyboru zwycięzcy nagrody MVP w lidze NBA z wykorzystaniem technik uczenia maszynowego. Po wykonaniu przygotowania wstępnego zgromadzonych danych nastąpiła ich analiza deskryptywna. Zawierała ona badanie współliniowości, korelacji, trendu, zmienności w czasie oraz zbalansowania zbioru. Na jej podstawie dokonano wyboru odpowiedniego podejścia do utworzenia modeli predykcyjnych. Spośród nich, zarówno klasyfikacyjnych, jak i regresyjnych wybrano dwa, które dały najlepsze rezultaty. Przewidziały one bezbłędnie zwycięzców nagrody MVP w latach 2010-2022, natomiast dla lat 2000-2022 otrzymały blisko 87 % skuteczności.

Zrealizowany projekt inżynierski nie wyczerpuje tematu i daje możliwość na dalsze ulepszenia. Przede wszystkim, z każdym rokiem zbiór danych powiększa się, co ma pozytywny wpływ na efektywność modeli. Tuning hiperparametrów oraz testowanie innych algorytmów są kolejnymi sposobami na poprawienie wyników. Wartą sprawdzenia jest także inżynieria cech (ang. feature engineering) i związana z nią możliwość utworzenia własnych dodatkowych zmiennych, które wprowadziłyby dodatkowe informacje.

Bibliografia

[1] Sourav Das: Top 10 Most Popular Sports In The World December 2022, https://sportsbrowser.net/most-popular-sports/ [dostęp 01.12.2022]

[2] Hartyáni Zsolt: History of Basketball, http://www.basketref.com/en/index.php/rules/rules-history [dostęp 01.12.2022]

[3] Redakcja polskikosz.pl: Krótka historia koszykówki, https://polskikosz.pl/krotka-historia-koszykowki/ [dostęp 01.12.2022]

[4] Steve Farrugia: THE 3-POINT LINE: HOW IT CHANGED THE GAME OF BASKETBALL, https://fieldinsider.com/the-3-point-line-how-it-changed-the-game-of-basketball/ [dostęp 01.12.2022]

[5] Redakcja thehoopsgeek.com: Basketball Court Dimensions – 25 Diagrams & All The Measurements, https://www.thehoopsgeek.com/basketball-court-dimensions [dostęp 01.12.2022]

[6] Redakcja History.com: NBA is born, https://www.history.com/this-day-in-history/nba-is-born [dostęp 01.12.2022]

[7] Redakcja daisydreams.com: HOW THE NBA IS STRUCTURED, https://daisydreams.net/how-the-nba-is-structured/ [dostęp 01.12.2022]

[8] Redakcja nba.com: NBA announces structure and format for 2020-21 season, https://www.nba.com/news/nba-announces-structure-and-format-for-2020-21-season [dostęp 01.12.2022]

[9] Michael Corvo: How voting is done for the NBA MVP and its evolution, https://clutchpoints.com/how-voting-is-done-for-the-nba-mvp-and-its-evolution [dostęp 01.12.2022]

[10] gswdgrinfelds: Diving into Stephen Curry’s Unanimous MVP Year, https://www.nba.com/warriors/news-blogs/stephen-curry-unanimous-mvp-20200624 [dostęp 01.12.2022]

[11] Redakcja nba.com: NBA MVP Award Winners, https://www.nba.com/news/history-mvp-award-winners [dostęp 01.12.2022]

[12] Omri Goldstein: NBA Players stats since 1950, https://www.kaggle.com/datasets/drgilermo/nba-players-stats?resource=download&select=Seasons_Stats.csv [dostęp 01.12.2022]

[13] Redakcja basketball-reference.com: Calculating PER, https://www.basketball-reference.com/about/per.html [dostęp 01.12.2022]

[14] Redakcja basketball-reference.com: NBA Win Shares, https://www.basketball-reference.com/about/ws.html [dostęp 01.12.2022]
[15] Daniel Myers: About Box Plus/Minus (BPM), https://www.basketball-reference.com/about/bpm2.html [dostęp 01.12.2022]
[16] Redakcja basketball-reference.com: Glossary, https://www.basketball-reference.com/about/glossary.html [dostęp 01.12.2022]
[17] Jason Brownlee: SMOTE for Imbalanced Classification with Python, https://machinelearningmastery.com/smote-oversampling-for-imbalanced-classification/ [dostęp 01.12.2022]
[18] Redakcja Datapred: The basics of backtesting, https://www.datapred.com/blog/the-basics-of-backtesting [dostęp 01.12.2022]
[19] Abolfazl Ravanshad: Gradient Boosting vs Random Forest, https://medium.com/@aravanshad/gradient-boosting-versus-random-forest-cfa3fa8f0d80 [dostęp 01.12.2022]
[20] Robert Felton: NBA: The Eight Most Controversial MVP Wins of All Time, https://bleacherreport.com/articles/573923-the-eight-most-controversial-nba-mvp-wins-of-all-time [dostęp 01.12.2022]