**PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU**
**RESEARCH PAPERS OF WROCŁAW UNIVERSITY OF ECONOMICS nr 469 • 2017**

Taksonomia 29
Klasyfikacja i analiza danych – teoria i zastosowania

ISSN 1899-3192
e-ISSN 2392-0041

**Paweł Sroka, Joanna Trzęsiok**
Uniwersytet Ekonomiczny w Katowicach
e-mails: pawel.sroka0@gmail.com; joanna.trzesiok@ue.katowice.pl

# CO OPOWIADAJĄ DRZEWA O TENISIE? PREDYKCJA WYNIKÓW SPOTKAŃ W TENISIE ZIEMNYM Z WYKORZYSTANIEM DRZEW KLASYFIKACYJNYCH

# WHAT DO TREES HAVE GOT TO SAY ABOUT TENNIS? PREDICTION OF PROFESSIONAL TENNIS MATCHES OUTCOMES USING CLASSIFICATION TREES

DOI: 10.15611/pn.2017.469.17
JEL Classification: C14

**Streszczenie:** W artykule przedstawiono problem dyskryminacji wyników spotkań w profesjonalnym tenisie ziemnym z wykorzystaniem metody *Random Forests*. Celem było zbudowanie modelu charakteryzującego się wyższą dokładnością predykcji meczów niż rynkowy model firm bukmacherskich. Analizy przeprowadzono na autorskich zbiorach danych rzeczywistych, zawierających wybrane charakterystyki opisujące mecze z turniejów tenisowych, jakie były rozegrane w 2015 r. przez zawodników notowanych w oficjalnych rankingach ATP i WTA. Przekształcenie wyniku każdego meczu, tak by przedstawić go w postaci zmiennej metrycznej, i utworzenie na tej podstawie dodatkowych zmiennych objaśniających, dających pełniejszy opis zarówno meczów, jak i zawodników, pozwoliło na zastosowanie metod klasyfikacji w sporcie w sposób, jaki nie był jeszcze przedstawiany w literaturze.

**Słowa kluczowe:** predykcja wyników meczów, tenis ziemny, model dyskryminacyjny, drzewa klasyfikacyjne.

**Summary:** The paper presents the classifications of matches outcomes in professional tennis, using Random Forests. The main goal is to build the model with higher prediction accuracy than bookmakers' model. The original real world data sets are analyzed. The objects in these data sets are the matches, played in 2015 by the players listed in the official ATP and WTA rankings. There are some special variables created based on a metric variable which is a quantitative representation of the match result. The paper presents a novelty use of classification trees in predicting tennis matches outcomes.

**Keywords:** prediction of tennis matches, tennis, discriminatory model, classification trees.

# 1. Wstęp

Niespotykana jak dotąd skala wzrostu ilości gromadzonych i przechowywanych informacji powoduje, iż coraz częściej konieczne do ich analizy jest wykorzystywanie odpowiednich metod eksploracji danych. Jednym z obszarów, w którym można odnotować znaczący wzrost zainteresowań tymi metodami, jest sport. W pracy zaprezentowany został problem dyskryminacji wyników spotkań w profesjonalnym tenisie ziemnym z wykorzystaniem modeli drzew klasyfikacyjnych.

Celem artykułu było przedstawienie modelu w postaci zagregowanych drzew klasyfikacyjnych, który charakteryzuje się wyższą dokładnością predykcji wyników meczów niż rynkowy model firm bukmacherskich, utożsamiany z kursami wystawianymi na zdarzenia sportowe i pozwalający na wyłonienie faworyta spotkania tenisowego.

Warto podkreślić, że model przedstawiony w pracy został zbudowany na autorskich zbiorach danych rzeczywistych, w których zmienne utworzono bazując wyłącznie na ogólnie dostępnych wynikach meczów tenisa ziemnego. Również koncepcja wykorzystania w badaniu zagregowanych drzew klasyfikacyjnych jest oryginalnym wkładem autorów, gdyż w literaturze nie spotkano dotychczas opracowania, w którym stosuje się takie podejście w analizie danych sportowych.

# 2. Charakterystyka zbiorów danych i opis metody

## 2.1. Źródła danych i materiał do badań

Jak już wspomniano, analizowano zbiory danych rzeczywistych, w których uwzględniono autorskie zmienne objaśniające. Obiektami w tych zbiorach były wszystkie możliwe mecze z turniejów tenisowych rozegranych w 2015 r. przez zawodników notowanych w oficjalnych rankingach ATP i WTA[1]. Przy czym, ze względu na odrębną specyfikę, osobno analizowano mecze kobiet i mężczyzn. Po odfiltrowaniu i usunięciu obiektów z brakującymi wartościami niektórych zmiennych, otrzymano zbiory o liczebnościach: 9573 dla mężczyzn (ATP) i 5855 dla kobiet (WTA).

Każdy mecz analizowany był z punktu widzenia faworyta, którego zidentyfikowano na podstawie wystawianych przez firmy bukmacherskie kursów, będących odwrotnością ilorazów szans (mówiących o szansach wystąpienia danego zdarzenia). Dlatego też zmienna objaśniana *wynik* przyjmowała 2 kategorie: wygra (*win*) i przegra (*lost*), a wskazania zbudowanego modelu odnosiły się zawsze do faworyta spotkania.

[1] Badaniom poddano wszystkie możliwe mecze z takich typów turniejów, jak: ATP Tour, WTA Tour, Davis Cup, Fed Cup, Challenger, ITF Futures. Oznacza to, że zbudowany model może zostać wykorzystany również do predykcji wyników spotkań zawodników nieznanych szerszej publiczności, grających w turniejach niższej rangi.

Większość zmiennych objaśniających, charakteryzujących badane mecze, obliczono w oparciu o zmienną metryczną dającą wgląd w rozmiar zwycięstwa czy porażki zawodnika. Zmienna ta ma bardzo duże znaczenie w badaniu, gdyż pozwala na pełniejsze analizowanie poszczególnych meczów, dlatego też w dalszej części pracy przedstawiono jej konstrukcję.

## 2.2. Konstrukcja zmiennych

Rezultat zdarzenia sportowego można przedstawić w dwojaki sposób. Przede wszystkim poprzez podanie nazwy zwycięzcy, bez szczegółów dotyczących wyniku. Jednak możliwe jest również skonstruowanie zmiennej metrycznej pozwalającej na określenie rozmiaru zwycięstwa czy porażki.

Dyscyplina sportowa, jaką jest tenis ziemny, posiada ten kluczowy atut, że w zdecydowanej większości przypadków[2] ma skończoną liczbę kombinacji możliwych wyników spotkań. Poszczególne spotkania, rozgrywane do 2 wygranych setów, mogą zakończyć się wynikiem 2:0 lub 2:1. Podobnie jest w poszczególnych setach, gdzie przy założeniu, iż w trakcie meczu żaden z zawodników nie podda meczu, możliwe są następujące kombinacje wyników: 6-0, 6-1, 6-2, 6-3, 6-4, 7-5, 7-6. Skończona liczba takich kombinacji umożliwia intuicyjne wypunktowanie wyników w pojedynczych setach. Taką propozycję przedstawiono w tabeli 1.

**Tabela 1.** Propozycja punktacji wyników gemowych w poszczególnych setach

| Wyniki w gemach | Punkty wygrywającego | Punkty przegrywającego |
|---|---|---|
| 6-0 | 1 | 0 |
| 6-1 | 0,9 | 0,1 |
| 6-2 | 0,8 | 0,2 |
| 6-3 | 0,7 | 0,3 |
| 6-4 | 0,6 | 0,4 |
| 7-5 | 0,55 | 0,45 |
| 7-6 | 0,52 | 0,48 |

Źródło: opracowanie własne.

Z racji, iż możliwe jest, że zawodnik wygrywający więcej punktów w całym spotkaniu nie zawsze zostaje triumfatorem meczu, propozycja zawarta w tabeli 1 niekoniecznie musi być optymalna[3]. Potwierdzają to analizy przeprowadzone w ar-

[2] Nie włączając w to niektórych spotkań rozgrywanych w tzw. Wielkich Szlemach.

[3] Rozważanym kolejnym etapem procesu punktacji wyników spotkań może być np. dodanie wagi uwzględniającej prestiż rozgrywanego turnieju bądź rundy, w której rozegrany został dany mecz. Niewątpliwie spotkania w Wielkich Szlemach mają niepodważalnie wyższą renomę aniżeli finał niejednego turnieju tenisowego.

tykule [Wright i in. 2013], gdzie spośród 61 tys. przeanalizowanych spotkań męskiego tenisa, w latach 1991-2011, 2794 (co daje 4,52%) spotkania zaklasyfikowano jako tzw. paradoks Simpsona, kiedy to triumfujący tenisista z wszystkich rozegranych podczas spotkania punktów zdobył mniej niż połowę.

Po nadaniu zawodnikowi punktów za rozegrany set kolejnym krokiem będzie uśrednienie uzyskanych w powyższy sposób ocen w zależności od liczby rozegranych setów. Zaletą tego typu podejścia wydaje się fakt, iż wartość końcowego wyniku dla całego spotkania znajduje się w przedziale (0, 1) co umożliwia dość łatwą i intuicyjną interpretację.

Należy jednak zwrócić uwagę, że uzyskany przez zawodnika, w różnych meczach, ten sam wynik może mieć dla niego zupełnie inną wartość w zależności od tego, z jak „silnym" rywalem rozgrywał ten mecz. Zdecydowanie ważniejsze jest dla niego zwycięstwo z zawodnikiem, który jest wyżej notowany w oficjalnych rankingach ATP czy WTA. Uwzględniając zatem pozycje rankingowe przeciwników, zaproponowano przyjęcie następujących wag dla wyników meczów:

$$w_i = \frac{N - ID}{N}, \tag{1}$$

gdzie: $i$ to numer meczu, $N$ – liczba wszystkich zawodników notowanych w rankingu, zaś $D$ – pozycja rankingowa przeciwnika.

Przykład zamieszony w tabeli 2 ilustruje sytuację, w której zawodnik wygrywa 2 mecze z tym samym wynikiem punktowym 0,77. Jednak po uwzględnieniu pozycji rankingowej rywali i tym samym odpowiedniej wagi, ostatecznie za pierwszy mecz otrzymuje 0,73 punktu, zaś za drugi – 0,36 punktu.

**Tabela 2.** Przykład pokazujący konstrukcję zmiennej metrycznej – wyniku ważonego

| Wynik meczu | Wynik jakośc. | Liczba setów | Punkty 1 set | Punkty 2 set | Punkty 3 set | Wynik średnia | Ranking rywala | Waga | Wynik ważony |
|---|---|---|---|---|---|---|---|---|---|
| 6-0, 4-6, 6-1 | wygra | 3 | 1 | 0,4 | 0,9 | 0,77 | 114 | 0,95 | 0,73 |
| 3-6, 6-0, 6-0 | wygra | 3 | 0,3 | 1 | 1 | 0,77 | 1195 | 0,47 | 0,36 |

Źródło: opracowanie własne

Na bazie tak skonstruowanej dodatkowej zmiennej metrycznej – wyniku ważonego, utworzono wiele autorskich zmiennych objaśniających. Zestawienie zmiennych wprowadzonych do modelu dyskryminacyjnego zawiera tabela 3.

**Tabela 3.** Zmienne wraz z opisem

| Nazwa zmiennej | Opis zmiennej |
|---|---|
| Wynik | Zmienna objaśniana o kategoriach: wygra (*win*) i przegra (*lost*) |
| Round | Runda rozgrywanego meczu w danym turnieju (zm. niemetryczna) |
| Surface | Nawierzchnia, na której jest rozgrywany mecz (zm. niemetryczna) |
| SameOPPointsAll | Różnica średnich wartości punktów uzyskanych przez danego zawodnika przeciwko tzw. wspólnym przeciwnikom podzielona przez liczbę wspólnych rywali |
| MeanPktAllthisYear | Różnica średnich wartości punktów spośród spotkań rozegranych w 2015 roku |
| MeanPktAlllastYear | Jw. od 2014 roku |
| MeanPktAllSamethisYear | Różnica średnich wartości punktów spośród spotkań rozegranych w 2015 roku w tym samym przedziale pozycji rankingowych przeciwników |
| MeanPktAllSamelastYear | Jw. od 2014 roku |
| MeanPktSurfacethisYear | Różnica średnich wartości punktów spośród spotkań rozegranych w 2015 roku na nawierzchni rozgrywanego spotkania |
| MeanPktSurfacelastYear | Jw. od 2014 roku |
| MeanPktSurfaceSamethisYear | Różnica średnich wartości punktów spośród spotkań rozegranych w 2015 roku na nawierzchni rozgrywanego spotkania w tym samym przedziale pozycji rankingowych przeciwników |
| MeanPktSurfaceSamelastYear | Jw. od 2014 roku |
| MeanPktAllWinthisYear | Różnica średnich wartości punktów spośród spotkań rozegranych w 2015 roku |
| MeanPktAllWinlastYear | Jw. od 2014 roku |
| DIFFBestWin | Różnica pozycji rankingowych najwyżej rozstawionego pokonanego rywala przez każdego z graczy |
| MeanPktLast45 | Różnica średnich wartości punktów spośród rozegranych spotkań w ostatnich 45 dniach, liczonych od dnia poprzedzającego analizowany mecz |
| MeanPktLast45P1 | Średnia wartość punktów spośród rozegranych spotkań w ostatnich 45 dniach, liczonych od dnia poprzedzającego analizowany mecz |
| MeanPktLast45P2 | Jw. nie faworyt |
| MeanPktLast45Same | Różnica średnich wartości punktów spośród rozegranych spotkań w ostatnich 45 dniach, liczonych od dnia poprzedzającego analizowany mecz, w tym samym przedziale pozycji rankingowych przeciwników |
| DIFFMeanLast45WeightPerfOPP | Różnica średnich wartości punktów spośród rozegranych spotkań w ostatnich 45 dniach, liczonych od dnia poprzedzającego analizowany mecz. Wartości punktów ważone formą, jaką prezentują przeciwnicy w ostatnich 45 dniach. Forma jest utożsamiana ze średnią wartością punktów uzyskanych w okresie ostatnich 45 dni |
| DIFFRegrAlfaSamelastYear | Różnica współczynników α w modelu regresji liniowej, gdzie zmienną objaśnianą są punkty, a zmienną objaśniającą pozycja rankingowa przeciwnika. Dane dotyczą spotkań rozegranych od 2014 roku w tym samym przedziale pozycji rankingowych przeciwników |
| DIFFIntegralSamelastYear | Różnica pól powierzchni pod krzywą regresji liniowej spośród wszystkich rozegranych spotkań w tym samym przedziale pozycji rankingowych przeciwników |

Źródło: opracowanie własne.

## 2.3. Metoda badań

Do budowy modeli dyskryminacyjnych zastosowano jedną z metod zagregowanych drzew klasyfikacyjnych – algorytm *Random Forests*, zaproponowany przez Breimana [2001].

Metoda *Random Forests* oparta jest na równoległym łączeniu wyników predykcji modeli składowych, które w tym przypadku są drzewami klasyfikacyjnymi. W metodzie tej dwukrotnie wykorzystuje się element losowania. Oprócz losowego doboru obserwacji do bootstrapowych prób uczących $U_1,\ldots,U_M$, na których budowane są modele składowe, losuje się również w każdym węźle drzewa zmienne objaśniające, spośród których algorytm wskazuje i wybiera najlepszą.

Kroki algorytm *Random Forests* przedstawiono w tabeli 4. Natomiast szczegółowo metoda ta została opisana m.in. w pracy [Gatnar 2008].

**Tabela 4.** Kroki algorytmu *Random Forests*

| | |
|---|---|
| 1. | Ustal liczbę modeli składowych $M$ oraz liczbę losowanych zmiennych $K$. |
| 2. | Dla $m = 1, ..., M$ wykonaj następujące kroki:<br>a) wylosuj próbę bootstrapową $U_m$ ze zbioru uczącego $U$,<br>b) zbuduj drzewo dyskryminacyjne $T_m$ na podstawie zbioru $U_m$, losując w każdym węźle drzewa $K$ zmiennych, spośród których do modelu $T_m$ wprowadzana jest tylko ta, która minimalizuje wartość przyjętej miary heterogeniczności. |
| 3. | Dokonaj predykcji na podstawie modelu zagregowanego, wykorzystując regułę majoryzacji. |

Źródło: opracowanie na podstawie [Gatnar 2008].

W zadaniu dyskryminacji proces oceny jakości modelu odbywa się poprzez zliczanie obserwacji poprawnie bądź błędnie zaklasyfikowanych. Proces ten można przedstawić za pomocą tzw. macierzy pomyłek (klasyfikacji). W przypadku gdy zmienna objaśniana ma tylko 2 kategorie i jedną z nich wyróżnimy jako interesującą nas klasę, to macierz pomyłek można zapisać w postaci tablicy kwadratowej o wymiarze $2 \times 2$, jak zaprezentowano w tabeli 5.

**Tabela 5.** Macierz pomyłek

| Stan przewidywany | Stan obserwowany | |
|---|---|---|
| | klasa wyróżniona (*P*) | klasa niewyróżniona (*N*) |
| Klasa wyróżniona (*P*) | *TP* (*true positives*)<br>obiekty z klasy wyróżnionej<br>poprawnie zaklasyfikowane | *FP* (*false positives*)<br>obiekty z klasy niewyróżnionej<br>błędnie zaklasyfikowane |
| Klasa niewyróżniona (*N*) | *FN* (*false negatives*)<br>obiekty z klasy wyróżnionej<br>błędnie zaklasyfikowane | *TN* (*true negatives*)<br>obiekty z klasy niewyróżnionej<br>poprawnie zaklasyfikowane |

Źródło: opracowanie własne na podstawie [Misztal 2014] .

Miarami najczęściej stosowanymi do oceny jakości modelu dyskryminacji, obliczanymi na podstawie macierzy pomyłek, są [Fielding 2007; Misztal 2014]:

- błąd klasyfikacji

$$ERR = \frac{FP + FN}{TP + FP + FN + TN}, \tag{2}$$

- dokładność modelu

$$ACC = \frac{TP + TN}{TP + FP + FN + TN}. \tag{3}$$

Niestety, miary te nie są najlepszymi miernikami oceny jakości klasyfikatora, chociażby w przypadku niezrównoważonych liczebnie klas. Jeśli jedna klasa ma zdecydowanie więcej elementów niż druga, to model, dążąc do minimalizacji błędu klasyfikacji (lub równoważnie – maksymalizacji dokładności), może w skrajnych przypadkach nawet wszystkie obiekty przydzielać do tej właśnie klasy większościowej. Dlatego też w literaturze zaproponowano inne miary pozwalające ocenić zdolność predykcyjną modelu. Są to między innymi [Fielding 2007; Misztal 2014]:

- czułość (*sensitivity*)

$$TPR = \frac{TP}{TP + FN}, \tag{4}$$

czyli odsetek poprawnie zidentyfikowanych obiektów z klasy wyróżnionej;

- specyficzność (*specificity*)

$$TNR = \frac{TN}{FP + TN}, \tag{5}$$

czyli odsetek poprawnie zidentyfikowanych obiektów klasy niewyróżnionej;

- dodatnia zdolność predykcyjna (*positive predictive value*)

$$PPV = \frac{TP}{TP + FP}, \tag{6}$$

określająca skuteczność modelu w predykcji klasy wyróżnionej;

- ujemna zdolność predykcyjna (*negative predictive value*)

$$NPV = \frac{TN}{FN + TN}, \tag{7}$$

określająca skuteczność modelu w predykcji klasy niewyróżnionej.

## 3. Opis i wyniki przeprowadzonej analizy

W przeprowadzonej analizie za pomocą algorytmu *Random Forests* budowano modele dyskryminacyjne dla meczów tenisa ziemnego, w podziale na tenis kobiecy i męski. Jak już wcześniej wspomniano, w zbiorze danych dla kobiet odnotowano 5855 obiektów, zaś w przypadku mężczyzn 9573. Odsetek klasy wyróżnionej (*win*) w zbiorze kobiet (WTA) wyniósł 67,68%, a w zbiorze mężczyzn (ATP) – 70,29%. Liczby te należy utożsamiać z dokładnością predykcji modelu rynkowego z punktu widzenia faworyta meczu. Wartości te oznaczają więc, iż w przypadku mężczyzn 70,29% meczów tenisa wygrywają faworyzowani przez bukmacherów zawodnicy. W tenisie kobiecym odsetek ten jest nieco niższy i wynosi 67,68%. Podane odsetki klasy wyróżnionej są ważne, ponieważ można je traktować jako wartości dodatniej zdolności predykcyjnej (*PPV*) modelu rynkowego.

Do budowy modeli wykorzystano funkcję `randomForest` z biblioteki o tej samej nazwie (z programu statystycznego R), przyjmując następujące wartości parametrów:

- liczbę drzew ($M$) w zagregowanym modelu ntree = 300
- liczbę losowanych zmiennych ($K$) w każdym węźle pojedynczego drzewa $\text{mtry} \approx \sqrt{L}$, gdzie $L = 21$ to liczba zmiennych objaśniających.

W celu oszacowania dokładności klasyfikacji oraz innych miar oceniających jakość modelu skorzystano z procedury sprawdzania krzyżowego, dzieląc każdy ze zbiorów danych na 10 losowych i w przybliżeniu równolicznych podzbiorów, utożsamianych ze zbiorami testowymi.

W pierwszym etapie analizy otrzymywano modele, które były poniekąd kopią modeli rynkowych, a ponadto klasyfikowały niemal wszystkie obiekty do klasy większościowej. Było to konsekwencją wykorzystania zbiorów uczących, w których występował problem niezrównoważonych liczebnie klas[4].

Jednym z możliwych rozwiązań takiej sytuacji była próba wyrównania frakcji klas w zbiorze uczącym poprzez wylosowanie podzbioru obiektów, o zadanej liczebności, z klasy większościowej (tzw. technika *down-sampling*) [Kuhn, Johnson 2013]. W tym przypadku zastosowano podejście polegające na zmniejszeniu liczebności klasy wygra (*win*). Przyjęto, że liczebność ta będzie równa liczbie elementów klasy *lost* w zbiorze uczącym. Zbiór testowy pozostał w swojej pierwotnej wersji, tzn. proporcje klas nie uległy zmianie. Zastosowanie opisanej procedury nie było problematyczne, ponieważ algorytm `randomForest` posiada parametr `sampSize`, który odpowiada za ustalenie liczebności każdej z klas zmiennej objaśnianej, która będzie wylosowana w sposób warstwowy z powtórzeniami z pierwotnego zbioru danych.

[4] Jak podano, 70,29% meczów mężczyzn i 67,68% meczów kobiet znajduje się w klasie wyróżnionej.

Stosując metodę próbkowania, wyrównującą liczebności klas (technikę *down--sampling*), otrzymano modele, dla których obliczono miary jakości klasyfikacji (wzory 3-7). Zestawienie wyników przedstawiono w tabeli 6.

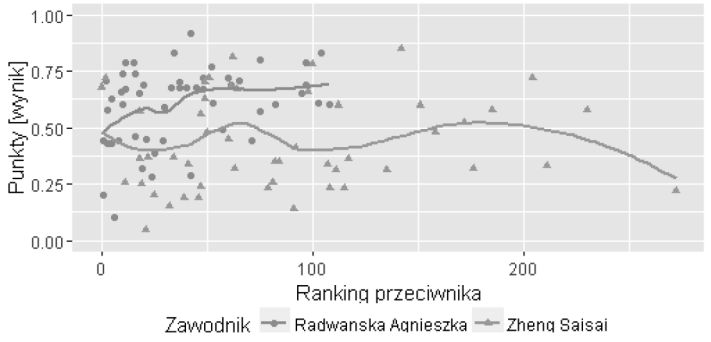
**Tabela 6.** Wartości miar jakości klasyfikacji uzyskanych modeli

| Miara | Model WTA | | Model ATP | |
|---|---|---|---|---|
| | średnia (%) | odchylenie standardowe (%) | średnia (%) | odchylenie standardowe (%) |
| ACC | 62,53 | 1,67 | 63,62 | 1,66 |
| PPV | 73,54 | 1,39 | 75,68 | 1,15 |
| NPV | 42,78 | 2,36 | 40,24 | 2,34 |
| TPR | 69,73 | 2,01 | 71,10 | 2,02 |
| TNR | 47,41 | 3,71 | 45,98 | 3,15 |
| Model rynkowy | 67,68 | | 70,29 | |

Źródło: opracowanie własne.

Jak już wspomniano, analizując wyniki otrzymane dla modeli utworzonych za pomocą metody *Random Forests*, należy odsetek klasy wyróżnionej, poprawnie zidentyfikowanej przez bukmacherów, porównać z dodatnią zdolnością predykcyjną (*PPV*) modeli WTA i ATP. Jak łatwo zauważyć, w obu przypadkach odnotowano wzrost skuteczności modeli w klasyfikacji wyników spotkań, w których wygra faworyt, o około 5 punktów procentowych w stosunku do modelu rynkowego.

Wartości miar czułości (*TPR*) oraz specyficzności (*TNR*) sugerują, iż otrzymano zrównoważone modele, mające zdolność do wykrywania spotkań zarówno wygranych, jak i przegranych przez faworyta. Zaobserwowano znaczącą poprawę




**Rys. 1.** Statystyki przedmeczowe (dane z ostatnich 11 miesięcy) spotkania Agnieszki Radwańskiej z Zheng Saisai podczas Letnich Igrzysk Olimpijskich 2016, w którym Agnieszka Radwańska przegrała 4:6, 5:7.

Źródło: opracowanie własne.

w stosunku do modeli zbudowanych bez zastosowania techniki próbkowania wyrównującej liczebności klas, szczególnie dla miary *TNR* (czterokrotny wzrost miary specyficzności dla mężczyzn i sześciokrotny dla kobiet). Tamte modele wykrywały tylko około 10% meczów, w których faworyt przegrywał. W tym przypadku udaje się zidentyfikować niemal połowę takich meczów.

Niewątpliwie wartością dodaną przeprowadzonych analiz jest również utworzona zmienna metryczna, będąca odpowiednikiem jakościowego wyniku meczu tenisowego, określająca rozmiary zwycięstwa bądź porażki jednego z zawodników. Przedstawienie wyniku w takiej postaci otwiera zupełnie nowe możliwości analizowania i wizualizowania poszczególnych spotkań (rys. 1).

## 4. Zakończenie

W artykule przedstawiono problem dyskryminacji wyników spotkań w profesjonalnym tenisie ziemnym z wykorzystaniem metody *Random Forests*. Celem było zbudowanie modelu charakteryzującego się wyższą dokładnością predykcji meczów, w których wygra faworyt, niż rynkowy model firm bukmacherskich.

Można powiedzieć, że cel został zrealizowany, gdyż otrzymano modele (osobno dla mężczyzn i kobiet), dla których wzrosła o 5 punktów procentowych dodatnia zdolność predykcyjna, a więc skuteczność w predykcji meczów, w których wygrywa faworyzowany zawodnik. Biorąc pod uwagę specyfikę analizowanego zbioru danych, występowanie obserwacji odstających oraz duże zróżnicowanie wartości zmiennych (ponieważ uwzględniono mecze wszystkich zawodników notowanych w oficjalnych rankingach WTA i ATP), należy uznać uzyskane wyniki za satysfakcjonujące.

## Literatura

Breiman L., 2001, *Random Forests*, Machine Learning, no. 45, s. 5-32.

Fielding A.H., 2007, *Cluster and Classification Techniques for the Biosciences*, Cambridge University Press, Cambridge.

Gatnar E., 2008, *Podejście wielomodelowe w zagadnieniach dyskryminacji i regresji*, Wydawnictwo Naukowe PWN, Warszawa.

Kuhn M., Johnson K., 2013, *Applied Predictive Modeling*, Springer, New York.

Misztal M., 2014, *Wybrane metody oceny jakości klasyfikatorów – przegląd i przykłady zastosowań*, Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu, nr 328, Taksonomia 23, s. 156-166.

Wright B., Rodenberg R.M., Sackman J., 2013, *Incentives in Best of N Contests: Quasi-Simpson's Paradox in Tennis*, www.papers.ssrn.com (11.12.2015).