

Określanie pozycji, na której powinien grać piłkarz z na podstawie jego cech z wykorzystaniem poznanych narzędzi oraz środowiska Microsoft Azure.

Marcin Kowalski

Adam Kwapisz

Cel Projektu .....	3
Analiza eksploracyjna .....	3
Podstawowe informacje o zbiorze .....	3
Omówienie atrybutów .....	3
Omówienie klas .....	4
Modyfikacje dokonane w zbiorze .....	4
Związek atrybutów z klasami .....	4
Atrybuty wyróżniające się w klasach .....	4
Wartości średnie atrybutów dla danych klas .....	6
Dodatkowe informacje o zbiorze .....	7
Problemy z pierwotnym zbiorem .....	8
Wnioski .....	9
Sposób rozwiązania projektu .....	9
Wybrane cechy .....	9
Współczynniki pozycji .....	9
Wspólne kategorie dla zawodników grających po bokach .....	10
Użyty algorytm .....	10
Ogólny opis działania programu .....	10
Analiza eksploracyjna .....	11
Model .....	11
Wyniki .....	11
Macierz pomyłek .....	11
Dokładność modelu ze względu na podział na zbiór treningowy i testowy oraz liczbę $n$ .....	12
Zaskakująco dobre wyniki dla split = 0.1 i $n = 1$ .....	14
Małe różnice w poprawności, mimo różnych współczynników podziału. ....	15
Wpływ połączenia pozycji po bokach na poprawność przewidywania .....	15
Najlepsza wartość $n$ .....	15
Wnioski końcowe .....	15
Wnioski finalne .....	15
Napastnicy też bronią, obrońcy też strzelają .....	15
Założenia początkowe, a wyniki .....	15
Wyróżniasz się na wykresie, łatwiej cię przyporządkować .....	16
Czy model nadaje się do użytku? .....	16
Źródła .....	16

## Cel Projektu

Celem projektu jest zbadanie zbioru danych ze statystykami piłkarzy z gry FIFA19, wydanej przez Electronic Arts, z trybu Ultimate Team i próba przewidzenia na ich podstawie pozycji, na której występuje piłkarz.

## Analiza eksploracyjna

### Podstawowe informacje o zbiorze

#### Omówienie atrybutów

W nieobrobionym zbiorze znajduje się 21 kolumn z czego 15 można uznać za atrybuty przydatne w ocenie piłkarza. Kolumny w pliku to:

Name – imię i nazwisko zawodnika

Ratings – ocena ogólna na karcie

Position – pozycja na jakiej występuje zawodnik

Version – rodzaj karty

PS – cena na rynku w grze

SKI – liczba gwiazdek sztuczek, czyli umiejętność ich wykonywania

WF – umiejętność gry słabszą nogą

WR – work rate, czyli zaangażowanie gracza w dany typ akcji

PAC – szybkość biegu

SHO – strzały

PAS – podania

DRI – drybling

DEF – umiejętności defensywne

PHY – siła fizyczna

Body – wzrost

Popularity – popularność w grze

BS – suma statystyk na karcie

IGS – suma statystyk w grze

Club – klub, w którym zawodnik występuje

League – liga, w której występuje klub zawodnika

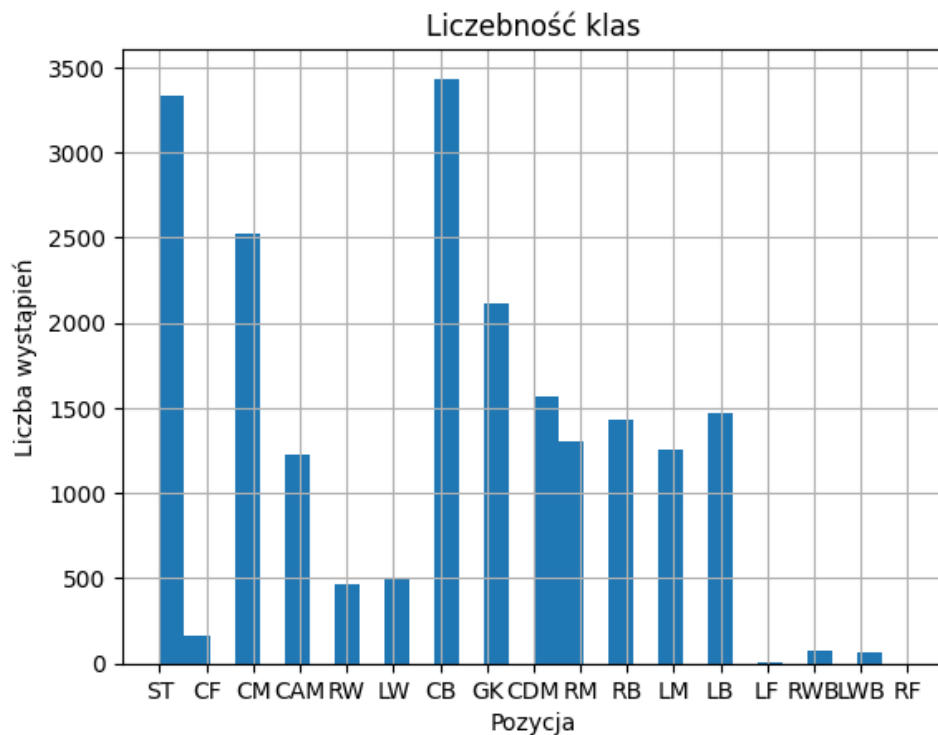
Country – kraj pochodzenia zawodnika

	Name	Ratings	Position	Version	PS	SKI	WF	WR	PAC	SHO	...	DRI	DEF	PHY	Body	Popularity	BS	IGS	Club	League	Country
0	Cristiano Ronaldo	99	ST	TOTY	7.19M	5	4	H \ L	97	99	...	98	50	95	187cm   6'2"	4833	534	2520	Juventus	Serie A TIM	Portugal
1	Lionel Messi	99	CF	TOTY	4.49M	4	4	M \ M	97	99	...	99	45	85	170cm   5'7"	3785	523	2507	FC Barcelona	LaLiga Santander	Argentina
2	Luka Modric	99	CM	TOTY	3.75M	4	4	H \ H	90	94	...	99	94	90	172cm   5'8"	956	566	2701	Real Madrid	LaLiga Santander	Croatia
3	Pelé	99	CAM	Icon	6.78M	5	4	H \ M	96	98	...	97	62	80	173cm   5'8"	1198	530	2563	Icons	Icons	Brazil
4	Lionel Messi	99	RW	CL TOTY	1.65M	4	4	M \ M	94	99	...	99	38	72	170cm   5'7"	678	500	2412	FC Barcelona	LaLiga Santander	Argentina

Do oceny pozycji piłkarza zdecydowaliśmy się zrezygnować z niektórych atrybutów, co zostało opisane w akapicie "modyfikacje dokonane w zbiorze".

### Omówienie klas

W projekcie jako klasy posłużą nam pozycje zawodników z gry FIFA19 (kolumna "Position"). Na poniższym histogramie przedstawiona jest liczebność poszczególnych klas.



Widzimy, że dla niektórych klas liczba reprezentantów jest na tyle niska, że nie da się z zadowalającą dokładnością określić ich pozycji, dlatego zdecydowaliśmy się zrezygnować z niektórych klas co zostało opisane w akapicie "modyfikacje dokonane w zbiorze". Podane klasy możemy podzielić na większe zbiory, zawodników linii ataku, pomocy i obrony. Taki podział ułatwi nam w dalszej analizie, dostrzeganie różnic między formacjami.

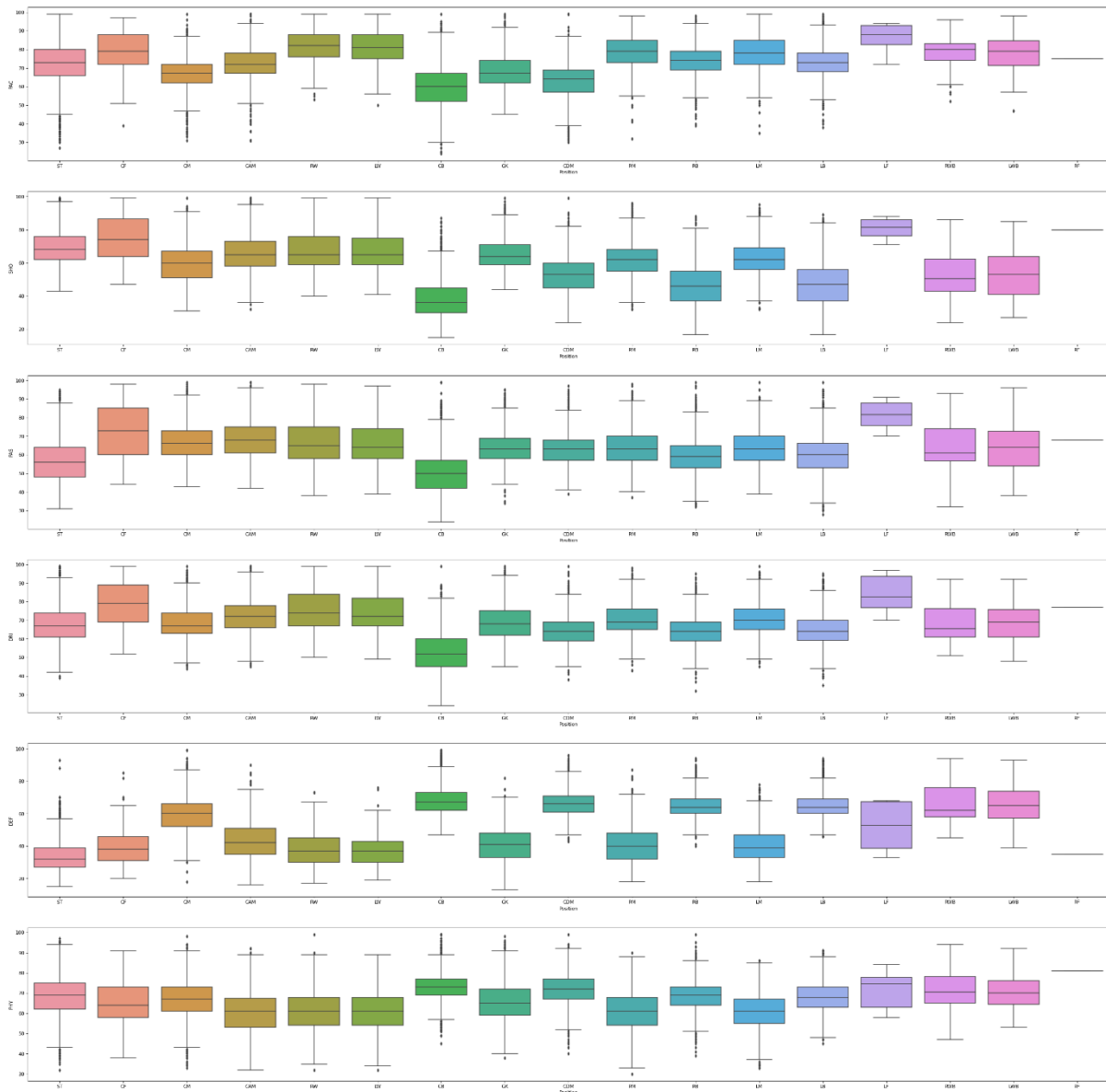
### Modyfikacje dokonane w zbiorze

Ze zbioru zostali usunięci zawodnicy występujący na pozycjach LF, RF z powodu małej liczebności oraz pozycji GK w miejscu podstawowych statystyk (dla graczy z pola) mają zapisane statystyki związane tylko z ich pozycją. Usunęliśmy także atrybuty, które nie będą przydatne w dalszej analizie są nimi: "Popularity", "BS", "IGS", "PS". Zmieniliśmy również format danych dla atrybutu "Body", żeby łatwiej pracowało się w nim w dalszej analizie. W celu ułatwienia konstrukcji modelu rozszerzyliśmy zbiór o dodatkowe kolumny co zostało opisane w akapicie Współczynnik pozycji.

### Związek atrybutów z klasami

#### Atrybuty wyróżniające się w klasach

W celu wstępnego wyróżnienia statystyk dla poszczególnych pozycji/grup pozycji użyliśmy wykresów pudełkowych.



- Największe wartości statystyki “PAC” cechują zawodników grających na bokach boiska, nazwy ich pozycji zaczynają się od R lub L, oraz na pozycji CF. Najniżej notowani w tej statystyce są zawodnicy z pozycji CB i CDM.
- Statystyka “SHO” przyjmuje największe wartości dla zawodników ofensywnych, a najmniejsze dla defensywnych.
- Atrybut “PAS” dla większości pozycji przyjmuje podobne wielkości. Można wyróżnić trzy pozycje ST i CB, dla których jest ona najniższa, oraz LF, dla której jest ona najwyższa (jest to spowodowane małą liczbą zawodników na tej pozycji).
- Największe wartości statystyki “DRI” mają zawodnicy grający na pozycjach CF, RW, LW, LF i RF, a najmniejsze “CB”.
- Dzięki statystyce “DEF” jesteśmy w stanie wyróżnić zawodników defensywnych, ze względu na wysoką jej wartość oraz zawodników ofensywnych mających niską jej wartość.
- Statystyka “PHY” pozwala odróżnić zawodników grających na środku boiska (wyższa wartość statystyki) od tych grających na jego bokach (niższa wartość statystyki).
- Zawodnicy mający najbardziej wyważone statystyki zajmują pozycje środkowych pomocników (CDM, CM, CAM).

### Wartości średnie atrybutów dla danych klas

Poniższe tabele przedstawiają wartości średnie statystyk w zależności od pozycji.

	Pozycje napastników			
Statystyka	ST	CF	RW	LW
Ratings	70,43	76,88	72,49	72,03
PAC	72,63	79,45	81,65	81,03
SHO	69,34	74,95	67,52	67,32
PAS	57,19	72,01	66,98	66,15
DRI	68,10	78,65	75,12	74,61
DEF	33,39	39,77	37,78	37,08
PHY	68,58	65,58	60,41	60,75
Body	182,71	177,83	176,23	176,95

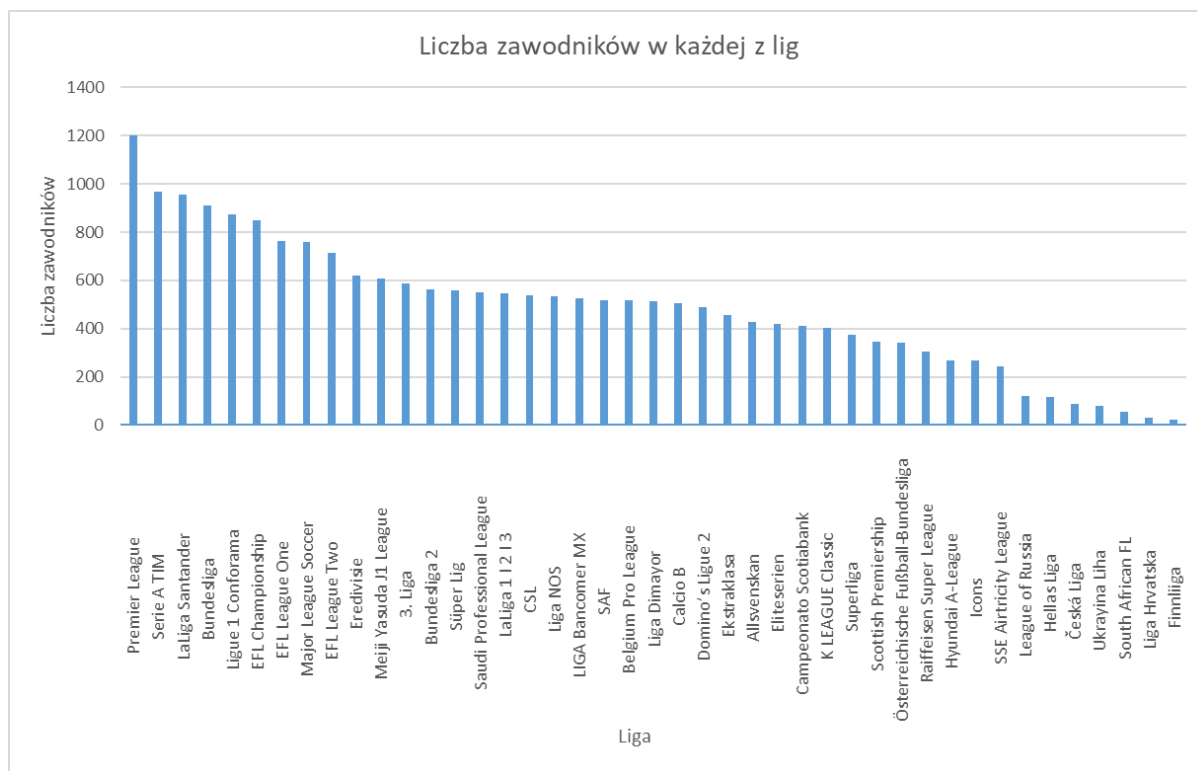
	Pozycje pomocników				
Statystyka	RM	LM	CM	CAM	CDM
Ratings	68,88	69,04	68,73	71,32	69,26
PAC	78,63	78,20	66,76	72,23	62,46
SHO	62,30	62,84	59,68	65,54	52,97
PAS	63,72	64,03	66,89	68,98	63,23
DRI	70,49	70,85	68,61	72,71	64,33
DEF	40,64	40,78	59,57	43,30	66,49
PHY	60,61	60,70	66,83	60,40	71,94
Body	176,73	176,86	179,10	176,80	180,93

	Pozycje obrońców				
Statystyka	CB	RB	LB	RWB	LWB
Ratings	69,41	67,74	67,67	70,93	70,89
PAC	59,13	73,86	73,45	77,79	77,47
SHO	38,13	46,36	47,11	52,38	52,31
PAS	50,39	59,56	60,02	64,24	64,01
DRI	52,48	64,37	64,69	68,74	68,37
DEF	68,75	64,91	64,72	66,34	65,97
PHY	73,10	68,70	67,76	71,16	70,54
Body	186,52	178,85	178,66	179,25	180,37

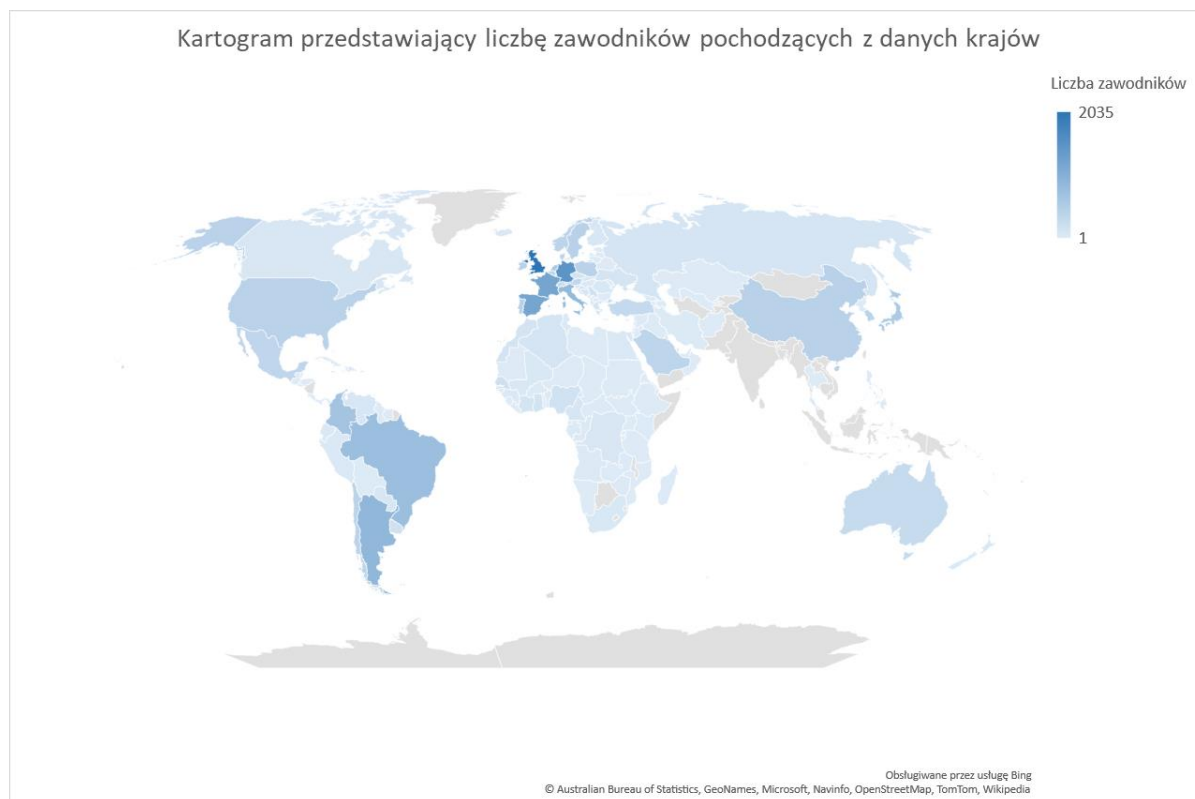
Wartości średnie potwierdzają wcześniej zauważone zależności pomiędzy statystykami a pozycjami na jakich występują zawodnicy. Warto również zwrócić uwagę na zależność pomiędzy wzrostem zawodnika (atrybut "Body") a pozycją, na której występuje. Można stwierdzić, że najczęściej zawodnicy niżsi występują po bokach boiska, a wyżsi w środku.

## Dodatkowe informacje o zbiorze

Poniżej znajduje się wykres przedstawiający liczbę zawodników w każdej z lig. Tak duże różnice pomiędzy najlepszymi ligami a tymi słabszymi spowodowane są możliwością występowania kilku różnych wersji karty zawodnika (w lepszych ligach występują zawodnicy, którzy najczęściej dostają specjalne karty). Są też ligi, z których liczba zawodników w zbiorze jest bardzo mała, jest to spowodowane brakiem licencji na ich pojawienie się w grze.



Na poniższym kartogramie możemy zauważyć rozłożenie narodowości zawodników znajdujących się w zbiorze. Powody różnic są bardzo podobne do opisanych powyżej dla tabeli przedstawiającej liczbę zawodników występujących w danych ligach. Warto zwrócić uwagę na obszary mające kolor szary (np. Indie), zbiór nie zawiera żadnych zawodników z tych krajów.



### Problemy z pierwotnym zbiorem

Pierwotny zbiór, na którego podstawie mieliśmy dokonać analizy okazał się być wadliwy. Zauważyliśmy to na podstawie zawodnika Liu Ruofan, którego statystyki nie pasowały do jego pozycji. Po porównaniu statystyk tego i kilku innych zawodników z ich oficjalnymi danymi z gry zdecydowaliśmy się na zmianę zbioru na taki, który jest zgodny z rzeczywistością.

Rodzaj danych	Name	Position	Ratings	PAC	SHO	PAS	DRI	DEF	PHY
Rzeczywiste	Ruofan Liu	CF	58	73	53	51	58	23	62
Błędne	Liu Ruofan	CF	58	71	29	41	48	55	66

Na powyższej tabeli można zauważyć wyraźne różnice w statystykach tego samego zawodnika z dwóch zbiorów. Podczas wstępnej analizy zwróciliśmy uwagę na bardzo niską statystykę strzałów “SHO” w porównaniu do statystyki obrony “DEF” dla zawodnika ofensywnego, u którego statystyka strzałów powinna znacznie przewyższać statystykę obrony.



## Wnioski

- Zawodnicy grający po bokach boiska mają inne atrybuty dominujące od tych, którzy grają w środku.
- Znaczne różnice w wielkościach atrybutów występują pomiędzy zawodnikami grającymi na różnej wysokości boiska (obrońcy, pomocnicy, napastnicy).
- Duże zróżnicowanie w ocenie ogólnej (atrybut "Ratings") powoduje duży rozstrzał wartości statystyk dla poszczególnych pozycji.
- Najbardziej ze wszystkich klas wyróżnia się pozycja CB.
- Wzrost zawodnika ma wpływ na jego umiejscowienie na boisku (boki lub środek).
- Niektóre z atrybutów są nieprzydatne w wyznaczaniu pozycji zawodnika.

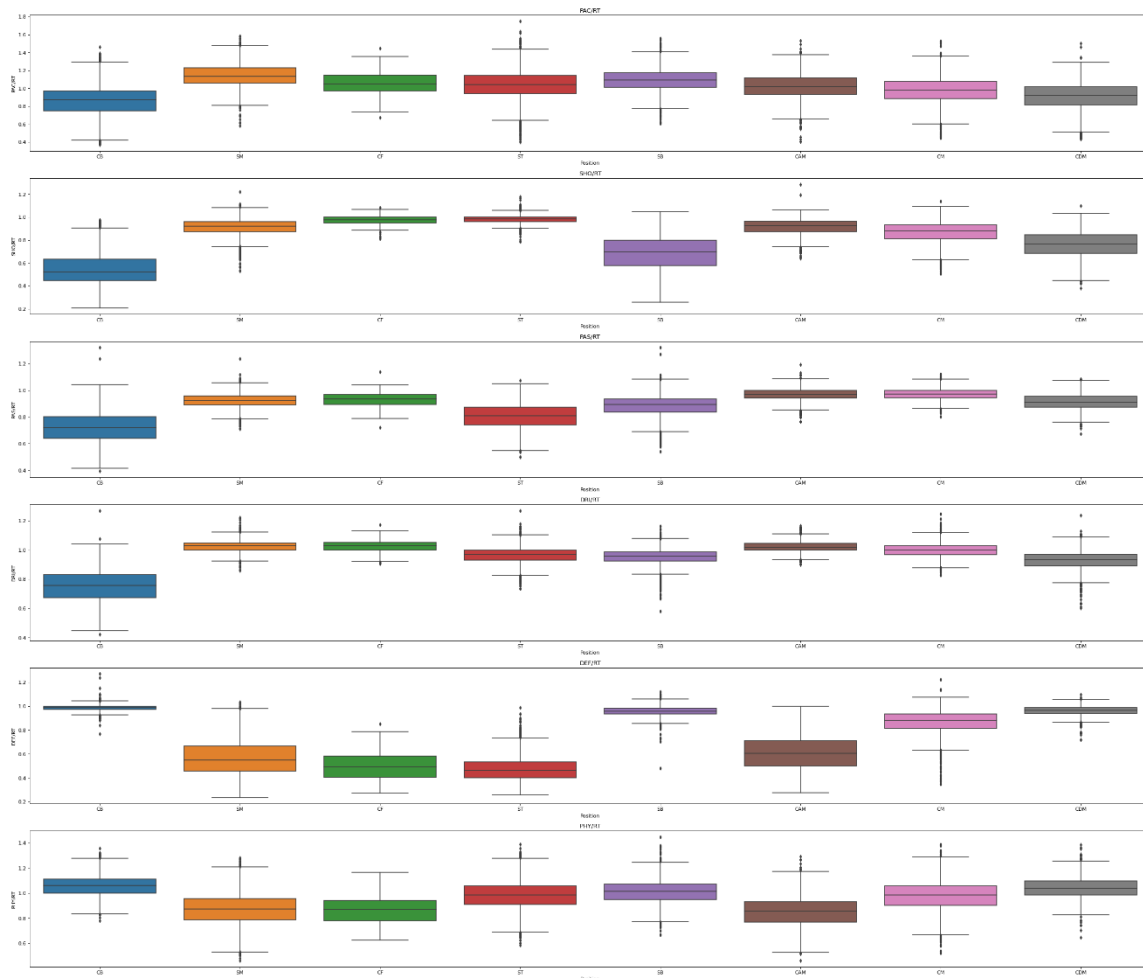
## Sposób rozwiązania projektu

### Wybrane cechy

Cechy, które wybraliśmy do dalszej analizy to: "PAC", "SHO", "PAS", "DRI", "DEF", "PHY", a konkretnie ich modyfikacje opisane w akapicie współczynniki pozycji. Będą one podstawowe w przyporządkowaniu pozycji. Są, to cechy opisujące umiejętności zawodników z każdej dziedziny związanej z grą w piłkę nożną. Zawodnicy występujący na różnych pozycjach, specjalizują się w innych elementach gry, co jest odwzorowane w wybranych cechach.

### Współczynniki pozycji

W wyniku analizy eksploracyjnej stwierdziliśmy, że w związku z różnicami w ocenie graczy problematyczne może się okazać przyporządkowywanie ich do pozycji w oparciu o bezwzględne wartości statystyk. Weźmy za przykład obrońcę światowej klasy, którego umiejętności gry obronnej są ocenione na najwyższym poziomie i zawodnika z polskiej ligi, którego umiejętności gry defensywnej osiągają wartość porównywalną z lepszymi pomocnikami. W takiej sytuacji odróżnienie gorszego obrońcy od pomocnika byłoby skomplikowane. Zauważyliśmy jednak korelację między oceną ogólną, a wartością kluczowych dla pozycji statystyk. Postanowiliśmy więc do zbioru dodać tzw. współczynniki pozycji, czyli stosunek każdej z 6 głównych cech do oceny ogólnej gracza (wartości w kolumnie "Rating"). Ma to na celu uwypuklić charakterystykę gracza, a co za tym idzie ułatwić przyporządkowanie go do pozycji.



Na wykresach pudełkowych przedstawiających rozkład współczynnika dla konkretnych pozycji, widzimy spłaszczenie względem rozkładu niezmodyfikowanych cech, co uwypukla cechy kluczowe dla danej pozycji.

### Wspólne kategorie dla zawodników grających po bokach

Zdecydowaliśmy się na zastosowanie wspólnych kategorii dla zawodników grających w bocznych sektorach boiska. Wynika to z tego, że w analizowanym zbiorze danych nie dysponujemy cechami pozwalającymi odróżnić zawodników grających po lewej stronie od tych grających po prawej. Charakterystyka zawodników grających po obu stronach boiska jest praktycznie identyczna. W analizie będziemy posługiwać kategoriami: boczny pomocnik dla zawodników z klas LW, RW, LM, RM; oraz boczny obrońca dla zawodników z klas LB, RB, LWB, RWB.

### Użyty algorytm

Do realizacji zadania wybraliśmy klasyfikator k- najbliższych sąsiadów. Klasyfikator k- najbliższych sąsiadów - rozwiązanie z wykorzystaniem uczenia nadzorowanego. Polega na przyporządkowaniu do grupy w oparciu o przyporządkowanie k najbliższych, względem rozpatrywanego punktu, punktów. Zdecydowaliśmy się na ten algorytm, bo uważamy, że zawodnicy grający na tych samych pozycjach mają podobną charakterystykę. Klasyfikator kNN właśnie na takim podobieństwie bazuje.

### Ogólny opis działania programu

Nasz program możemy podzielić na dwie części: analiza eksploracyjna oraz część odpowiedzialna za model.

## Analiza eksploracyjna

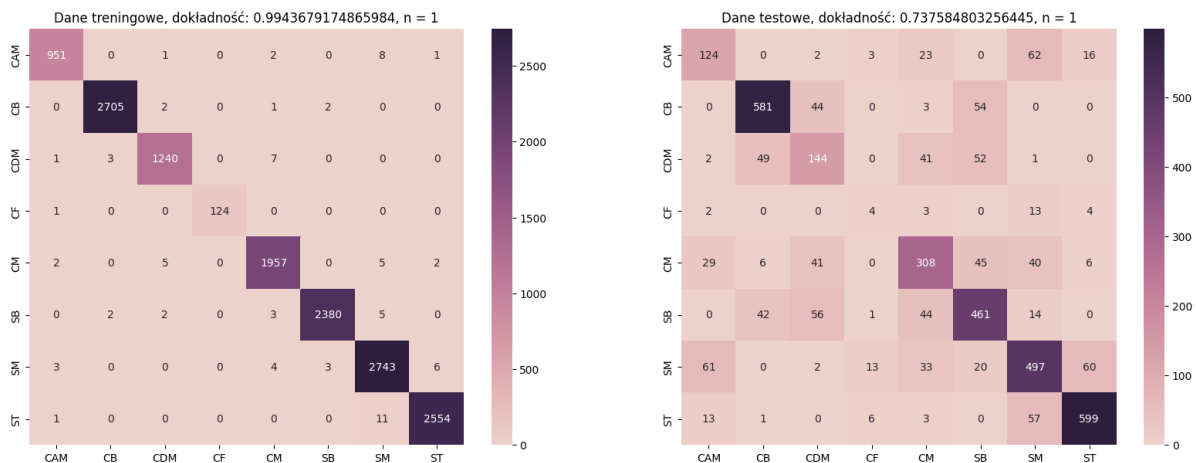
Program rozpoczyna się od zaimportowania bibliotek potrzebnych do przeprowadzenia analizy i wczytania danych z dwóch zbiorów (poprawnego i błędnego). Następnie porównujemy dane jednego z zawodników w celu pokazania problemów z jednym ze zbiorów. Kolejnym krokiem jest wyświetlenie atrybutów zawartych w zbiorze. Wyświetlamy wykresy pudełkowe dla konkretnych statystyk i pozycji. Zliczamy zawodników występujących na każdej z pozycji i wyświetlamy histogram liczebności pozycji. Usuujemy ze zbioru niepotrzebne kolumny i problematyczne obiekty. Następnie wyświetlamy wartości średnie dla całego zbioru i poszczególnych pozycji oraz mediany z atrybutów dla całego zbioru. Na koniec analizy eksploracyjnej wyświetlamy liczebność zawodników w zależności od ligi, w której występują oraz państwa, z którego pochodzą.

## Model

Na początku części poświęconej na model dokonujemy zmian w zbiorze danych. Pozbywamy się skrajnych centyli, wyliczamy współczynnik pozycji oraz zostawiamy tylko te kolumny, które będą nam potrzebne w przewidywaniu pozycji. Decydujemy się na zostawienie tylko kolumn ze współczynnikami pozycji oraz kolumny z etykietami. Na końcu tworzenia zbioru dla modelu ujednolicamy pozycje po bokach boiska. W dalszej części generujemy wykresy, żeby ocenić wpływ współczynników pozycji na rozróżnialność pomiędzy pozycjami. Kolejnym krokiem jest zaimportowanie niezbędnych do przeprowadzenia klasyfikacji kNN bibliotek. Na tym etapie przechodzimy do tworzenia modelu. Definiujemy funkcję, której zadaniem jest trening i ocena jakości modelu a także, jeżeli użytkownik potrzebuje, wyświetlanie macierzy pomyłek. Następnie definiujemy funkcję mającą wyświetlać wykresy dokładności. Przechodzimy do testowania klasyfikatora dla różnych wartości współczynnika podziału oraz liczby najbliższych sąsiadów. Sprawdzamy dokładność klasyfikatora dla konkretnego przypadku split = 0.1 i n=1. Na końcu sprawdzamy, jak model poradzi sobie gdybyśmy nie ujednoliciли pozycji po bokach boiska.

## Wyniki

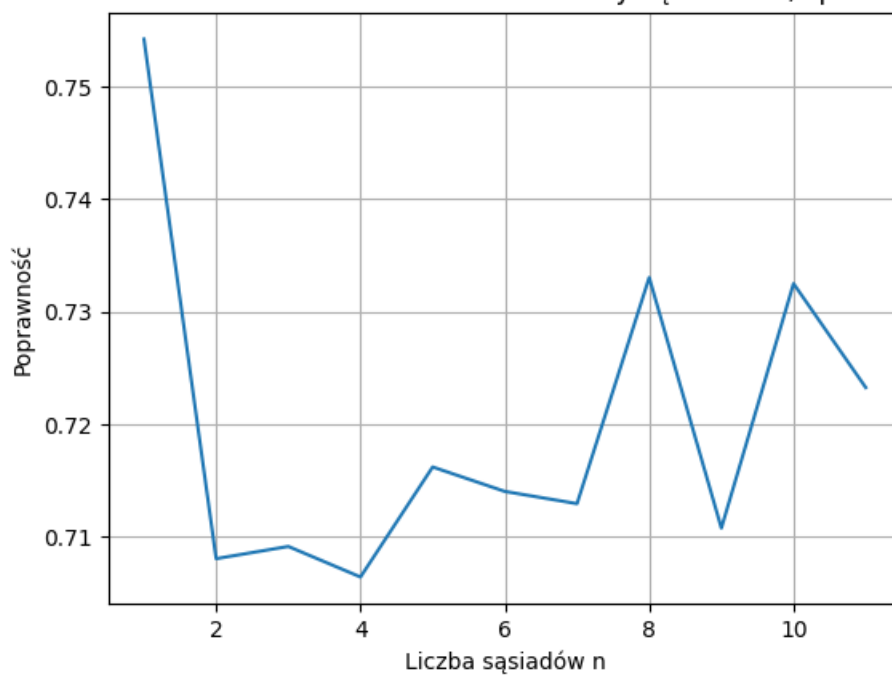
### Macierz pomyłek



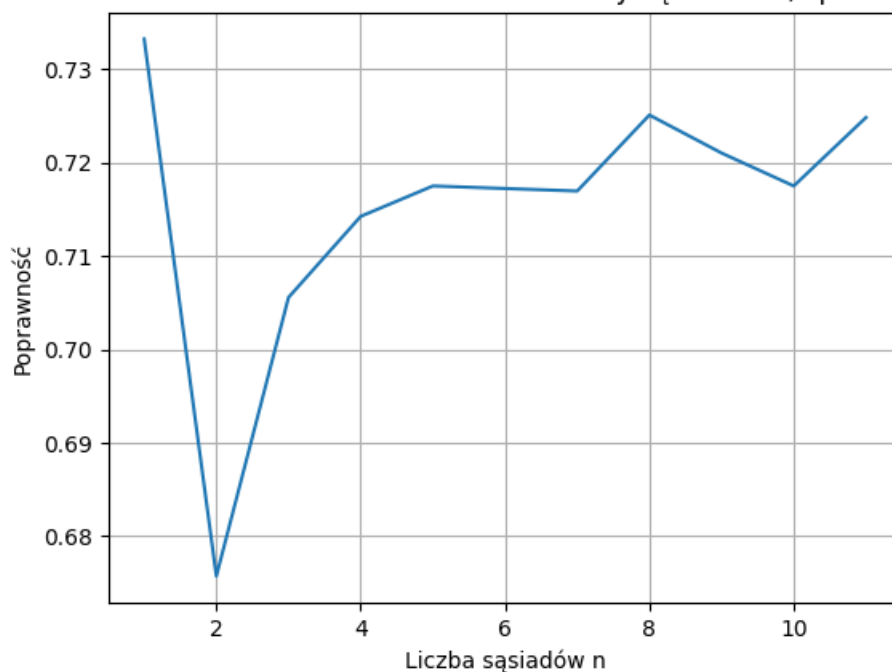
Na powyższym zdjęciu pokazana jest macierz pomyłek dla  $n = 1$  i współczynnika podziału = 0.2. Są to wartości, dla których model uzyskuje dobre rezultaty. Wybór akurat takiej wartości został uzasadniony w kolejnych akapitach. W macierzy można zauważyć, że niektóre pozycje są łatwiejsze do zakwalifikowania. Są to np. CB (środkowy obrońca) i ST (środkowy napastnik). Widać też, że część pozycji jest problematyczna dla modelu. Tutaj negatywnie wyróżnia się CF (cofnięty napastnik), gdzie liczba błędnych przypisań kilkukrotnie przewyższa liczbę właściwych. Może to wynikać, z małej liczby przykładów w zbiorze oraz podobieństwa do innych ofensywnych pozycji.

Dokładność modelu ze względu na podział na zbiór treningowy i testowy oraz liczbę  $n$

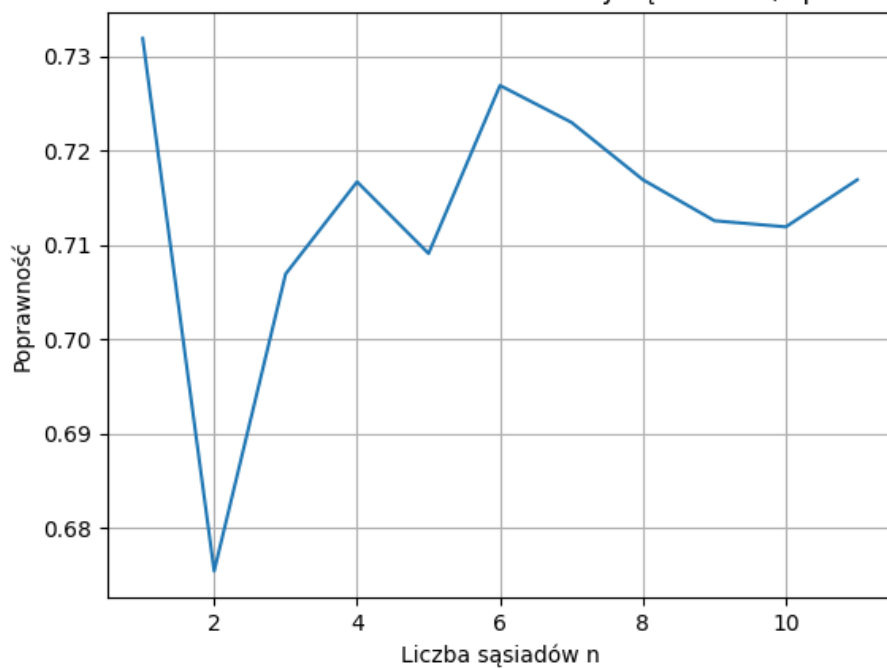
Dokładność modelu w zależności od liczby sąsiadów  $n$ , split = 0.1



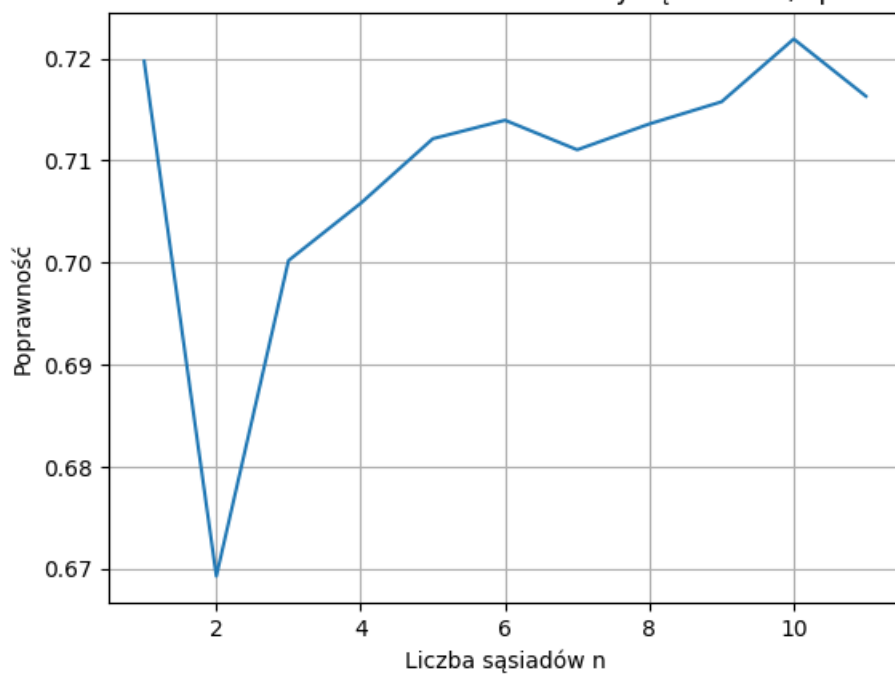
Dokładność modelu w zależności od liczby sąsiadów  $n$ , split = 0.2

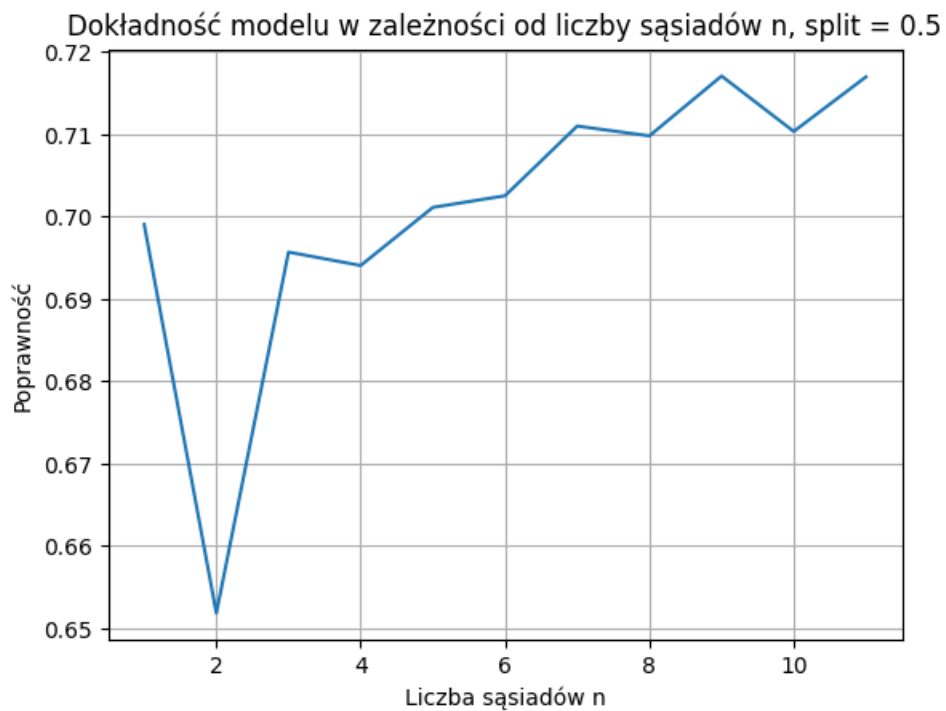
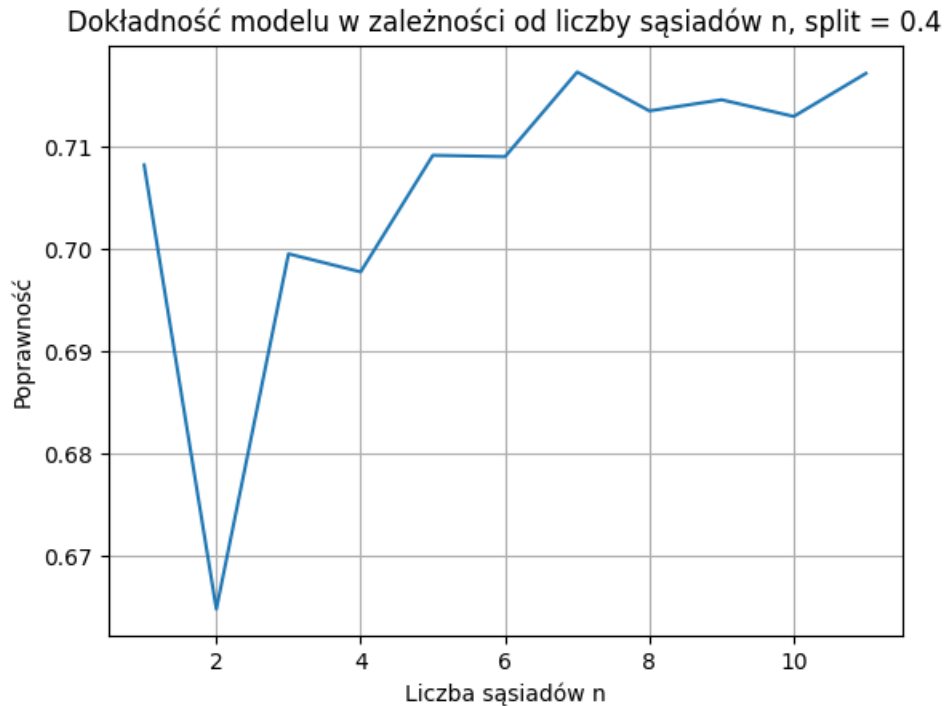


Dokładność modelu w zależności od liczby sąsiadów  $n$ , split = 0.25



Dokładność modelu w zależności od liczby sąsiadów  $n$ , split = 0.3





Na wykresach widać, że bez względu na wartość współczynnika podziału na zbiór treningowy i testowy, model uzyskuje bardzo dobre wyniki dla  $n = 1$ . Inne powtarzające się tendencje, to słabsze wyniki dla parzystych wartości  $n$  oraz lepsze dla nieparzystych  $n$ .

Zaskakująco dobre wyniki dla split = 0.1 i  $n = 1$

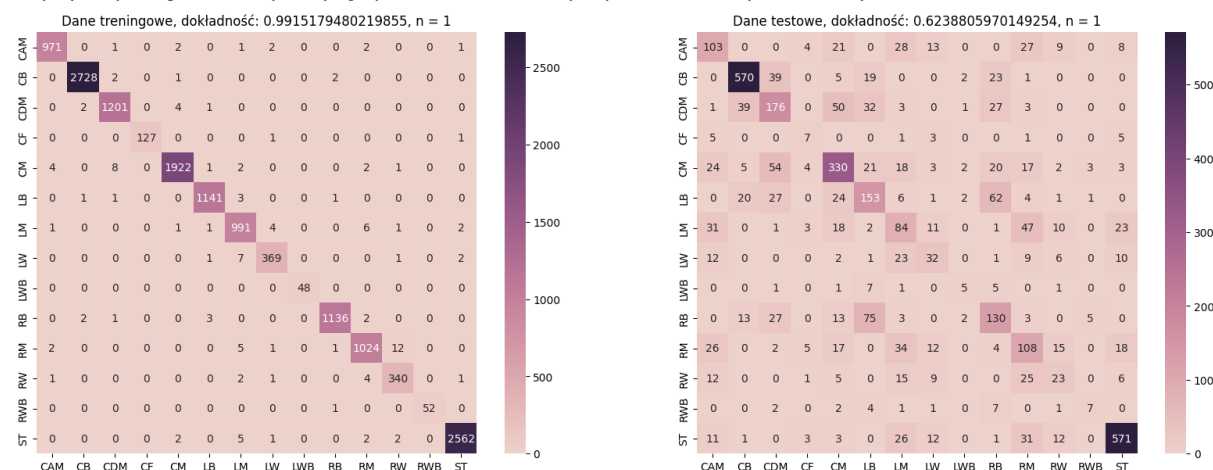
Co ciekawe model najlepszą poprawność przewidywania uzyskał dla współczynnika podziału=0.1 i  $n=1$ . Jest to zastanawiające, ponieważ taka wartość współczynnika podziału często nie daje zadowalających rezultatów. Prawdopodobnie, mamy tutaj do czynienia z nadmiernym dopasowaniem modelu do

zbioru danych. Model w tej sytuacji trenuje na większej liczbie przykładów, co ułatwia mu identyfikację danych testowych z tego zbioru.

Małe różnice w poprawności, mimo różnych współczynników podziału.

Można zaobserwować, że różnice w poprawności są niewielkie, pomimo że różnice w wielkości zbioru treningowego, są duże. Może to świadczyć o dużym podobieństwie między danymi w zbiorze.

### Wpływ połączenia pozycji po bokach na poprawność przewidywania



Po wytrenowaniu modelu i wykonaniu przewidywania z jego pomocą, widać, że dla zbioru bez połączonych pozycji liczba poprawnych przyporządkowań spadła o około 12 punktów procentowych. W macierzy pomyłek widać, że dla pozycji z boku boiska występuje najwięcej nieprawidłowych przyporządkowań. Dla każdej z pozycji z boku boiska, błędy wynikające z przyporządkowania do pozycji podobnej stanowią niemal tyle co poprawne przyporządkowania.

### Najlepsza wartość n

Po analizie otrzymanych wykresów dokładności ciężko jest wskazać bezwzględnie najlepszą wartość. Uważamy, że dla tego zbioru najrozsądniejszym wyborem jest  $n = 1$ . Dla  $n = 1$ , liczba poprawnych przyporządkowań, bez względu na podział zbioru, wynosiła około 70% (dla bardziej standardowych wartości podziału tj. z przedziału 0.2 - 0.3 liczba poprawnych przyporządkowań wahała się w okolicy 72%). Dla innych nieparzystych wartości liczba poprawnie dobranych pozycji, wahała się mocno w zależności od podziału zbioru.

## Wnioski końcowe

### Wnioski finalne

#### Napastnicy też bronią, obrońcy też strzelają

W obecnych czasach coraz bardziej pożądanymi są zawodnicy ogólnie uzdolnieni. W związku z tym, coraz większa ich liczba decyduje się na rozwój wielu umiejętności na raz, a nie tak jak kiedyś, specjalizacja w jednej konkretnej. Widać to w zbiorze danych i w wynikach. Przed rozpoczęciem projektu założyliśmy, że będziemy zadowoleni, jeżeli poprawność modelu osiągnie ok. 70%. W związku z powyższym jesteśmy usatysfakcjonowani uzyskanym wynikiem.

#### Założenia początkowe, a wyniki

Większość poczynionych przez nas na początku założeń sprawdziła się. Wyniki pokazały, że dysponując jedynie umiejętnościami nie jesteśmy w stanie z zadowalającą poprawnością odróżnić zawodników grających po lewej stronie od tych grających po prawej. Podobna sytuacja wystąpiła w przypadku zawodników grających na pozycji (CAM), których cechy okazały się podobne do cech pomocników

grających na bokach boiska i w mniejszym stopniu na pozycji (CM). Kolejnym przykładem pozycji niejednoznacznej na podstawie cech jest (CDM), ponieważ pozycje obrońców okazały się mieć podobne cechy do niej oraz, jak w przypadku pozycji (CAM), środkowi pomocnicy (CM) cechują się podobnymi statystykami. Założenie o podobnych cechach zawodników grających na tej samej pozycji sprawdziło się tylko po części. Zawodnicy są podobni, natomiast, z łatwością możemy znaleźć wielu zawodników grających na różnych pozycjach, a mimo to mających podobny zestaw cech. Świadczy o tym poprawność modelu dla  $n = 1$  i małe różnice w poprawności dla większych wartości  $n$ , gdzie jednak powinniśmy się spodziewać większej poprawności.

#### Wyróżniasz się na wykresie, łatwiej cię przyporządkować

Model najmniej błędów popełniał dla pozycji, które na największej liczbie wykresów pudełkowych wyróżniały się od innych. Są to m. in. napastnik (ST) i środkowy obrońca (CB). Już na etapie analizy eksploracyjnej łatwo można było zauważyć ich różnice odpowiednio w atrybutach strzałów (SHO) i obrony (DEF). Poskutkowało to bardzo małą (nieprzekraczającą kilku procent) liczbą błędnych przyporządkowań.

#### Czy model nadaje się do użytku?

Poprawność przypisania do pozycji na poziomie 70% dyskwalifikuje ten model z poważnego przewidywania. Model może natomiast sprawdzić się dobrze w sytuacji, kiedy chcemy określić, czy zawodnik jest ukierunkowany bardziej ofensywnie, czy defensywnie. Błędy w przyporządkowaniu, polegały głównie na przypisaniu do pozycji pokrewnych (np. napastnik i cofnięty napastnik). Nie były to losowe wybory, jak np. oznaczenie obrońcy jako napastnika. Z tych powodów model może sprawdzić się, jeżeli ktoś chce sobie wyrobić intuicję na temat ogółu kart w grze.

#### Źródła

<https://www.kaggle.com/datasets/mohammedessam97/fifa-1020-fut-players-dataset?select=Fifa+19+Fut+Players.csv>