Metody Odkrywania Wiedzy

Przewidywanie spożycia alkoholu przez studentów przy użyciu lasu losowego

Dokumentacja końcowa

Joanna Kiesiak, Julia Kłos

17 maja 2020 r.

1 Opis projektu

Celem badań jest porównanie dwóch modeli predykcyjnych - jeden model zbudowany z lasu losowego oraz drugi model zbudowany z klasyfikatora bayesowskiego. W tym celu zbiór danych zostanie podzielony na zbiór treningowy oraz testowy. Na zbiorze treningowym zostanie wykonane uczenie wspomnianych modeli, a na zbiorze testowym będziemy sprawdzać jak bardzo uzyskany wynik różni się od wartości rzeczywistej.

2 Przeprowadzone badania

W ramach projektu przeprowadzine zostały eksperymenty:

- badanie wpływu ilości drzew, składających się na las losowy na jakość klasyfikacji (w tym las składający się z jednego drzewa);
- ilości zmiennych losowanych w każdym węźle - weryfikacja tezy, że najkorzystniejsze jest wykorzystywanie $\sqrt{iloscatrybutow}$;
- maksymalna głębokości pojedynczego drzewa;
- wpływ losowania przypadków do modelu ze zwracaniem i bez;
- wprowadzenie wag danych klas[1];
- dla naiwnego klasyfikatora bayesowskiego spawdzony zostanie wpływ wygladzenia Laplace'a;

Część danych (ok. 10-20%) zostanie wydzielona jako zbiór testujący, pozostała część będzie stanowiła zbiór uczący. Pozwoli to ocenić, jak dobrze model radzi sobie z uogólnianiem pojęcia.

Ocena modelu predykcyjnego odbędzie się na podstawie poniższych parametrów:

- sprawdzenie dokładności modelu - czyli liczba poprawnie predykowanych próbek do wszystkich próbek;
- sprawdzenie confusion matrix - czyli informacja o klasyfikacji próbek fałszywie i prawdziwie sklasyfikowanych dla poszczególnych klas, co pozwoli ocenić jak model radzi sobie z niezbalansowanymi klasami;
- sprawdzenie dokładności modelu przez użycie walidacji krzyżowej;
- w modelu zastosujemy bootstrapping, czyli K-krotne użycie pierwotnego zbioru danych do uczenia modelu, w tym wypadku zbadamy zagregowany błąd predykowany w agregacie do zagregowanego błędu rzeczywistego;

[1] opcjonalne

3 Las losowy

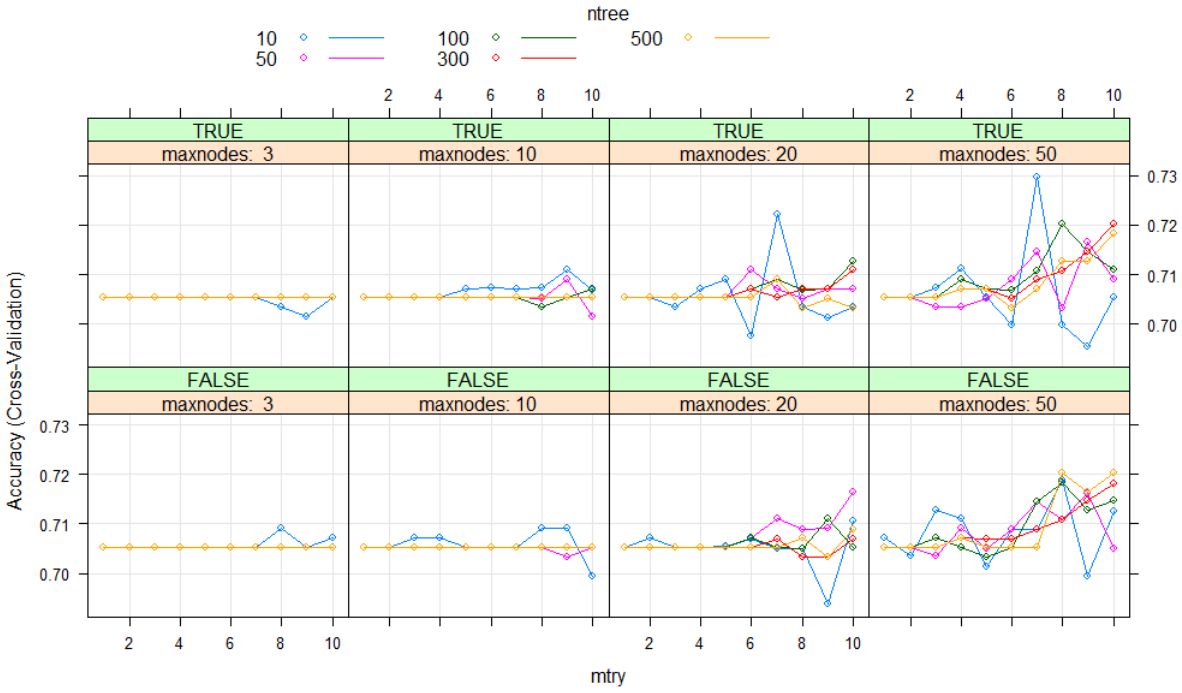
Działanie lasów losowych polega na klasyfikacji za pomocą grupy drzew decyzyjnych. Parametrem, który odpowiada za finalną decyzję jest średnia, gdy przewidywana jest wartość liczbowa lub wynik głosowania dla analizowanej przynależności do klasy. Każde z drzew w lasie losowym jest tworzone w oparciu o próbę, powstałą przez wylosowanie N obiektów ze zbioru uczącego. W każdym węźle danego drzewa podział jest dokonywany na podstawie części losowo wybranych cech, których liczba jest zazwyczaj mniejsza od liczby wszystkich cech. Ma to pozwolić na uzyskanie jak największej niezależności poszczególnych węzłów, czyli zmniejszenie wariancji modelu. Błąd klasyfikacji może być szacowany na podstawie obiektów nie włączonych do próby.

Implementacja algorytmu lasu losowego została wykorzystana przez pakiet *RandomForest*. Przy pierwotnym budowaniu lasu losowego wzięto pod uwagę trzy parametry takie jak:

- ilość drzew (*ntree* = 1, 10,50,100,300,500)
- głębokość pojedynczego drzewa (*maxnodes* = 3,10,20,50)
- ilość atrybutów rozpatrywana przy tworzeniu węzła (*mtry* = 1,2,3,4,5,6,7,8,9,10)
- losowanie próbek do modelu ze zwracaniem i bez (*replace* = TRUE/FALSE)

Do zbadania 240 modelów z wymienionymi parametrami użyto funkcji *train*. Funkcja *train* zwraca dwie ważne wartości: dokładność modelu (Accuracy) oraz współczynnik Kappa. Dokładność modelu to nic innego jak ilość poprawnie predykowanych próbek do ilości wszystkich badanych próbek. Współczynnik Kappa parametr, który zawiera informacje o odtwarzalności lub powtarzalności pomiaru zmiennej w różnych warunkach.

Uzyskane wyniki są przedstawione na rys. 1:



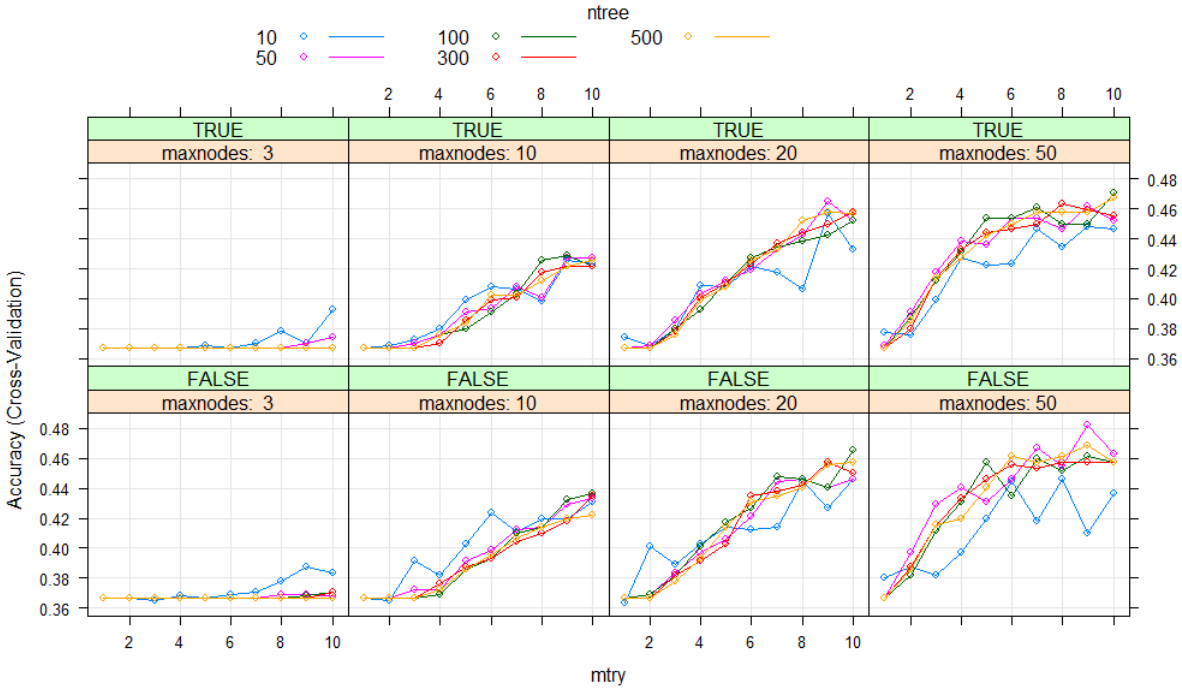
Rysunek 1: Dokładność predykcji modelu dla klasy spożywającej alkohol w tygodniu (**Dalc**) w zależności od różnych kombinacji parametrów.

Na podstawie przeprowadzonych badań dla spożywania alkoholu w tygodniu (**Dalc**) okazuje się, że najlepszy jest w tym przypadku model zbudowany z zaledwie 10 drzew oraz 7 atrybutów wykorzystywanych do wyboru podziału na każdym poziomie, maksymalnie 50 węzłów w drzewie oraz wykorzystywania losowania ze zwracaniem.

W tabeli 1 zestawiono wyniki uzyskane przez ten model. Całkowita wartość błędu **out-of-bag** to 29.62%, co oznacza że dokładnie 70.38% próbek zostało poprawnie sklasyfikowanych. Warto jednak zwrócić uwagę, że na ten wynik składa się głównie poprawna klasyfikacja dla najliczniejszej klasy 1. Dla pozostałych klas model prawie zawsze się mylił.

Tabela 1: Wyniki predykcji dla najlepszego modelu lasu losowego **Dalc**

klasa prawdziwa	**predykcja**					**błąd klasy [%]**
	1	**2**	**3**	**4**	**5**	
1	343	14	5	0	2	5,77
2	68	19	6	0	0	79,57
3	22	10	3	0	2	91,89
4	5	3	2	0	1	100
5	10	0	3	0	1	92,86



Rysunek 2: Dokładność predykcji modelu dla klasy spożywającej alkohol w weekend (**Walc**) w zależności od różnych kombinacji parametrów.

Na podstawie przeprowadzonych badań dla spożywania alkoholu w weekend (**Walc**) wyznaczono jako najlepszy model ten zbudowany z 50 drzew oraz 9 atrybutów wykorzystywanych do budowania podziału oraz maksymalnie 50 węzłów w drzewie oraz niewykorzystywania losowania ze zwracaniem (rys. 2).

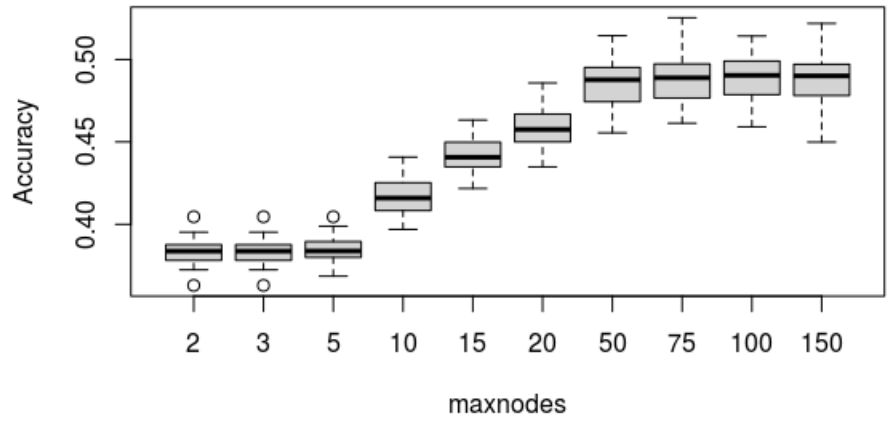
Łączna wartość błędu **out-of-bag** to 55.77%, co oznacza że dokładnie 44.23% próbek zostało poprawnie sklasyfikowanych. W tym przypadku błąd jest znacznie większy niż dla opisującej spożycie alkoholu w ciągu tygodnia, wynika to liczniejszej reprezentacji poszczególnych klas, dla których błędy wynoszą 70-90%. Klasa 1 wciąż klasyfikowana jest małym błędem.

Tabela 2: Wyniki predykcji dla najlepszego modelu lasu losowego **Walc**

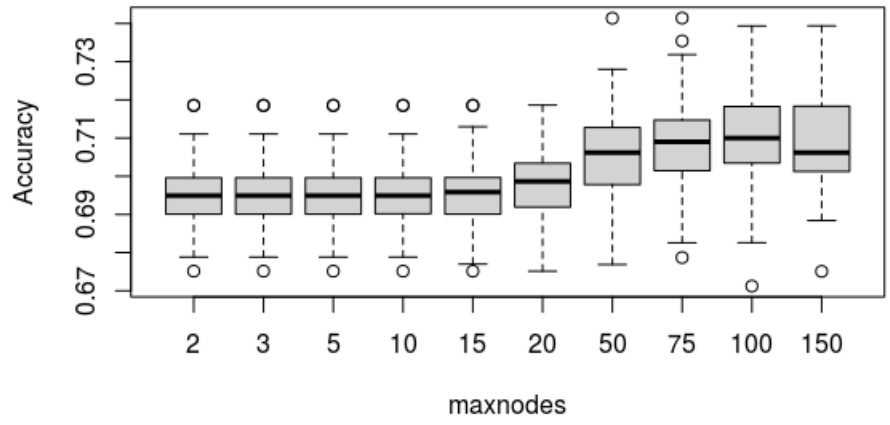
klasa prawdziwa	**predykcja**					**błąd klasy [%]**
	1	**2**	**3**	**4**	**5**	
1	179	7	7	0	1	7,73
2	88	13	15	7	0	89,43
3	50	18	12	22	1	88,43
4	25	9	21	22	1	71,79
5	7	3	5	8	8	74,19

3.1 Wpływ zmiennych na jakość modelu

W celu całkowitego wyczerpania tematu - sprawdzono jak wygląda predykcja modelu przy zachowaniu 2 stałych parametrów a zmieniając trzeci. Dwa ustalone parametry były dobierane na podstawie wcześniejszych eksperymentów. Liczba drzew została ustalona na 100, tak aby wariancja pojedynczego drzewa nie wpływała znacząco wynik skuteczności. Założono, że maksymalna głębokość pojedynczego drzewa będzie wynosiła 50, a ilość zmiennych analizowanych przy pojedynczym podziale będzie równa 7. W celu uzyskania wykresów pudełkowych każdy typ modelu został utworzony 50 razy.

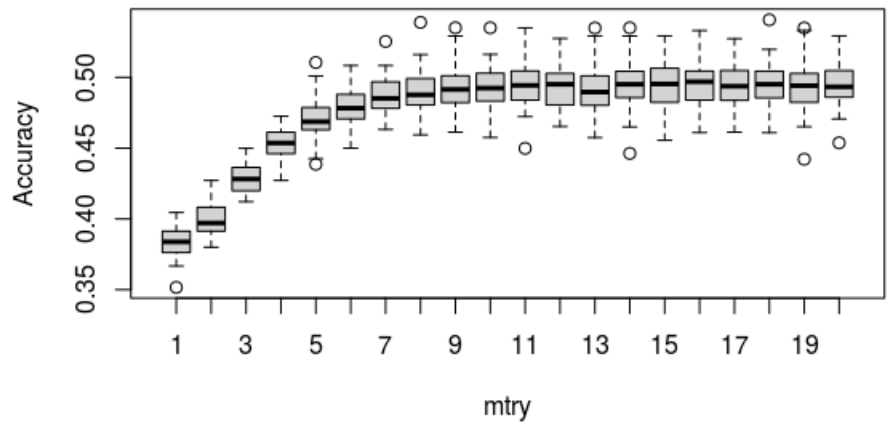


Rysunek 3: Zależność dokładności od maksymalnej ilości węzłów dla mtry = 7, ntree = 100: **Walc**

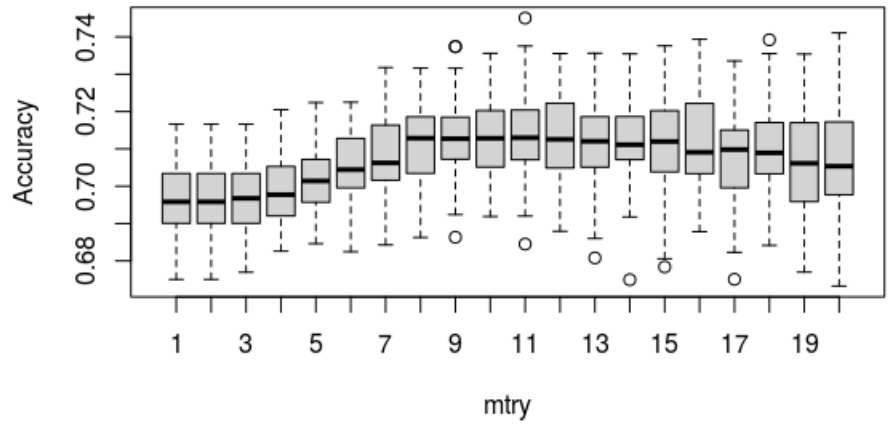


Rysunek 4: Zależność dokładności od maksymalnej ilości węzłów dla mtry = 7, ntree = 100: **Dalc**

Zwiększając ilość możliwych poziomów w pojedynczym drzewie zwiększana jest dokładność modelu. Pojedyncze drzewo z tak późno zastosowanym kryterium stopu, zastosowane samodzielnie jako klasyfikator zostałoby najprawdopodobniej uznane jako nadmiernie dopasowane. Dzięki temu zabiegowi dla elementów lasu możliwe jest uzyskanie większego zróżnicowania drzew, a w wyniku tego model zbiorowy cechuje większa skuteczność klasyfikacji. Większy wpływ tego zjawiska widoczny jest dla danych dotyczących spożycia alkoholu w weekend **Walc** (rys. 3), gdzie rozkład klas jest bardziej równomierny.

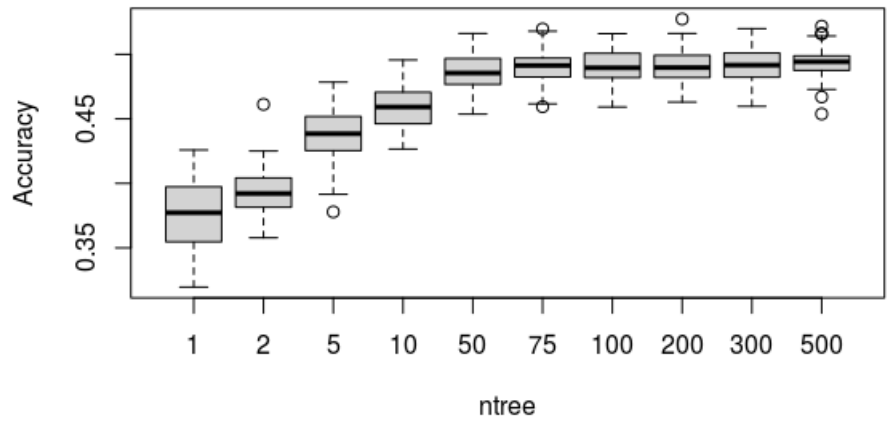


Rysunek 5: Zależność dokładności od mtry dla maxnodes = 50, ntree = 100: **Walc**

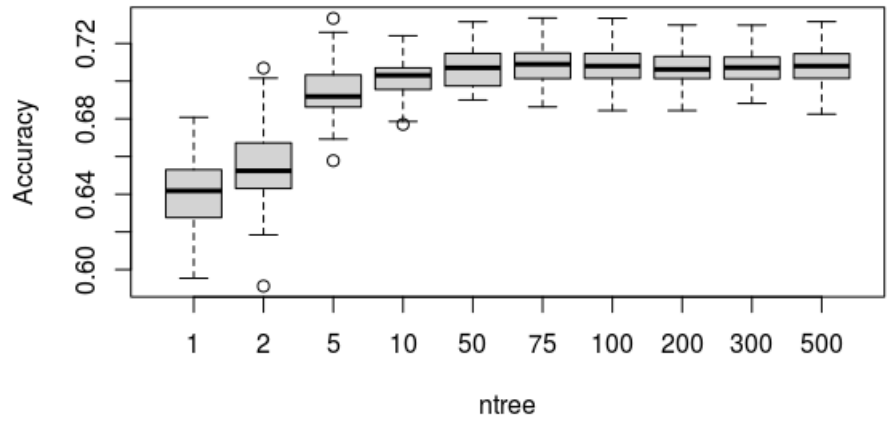


Rysunek 6: Zależność dokładności od mtry dla maxnodes = 50, ntree = 100: **Dalc**

Wybór podziału w każdym węźle wykonywany jest na podstawie analizy losowego podzbioru atrybutów. Zwiększając rozmiar tego podzbioru można poprawić jakość klasyfikacji. Stosowana często wartość pierwiastka z liczby atrybutów (w tym przypadku liczba atrybutów to 32, czyli wartość *mtry* powinna wynosić 5–6) znajduje pewne uzasadnienie, pozwala uzyskać lepsze wyniki niż w przypadku zupełnie losowego wyboru. Na każdym poziomie pula analizowanych atrybutów jest inna, co wpływa na większe zróżnicowanie drzew. Na podstawie rys. 6 można zauważyć, że zbyt duża ilość atrybutów nieznacznie pogarsza jakość klasyfikacji. Prawdopodobnie drzewa stały się do siebie zbyt podobne.



Rysunek 7: Zależność dokładności od ilości drzew dla mtry = 7, maxnodes = 50: **Walc**



Rysunek 8: Zależność dokładności od ilości drzew dla mtry = 7, maxnodes = 50: **Dalc**

Porównując pojedyncze drzewo z modelem zbiorowym obserwowana jest znacząca poprawa dokładności klasyfikacji. Wystarczająco wiele (w tym przypadku już ok. 50) drzew pozwala poprawić jakość i rozrzut predykcji.

4 Klasyfikator Bayesowski

Naiwny klasyfikator bayesowski jest modelem probabilistycznym opartym na klasycznym twierdzeniu Bayesa. Opisuje ono relację pomiędzy prawdopodobieństwem warunkowym pewnego zdarzenia oraz jego prawdopodobieństwem bezwarunkowym:

$$P(A|B) = \frac{P(A) \times P(B|A)}{P(B)}. \tag{1}$$

Wykorzystując wzór (1), można znaleźć prawdopodobieństwo zdarzenia A, zakładając, że zdarzenie B wystąpiło. Zdarzenie B można potraktować jako dowód, a zdarzenie A jako hipotezę.

Implementacja algorytmu klasyfikatora bayesowskiego została wykorzystana przez pakiet e1071. Przy pierwotnym budowaniu modelu wykorzystano funkcję tune do przybliżenia jak najlepszych parametrów modelowi. Następnie przestestowano hipotezę, która miała na celu sprawdzenie wpływu wygładzenia Laplace'a.

Tabela 3: Wyniki dla atrybutu **Dalc** dla naiwnego klasyfikatora bayesowskiego

klasa prawdziwa	**predykcja**					**błąd klasy [%]**
	1	**2**	**3**	**4**	**5**	
1	307	50	8	3	1	16,8
2	38	41	5	4	0	53,4
3	8	10	23	2	0	46,51
4	2	4	2	7	2	58,82
5	3	0	0	0	9	25,00

Tabela 4: Wyniki dla atrybutu **Dalc** dla naiwnego klasyfikatora bayesowskiego przy zastosowaniu wygładzenia Laplace'a

klasa prawdziwa	**predykcja**					**błąd klasy [%]**
	1	**2**	**3**	**4**	**5**	
1	286	39	5	1	1	13,86
2	29	29	3	2	0	53,97
3	11	10	18	2	0	56,1
4	16	18	7	11	2	79,63
5	16	9	5	0	9	76,93

Całkowity błąd naiwnego klasyfikatora Bayesa dla atrybutu **Dalc** wynosi 35,93% (dla zbioru testowego: 22,56%), natomiast przy użyciu wygładzenia Laplace'a 41,39% (dla zbioru testowego 30,0%). W tym przypadku wygładzenie Laplace'a znacząco pogarsza klasyfikację klas, które nie są licznie reprezentowane. Całkowita wartość błędu tego algorytmu jest porównywalna z wartością lasu losowego (tab. 1), jednak rozkład błędów dla poszczególnych klas wypada na korzyść prostszego modelu- naiwnego Bayesa.

Tabela 5: Wyniki dla atrybutu **Walc** dla naiwnego klasyfikatora bayesowskiego

Klasa prawdziwa	**Predykcja**					**błąd klasy [%]**
	1	**2**	**3**	**4**	**5**	
1	150	41	25	10	1	33,92
2	29	49	27	11	3	58,82
3	15	13	24	9	1	61,29
4	3	11	22	37	9	54,88
5	1	2	6	8	22	43,59

Tabela 6: Wyniki dla atrybutu **Walc** dla naiwnego klasyfikatora bayesowskiego przy zastosowaniu wygładzenia Laplace'a 62,39

klasa prawdziwa	**predykcja**					**błąd klasy [%]**
	1	**2**	**3**	**4**	**5**	
1	140	33	21	5	2	33,33
2	35	56	30	11	2	58,22
3	14	13	25	9	0	59,02
4	7	12	21	41	8	53,94
5	2	2	7	9	24	45,46

Dla atrybutu **Walc** całkowity błąd wyniósł 58,03% (dla zbioru testowego 57,31%), a przy zastosowaniu wygładzenia Laplace'a 62,39% (zbiór testowy 57,89%). W tym przypadku nie zauważono znaczącego wpływu wykorzystania wygładzenia. Ponownie błąd tego typu modelu jest porównywalny z lasem losowym, ale również rozkład błędów w zależności od klasy jest korzystniejszy dla naiwnego Bayesa. Ten algorytm lepiej radzi sobie z nierównomiernym rozkładem klas, podczas gdy las losowy wykorzystujący głosowanie większościowe preferuje klasy silniej reprezentowane.

5 Podsumowanie

W ramach projektu przewidywania spożycia alkoholu przez studentów przy użyciu lasu losowego oraz klasyfikatora Bayesa przeprowadzono szereg badań, aby osiągnąć jak największą predykcję. Utrudnieniem w procesie modelowa-

nia było niezbalansowanie klas, które miało znaczący wpływ na odpowiednie dobranie parametrów oraz ostateczną predykcję.

W przypadku lasu losowego zbadano który parametr zwiększa detekcję, czy wpływ na model ma losowanie ze zwracaniem próbek. Udało się ustalić że predykcja w przypadku spożywania alkoholu w weekend jest wyższa niż w tygodniu. Jest to spodziewany wynik, ponieważ więcej studentów spożywa alkohol w weekend. Lepsze wyniki dla poszczególnych klas, bardziej równomierny rozkład błędu uzyskano przy zastosowaniu naiwnego klasyfikatora bayesowskiego. Znacząco lepiej poradził sobie z rozdzielaniem klas, podczas gdy las losowy preferował najsilniej reprezentowaną grupę.

W obu podejściach klasyfikacyjnych problem z którym się spotkano - to znacząca klasyfikacja próbek do niepoprawnych klas, w większość przypisywanie próbki do klasy pierwszej, czyli klasy która miała największy udział proporcji w zestawie danych. Wskazuje to jak istotne jest zebranie reprezentatywnych danych.