Wiadomości Statystyczne. The Polish Statistician, 2022, vol. 67, 12, 90–93
Dyskusje. Recenzje. Informacje / Discussions. Reviews. Information

XXVI Konferencja Naukowa „Zastosowania statystyki i data mining w badaniach naukowych"

The 26th Scientific Conference 'Applications of statistics and data mining in scientific research'

19 października 2022 r. odbyła się XXVI Konferencja Naukowa „Zastosowania statystyki i data mining w badaniach naukowych" zorganizowana przez firmę StatSoft Polska. Patronat nad konferencją objęło Polskie Towarzystwo Statystyczne, a patronami medialnymi zostały: „Wiedza i Życie", „Świat Nauki", „Forum Akademickie" i „Wiadomości Statystyczne. The Polish Statistician". Spotkanie miało posłużyć jako forum wymiany myśli i doświadczeń, a także przyczynić się do integracji środowiska polskich naukowców zajmujących się analizą danych. Interdyscyplinarny charakter konferencji dał uczestnikom możliwość szerszego spojrzenia na poprawne zaplanowanie badania oraz dobór odpowiednich metod analizy danych empirycznych. Do udziału w wydarzeniu, odbywającym się w formie zdalnej, zgłosiło się ponad tysiąc osób.

Prelegenci – wykładowcy renomowanych polskich ośrodków naukowych – wskazywali, w jaki sposób metody analizy danych wykorzystuje się w następujących obszarach: neuromonitorowanie pacjentów z urazami mózgu; analiza wyników testów, ankiet i kwestionariuszy w badaniach społecznych; wpływ surowca i technologii na kształtowanie jakości produktów (na przykładzie miodów pitnych); odkrycia naukowe z wykorzystaniem zasobów big data i narzędzi sztucznej inteligencji oraz ocena wybranych czynników rokowniczych w rozwoju inwazyjnego nowotworu piersi. Przedstawiciel firmy StatSoft Polska omówił nowości wprowadzane do zestawów Statistica – narzędzi stworzonych w odpowiedzi na konkretne potrzeby badaczy, które ułatwiają i przyspieszają proces analizy danych.

W wystąpieniu *Statystyczna analiza danych w neuromonitorowaniu pacjentów po urazach mózgu* dr hab. inż. Magdalena Kasprowicz (Politechnika Wrocławska) przedstawiła wyniki swoich badań nad urazowym uszkodzeniem mózgu (ang. *traumatic brain injury* – TBI), które jest główną przyczyną zgonów i niepełnosprawności na całym świecie. TBI często prowadzi do wzrostu ciśnienia wewnątrzczaszkowego (ang. *intracranial pressure* – ICP) powstającego w wyniku narastania dodatkowej, patologicznej objętości wewnątrzczaszkowej, co może skutkować wtórnym uszkodzeniem mózgu z powodu niedokrwienia lub wklinowania struktur mózgowych w obrębie sklepienia czaszki. W badaniach zwrócono uwagę na podatność mózgową, która określa zdolność układu czaszkowo-rdzeniowego do kompensowa-

nia wzrostów objętości mózgowej. Pacjenci z obniżoną podatnością są narażeni na nieproporcjonalny wzrost ICP, dlatego monitorowanie nie tylko ICP, ale także podatności systemu może mieć wartość kliniczną w postępowaniu z pacjentami, którzy doznali TBI. Podczas wystąpienia prelegentka omówiła techniki oceny stanu podatności mózgowej bazujące na analizie kształtu pulsacji tętniczopochodnej ICP w dziedzinie częstotliwości i czasu oraz z użyciem technik uczenia maszynowego. Przedstawiła wyniki najnowszych badań dotyczących związku wymienionych parametrów z wynikami leczenia i predykcją śmiertelności pacjentów z TBI.

Dr Justyna Brzezińska (Uniwersytet Ekonomiczny w Katowicach), autorka referatu *Modele odpowiedzi na pozycje testowe w badaniach społecznych*, zaprezentowała modele odpowiedzi na pozycje testowe (ang. *item response theory models* – modele IRT) wykorzystywane w badaniach testowych, kwestionariuszach oraz badaniach edukacyjnych, w których sprawdza się stosunek respondenta do różnych kategorii lub pojęć albo bada się jego poziom umiejętności. Prelegentka przybliżyła słuchaczom pojęcie zmiennej ukrytej – pozycji testowej (pytania, zadania, stwierdzenia), względem której respondent wyraża swoje nastawienie. Kluczową rolę w metodzie odgrywa model probabilistyczny, który przedstawia prawdopodobieństwo udzielenia odpowiedzi w zależności od badanej cechy ukrytej (umiejętności). Szczególnie istotną funkcję w modelach IRT pełni krzywa charakterystyczna pozycji testowej (ang. *item characteristic curve* – ICC), która przedstawia związek między poziomem umiejętności a prawdopodobieństwem udzielenia poprawnej odpowiedzi. Obliczenia oparte na rzeczywistych zbiorach danych zostały wykonane w środowisku R.

Przedmiotem referatu dr Marty Bednarek (Uniwersytet Przyrodniczy w Poznaniu) *Uwarunkowania surowcowe i technologiczne w kształtowaniu jakości miodów pitnych* był wpływ surowca i kluczowych etapów procesu technologicznego na wydajność fermentacji oraz wyróżniki jakościowe miodów pitnych typu trójniak. Do zaplanowania niektórych doświadczeń w swoim badaniu wykorzystała takie metody statystyczne, jak plan dwuwartościowy kompletny i plan centralny kompozycyjny z wyznaczeniem płaszczyzny odpowiedzi. Zastosowaną metodą badawczą była analiza skupień. Przeprowadzono grupowanie metodą k-średnich z wykorzystaniem algorytmu maksymalizacji wartości oczekiwanej z v-krotnym sprawdzianem krzyżowym w celu uzyskania grup jednorodnych i oceny mocy dyskryminacyjnej zmiennych zależnych. Aby wykazać korelację pomiędzy kilkoma zmiennymi i przypadkami, zastosowano analizę głównych składowych (PCA), regresję przeprowadzoną metodą cząstkowych najmniejszych kwadratów (PLS) oraz analizę regresji wielorakiej w układach z jedną zmienną zależną.

Dr Marcin Braun (Uniwersytet Medyczny w Łodzi) w wystąpieniu *Znaczenie kliniczne osi FGFR2-PR w inwazyjnym raku piersi* przybliżył słuchaczom wyniki

badania przeprowadzonego wśród pacjentek z inwazyjnym luminalnym rakiem piersi. Niezależny od hormonów steroidowych wzrost guza i wykształcenie oporności na terapię hormonalną to jedne z głównych przyczyn progresji nowotworu i niepowodzeń terapeutycznych wśród tych pacjentek. Badania przedkliniczne wykazały, że interakcja pomiędzy sygnalizacjami od receptora czynnika wzrostu dla fibroblastów 2 (ang. *fibroblast growth factor receptor 2* – FGFR2) i od receptorów hormonów steroidowych (ang. *estrogen and progesterone receptors* – ER i PR) może warunkować oporność na hormonoterapię. Głównym celem omówionego badania było określenie znaczenia prognostycznego ekspresji białka i mRNA genu FGFR2 w guzach pacjentek z inwazyjnym rakiem piersi w kontekście standardowo ocenianych czynników prognostycznych i predykcyjnych, ze szczególnym uwzględnieniem statusu receptorów steroidowych i sygnatury molekularnej związanej z hiperaktywacją PR. Celem pobocznym było dokonanie oceny – na przykładzie immunohistochemicznego oznaczenia białka FGFR2 za pomocą sztucznej inteligencji (ang. *artificial intelligence* – AI) – przydatności AI i uczenia maszynowego w procesie zrewolucjonizowania patomorfologii. Wyniki badania nie potwierdziły przyjętej hipotezy (opartej na badaniach przedklinicznych sugerujących związek pomiędzy podwyższonym poziomem ekspresji FGFR2 i złym rokowaniem u pacjentek z inwazyjnym luminalnym rakiem piersi), przeciwnie – wykazały, że istnieje związek niskich poziomów białka FGFR2 ze słabym zróżnicowaniem i większą agresywnością guzów, a więc gorszym rokowaniem w tej grupie pacjentek. Opisane efekty odnotowano jedynie wśród pacjentek ER+PR+, a wszystkie obserwacje zostały potwierdzone na poziomie mRNA w niezależnych zewnętrznych bazach danych. Uzyskane wyniki świadczą o złożoności roli FGFR2 w patogenezie inwazyjnego raka piersi i zależności znaczenia prognostycznego FGFR2 od kontekstu biologicznego. Mogą też tłumaczyć uzyskiwane dotychczas wątpliwe efekty terapeutyczne celowanego hamowania szlaku FGFR2 w badaniach klinicznych. Zanik złego efektu niskiego poziomu FGFR2 w grupie pacjentek ER+PR– pokazuje jednocześnie, że jest grupa pacjentek, które potencjalnie mogą odnieść korzyść terapeutyczną w wyniku hamowania sygnalizacji FGFR2-zależnej. Przeprowadzone badanie wykazało ponadto, że pomimo niewątpliwej użyteczności w pewnej grupie przypadków wykorzystanie AI jako narzędzia w patologii wymaga nadzoru specjalisty patologa.

Prof. Ryszard Tadeusiewicz (Akademia Górniczo-Hutnicza w Krakowie) wprowadził słuchaczy w zagadnienie *Big data i data mining w badaniach naukowych*. Prelegent zwrócił uwagę, że big data to obszar silnie determinujący obraz współczesnej informatyki, i krótko omówił jego najważniejsze cechy. Następnie skupił się na możliwościach dokonywania nowych odkryć naukowych po zastosowaniu do zasobów big data nowoczesnych narzędzi sztucznej inteligencji składających się na me-

todykę określaną jako data mining. Podał dwa konkretne przykłady uzyskania bardzo wartościowych i oryginalnych wyników dzięki eksplorowaniu technikami data mining danych, które są publicznie dostępne, ale ich ogromna ilość, zmienność w czasie i różnorodność utrudniała wnioskowanie. Osoby, które stawiły czoła charakterystycznemu dla big data „problemowi 3 V" (*volume*, *velocity*, *variety*), uzyskały oryginalne i wartościowe wyniki. Prelegent zachęcał słuchaczy, aby stosowali metody data mining do analizy wyników naukowych zgromadzonych przez różnych badaczy podczas prowadzenia badań zmierzających do określonych – z góry zdefiniowanych – celów naukowych, które w kontekście tych celów całkowicie wyeksploatowano. Jest bardzo prawdopodobne, że „hałdy naukowych odpadów" zawierają cenne informacje prowadzące do zupełnie nowych odkryć naukowych, nieprzewidzianych i nieanalizowanych przez naukowców. Ilość pobocznych informacji w rozważanych wynikach badań empirycznych jest oczywiście znacznie mniejsza niż ilość informacji, dla uzyskania których eksperymenty te planowano i przeprowadzano. Dzięki dużej ilości takich danych (big data) i potężnemu narzędziu analitycznemu (data mining) można jednak dokonywać odkryć naukowych bez przeprowadzania kosztownych i pracochłonnych eksperymentów.

Konferencja stała się także okazją do przedstawienia nowości, które wkrótce pojawią się wśród narzędzi analizy danych oferowanych przez StatSoft Polska. Mgr Paweł Januszewski, analityk reprezentujący organizatora, na konkretnych przykładach zaprezentował m.in.: testy post hoc w jednoczynnikowej analizie wariancji przy założeniu braku jednorodności wariancji, obliczanie ilorazu szans wraz z przedziałem ufności metodą Garta oraz Fagerlanda-Newcombe'a, nieparametryczny przedział ufności dla mediany i ogólnie kwantyla dowolnego rzędu, analizę post hoc dla serii doświadczeń przyrodniczych, test jednorodności macierzy kowariancji w analizie dyskryminacyjnej oraz kilka innych użytecznych metod analizy sporządzania wykresów.

Prezentacje wygłoszone podczas konferencji i nagrania wystąpień są dostępne na stronie www.statsoft.pl.

Agnieszka Jabłońska
StatSoft Polska