

Studium przypadku

8 września 2020

Autor **Franciszek Sapikowski** f.sapikowski@gmail.com

Oświadczam, że niniejsze sprawozdanie zostało przygotowane wyłącznie przez powyższego autora, a wszystkie elementy pochodzące z innych źródeł zostały odpowiednio zaznaczone i są cytowane w bibliografii.

# 1 Wstęp

Poniższe sprawozdanie przedstawia realizację zadania, które polegało na wyborze zbioru danych, jego analizie, wykryciu ciekawych zależności, wyciągnięciu wniosków oraz na wykorzystaniu trzech algorytmów uczenia maszynowego. Zadanie podzieliłem na dwa etapy: analizę zbioru danych (rozdział 2) oraz wykorzystanie algorytmów uczenia maszynowego (rozdział 3).

## 2 Analiza zbioru danych

Wybranym przez mnie zbiorem jest opis zawodników z gry FIFA [1]. Dane zawierają specyfikację ponad 18 tysięcy zawodników, którzy są dostępni w grze FIFA 19. Każdy piłkarz jest opisany przez szereg atrybutów, które można podzielić na kategorie:

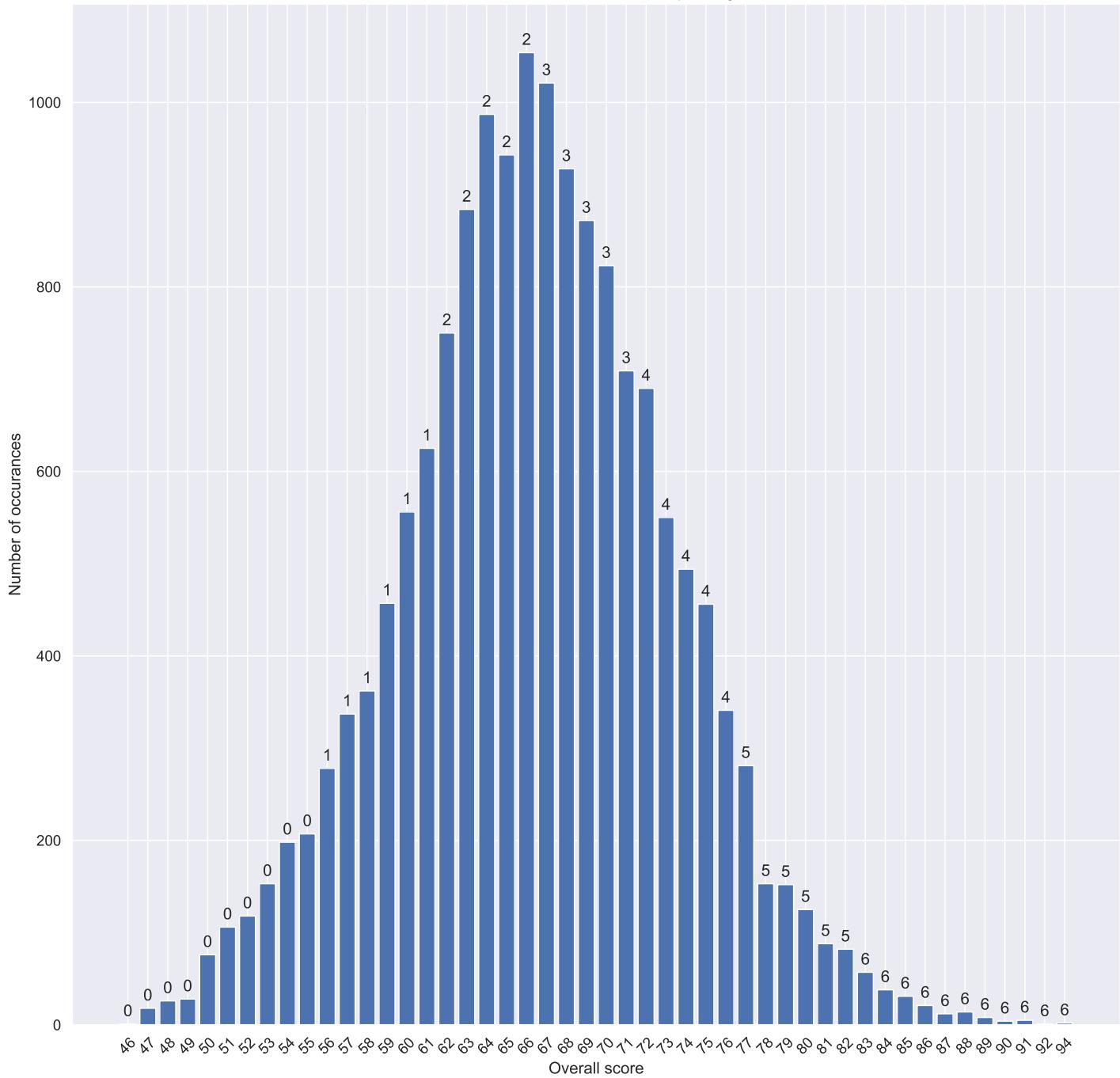
- atrybuty identyfikacyjne np. imię i nazwisko, kraj pochodzenia, klub
- atrybuty zarobkowe np. wartość zawodnika, zarobki
- atrybuty opisujące umiejętności oraz fizjologię
- pozostałe atrybuty takie jak: numer na koszulce, data wygaśnięcia kontraktu, informacja czy w grze została użyta prawdziwa twarz zawodnika, pozycja

Interesujące nas atrybuty to te opisujące umiejętności i fizjologię graczy, gdyż to one definiują jak dobry jest dany gracz. Jest ich 69. Są to: wiek, wzrost, waga, ocena ogólna, potencjał, reputacja, umiejętność gry słabszą nogą, triki, 26 cech opisujących jak dany gracz radzi sobie na konkretnej pozycji na boisku (wszystkie są opisane tutaj [2]) oraz 35 umiejętności piłkarskich np. drybling, reakcja, balans, strzelanie głową, z których 5 należy tylko do oceny umiejętności bramkarskich. Dwie ostatnie grupy oraz ocena ogólna i potencjał opisane są w skali interwałowej od 10 do 100. Reputacja i gra słabszą nogą w skali od 1 do 5. Ocena ogólna, po odpowiedniej dyskretyzacji, będzie wykorzystywana jako klasa dla przyszłych modeli klasyfikacji i nie będzie brana pod uwagę w analizie.

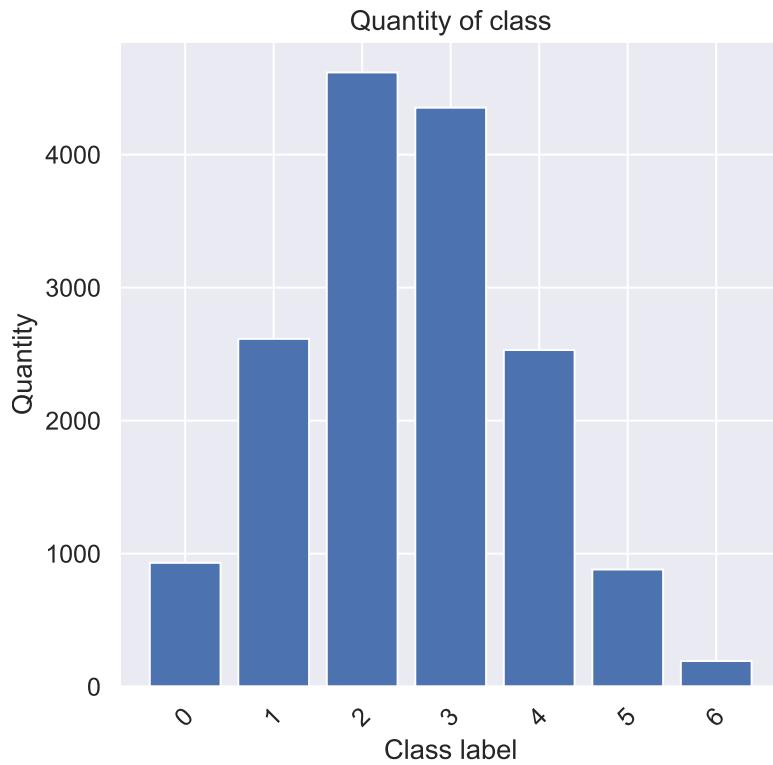
Pierwszym krokiem było usunięcie wartości *NaN* ze zbioru danych. Jak się okazało bramkarze nie mają ocen na atrybutach związanych z pozycjami. Wychodzi na to, że w FIFA nie można brać bramkarzem np. na pozycji napastnika jednak można grać napastnikiem na pozycji bramkarza. Ponadto, niektórzy piłkarze, niebędący bramkarzami, również nie mieli ocen na niektórych pozycjach. Oddzieliłem zatem bramkarzy od reszty zawodników i usuваłem ze zbioru tych piłkarzy, którym brakowało pewnych ocen umiejętności. W rezultacie uzyskałem 2025 przykładów bramkarzy oraz 16122 przekłady „piłkarzy-niebramkarzy”. Zbiór ten po wydzieleniu kolumny oceny ogólnej stał się zbiorem uczącym, który poddany został dalszej analizie.

Przed samą analizą musiałem podać dyskretyzacji ocenę ogólną piłkarzy. Skorzystałem z funkcji *KBinsDiscretizer* z parametrami *n\_bins = 7, encode = 'ordinal', strategy = 'kmeans'*. W wyniku działania algorytmu uzyskałem 7 klas decyzyjnych. Dystrybucja klas przedstawiona została na wykresach 1, 2.

Overall score distribution with corresponding class

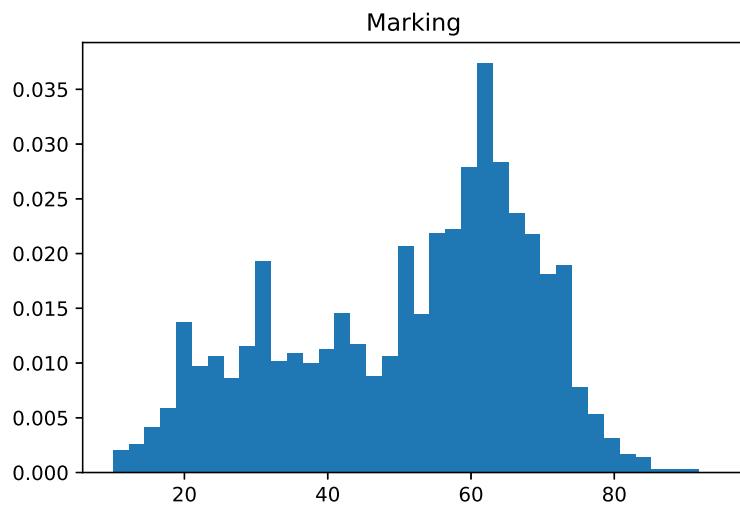


Wykres 1: Wykres rozkładu oceny ogólnej piłkarzy wraz z zaznaczoną klasą decyzyjną.

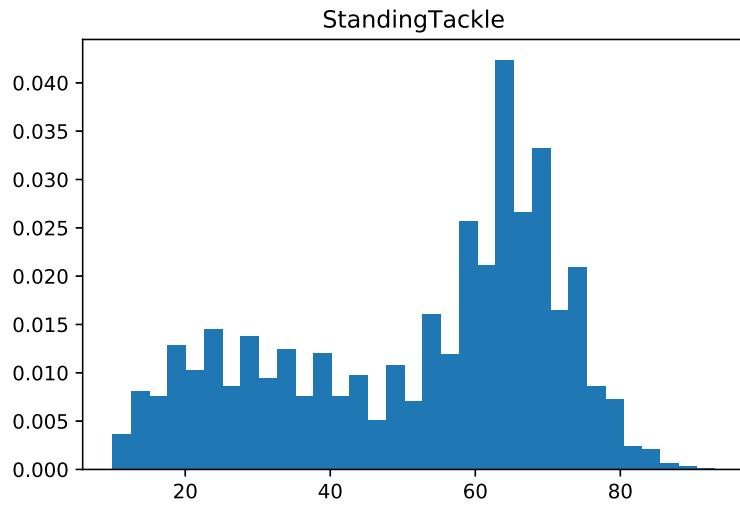


Wykres 2: Wykres rozkładu klas decyzyjnych.

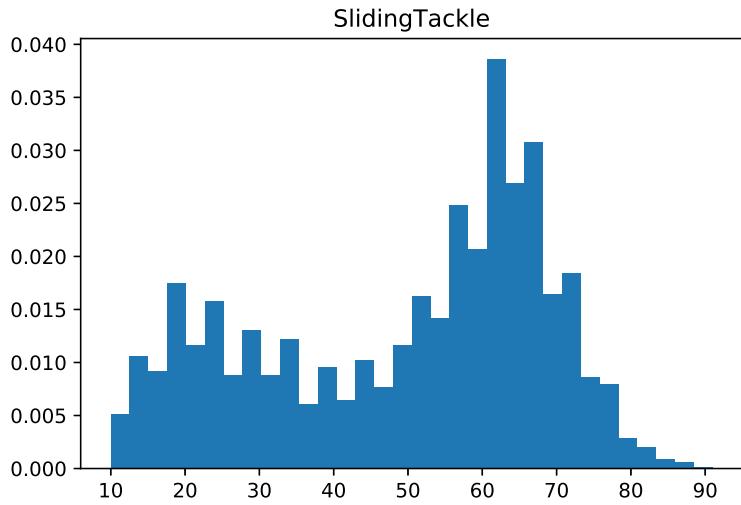
Następnie przedstawiłem na wykresach rozkład wartości dla kolejnych atrybutów. Prawie wszystkie rozkłady przypominają krzywe gaussowskie. Mówiąc ogólnie o większości atrybutów, to najczęściej jest średnich zawodników. Drugie kwartyle zazwyczaj są bardzo bliskie średniej, a kwartyle pierwszy i trzeci mieszczą się nie dalej od średniej niż wartość odchylenia standardowego. Wyjątkowymi atrybutami są umiejętności osłaniania gracza (marking), odbiór piłki na stojąco (standing tackle) i odbiór piłki wslizgiem (sliding tackle). Rozkład jest dość równy, płaski dla wartości od 10 do 60. Dla tych umiejętności trzeci kwartyl wynosi kolejno 65, 67, 65., gdzie średnia wynosi odpowiednio 51, 51, 49. Oznacza to, że 75% graczy ma ocenę tej umiejętności niższą niż podana wartość. Może to wskazywać na trudność w doskonaleniu tych umiejętności i prawdopodobnie niewielu zawodników osiąga perfekcję w tej dziedzinie. Wykresy rozkładów wspomnianych umiejętności przedstawiam na wykresach 3, 4, 5.



Wykres 3: Wykres rozkładu oceny umiejętności osłaniania pośród piłkarzy w grze FIFA.

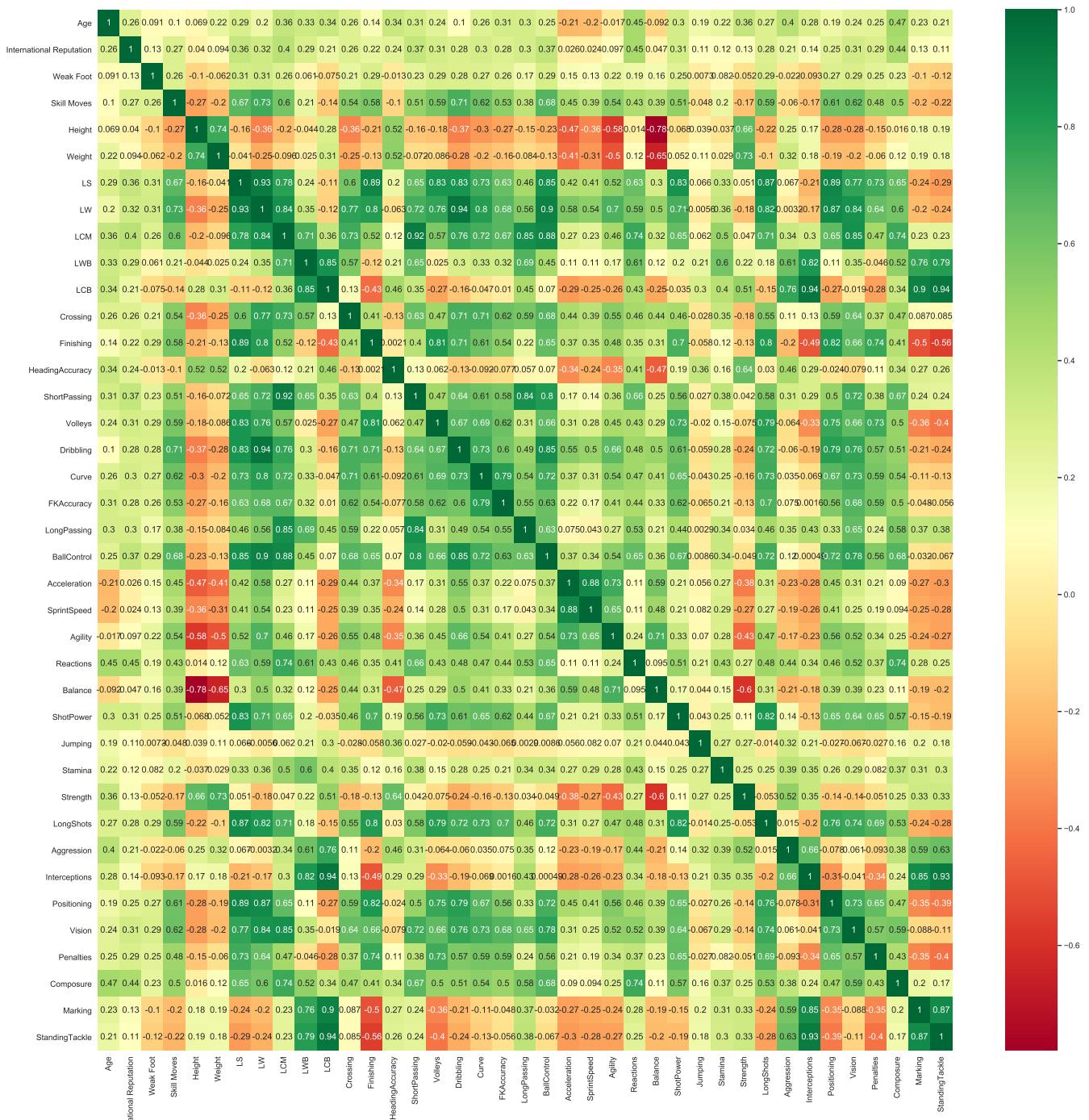


Wykres 4: Wykres rozkładu oceny umiejętności odbioru piłki stojąc pośród piłkarzy w grze FIFA.



Wykres 5: Wykres rozkładu oceny umiejętności odbioru piłki w ślimakiem pośród piłkarzy w grze FIFA.

Następnym krokiem była analiza korelacji. Niektóre atrybuty są silnie ze sobą skorelowane. Wynika to z tego, że niektóre pozycje piłkarskie, pomimo innej funkcji na boisku, wymagają podobnych umiejętności. Ponadto, istnieje pewne duże skorelowanie pomiędzy ocenami pozycji a poszczególnymi umiejętnościami np. ocena lewego napastnika LS-left striker, silnie koreluje z umiejętnością kontroli piłki co jest całkiem logiczne, gdyż napastnik musi dobrze kontrolować piłkę. Dla próby usunąłem atrybuty silnie skorelowane (posiadające wartość bezwzględną współczynnika korelacji większą niż 0.95), aby sprawdzić później jak algorytmy poradzą sobie na zmniejszonych danych. Macierz przedstawiam na wykresie 6.



W dalszym ciągu zająłem się znalezieniem zbioru najważniejszych atrybutów w zbiorze wszystkich atrybutów. Wykorzystałem algorytm *SelectKBest* z testem *chiwadrat*, algorytm rekursywny z liniowym klasyfikatorem *SVC*, podejście embedded z klasyfikatorami *LogisticRegression* oraz *RandomForest*. Algorytmy uruchomiłem na całym zbiorze przykładów oraz na zbiorze z usuniętymi atrybutami silnie ze sobą skorelowanymi. Następnie utworzyłem ranking 30 najbardziej znaczących cech.

Feature	Chi-2	RFE	Logistics	Random Forest	Total
ST	True	True	True	True	4
Reactions	True	True	True	True	4
RS	True	True	True	True	4
LS	True	True	True	True	4
Skill Moves	True	True	True	False	3
RWB	True	True	False	True	3
RDM	True	True	False	True	3
RB	True	True	False	True	3
LongShots	True	True	True	False	3
LWB	True	True	False	True	3
LDM	True	True	False	True	3
LB	True	True	False	True	3
International Reputation	True	True	True	False	3
Composure	True	True	True	False	3
CDM	True	True	False	True	3
BallControl	True	True	True	False	3
Age	True	True	True	False	3
Vision	False	True	True	False	2
StandingTackle	False	True	False	True	2
ShortPassing	False	True	True	False	2
RM	True	False	False	True	2
RF	True	False	False	True	2
RCB	False	True	False	True	2
RAM	True	False	False	True	2
Positioning	False	True	True	False	2
LongPassing	True	True	False	False	2
LM	True	False	False	True	2
LF	True	False	False	True	2
LCB	False	True	False	True	2
LAM	True	False	False	True	2

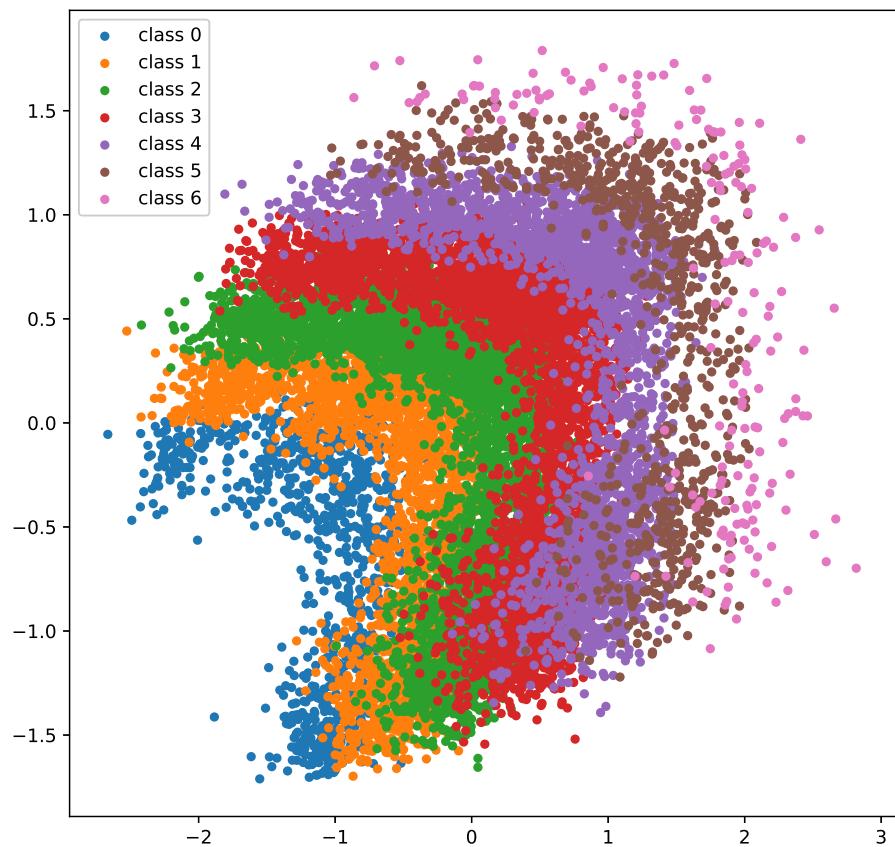
Tabela 1: Ranking 30 najważniejszych atrybutów wyznaczonych z całego zbioru danych. Wartość *True* w odpowiedniej kolumnie oznacza, że dany atrybut znalazł się w zbiorze 30 najważniejszych cech wyznaczonych przez konkretny algorytm.

Feature	Chi-2	RFE	Logistics	Random Forest	Total
ShortPassing	True	True	True	True	4
Reactions	True	True	True	True	4
LWB	True	True	True	True	4
LS	True	True	True	True	4
LCM	True	True	True	True	4
LCB	True	True	True	True	4
HeadingAccuracy	True	True	True	True	4
Composure	True	True	True	True	4
BallControl	True	True	True	True	4
StandingTackle	True	True	False	True	3
Skill Moves	True	True	True	False	3
Marking	True	True	False	True	3
LW	True	True	False	True	3
International Reputation	True	True	True	False	3
Dribbling	True	True	False	True	3
Age	True	True	True	False	3
Volleys	True	True	False	False	2
Vision	True	True	False	False	2
Stamina	True	True	False	False	2
ShotPower	True	True	False	False	2
Positioning	True	True	False	False	2
LongShots	True	True	False	False	2
LongPassing	True	True	False	False	2
Interceptions	True	False	False	True	2
Finishing	True	True	False	False	2
Crossing	True	True	False	False	2
Weight	False	True	False	False	1
Strength	False	True	False	False	1
SprintSpeed	False	True	False	False	1
Penalties	True	False	False	False	1

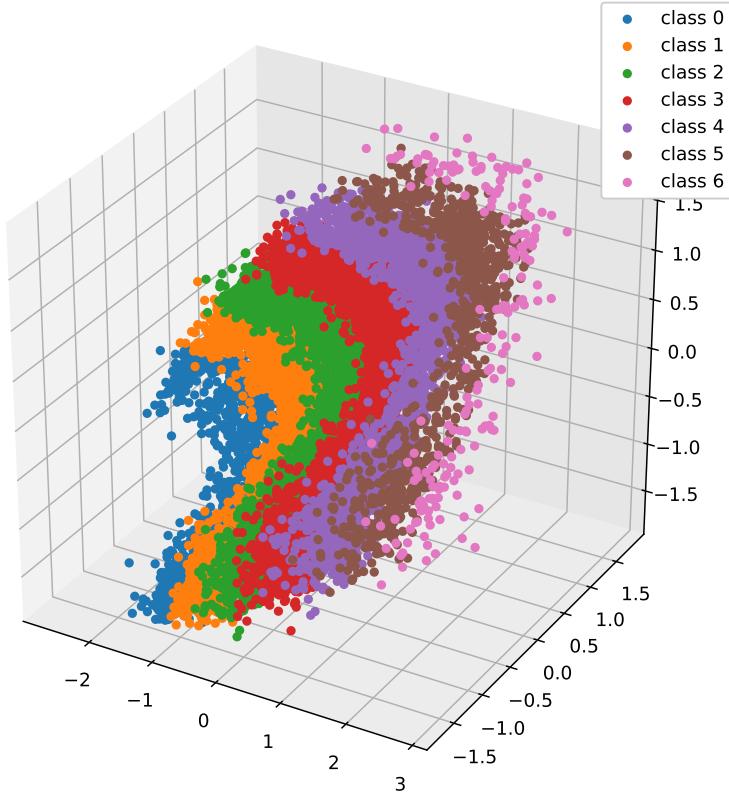
Tabela 2: Ranking 30 najważniejszych atrybutów wyznaczonych z redukowanego zbioru danych. Wartość *True* w odpowiedniej kolumnie oznacza, że dany atrybut znalazł się w zbiorze 30 najważniejszych cech wyznaczonych przez konkretny algorytm.

Różnice w rankingu wynikają z faktu, że pewne znaczące cechy w pierwszym rankingu są silnie ze sobą skorelowane więc mają, z punktu patrzenia algorytmów, wysokie znaczenie, pomimo że mogą dawać podobne informacje. Jednak widać pewne cechy, które widnieją w obu rankingach. Są to: reakcje, triki (skill moves), długie podania, oceny na pozycjach: lewy skrzydłowy z powrotem (LWB), lewy centralny z powrotem (LCB), lewy napastnik (LS). Wyniosek jaki nasuwa się po tej analizie to, że nie ma jednej kluczowej umiejętności, która definiuje najlepszego piłkarza. Na dobrego piłkarza wpływ wiele składowych.

Ostatnim etapem było wykorzystanie algorytmu *PCA*. W wyniku jego zastosowania uzyskaliśmy wykresy dla 2 i 3 składowych (odpowiednio wykresy 7 i 8)



Wykres 7: Wizualizacja zbioru danych w 2D dla osi posiadających największą wariancję.



Wykres 8: Wizualizacja zbioru danych w 3D dla osi posiadających największą wariancję.

Na wykresach 7 i 8 widzimy, że klasy nie są separowalne liniowo jednak da się rozróżnić konkretne klasy. Część przykładów przenika od innych klas co wpłynie negatywnie na jakość klasyfikacji. Dla 2 osi suma wariancji przez nie opisywanych wyniosła 73.34%, dla 3 osi: 79.83%. Widzimy więc, że utraciliśmy część informacji. Potwierdza to tezę, że do opisu piłkarza wymaganych jest wiele cech.

### 3 Wykorzystanie algorytmów uczenia maszynowego

Eksperymenty podzieliłem na pięć etapów:

- etap 1: sprawdzenie trafności klasyfikacji dla kilku podstawowych algorytmów z domyślnymi parametrami wraz ze standaryzacją danych, z 10-krotną stratified CV,

- etap 2: sprawdzenie trafności klasyfikacji dla kilku podstawowych algorytmów z domyślnymi parametrami wraz ze standaryzacją danych oraz przekształceniem atrybutów algorytmem *PCA* z parametrem  $n\_components = 15$ , z 10-krotną stratified CV,
- etap 3: sprawdzenie trafności klasyfikacji dla kilku podstawowych algorytmów z domyślnymi parametrami wraz ze standaryzacją zredukowanych danych, z 10-krotną stratified CV,
- etap 4: dla trzech najlepiej spisujących się klasyfikatorów poszukanie najlepszych parametrów za pomocą algorytmu *GridSearch* wraz ze standaryzacją danych, z 10-krotną stratified CV,
- etap 5: uczenie i testowanie najlepszych modeli wraz ze standaryzacją danych odpowiednio na niezredukowanym zbiorze uczącym i testowym (podzielonym w proporcji 8 : 2).

Wykorzystane przeze mnie algorytmy to:

- *DecisionTreeClassifier*,
- *QuadraticDiscriminantAnalysis*
- *SVC*, podejście *OneVsRest*,
- *SVC*, podejście *OneVsOne*,
- *KNeighborsClassifier*
- *GaussianNB*.

Użyty algorytm standaryzacji: *MinMaxScaler*. Mógł zostać również użyty algorytm *StandardScaler* ale ze względu na dziwny rozkład atrybutów wymienionych w sekcji 2 pozostał przy *MinMaxScaler*. Ze względu na brak przykładów odstających nie użyłem *RobustScaler*.

Wyniki pierwszych trzech etapów przedstawiono w tabeli 3.

Algorytm	Metryka	Etap 1	Etap 2	Etap 3
<i>DecisionTreeClassifier</i>	Accuracy	0.802	0.678	0.782
	G-mean	0.776	0.676	0.764
<i>QuadraticDiscriminantAnalysis</i>	Accuracy	0.444	0.61	0.558
	G-mean	0.39	0.6	0.51
<i>SVC, OneVsRest</i>	Accuracy	0.769	0.784	0.747
	G-mean	0.612	0.647	0.544
<i>SVC, OneVsOne</i>	Accuracy	0.759	0.769	0.811
	G-mean	0.821	0.711	0.737
<i>KNeighborsClassifier</i>	Accuracy	0.736	0.745	0.685
	G-mean	0.643	0.643	0.588
<i>GaussianNB</i>	Accuracy	0.501	0.46	0.545
	G-mean	0.559	0.56	0.596

Tabela 3: Wyniki klasyfikacji dla pierwszych trzech etapów.

Najlepszym klasyfikatorem okazały się drzewa decyzyjne i klasyfikator *SVC*. Dobre rezultaty, po zastosowaniu algorytmu PCA, dał również klasyfikator *KNN*. Właśnie dla tych klasyfikatorów zastosowałem dostrajanie parametrów (dla klasyfikatora *SVC* wybrałem podejście *OneVsOne*). Wyniki uzyskane przedstawiane zostały w tabeli 3.

Algorytm	Metryka	Etap 4	Etap 5
<i>DecisionTreeClasifier</i>	Accuracy	0.8	0.832
<i>DecisionTreeClasifier</i>	G-mean	0.79	0.832
<i>SVC, OneVsOne</i>	Accuracy	0.794	0.827
<i>SVC, OneVsOne</i>	G-mean	0.81	0.846
<i>KNeighborsClassifier</i>	Accuracy	0.759	0.79
<i>KNeighborsClassifier</i>	G-mean	0.7	0.731

Tabela 4: Wyniki klasyfikacji dla czwartego i piątego etapu.

## 4 Wnioski

Po analizie danych, w szczególności analizie *PCA*, to co możemy zauważyć to fakt, że nie ma ściśle określonej granicy klas decyzyjnych. Istnieją przykłady z różnych klas decyzyjnych, posiadające podobne wartości atrybutów co wpływa negatywnie na proces klasyfikacji. Wykonanie to z faktu, że stworzenie jednej oceny dla piłkarza, na którego grę wpływa tak wiele składowych jest po prostu niemożliwe. Dodatkowym problemem jest brak wiedzy jak taka ocena była konstruowana. Czy tworzył ją algorytm czy może fizyczna osoba, która przecież mogła się przejawiać swoimi własnymi opiniami na temat piłkarzy. Maksymalny uzyskany przez mnie wynik klasyfikatora to 0.846 dla metryki *G – mean* dla klasyfikatora *SVC* oraz 83.2% trafności dla drzew. Trudno mi z tego wyciągnąć jakieś wnioski. Po wstępnej analizie można się było spodziewać, że dla tych danych klasyfikator *SVC* i *DTC* spiszą się całkiem dobrze. Biorąc pod uwagę przenikanie klas to *KNN* również spisał się nieźle. Dla tego zbioru danych warto by było sprawdzić działanie metod radzenia sobie z niebilansowanymi klasami lub klasyfikatory złożone. Trudno mi o więcej wniosków.

## Literatura

- [1] Fifa 19 complete player dataset. <https://www.kaggle.com/karangadiya/fifa19>.
- [2] Soccer positions. [https://fifafotballvideogames.fandom.com/wiki/Soccer\\_positions](https://fifafotballvideogames.fandom.com/wiki/Soccer_positions).