AKADEMIA GÓRNICZO-HUTNICZA IM. STANISŁAWA STASZICA W KRAKOWIE
WYDZIAŁ ZARZĄDZANIA

Samodzielna Pracownia Zastosowań Matematyki w Ekonomii

Praca dyplomowa licencjacka

Badanie czynników wspomagających wybór klasy sportowej

The analysis of factors influencing the choice of a sport class

Autor: *Olga Maria Chwedoruk*
Kierunek studiów: Informatyka i Ekonometria
Opiekun pracy: dr Łukasz Lach

Kraków, 2020

Spis treści

Wstęp

Od pierwszych lat życia dzieciom wpaja się, że sport jest bardzo ważną częścią życia. Aktywność fizyczna wspomaga rozwój somatyczny, a dzięki rozwojowi mięśni, zwiększana jest także wytrzymałość kości. Ponadto sport poprawia koordynację ruchową oraz wspomaga proporcjonalny przyrost masy ciała. Skupiając się na rozwoju psycho-emocjonalnym, poprawia pamięć, co może przekładać się na lepsze wyniki w nauce. Odnotowano także, że sport obniża stany depresyjne i lękowe. Przyglądając się aktywności fizycznej od strony rozwoju społecznego, stwierdzono znaczną poprawę stosunków między ludzkich, sport uczy współpracy oraz samokontroli. Ruch wpływa także pozytywnie na nasze zdrowie, od chorób przewlekłych takich jak cukrzyca czy niektóre nowotwory, po zdrowie psychiczne.[1] Sport pomaga także eliminować nałogi, czego przykładem może być Jerzy Górski, który przez 14 lat był uzależniony od narkotyków, a dzięki sportowi uwolnił się od nich i w 1990 roku stał się mistrzem świata w jednych z najbardziej prestiżowych zawodów triathlonowych na świecie – Double Iron Triathlon.

W ostatnich latach chęć uprawiania sportu wśród młodzieży systematycznie maleje. Dzieci dużo więcej czasu spędzają przed komputerem, przez co pogarsza się ich stan zdrowia. Tą tendencję widać także w klasach sportowych, gdzie dzieci cały czas ubywa, co zostanie dokładniej przedstawione w rozdziale 2. Chcąc przekonać młodych ludzi do pójścia do klas lub szkół sportowych, warto najpierw dowiedzieć się, co skłania ich do tego wyboru. W niniejszej pracy celem jest zbadanie, które czynniki wpływają na pozytywną decyzję o wybraniu klasy sportowej. Ponadto sprawdzona zostanie kwestia zdrowia, czy miała duży wpływ na decyzję, czy może większe znaczenie miało otoczenia, czyli rodzice i rówieśnicy dziecka.

W badaniu wykorzystana zostanie technika drzew decyzyjnych. Praca podzielona jest na trzy rozdziały. W rozdziale pierwszym zostanie przedstawiona część teoretyczna badania, niezbędne wzory oraz algorytm budowania drzew decyzyjnych. W drugim rozdziale zaprezentowany będzie zbiór danych, statystyczne przedstawienie częstotliwości uprawianego sportu przez dzieci oraz charakterystyka zmiennych. Trzeci rozdział poświęcony będzie przeprowadzeniu badania empirycznego oraz zwięźle przedstawione

[1] Instytut Matki i Dziecka, Zakład Zdrowia Dzieci i Młodzieży, „Aktywność fizyczna młodzieży szkolnej w wieku 9-17 lat", Warszawa 2013, s.11-12

zostanie porównanie czterech metod – drzewa regresyjnego, klasyfikacyjnego, lasów losowych oraz boosting. Część empiryczna została wykonana w języku R, w środowisku RStudio.

1. Opis metod badawczych

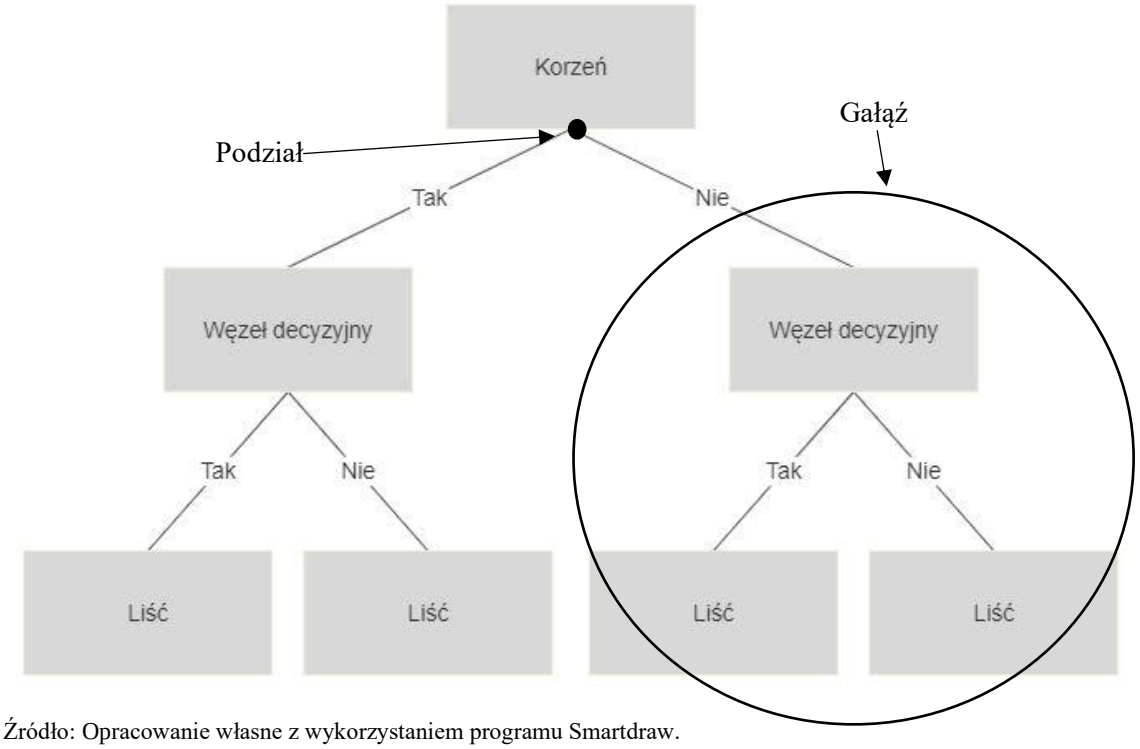
Badanie empiryczne będzie przeprowadzone techniką drzew decyzyjnych. Zastosowane będą metody drzewa regresyjnego, klasyfikacyjnego, lasów losowych oraz boosting. Do zbadania poprawności prognoz, posłużono się błędami predykcji RMSE, MAE i MSE.

1.1 Drzewo decyzyjne

Celem pracy jest zbadanie, które czynniki najbardziej wspomagają wybór klasy sportowej przez młodzież. Jedną z najbardziej znanych technik, która pomaga w podejmowaniu decyzji, jest drzewo decyzyjne. Jest ono graficzną metodą przedstawienia decyzji oraz ich konsekwencji. W celu stworzenia drzewa, dane dzieli się na dwa zbiory. Na danych treningowych budowany jest model, a na danych testowych sprawdzana jest jego skuteczność.

Poniżej przedstawiono ideową strukturę drzewa.

Rysunek 1. Struktura drzewa.



Źródło: Opracowanie własne z wykorzystaniem programu Smartdraw.

- Korzeń – przedstawia całą populację lub jej część. W kolejnym kroku następuje podział na dwa lub więcej jednorodnych zbiorów.
- Węzeł decyzyjny – moment, w którym pod węzły dzielą się na dalsze podwęzły.

- Liść – węzeł, który nie ma podziału.
- Podział – proces dzielenia węzła na dwa lub więcej podwęzłów.
- Gałąź – podsekcja drzewa[2]

Drzewa decyzyjne znajdują praktyczne zastosowanie w problemach związanych z klasyfikacją pojęć np. w diagnostyce medycznej, zagadnieniu udzielania kredytów czy prognozowaniu wydajności.[3]

Największymi zaletami drzew decyzyjnych jest ich łatwość w wykonaniu oraz - dzięki graficznemu przedstawieniu - prostota w interpretacji. Ponadto nie wymagają normalizacji danych, czy usuwania pustych zmiennych. Drzewa pozwalają działać na danych numerycznych oraz kategorycznych, co jest rzadkością w przypadku innych technik.[4]

Niestety jak każda technika, podejście to ma swoje wady. Niemożliwe jest uaktualnienie drzewa, poprzez dodanie obserwacji, ponieważ nawet najmniejsza modyfikacja może prowadzić do dużej zmiany w ostatecznej postaci drzewa.[5]

1.1.1 Drzewo regresyjne

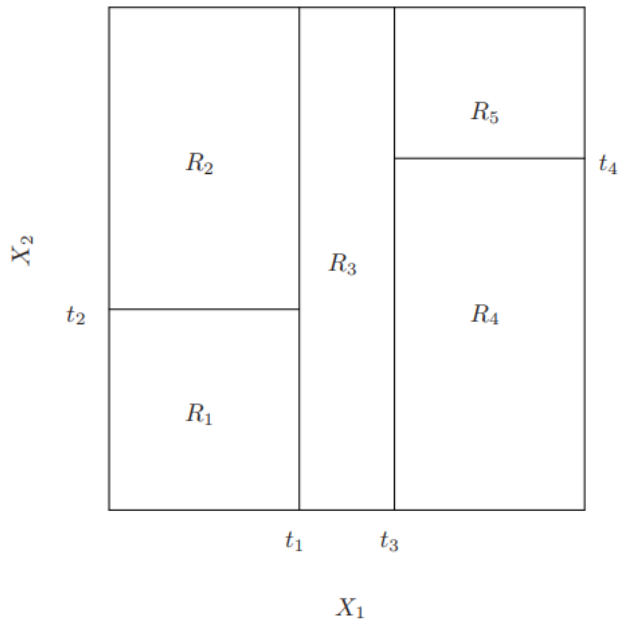
Budowę drzewa regresyjnego można podzielić na 2 etapy:

1. Predykcyjna przestrzeń dzielona jest na J odrębnych, nienakładających się regionów $R_1, R_2, \ldots, R_J$.
2. Dla każdej obserwacji należącej do regionu R_J dokonuje się tej samej prognozy, która jest średnią wartości odpowiedzi dla obserwacji treningowych w R_J.

Regiony $R_1, R_2, \ldots, R_J$ mogłyby mieć dowolny kształt, jednakże dla ułatwienia interpretacji, przestrzeń dzieli się na prostokąty lub pudła (Rysunek 2).

[2] J. Le, *Decision Trees in R,* „Data Camp", 19.06.2018 https://www.datacamp.com/community/tutorials/decision-trees-R [dostęp: 09.06.2020]
[3] P. Marynowski, *Drzewa decyzyjne,* http://home.agh.edu.pl/~pmarynow/pliki/iwmet/drzewa.pdf [dostęp: 10.06.2020]
[4] Bujak Ł., *Drzewa decyzyjne,* Uniwersytet Mikołaja Kopernika, Toruń 2008, http://www.is.umk.pl/~duch/Wyklady/CIS/Prace%20zalicz/08-Bujak.pdf [dostęp: 10.06.2020]
[5] M. Demichowicz , P. Mazur, *Drzewa decyzyjne,* „Automatyczne pozyskiwanie wiedzy", 18.05.2003 https://www.ii.pwr.edu.pl/~kwasnicka/tekstystudenckie/apw/decyzyjne.htm [dostęp: 10.06.2020]

Rysunek 2: Przestrzeń podzielona na prostokąty.



Źródło: James G., Witten D., Hastie T., Tibshirani R., *Tree-Based Methods*, „An Introduction to Statistical Learning" Springer Science + Business Media New York 2013

Celem jest znalezienie takiego prostokąta, który zmniejsza resztową sumę kwadratów (RSS). Przedstawia się ją wzorem:

$$\sum_{j=1}^{J} \sum_{i \in R_j} \left(yi - \hat{y}_{R_j} \right)^2 \tag{1.1}$$

gdzie:

$\hat{y}_{R_j}$ – średni wynik dla obserwacji w zbiorze treningowym.

yi – rzeczywista wartość obserwacji.

J – liczba prostokątów

W praktyce używane jest rekurencyjne dzielenie binarne. Algorytm rozpoczyna się od korzenia, a następnie dzieli przestrzeń predykatora na regiony. W każdym procesie jest wybierany najlepszy podział, dzięki czemu model automatycznie staje się lepszy.

Chcąc wykonać taki podział, należy najpierw wybrać predykator X_j oraz punkty odcięcia s, w taki sposób, gdzie otrzymany podział przestrzeni na regiony $\{X|Xj < s\}$ i $\{ X|Xj \geq s\}$ prowadzi do zminimalizowania RSS. Rozważane są wszystkie predykatory X_1, …, X_p oraz wszystkie możliwe wartości punktów odcięcia dla każdego z nich, a następnie

wybierane są te wartości, dla których RSS jest najmniejsze. Etapy powtarzane są dla wszystkich regionów i trwają, dopóki nie zostanie spełniony warunek końcowy. Takim przykładem może być kontynuacja tworzenia regionów, dotąd aż w regionach będzie nie więcej niż sześć obserwacji. Końcowym etapem jest przewidywanie wyniku na podstawie średniego wyniku obserwacji treningowych dla danego regionu.

Proces ten może spowodować przeuczenie się drzewa, co będzie skutkowało słabymi wynikami w zbiorze testowym. Im mniejsza liczba regionów, tym mniejsza wariancja, więc i lepsze wyniki w zbiorze testowym. Jedną z metod ulepszenia modelu jest przycięcie drzewa. Konstruowane jest drzewo z dużą liczbą gałęzi, a następnie przycina się je. Budowa drzewa trwa tak długo, jak długo spadek RSS przekracza pewną, wysoką wartość. Ta strategia ma swoje wady, ponieważ podczas dzielenia drzewa, można dojść do takiego podziału, gdzie RSS znacznie spadnie. Lepszym działaniem jest wyhodowanie bardzo dużego drzewa, a następnie przycięcia go z powrotem, aby uzyskać poddrzewo z mniejszą wartością błędu. Prosta walidacja krzyżowa pozwala na podział próby na zbiór treningowy i testowy. W przypadku K-krotnej walidacji próba dzieli się na K podzbiorów, gdzie kolejno jeden z nich traktowany jest jako zbiór testowy a pozostałe, wspólnie jako treningowe. Ten algorytm wykonywany jest K razy, a następnie wyniki zostają uśrednione i otrzymany zostaje jeden, lepszy wynik.[6]

1.1.2 Drzewo klasyfikacyjne

Różnica między drzewem regresyjnym, a klasyfikacyjnym jest taka, że w tym wypadku drzewo służy do predykcji odpowiedzi jakościowej, a nie ilościowej. Budowanie drzewa jest podobne jak w przypadku drzewa regresyjnego, z wyjątkiem RSS, który wykorzystywany jest jako kryterium potrzebne do tworzenia podziałów binarnych. Zamiast tego, stosuje się poziom błędu klasyfikacji, który jest określany wzorem:

$$E = 1 - max(\hat{p}_{mk}) \tag{1.2}$$

gdzie:

$\hat{p}_{mk}$ – odsetek obserwacji treningowych w m-tym regionie.
k – liczba klas.

[6] G. James, D. Witten, T. Hastie, R. Tibshirani, *An Introduction to Statistical Learning,* Springer Science + Business Media New York 11.02.2013, s.308-311

W praktyce stosowane są także dwie inne miary pozwalające wyhodować lepsze drzewo. Pierwsza z nich to indeks Giniego określany wzorem:

$$G = \sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk}) \tag{1.3}$$

Indeks przyjmuje małe wartości, jeśli wszystkie $\hat{p}_{mk}$ są bliskie zeru bądź jedności. Drugą miarą jest entropia krzyżowa, którą można zapisać jako:

$$D = -\sum_{k=1}^{K} \hat{p}_{mk} \log \hat{p}_{mk} \tag{1.4}$$

Entropia krzyżowa przyjmuje wartości bliskie zeru, jeśli $\hat{p}_{mk}$ znajduję się w przedziale od 0 do 1. Indeks Giniego oraz entropia są podobne do siebie liczbowo. Podczas budowy drzewa stosuje się właśnie te dwie miary służące ocenie jakości konkretnego podziału, z uwagi na to, że są bardziej wrażliwe od poziomu błędu klasyfikacji. Natomiast podczas przycinania drzewa, warto zastosować wskaźnik błędu klasyfikacji, jeśli celem jest dokładność przewidywań końcowego przyciętego drzewa.[7]

1.1.2 Lasy losowe

Lasy losowe dożą do zmniejszenia wariancji pojedynczych drzew i ich dekorelacji. Jeśli w zbiorze danych wystąpi jeden mocny predykator, lasy losowe nie wykorzystają wszystkich zmiennych objaśniających. W innym wypadku, podobieństwo drzew będzie duże, a ich prognozy silnie skorelowane. Co skutkuje także sporym wynikiem wariancji. Algorytm lasów losowych można podzielić na 4 etapy:

1. Wykorzystując zbiór treningowy losowane są różne podzbiory B ze zwracaniem.
2. Spośród liczby zmiennych objaśniających p losowana jest wartość m, gdzie często $m \approx \sqrt{p}$ (zastosowanie małej wartości m, będzie dobrym rozwiązaniem, gdy w zbiorze występuje duża liczba skorelowanych predykatorów).
3. Następnie tworzone jest drzewo dla każdego podzbioru B i przeprowadzana jest prognoza.
4. Finalna predykcja jest uśrednieniem predykatorów, który zostały utworzone z podzbiorów.[8]

[7] Tamże, s.314-318.
[8] Tamże, s.324-325.

Algorytm drzew losowych umożliwia także stworzenie rankingu zmiennych pod względem ich istotności, skupiając się na ich losowych permutacjach oraz analizy danych cech, które skutkują obniżeniu wartości współczynnika różnorodności węzłów, podczas tworzenia drzew decyzyjnych.[9]

1.1.3 Boosting

Ostatnią omawianą metodę jest boosting. Metoda ta jest ogólnym podejściem, które zwiększa skuteczność drzew decyzyjnych. Drzewa rosną sekwencyjnie, czyli każde drzewo opiera się na informacji z wcześniej zbudowanych drzew. Poniżej przedstawiony zostanie algorytm boosting:

1. Ustalana jest $\hat{f}(x) = 0$ i $r_i = y_i$ dla każdego i w zbiorze treningowym
2. Dla $b = 1, 2, \dots, B$ powtarza się:
   a) Dopasowywane jest drzewo $\hat{f}^b$ z d podziałami ($d + 1$ końcowych węzłów) w zbiorze treningowym (X, r)
   b) Następuje aktualizacja $\hat{f}$ poprzez dodanie skurczonej wersji nowego drzewa:

$$\hat{f}(x) \Leftarrow f(x) + \lambda \hat{f}^b(x) \tag{1.5}$$

   c) Aktualizacja reszty:

$$r_i \Leftarrow r_i - \lambda \hat{f}^b(x_i) \tag{1.6}$$

3. Model finalny:

$$\hat{f}(x) = \sum_{b=1}^{B} \lambda \hat{f}^6(x) \tag{1.7}$$

gdzie:

$\hat{f}(x)$ – model finalny

r_i – reszta poszczególnej obserwacji

$\hat{f}^b$ – model drzewa przed aktualizacją

d – liczba węzłów w drzewie

λ – parametr kurczenia się

[9] J. Pociecha, *Współczesne zmiany narzędzi badań statystycznych,* Zeszyty Naukowe, Uniwersytet Ekonomiczny w Krakowie, Kraków

Boosting posiada trzy ważne parametry:

1. Ilość drzew B – w przypadku lasów losowych, gdy ta wartość jest wysoka, może dojść do przeuczenia się modelu, natomiast w metodzie boosting używana jest walidacja krzyżowa w celu wybrania odpowiedniej wartości B.
2. Parametr kurczenia się λ – mała, dodatnia liczba, która kontroluje szybkość uczenia się. Najbardziej popularne wartości to 0,01 bądź 0,001, ich wybór zależy od konkretnego problemu. Bardzo mała wartość λ, może spowodować użycie dużej liczby B, w celu osiągniecia lepszych rezulatatów.
3. Liczba węzłów d – kontroluje złożoność algorytmu. Jeśli $d = 1$ to w takim przypadku tworzone są modele addytywne. Najczęściej d jest równe głębokości iteracji i kontroli kolejności interakcji modelu.[10]

1.2 Błędy predykcji

Po zbudowaniu modeli drzew nastąpi ich porównanie, do tego celu posłużą miary błędów.

1.2.1 RMSE

RMSE (ang. *root mean squared error*), oznacza pierwiastek błędu średniokwadratowego pomaga w ocenie prognozy ex-post. Błąd ten oblicza się ze wzoru:

$$RMSE = \sqrt{\frac{1}{S}\sum_{\tau=1}^{S}(y_\tau - y_\tau^P)^2} \tag{1.8}$$

gdzie:

S – horyzont prognozy

y_τ – prognozowana zmienna

y_τ^P – prognozy wygasłe, czyli prognozy przewidywanych zmiennych

Im mniejsza wartość błędu RMSE, tym model jest bardziej optymalny.

[10] G. James, D. Witten, T. Hastie, R. Tibshirani, *An Introduction to Statistical Learning*, s.325-328

1.2.2 MAE

Błąd bezwzględny MAE (ang. *mean absolute error*) podaje wartość błędu miary, czyli informuje o ile będzie się różnić odchylenie od wartości rzeczywistej. Błąd określony jest wzorem:

$$MAE = \frac{1}{S}\sum_{\tau=1}^{S}|y_\tau - y_\tau^P| \tag{1.9}$$

Duża różnica między MAE a RMSE może świadczyć, że w okresach prognozy występowały znaczne błędy.

1.2.3 MSE

Błąd MSE (ang. mean squared error), czyli błąd średniokwadratowy opisywany jest wzorem[11]:

$$MSE = \frac{1}{S}\sum_{\tau=1}^{S}(y_\tau - y_\tau^P)^2 \tag{1.10}$$

Tak jak w poprzednich błędach, im mniejsza wartość, tym lepszy model.

[11] R. Hyndman, A. Koehler, *Another look at measures of forecast accuracy*, „International Journal of Forecasting", s.679-688, 2006

2. Wykorzystany zbiór zmiennych

W Polsce, według rozporządzenia ministra edukacji, za szkołę sportową uznaje się szkołę, w której prowadzone są szkolenia sportowe w jednym lub kilku sportach, w co najmniej dwóch klasach oraz liczących co najmniej 15 uczniów każda. W oddziałach sportowych liczba godzin przeznaczonych na treningi, z wyłączeniem ćwiczeń wychowania fizycznego, wynosi 10 godzin, a w szkołach sportowych 16. Ważnym elementem takich klas, jest stworzenie optymalnego planu zajęć, dzięki któremu, uczniowie będą w stanie godzić zajęcia sportowe z innymi zajęciami edukacyjnymi.[12] W kraju w 2018 roku istniało 77 szkół sportowych. Najwięcej odnotowano w województwie pomorskim, bo aż 14, a na drugim miejscu uplasowało się województwo mazowieckie z liczbą 12 placówek. Niestety w porównaniu z 2016 rokiem liczba szkół zmniejszyła się prawie o połowę, a główną przyczyną była likwidacja gimnazjów. Do takich szkół sportowych, w 2018 roku, uczęszczało 47811 uczniów, co jest sporym spadkiem w porównaniu do 2015 roku, gdzie ta liczba wynosiła 56861. Świadczyć to może o mniejszym zainteresowaniu szkołami sportowymi wśród dzieci.[13]

Problem ten powiększa także pandemia koronawirusa. Jak podaje gazeta sportowa *„Przegląd Sportowy”* urząd miejski w Łodzi zapowiedział brak rekrutacji do takich szkół na rok 2020/21. Powodem był problem z przeprowadzeniu bezpiecznych testów sprawnościowych młodzieży. Pod wpływem rodziców, decyzja ta została zmieniona, jednak do podobnych sytuacji dochodzi w innych miastach. W Białymstoku prezydent miasta Tadeusz Truskolaski, tłumaczy, że główną przyczyną wstrzymania naborów są cięcia budżetowe. „Czeka nas rewolucja, jeśli chodzi o finanse i wcale nie zakończy się to dobrze. (...) Podam przykład: miasto będzie płacić nauczycielom pensje, ale pieniędzy na klasy sportowe może już nie wystarczyć”. Amerykańscy ekonomiści już kilka miesięcy temu mówili, że dzieci będą ofiarami pandemii, w związku z utratą pracy przez rodziców, którzy zapewniali im wychowanie sportowe.[14] Amerykańska agencja prasowa

[12] Rozporządzenie Ministra Edukacji Narodowej z dnia 27 marca 2017 r. w sprawie warunków tworzenia, organizacji oraz działania oddziałów sportowych, szkół sportowych oraz szkół mistrzostwa sportowego (Dz.U. z 2017 r. poz. 671)
[13] Ośrodek Statystyki Sportu i Turystyki oraz Podkarpacki Ośrodek Badań Regionalnych, Kultura Fizyczna w Polsce w latach 2017 i 2018, GUS, Departament Badań Społecznych i Warunków Życia, Warszawa-Rzeszów 2019, s. 54.
[14] M. Chmielewski, *Przyszedł kryzys, likwidują klasy sportowe*, „Przegląd Sportowy” 20.05.2020

Associated Press porównała epidemię do kryzysu z 2008 roku, w którym to w skali świata aż kilkadziesiąt procent dzieci zrezygnowało z sekcji sportowych.[15]

Tabela 1: Procent 13-latków uprawiających sport przynajmniej 4 razy w tygodniu w krajach UE.

Państwo	2014		2018	
	Dziewczęta	Chłopcy	Dziewczęta	Chłopcy
Finlandia	57	61	55	62
Irlandia	46	66	49	59
Słowenia	38	53	44	59
Austria	37	53	41	60
Grecja	32	54	43	57
Słowacja	47	61	42	58
Bułgaria	38	55	39	60
Łotwa	36	53	35	53
Węgry	32	49	35	50
Holandia	43	60	34	49
Czechy	37	46	35	47
Malta	36	45	35	47
Niemcy	39	54	30	51
Estonia	36	43	35	45
Belgia	32	49	32	46
Chorwacja	29	47	29	48
Litwa	35	53	31	46
Hiszpania	25	54	25	52
Rumunia	38	56	29	47
Luksemburg	40	58	25	50
Szwecja	36	46	33	41
Polska	31	49	27	37
Francja	23	43	20	43
Dania	37	48	29	33
Włochy	21	39	22	40
Portugalia	19	52	20	38

Źródło: *Spotlight on adolescent health and well-being*, WHO 2020.
https://apps.who.int/iris/bitstream/handle/10665/332104/9789289055017-eng.pdf [dostęp: 08.06.2020]

[15] S. Dixon, *Pandemic costing youth sports millions, creating uncertainty*, „Associated Press", 18.04.2020 https://apnews.com/e1f97b5e44ec5d7a742e0845a5eca7c6 [dostęp: 08.06.2020]

Tabela 1 przedstawia odsetek dzieci, mieszkających w krajach Unii Europejskiej, które przejawiają aktywność fizyczną poza szkołą, co najmniej cztery razy w tygodniu. Można zauważyć, że Polska znajduje się pod koniec spisu, wyprzedzając tylko 4 kraje. Zaobserwowano duża rozpiętość procentową między krajami. W 2018 roku w Finlandii odnotowano, że 55% dziewcząt często ćwiczy, gdzie w Portugalii ten wynik to zaledwie 25%. U chłopców ponownie na czele występuję Finlandia, której wynik to 62%, natomiast najniższy zaobserwowano w Danii, tylko 33% nastolatków uprawiało sport co najmniej 4 razy w tygodniu. W porównaniu z rokiem 2014, wśród chłopców we wszystkich krajach odnotowano znaczny średni spadek, bo aż o 12 punktów procentowych. Niestety u dziewcząt także zaobserwowano ogólny spadek, jednakże znacznie mniejszy, bo tylko o 4 pp. Średnia aktywności sportowej wśród młodzieży, we wszystkich krajach UE w 2014 wyniosła 43,6%, natomiast w 2018 41,4%. Zatem można zauważyć, że dzieci coraz mniej czasu poświęcają na ćwiczenia fizyczne. Poprzez mniejszy ruch, nastolatkowie narażają się na problemy z otyłością czy na wady ortopedyczne. Jak mówią eksperci, sport oddziałuje także na nasz umysł, wpływa na to jak myślimy i jak czujemy. Klasy i szkoły sportowe, próbują wprowadzić sport w rutynę dzieci, dzięki czemu w przyszłości, być może unikną różnych schorzeń czy nadwagi.

2.1 Charakterystyka zmiennych

Dane wykorzystane w badaniu zostały uzyskane przez ankietę. Została ona przeprowadzona w klasach sportowych 7 i 8 szkół podstawowych. Niestety liczba obserwacji nie jest duża, ze względu na wybuch epidemii koronawirusa w trakcie zbierania danych.

Młodzież otrzymała ankiety z 12 pytaniami, gdzie 11 z nich było zmiennymi objaśniającymi, a ostatnie, 12 pytanie, zmienną objaśnianą. Odpowiedzi na pytania są w skali od 0 do 10, z wyjątkiem 4 pytania, będącego w przedziale od 0 do 4. Utworzony zbiór danych zawiera 60 obserwacji. Aby móc sprawdzić, jaka dokładnie liczba, będzie odpowiadała za wynik w ostatnim pytaniu, posłużono się zmiennymi kategorycznymi.

Zmienna objaśniana:

- Pytanie 12 – W jakim stopniu dalej chcesz kontynuować naukę w klasie sportowej?

Zmienne objaśniające:

- Pytanie 1 – Płeć.
- Pytanie 2 – Jak bardzo lubisz uprawiać sport?
- Pytanie 3 – Jak bardzo interesujesz się sportem uprawianym przez Ciebie? (Czy czytasz wiadomości sportowe, śledzisz ulubionych sportowców i ich karierę?)
- Pytanie 4 – Jak mocno rodzice zachęcali Cię do pójścia do klasy sportowej?
- Pytanie 5 – Czy masz wadę postawy?
- Pytanie 6 – Jak duży wpływ miał twój stan zdrowia na wybór klasy sportowej?
- Pytanie 7 – Jak bardzo twoi znajomi wpłynęli na twoją decyzję o klasie sportowej? – Bada czy na decyzje wpływ mieli rówieśnicy.
- Pytanie 8 – Jak bardzo wiążesz swoją przyszłość z karierą sportowca?
- Pytanie 9 – W jakim stopniu uważasz, że wybranie klasy sportowej przybliży Cię do zdobycia sławy?
- Pytanie 10 – W jakim stopniu uważasz, że sportowcy są bogaci?
- Pytanie 11 – W jakim stopniu sądzisz, że dzięki klasie sportowej zyskasz atletyczną figurę?

Pytania badają po kolei, jakie czynniki wpływają na wybór klasy sportowej. Pytanie 1 sprawdza, czy chłopcy lub dziewczęta wykazują większą chęć do uczęszczania do klas sportowych. Pytanie 2 bada chęć młodzieży do ćwiczeń, sportu, gdzie pytanie 3 bada ich zainteresowanie, czyli śledzenie wydarzeń sportowych lub profili swoich ulubionych sportowców. Pytanie 4 analizuje, czy dziecko było poddawane naciskom ze strony rodziców. Pytanie 5 sprawdza, czy wada ortopedyczna miała znaczenie na wybór klasy, ponieważ istnieją sporty takie jak pływanie czy gimnastyka, które pomagają częściowo wyleczyć, lub przynajmniej zahamować rozwój wady. Kolejne pytanie analizuje czy powodem wyboru klasy sportowej były choroby, ogólny stan zdrowia (np. cukrzyca). Pytanie 7 bada, czy na decyzję ucznia wpływ mieli rówieśnicy, którzy także poszli do tej klasy. Pytanie 8 odnosi się do przyszłego, planowanego zawodu sportowca, być może właśnie udział w olimpiadzie zachęca młodych ludzi do przystąpienia do klasy sportowej. Następne pytanie sprawdza, czy może sława, duża rozpoznawalność sportowców zaintrygowała młodzież do spróbowania swoich sił w tej klasie. Pytanie 10 bada, czy korzyści materialne miały wpływ na decyzję. Ostatnie pytanie analizuje czy chęć zdobycia atletycznej figury ma znaczenie. Wszystkie pytania zawarte w ankiecie są

zmiennymi, które objaśniają pytanie, czyli czy młodzież zamierza kontynuować naukę w klasie sportowej.

3. Badanie empiryczne

Jak już wcześniej wspomniano, technika drzew sprawdza się także dla danych kategorycznych, dzięki czemu można ją wykorzystać do badania powyższych zmiennych.

3.1 Przygotowanie danych

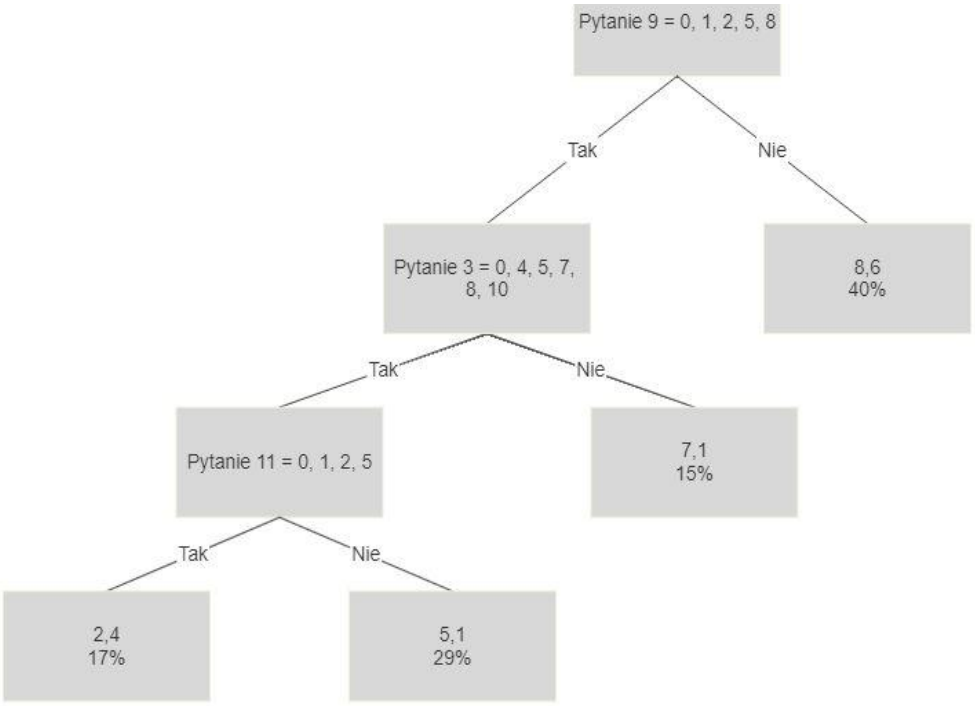
Rozpoczynając analizę, należy najpierw podzielić zbiór danych na treningowe oraz testowe. Sporządzony został losowy podział obserwacji, gdzie dane treningowe, które posłużą do nauki drzewa, będą zawierały 80% zbioru, a testowe służące do sprawdzenia błędu predykcji 20%. W zbiorze uczącym drzewo znalazło się 48 obserwacji, a w zbiorze testowym 12.

3.2 Drzewo regresyjne

Przystępując do badania, najpierw zostanie przedstawione pierwsze drzewo decyzyjne – regresyjne. Za pomocą pakietu statystycznego RStudio, został zaprezentowany model drzewa wykorzystujący dane treningowe.

Jak wynika z rysunku 3, zmiennymi decydującymi okazały się pytania 9, 3 i 11. Czyli kluczowymi pytaniami są kolejno pytania o sławę, zainteresowanie sportem oraz atletyczną figurę. Wynik znajdujący się w liściach drzewa to średnia odpowiedzi oddanych na ostatnie, dwunaste pytanie, czyli zmienną objaśnianą oraz odsetek młodzieży, która w ten sposób odpowiedziała. Zatem, jeśli nastolatkowie udzielali wysokich odpowiedzi na pytanie, w jakim stopniu pójście do klasy sportowej przybliży ich do zdobycia sławy, to pytaniu dwunastemu przyznawali średnio 8,6 punktu oraz ich odsetek wyniósł 40%. W związku z niedostateczną liczbą obserwacji odpowiedź 8 przy pytaniu 9, może wynikać z przypadkowej odpowiedzi. Pytanie 3, odnoszące się do zainteresowania się sportem, może natomiast wykazywać, że udzielając niższych odpowiedzi, 15% nastolatków udzieli odpowiedzi na ostatnie pytanie w granicach 7,1. Przy tym pytaniu, można zauważyć, znaczną niedoskonałość modelu. Ostatnia zmienna decydująca, określająca istotność zgrabnej figury, sugeruje, że im wyższe odpowiedzi ankietowanych, tym większa skłonność do zaznaczenia odpowiedzi o wartości 5. Jeżeli młodzież udzielała kolejno wysokich, niskich i wysokich odpowiedzi na pytania 9, 3 i 11 to zmienna objaśniana przyjmie wartość 2,4. Procenty powinny się sumować do 100%, jednakże z powodu niedokładnego przybliżenia, ich suma wynosi 101%.

Rysunek 3: Model drzewa regresyjnego



Źródło: Opracowanie własne z wykorzystaniem programu SmartDraw.

Kolejnym krokiem, będzie przedstawienie wpływu zmiennych na model.

Tabela 2: Wpływ zmiennych na model

Zmienna	**Wartość**
Pytanie 9	0,8945
Pytanie 11	0,8722
Pytanie 6	0,4159
Pytania 10	0,3815
Pytanie 8	0,3525
Pytanie 3	0,3168
Pytanie 4	0,2116
Pytanie 2	0,1780
Pytanie 1	0,0000
Pytanie 5	0,0000
Pytanie 7	0,0000

Źródło: Opracowanie własne

W tabeli 2 można zauważyć, że pytania 9 i 11, pojawiające się w drzewie regresyjnym, mają największy wpływ. Następne w tabeli jest pytanie nr 6, które bada wpływ stanu zdrowia na decyzję o pójściu do klasy sportowej. Zaskakujący jest to, że niewielki wpływ na decyzję o wyborze klasy sportowej ma 4 pytanie, które odnosi się do wpływu rodziców. Oczywistym jest fakt, że zmienna decydująca, pytanie 2, będzie miała znikomy wpływ na model, ponieważ odnosi się do czerpania radości z uprawianego sportu. Naturalne jest to, że odpowiedzi będą wysokie, ponieważ osoby będące w klasie sportowej, w większości przypadków lubią uprawiać sport. Pytania 1, 5 i 7, odnoszące się kolejno do płci, wady postawy i sugestii znajomych nie mają żadnego wpływu na model.

W kolejnym etapie należy zbadać skuteczność modelu bazującego na danych treningowych dla zbioru testowego. Dokonując predykcji, napotkano błąd, który wskazuje na to, że pytanie 7 w danych treningowych zawiera odpowiedzi, których brakuje w zbiorze testowym. Podjęto decyzję o usunięciu zmiennej „Pytanie 7", zamiast usunięcia obserwacji, które zawierały takie odpowiedzi, ponieważ ilość obserwacji jest niewielka, a wpływ zmiennej na model jest zerowy. Następnie, taki sam błąd pojawił się przy pytaniu 8, w tym wypadku wystarczające było usunięcie jednej odpowiedzi.

Po zlikwidowaniu błędów przystąpiono do predykcji. Dokładność wyników, pozwolą sprawdzić błędy RMSE, czyli pierwiastek błędu średniokwadratowego, MAE, średni błąd bezwzględny, oraz MSE, czyli średni błąd kwadratowy. Wyniki zostały przedstawione w tabeli.

Tabela 3: Błędy predykcji

RMSE	**MAE**	**MSE**
3,809	3,012	14,508

Źródło: Opracowanie własne.

RMSE wynosi 3,809, co oznacza, że prognozy różnią się od wartości rzeczywistych średnio o 3,809. Praktyczną właściwością tego błędu, jest to, że występuje w tych samych jednostkach co zmienna odpowiedzi. W tym przypadku oznacza to, że wynik odpowiedzi na pytanie 12 może się różnić o 3,809. Widoczna jest różnica między błędem RMSE i MAE, co także wskazuje na zawodność modelu. Zauważalny jest także wysoki błąd MSE.

Podjęto próbę zmniejszenia błędu RMSE, co poprawiłoby jakość modelu. Aby tego dokonać, należy znaleźć optymalne parametry Cp, który jest bezpośrednim wskaźnikiem zdolności procesu, Minsplit, czyli najmniejsza liczba obserwacji w węźle nadrzędnym, którą można podzielić dalej, oraz Maxdepth, który zapobiega wzrostowi drzewa powyżej określonej wysokości. Kod wyliczający parametry został zaprezentowany w dodatku. W tabeli zostały przedstawione optymalne parametry.

Tabela 4: Optymalne parametry

Cp	**Minsplit**	**Maxdepth**
0	8	2

Źródło: Opracowanie własne.

Po wyliczeniu parametrów i ich zastosowaniu ponownie zostały wyliczone błędy predykcji. Wyniki zostały przedstawione w poniższej tabeli.

Tabela 5: Błędy predykcji

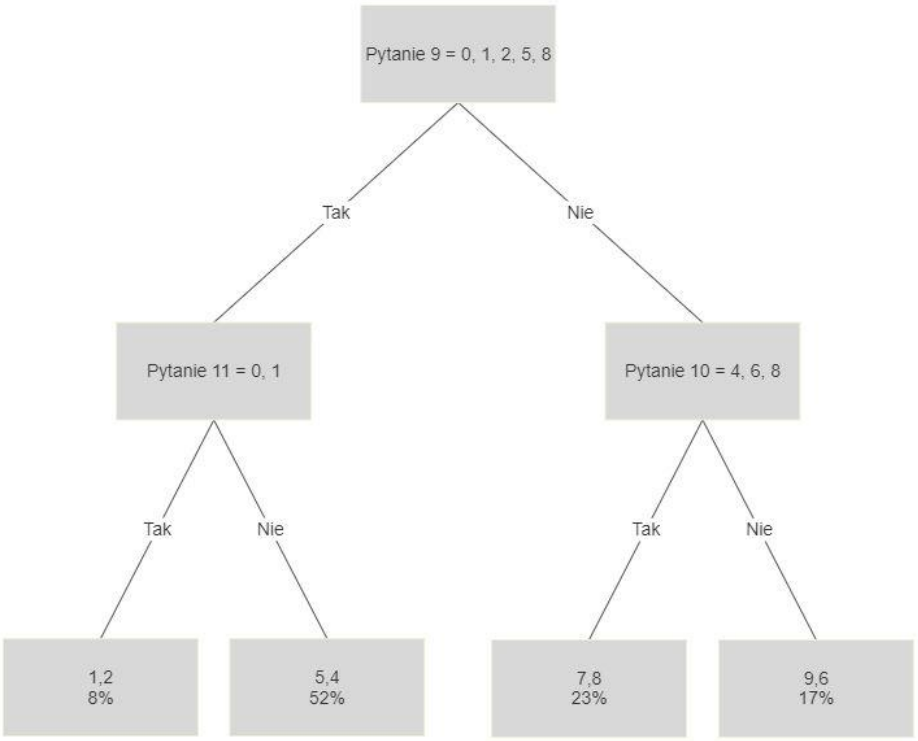
RMSE	**MAE**	**MSE**
3,224	2,633	10,395

Źródło: Opracowanie własne.

Błędy zostały znacznie zmniejszone. Teraz wynik odpowiedzi na pytanie 12 może się różnić o 3,224, a różnica między RMSE a MAE zmalała o 0,206. Parametr Maxdepth wskazuje na to, że optymalna wysokość drzewa to dwa poziomy. Na rysunku 4 przedstawiono drzewo regresyjne, zawierający optymalne parametry.

Można zauważyć pojawienie się nowej zmiennej decydującej – pytanie 10, które bada, czy młodzież sugerowała się przyszłym bogactwem, wybierając klasę sportową. Skrajnie niskie i skrajnie wysokie odpowiedzi wybrane w dziesiątym pytaniu, oznaczają, że młodzież udzieli odpowiedzi na pytanie dwunaste w granicy 9,6. Odsetek takich osób jest równy 17%. Poniżej zostanie przedstawiony wpływ zmiennych objaśniających na model.

Rysunek 4: Model drzewa regresyjnego z optymalnymi parametrami



Źródło: Opracowanie własne z wykorzystaniem programu SmartDraw.

Poniżej zostanie przedstawiony wpływ zmiennych objaśniających na model.

Tabela 6: Wpływ zmiennych na model

Zmienna	**Wartość**
Pytanie 9	0,9611
Pytanie 10	0,7361
Pytanie 3	0,6359
Pytanie 11	0,6017
Pytanie 8	0,5398
Pytanie 6	0,4159
Pytanie 2	0,3027
Pytanie 1	0,0000
Pytanie 4	0,0000
Pytanie 5	0,0000

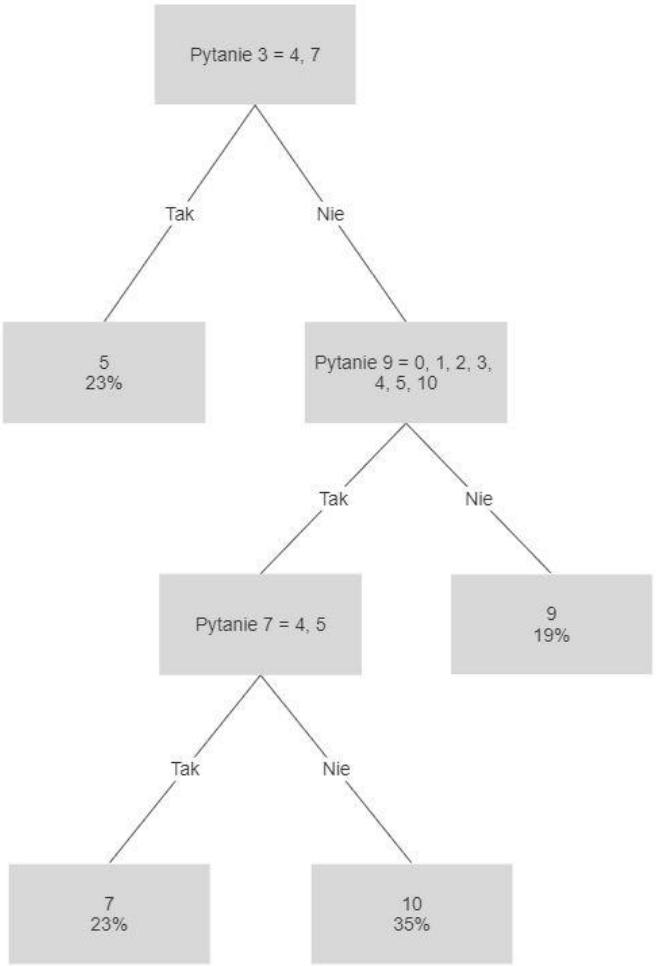
Źródło: Opracowanie własne.

Porównując wpływ zmiennych poprzedniego modelu z nowym modelem, można zauważyć, że zmienna pytanie 9, nadal ma największy wpływ na model. Dodatkowo wpływ ten się zwiększył. Następne miejsce w tabeli zajęła zmienna pytanie 10, pojawiła się ona także w drzewie regresyjnym. Jej wpływ zwiększył się o 0,3546 w porównaniu z poprzednim. Zaskakujący jest spadek pytania 11, odnoszącego się do figury sportowca, jednakże wciąż ta zmienna pojawia się w drzewie regresyjnym. W nowym modelu, zmienna pytanie 4 ma zerowy wpływ. Odnosi się ona do sugestii rodziców na temat klasy.

3.3 Drzewo klasyfikacyjne

Drugą omawianą przeze mnie metodą jest drzewo klasyfikacyjne. Za pomocą środowiska RStudio, został zaprezentowany model drzewa wykorzystujący dane treningowe.

Rysunek 5: Model drzewa klasyfikacyjnego.



Źródło: Opracowanie własne.

Zmiennymi decydującymi o podziale zostały pytania 3, 9 i 7. Pierwsze z nich odnosi się do zainteresowania sportem w kontekście śledzenia wydarzeń sportowych czy ich oglądania. Jeżeli odpowiedzi młodzieży na to pytanie wynoszą 4 lub 7 to dwunastemu pytaniu przyznają 5. Jeśli ich odpowiedź będzie inna, to następną zmienną, która zadecyduje o kolejnym podziale, będzie pytanie 9, odnoszące się do sławy. Wysokie odpowiedzi, będą odpowiadały za zakreślenie 9 przy dwunastym pytaniu. Sądzi się, że dziesiątka podana przy tym pytaniu, może być błędna z uwagi na małą ilość obserwacji. Ostatnią zmienną, pojawiającą się w modelu, jest pytanie 7, dotyczące wpływu znajomych. Co ciekawe, wybór środkowych odpowiedzi determinuje wybór 7 przy ostatnim pytaniu, a skrajnych 10.

Przedstawiony teraz zostanie wpływ zmiennych na model.

Tabela 7: Wpływ zmiennych na model.

Zmienna	**Wartość**
Pytanie 9	11,2398
Pytanie 7	10,1511
Pytanie 4	9,4612
Pytanie 8	5,8207
Pytanie 3	4,4188
Pytanie 11	3,1159
Pytanie 6	2,7214
Pytanie 10	2,5398
Pytanie 1	0,0000
Pytanie 2	0.0000
Pytanie 5	0,0000

Źródło: Opracowanie własne.

Wyraźnie, największy wpływ na model mają pytania 9 i 7, które także pojawiają się w modelu drzewa. Następnie, dość wysokie znaczenie mają pytania 4 i 8 odnoszące się do zachęceń ze strony rodziców oraz wiązania przyszłości z karierą sportowca. Oba zagadnienia nie znalazły się w końcowym modelu. Na pozycji piątej pod względem wpływu zmiennych na model, uplasowało się pytanie 3, które pojawiło się także

w modelu drzewa klasyfikacyjnego. Zmiennymi o małym oddziaływaniu na model okazały się pytania 11, 6 i 10. Natomiast pozostałe 3 pytania: 1, 2 i 5, nie mają żadnego wpływu na model.

W następnym etapie należy przejść do zbadania skuteczności modelu na zbiorze testowym. Podobnie jak w przypadku drzewa regresyjnego podczas badania predykcji napotkano błąd, w pytaniu 7. Tym razem to pytanie ma znaczny wpływ na model, w związku z czym zdecydowano o usunięciu 4 obserwacji z modelu, w których występował błąd.

Po wyeliminowaniu błędów zbadano błędy: RMSE, MAE oraz MSE, które posłużą do sprawdzenia dokładności predykcji. Wyniki zaprezentowano w tabeli.

Tabela 8: Błędy predykcji

RMSE	**MAE**	**MSE**
4,307	3,950	18,547

Źródło: Opracowanie własne.

Z tabeli wynika, że wyniki zboru treningowego i testowego mogą się różnić o 4,308, co oznacza, że o taką wartość może różnić się ostateczny wynik w pytaniu 12. Wskazuje na to błąd RMSE. Różnica między tym właśnie błędem, a MAE nie jest, aż tak duża, natomiast wynik błędu MSE jest znaczący.

W celu poprawienia modelu ponownie podjęto próbę zmniejszenia błędu RMSE. Znalezione zostały parametry Cp, Minsplit oraz Maxdepth.

Tabela 9: Optymalne parametry modelu.

Cp	**Minsplit**	**Maxdepth**
0	1	4

Źródło: Opracowanie własne.

Obliczone zostały optymalne parametry. Najmniejsza liczba obserwacji w węźle nadrzędnym wynosi 1, a wysokość drzewa to 4 poziomy. Po ich zastosowaniu w modelu ponownie wyliczono błędy.

Tabela 10: Błędy predykcji.

RMSE	MAE	MSE
4,257	3,835	18,126

Źródło: Opracowanie własne.

Błędy nie zostały znacznie zmniejszone, co powoduje, że model wciąż nie jest idealny. Błąd RMSE zmalał jedynie o 0,05, co oznacza, że prognoza może się mylić aż o 4,257. Różnica między błędem RMSE, a MAE zwiększyła się o 0,065.

Poniżej przedstawiono model drzewa.

Rysunek 6: Model drzewa klasyfikacyjnego z optymalnymi parametrami.

Źródło: Opracowanie własne.

Pytania wpływające na model to 3, 9, 7, 6 i 11. Trzy pierwsze pytania pojawiły się także w pierwszym modelu drzewa klasyfikacyjnego. Zmienną determinującą pierwszy podział jest pytanie 3, a następnie decydujące jest pytanie 9, które zawiera się w obu ścieżkach drzewa. Zaskakujący jest także fakt, że aby uzyskać odpowiedź 10, przy pytaniu 12, można skorzystać z dwóch ścieżek. Pierwszą z nich jest wybranie odpowiedzi 4 lub 7 przy pytaniu 3, zaznaczenie wyższej niż połowa odpowiedzi na 9 pytanie oraz wybranie większej odpowiedzi niż 2 na pytanie 7. Druga ścieżka rozpoczyna się udzieleniem przeciwnej odpowiedzi na pytanie 3 niż w przypadku pierwszej drogi, a następnie zaznaczeniem w pytaniu 9 odpowiedzi 0, 1, 3, 4 lub 10. Końcową zmienną determinującą podział jest pytanie 11, gdzie należy wybrać odpowiedzi większe niż 5 z wyjątkiem 8, aby uzyskać wartość 10 w pytaniu 12.

Tabela 11: Wpływ zmiennych na model.

Zmienna	**Wartość**
Pytanie 9	17,422
Pytanie 8	15,367
Pytanie 10	13,686
Pytanie 7	13,451
Pytanie 11	13,194
Pytanie 3	11,036
Pytanie 6	8,806
Pytanie 4	7,706
Pytanie 2	2,836
Pytanie 1	0,000
Pytanie 5	0,000

Źródło: Opracowanie własne.

W nowym modelu, w przypadku większości zmiennych wpływ na model się zwiększył. Pytanie 9, wciąż przyjmuje największą wartość. Na następnych pozycjach, ukazują się pytania 8 i 10, które nie znalazły się w końcowym modelu natomiast następne 4 zmienne już tak. Jedyny spadek znaczenia zmiennej, odnotowano w przypadku pytania 4, odnoszącego się do wpływu rodziców. Pytanie 2, przy optymalnych parametrach, minimalnie wpływa na model. Pytania 1 i 5, wciąż nie mają znaczenia w modelu.

3.4 Lasy losowe

Kolejną opisywaną metodą, będzie metoda lasów losowych. Wykonano prognozę, a następnie obliczono błędy RMSE, MAE i MSE. Wyniki przedstawiono w tabeli.

Tabela 12: Błędy predykcji.

RMSE	MAE	MSE
2,466	3,500	17,167

Źródło: Opracowanie własne.

Porównując poprzednie metody, błędy te nie są duże, widoczna jest jednak różnica między błędem RMSE a MAE. Błąd MSE jest dość niski, ale optymalizacja parametrów może go jeszcze obniżyć. W tabeli 13 zostały przedstawione poprawione domyślne parametry modelu *mrty* oznaczający liczbę zmiennych, które zostały losowo próbkowane jako kandydatów w każdym podziale, *nodesize*, czyli minimalny rozmiar węzła i *ntree* określający liczbę drzew do wzrostu.

Tabela 13: Optymalne parametry modelu.

Mtry	Nodesize	Ntree
7	3	100

Źródło: Opracowanie własne.

Stworzony zostanie model z powyższymi parametrami i na jego podstawie dokonana predykcja na zbiorze testowym. Błędy zostały wyliczone i przedstawione w tabeli.

Tabela 14: Błędy predykcji.

RMSE	MAE	MSE
1,826	1,333	3,333

Źródło: Opracowanie własne.

Wartość błędów się zmniejszyła i jak dotąd jest ona najniższa spośród przedstawionych wcześniej metod. Wartość błędu RMSE wynosi 1,826, czyli odpowiedź dziecka może różnić się średnio o 1,8, co jest już dość zadowalającym wynikiem. Wciąż widoczna jest różnica między RMSE i MAE, jednak nieznacznie się zmniejszyła. MSE zmalało do 3,333.

Tabela 15 przedstawia znaczenie zmiennych na model. Największy wpływ na model ma pytanie 3 oznaczające zainteresowanie się sportem. Następną w kolejności jest zmienna

pytanie 11, która odnosi się do sportowej figury. Na trzeciej pozycji uplasowała się zmienna pytanie 9, która w poprzednich metodach była najbardziej wpływowa, a zaraz za nią pytanie 8 dotyczące przyszłej kariery. Pozostałe zmienne nie wpływają znacząco na model.

Tabela 15: Wpływ zmiennych na model.

Zmienna	Wartość
Pytanie 3	6,320
Pytanie 11	6,165
Pytanie 9	5,865
Pytanie 8	5,199
Pytanie 10	4,033
Pytanie 7	3,448
Pytanie 4	2,367
Pytanie 6	2,154
Pytanie 2	1,793
Pytanie 1	0,441
Pytanie 5	0,154

Źródło: Opracowanie własne.

3.5 Boosting

Ostatnią omawianą metodą będzie metoda boosting. Dokonano predykcji na zbiorze testowym, a następnie wyliczono błędy RMSE, MAE oraz MSE.

Tabela 16: Błędy predykcji.

RMSE	MAE	MSE
8,753	8,267	76,616

Źródło: Opracowanie własne.

W porównaniu do drzewa regresyjnego i klasyfikacyjnego, błędy są bardzo duże. Błąd RMSE wynosi 8,7531, co jest największym jak dotąd wynikiem. Odpowiedzi na pytania są w granicy od 0 do 10, więc wynik tego błędu jest znaczący. Dla przykładu, jeśli dziecko w dwunastym pytaniu zaznaczy 10, to według modelu jego odpowiedź może być także

równa 2. Wartość błędu MSE, także jest bardzo wysoka. Natomiast różnica między RMSE, a MAE jest stosunkowo niewielka.

W celu zoptymalizowania modelu, obliczono parametry *shrinkage* oznaczający szybkość uczenia się, bądź zmniejszenie rozmiaru kroku, *interaction.depth*, wskazujący najwyższy dozwolony poziom interakcji zmiennych, *cv.folds*, który wskazuje na liczbę zagęszczenia krzyżowego do wykonania, *n.minobsinnode* oznaczający ilość obserwacji w końcowych węzłach drzew oraz *n.trees* określający liczbę drzew do dopasowania. Kod został zaprezentowany w dodatku. Poniżej przedstawiono tabelę z optymalnymi wartościami parametrów.

Tabela 17: Optymalne parametry modelu.

Shrinkage	**Interaction.depth**	**Cv.foldes**	**N.minobsinnode**	**N.trees**
0,001	1	0	5	100

Źródło: Opracowanie własne.

Na podstawie powyższych parametrów ponownie obliczono błędy RMSE, MAE oraz MSE i ich wyniki przedstawiono w tabeli.

Tabela 18: Błędy predykcji.

RMSE	**MAE**	**MSE**
2,825	2,407	7,982

Źródło: Opracowanie własne.

Porównując tabele 16 i 18, można zauważyć istotną poprawę błędów. Błąd RMSE poprawił się aż czterokrotnie. MSE zmalało do 7,982, a różnica między RMSE i MAE się zmniejszyła.

W tabeli 19 przedstawiono wpływ zmiennych na model. Podobnie jak w pierwszych dwóch metodach, największy wpływ na końcowy wynik modelu ma pytanie 9 dotyczące zdobycia sławy przez młodzież. Pytanie 3 i 8 odnoszące się kolejno do zainteresowania sportem i przyszłą, ewentualną karierą sportowca, także, choć w mniejszym stopniu, wpływają na model. Pytanie 11, 10, 7, 4 i 6 są mniej istotne, a pytania 1, 2, 5 nie mają żadnego wpływu.

Tabela 19: Wpływ zmiennych na model.

Zmienna	Wartość
Pytanie 9	20,284
Pytanie 3	4,634
Pytanie 8	4,476
Pytanie 11	1,600
Pytanie 10	1,522
Pytanie 7	1,322
Pytanie 4	1,091
Pytanie 6	1,071
Pytanie 1	0,000
Pytanie 2	0,000
Pytanie 5	0,000

Źródło: Opracowanie własne.

3.6 Porównanie metod

Powyżej zostały zaprezentowane metody badania mającego na celu sprawdzenie, które z czynników wpływają na wybór klasy sportowej. Wykorzystaną techniką były drzewa decyzyjne, a metodami, jakimi się posłużono były drzewo regresyjne, drzewo klasyfikacyjne, lasy losowe oraz boosting. Do porównania metod konieczna była ocena pomiarów błędów. Poniżej przedstawiono tabelę z wynikami błędów RMSE, MAE oraz MSE.

Tabela 20: Porównanie błędów predykcji.

	RMSE	MAE	MSE
Drzewo regresyjne	3,224	2,633	10,395
Drzewo klasyfikacyjne	4,257	3,835	18,126
Lasy losowe	1,826	1,333	3,333
Boosting	2,825	2,407	7,982

Źródło: Opracowanie własne.

Zdecydowanie, najlepsza okazała się metoda lasów losowych. Wszystkie wartości błędów są niższe od pozostałych. Wyniki tej metody są dość zadawalające. Odpowiedzi młodzieży na 12 pytanie mogą się różnić o 1,826, czyli wybierając odpowiednią ścieżkę

odpowiedzi, zamiast wyniku 7 przy zmiennej objaśnianej sugerującej, że dosyć prawdopodobne jest wybranie klasy sportowej, można się spodziewać, także zaznaczenie odpowiedzi w granicach 5, co oznaczałoby, że zainteresowanie się tym profilem klasy jest średnie. Zaraz za metodą lasów losowych, uplasowała się metoda boosting, której wyniki również można uznać za wystarczające. Pozostałe dwa modele mają słabszą moc predykcyjną i ich wyniki nie są miarodajne.

Zmienną, której wpływ na modele był najbardziej zróżnicowane, jest pytanie 7. Była ona istotna dla modelu drzewa klasyfikacyjnego zarówno przed jak i po zastosowaniu optymalnych parametrów (tabela 7 i 11), natomiast dla modelu drzewa regresyjnego zmienna ta jest nieistotna (tabela 2). Pytania 9, 3 i 10 są na wysokiej pozycji pod względem istotności we wszystkich modelach.

Podsumowanie

Praca została podzielona na trzy części. W pierwszym z nich została przedstawiona budowa i definicja drzewa decyzyjnego, oraz cztery metody – drzewo regresyjne, klasyfikacyjne, lasy losowe i boosting. Zwięźle zostały opisane błędy prognozy. W kolejnym rozdziale przedstawiona została obecna sytuacja klas sportowych, zaangażowanie dzieci sportem oraz charakterystyka zmiennych. Trzeci rozdział został poświęcony na badanie empiryczne oraz porównanie technik.

Celem pracy było zbadanie, które czynniki wpływają na wybór klasy sportowej przez młodzież. Z przeprowadzonego badania wynika, że metoda lasów losowych okazała się być najbardziej miarodajna. Wynika z niej, że najistotniejszy wpływ miało pytanie o zainteresowanie młodzieży sportem. Wysoka pozycja tego czynnika cieszy, ponieważ uprawianie sportu, którego w pewnej mierze jest się fanem, pomaga osiągać lepsze wyniki i sprawiać większą satysfakcje. Kolejnymi istotnymi czynnikami są atletyczna figura, popularność oraz przyszła kariera jako sportowiec. Najmniejszą istotność mają zmienne dotyczące czerpania radości dla sportu, płci i wady postawy.

W modelu drzewa regresyjnego i klasyfikacyjnego zdecydowanie największy wpływ na model ma pytanie o sławę. Im pragnienie zdobycia popularności jest większe, tym częściej młodzi ludzie wybierają się do klasy o takim profilu. Pojęcie sławy i pieniędzy często się łączą, w tym przypadku także można zauważyć związek, ponieważ pytanie o majętność wpływa dość znacząco na decyzję dzieci. Niestety płace sportowców bardzo się różnią w szczególności, gdy porówna się pensję piłkarzy z najwyższych klas rozgrywkowych do wynagrodzenia lekkoatletów, ale dzieci mogą sobie nie zdawać z tego sprawy.

Stan zdrowia oraz wpływ otoczenia, czyli rówieśników i rodziców, nie były istotnymi elementami podczas dokonywania wyboru klasy sportowej przez dzieci. Zdawać by się mogło, że właśnie ta druga zmienna będzie miała znaczący wpływ na decyzję młodzieży, ponieważ uczęszczanie na zajęcia sportowe w grupie znajomych, jest przyjemniejsze, a nawet pożyteczniejsze, ze względu na fakt, że dzieci uczą się współzawodnictwa.

Bibliografia

1. Bujak Ł., Drzewa decyzyjne, Uniwersytet Mikołaja Kopernika, Toruń 2008, http://www.is.umk.pl/~duch/Wyklady/CIS/Prace%20zalicz/08-Bujak.pdf [dostęp: 10.06.2020]
2. Chmielewski M., *Przyszedł kryzys, likwidują klasy sportowe*, „Przegląd Sportowy" 20.05.2020
3. Demichowicz M., Mazur P., Drzewa decyzyjne, „Automatyczne pozyskiwanie wiedzy", 18.05.2003 https://www.ii.pwr.edu.pl/~kwasnicka/tekstystudenckie/apw/decyzyjne.htm [dostęp: 10.06.2020]
4. Dixon S., Pandemic costing youth sports millions, creating uncertainty, „Associated Press", 18.04.2020 https://apnews.com/e1f97b5e44ec5d7a742e0845a5eca7c6 [dostęp: 08.06.2020]
5. Hyndman R., Koehler A., Another look at measures of forecast accuracy, 2005
6. Instytut Matki i Dziecka, Zakład Zdrowia Dzieci i Młodzieży, „Aktywność fizyczna młodzieży szkolnej w wieku 9-17 lat", Warszawa 2013
7. James G.., Witten D., Hastie T., Tibshirani R., *Tree-Based Methods,* „An Introduction to Statistical Learning" Springer Science + Business Media New York 11.02.2013
8. Le J., Decision Trees in R, „Data Camp", 19.06.2018 https://www.datacamp.com/community/tutorials/decision-trees-R [dostęp: 09.06.2020]
9. Marynowski P., Drzewa decyzyjne, http://home.agh.edu.pl/~pmarynow/pliki/iwmet/drzewa.pdf [dostęp: 10.06.2020]
10. Ośrodek Statystyki Sportu i Turystyki oraz Podkarpacki Ośrodek Badań Regionalnych, Kultura Fizyczna w Polsce w latach 2017 i 2018, GUS, Departament Badań Społecznych i Warunków Życia, Warszawa-Rzeszów 2019.
11. Pociecha J., *Współczesne zmiany narzędzi badań statystycznych*, Zeszyty Naukowe, Uniwersytet Ekonomiczny w Krakowie, Kraków
12. Rozporządzenie Ministra Edukacji Narodowej z dnia 27 marca 2017 r. w sprawie warunków tworzenia, organizacji oraz działania oddziałów sportowych, szkół sportowych oraz szkół mistrzostwa sportowego (Dz.U. z 2017 r. poz. 671)

13. Spotlight on adolescent health and well-being, WHO 2020. https://apps.who.int/iris/bitstream/handle/10665/332104/9789289055017-eng.pdf [dostęp: 08.06.2020]

Spis tabel

Spis rysunków

Załączniki

Załącznik 1. Link źródłowy