UMA Projekt- dokumentacja końcowa

Andrii Gamalii
Wiktor Topolski

Treść zadania

Algorytm analogiczny do lasu losowego z użyciem zbiorów reguł zamiast drzew decyzyjnych.

Interpretacja zadania

Las losowy z użyciem zbiorów reguł uporządkowanych zamiast drzew decyzyjnych. Indukcja reguł za pomocą specjalizacji AQ.

Zbiory danych

1. http://archive.ics.uci.edu/dataset/73/mushroom

Klasyfikacja czy grzyb jest trujący na podstawie 20 atrybutów dyskretnych.

8124 przykłady - 4208 jadalnych grzybów, 3916 niejadalnych.

Zbiór trenujący wybieramy losowo.

2. https://www.kaggle.com/competitions/titanic/data

Dane o pasażerach titanica – klasyfikacja czy pasażer przeżył czy nie

891 przykładów - 549 nie przeżyło, 342 tak

3. https://archive.ics.uci.edu/dataset/697/predict+students+dropout+and+academic+success

Informacje o studentach – klasyfikacja czy student zdał studia, zrezygnował/został wyrzucony czy dalej się uczy

4424 przykładów. Zdał: 2209, Nie zdał : 1421, W trakcie uczenia się: 794

Eksperymenty

Hiperparametry

Użyjemy przeszukiwania losowego wartości hiperparametrów.

- Rozmiar zbioru trenującego (dla całego algorytmu), np. 80%, 90%, 95%
- Rozmiar zbiorów trenujących dla poszczególnych zbiorów reguł, np. 500 przykładów
- Maksymalna ilość zbiorów reguł, np. 100, 500, 1400
- Maksymalna ilość reguł w zbiorze reguł, np. 2, 10, 250

Domyślne wartości parametrów zastosowane w eksperymentach

Jeśli eksperyment nie sprawdza wpływu parametru na wynik to jego wartość to:

- 25 - iteracji

- 100 - zbiorów reguł
- 100 – wielkość podzbioru trenującego dla każdego zbioru reguł
- 10 - reguł w każdym zbiorze reguł
- 5 – maksymalna ilość reguł w jednym zbiorze reguł w czasie specjalizacji
- 0.2 - część zbioru danych przeznaczona na zbiór testowy
- Pokrycie – metoda oceniania reguł w trakcie specjalizacji AQ
- Pierwiastek kwadratowy wszystkich kolumn - ilość kolumn w danych testowych dla pojedynczego zbioru reguł

Miary jakości

- Dokładność
- Precyzja
- F1-score

Zrezygnowaliśmy ze zmiennej: sposoby wyboru ziarna np. losowy, pierwszy przykład, losowy z klasy dominującej, ponieważ uznaliśmy, że losowe ziarna będą najlepsze.

Zmienne algorytmów - Sposób oceniania reguł

Sposoby oceny kompleksów, np. pokrycie, dokładność, dominacja klasy z uwzględnieniem pokrycia.

Statystyki wartości metryk

- Zbiór danych nr. 1 – grzyby
    - Pokrycie

	Średnia	Odchylenie standardowe	Wartość minimalna	Wartość maksymalna
Dokładność	0,95	0,02	0,91	0,99
Precyzja	0,96	0,02	0,92	0,99
F1	0,96	0,02	0,91	0,99

    - Dokładność

	Średnia	Odchylenie standardowe	Wartość minimalna	Wartość maksymalna
Dokładność	0,97	0,02	0,92	0,99
Precyzja	0,97	0,01	0,94	0,99
F1	0,97	0,02	0,92	0,99

    - Dominacja klasy

	Średnia	Odchylenie standardowe	Wartość minimalna	Wartość maksymalna
Dokładność	0,97	0,01	0,94	0,99
Precyzja	0,97	0,01	0,95	0,99
F1	0,97	0,01	0,94	0,99

- Zbiór danych nr. 2 – titanic

- Pokrycie

	Średnia	Odchylenie standardowe	Wartość minimalna	Wartość maksymalna
Dokładność	0,67	0,04	0,57	0,74
Precyzja	0,76	0,07	0,41	0,81
F1	0,59	0,07	0,45	0,71

- Dokładność

	Średnia	Odchylenie standardowe	Wartość minimalna	Wartość maksymalna
Dokładność	0,71	0,06	0,57	0,80
Precyzja	0,77	0,02	0,71	0,84
F1	0,64	0,08	0,43	0,79

- Dominacja klasy

	Średnia	Odchylenie standardowe	Wartość minimalna	Wartość maksymalna
Dokładność	0,73	0,04	0,63	0,81
Precyzja	0,77	0,03	0,70	0,83
F1	0,68	0,06	0,53	0,80

- Zbiór danych nr. 3 – studenci
  - Pokrycie

	Średnia	Odchylenie standardowe	Wartość minimalna	Wartość maksymalna
Dokładność	0,65	0,03	0,60	0,71
Precyzja	0,59	0,05	0,51	0,76
F1	0,57	0,03	0,51	0,63

  - Dokładność

	Średnia	Odchylenie standardowe	Wartość minimalna	Wartość maksymalna
Dokładność	0,66	0,02	0,62	0,71
Precyzja	0,62	0,08	0,52	0,76
F1	0,59	0,03	0,54	0,64

  - Dominacja klasy

	Średnia	Odchylenie standardowe	Wartość minimalna	Wartość maksymalna
Dokładność	0,66	0,03	0,59	0,71
Precyzja	0,59	0,07	0,53	0,76
F1	0,59	0,03	0,50	0,65

Wnioski

Powyższe statystyki pokazują, że wszystkie metody oceniania reguł wykazały się mniej więcej jednakowo dobrze na każdym ze zbiorów danych. A więc nie ma większego znaczenia jakiej metody użyć.

Zmienne algorytmów - Ilość zbiorów reguł w lesie

- Zbiór danych nr.1 - grzyby
    - 30 zbiorów reguł

	Średnia	Odchylenie standardowe	Wartość minimalna	Wartość maksymalna
Dokładność	0,95	0,02	0,89	0,98
Precyzja	0,95	0,02	0,91	0,98
F1	0,95	0,02	0,89	0,98

    - 70 zbiorów reguł

	Średnia	Odchylenie standardowe	Wartość minimalna	Wartość maksymalna
Dokładność	0,96	0,02	0,91	0,99
Precyzja	0,96	0,02	0,93	0,99
F1	0,96	0,02	0,91	0,99

    - 110 zbiorów reguł

	Średnia	Odchylenie standardowe	Wartość minimalna	Wartość maksymalna
Dokładność	0,95	0,02	0,93	0,99
Precyzja	0,96	0,01	0,93	0,99
F1	0,95	0,02	0,93	0,99

    - 150 zbiorów reguł

	Średnia	Odchylenie standardowe	Wartość minimalna	Wartość maksymalna
Dokładność	0,96	**0,01**	0,93	0,99
Precyzja	0,96	**0,01**	0,94	0,99
F1	0,96	**0,01**	0,93	0,99

- Zbiór danych nr.2 - titanic
    - 30 zbiorów reguł

	Średnia	Odchylenie standardowe	Wartość minimalna	Wartość maksymalna
Dokładność	0,69	0,05	0,56	0,78
Precyzja	0,74	0,04	0,64	0,79
F1	0,64	0,08	0,42	0,77

    - 70 zbiorów reguł

	Średnia	Odchylenie standardowe	Wartość minimalna	Wartość maksymalna
Dokładność	0,71	0,06	0,60	0,81
Precyzja	0,77	0,04	0,65	0,84
F1	0,67	0,08	0,49	0,80

- 110 zbiorów reguł

	Średnia	Odchylenie standardowe	Wartość minimalna	Wartość maksymalna
Dokładność	0,70	0,06	0,60	0,80
Precyzja	0,77	0,04	0,70	0,84
F1	0,65	0,09	0,50	0,79

- 150 zbiorów reguł

	Średnia	Odchylenie standardowe	Wartość minimalna	Wartość maksymalna
Dokładność	0,71	0,04	0,64	0,80
Precyzja	0,77	0,03	0,72	0,84
F1	0,67	0,05	0,56	0,79

- Zbiór danych nr.3 - studenci
  - 30 zbiorów reguł

	Średnia	Odchylenie standardowe	Wartość minimalna	Wartość maksymalna
Dokładność	0,63	0,03	0,56	0,70
Precyzja	0,63	0,07	0,51	0,77
F1	0,56	0,05	0,47	0,63

- 70 zbiorów reguł

	Średnia	Odchylenie standardowe	Wartość minimalna	Wartość maksymalna
Dokładność	0,64	0,02	0,61	0,68
Precyzja	0,59	0,07	0,54	0,76
F1	0,57	0,02	0,53	0,60

- 110 zbiorów reguł

	Średnia	Odchylenie standardowe	Wartość minimalna	Wartość maksymalna
Dokładność	0,65	0,07	0,60	0,70
Precyzja	0,58	0,03	0,53	0,68
F1	0,57	0,02	0,52	0,63

- 150 zbiorów reguł

	Średnia	Odchylenie standardowe	Wartość minimalna	Wartość maksymalna

Dokładność	0,65	0,02	0,61	0,70
Precyzja	0,58	0,03	0,53	0,68
F1	0,57	0,02	0,52	0,63

Wnioski

Średnie metryk dla wszystkich zbiorów różnią się na tyle mało, że nie można wybrać najlepszej ilości zbiorów reguł. Jedyna różnica to zmniejszenie odchylenia standardowego wraz ze wzrostem ilości zbiorów - jest to logiczne – im większa próba, tym bliżej wartości oczekiwanej jesteśmy - w zamian zwiększa się ilość obliczeń.

Zmienne algorytmów - Maksymalna ilość reguł w zbiorze reguł

- Zbiór danych nr.1 - grzyby
    - 1 reguła

	Średnia	Odchylenie standardowe	Wartość minimalna	Wartość maksymalna
Dokładność	0,50	0,02	0,47	0,53
Precyzja	0,25	0,02	0,22	0,28
F1	0,33	0,02	0,30	0,37

    - 2 reguły

	Średnia	Odchylenie standardowe	Wartość minimalna	Wartość maksymalna
Dokładność	0,95	0,03	0,89	0,99
Precyzja	0,95	0,03	0,91	0,99
F1	0,95	0,03	0,89	0,99

    - 10 reguł

	Średnia	Odchylenie standardowe	Wartość minimalna	Wartość maksymalna
Dokładność	0,95	0,02	0,90	0,99
Precyzja	0,96	0,02	0,92	0,99
F1	0,95	0,03	0,90	0,99

    - 15 reguł

	Średnia	Odchylenie standardowe	Wartość minimalna	Wartość maksymalna
Dokładność	0,96	0,02	0,92	0,99
Precyzja	0,96	0,02	0,93	0,99
F1	0,96	0,02	0,92	0,99

- Zbiór danych nr.2 - titanic
    - 1 reguła

	Średnia	Odchylenie standardowe	Wartość minimalna	Wartość maksymalna

Dokładność	0,59	0,03	0,53	0,64
Precyzja	0,35	0,04	0,28	0,41
F1	0,44	0,04	0,37	0,50

- 2 reguły

	Średnia	Odchylenie standardowe	Wartość minimalna	Wartość maksymalna
Dokładność	0,69	0,08	0,55	0,83
Precyzja	0,75	0,09	0,35	0,83
F1	0,63	0,12	0,42	0,83

- 10 reguł

	Średnia	Odchylenie standardowe	Wartość minimalna	Wartość maksymalna
Dokładność	0,70	0,04	0,62	0,77
Precyzja	0,77	0,03	0,67	0,80
F1	0,65	0,05	0,51	0,75

- 15 reguł

	Średnia	Odchylenie standardowe	Wartość minimalna	Wartość maksymalna
Dokładność	0,70	0,05	0,61	0,81
Precyzja	0,77	0,03	0,70	0,82
F1	0,65	0,07	0,48	0,80

- Zbiór danych nr.3 - studenci
  - 1 reguła

	Średnia	Odchylenie standardowe	Wartość minimalna	Wartość maksymalna
Dokładność	0,50	0,02	0,45	0,53
Precyzja	0,24	0,01	0,20	0,28
F1	0,33	0,02	0,28	0,37

- 2 reguły

	Średnia	Odchylenie standardowe	Wartość minimalna	Wartość maksymalna
Dokładność	0,55	0,06	0,47	0,66
Precyzja	0,48	0,12	0,22	0,61
F1	0,43	0,09	0,30	0,59

- 10 reguł

	Średnia	Odchylenie standardowe	Wartość minimalna	Wartość maksymalna
Dokładność	0,64	0,03	0,59	0,69

Precyzja	0,58	0,05	0,52	0,75
F1	0,56	0,03	0,50	0,62

- 15 reguł

	Średnia	Odchylenie standardowe	Wartość minimalna	Wartość maksymalna
Dokładność	0,64	0,02	0,61	0,68
Precyzja	0,85	0,05	0,54	0,76
F1	0,57	0,02	0,53	0,60

Wnioski

1 lub 2 reguły na zbiór reguł daje gorsze wyniki niż 10 czy 15, jednak 2 reguły nie odbiegają zbytnio wartością od 10 czy 15. Ponownie wraz ze zwiększeniem ilości reguł, spada odchylenie standardowe – znowu związane większą ilością prób.

Zmienne algorytmów - Wielkość zbioru do trenowania całego lasu

Sposoby oceny kompleksów, np. pokrycie, dokładność, dominacja klasy z uwzględnieniem pokrycia.

Statystyki wartości metryk

- Zbiór danych nr. 1 – grzyby
  - 0,5

	Średnia	Odchylenie standardowe	Wartość minimalna	Wartość maksymalna
Dokładność	0,95	0,02	0,91	0,99
Precyzja	0,96	0,02	0,93	0,99
F1	0,95	0,02	0,91	0,99

  - 0,8

	Średnia	Odchylenie standardowe	Wartość minimalna	Wartość maksymalna
Dokładność	0,96	0,02	0,92	0,98
Precyzja	0,96	0,02	0,93	0,98
F1	0,96	0,02	0,92	0,98

  - 0,95

	Średnia	Odchylenie standardowe	Wartość minimalna	Wartość maksymalna
Dokładność	0,96	0,02	0,92	0,99
Precyzja	0,96	0,02	0,93	0,99
F1	0,96	0,02	0,92	0,99

- Zbiór danych nr. 2 – titanic
  - 0,5

	Średnia	Odchylenie standardowe	Wartość minimalna	Wartość maksymalna
Dokładność	0,68	0,05	0,58	0,81
Precyzja	0,77	0,03	0,71	0,83
F1	0,61	0,09	0,44	0,80

- 0,8

	Średnia	Odchylenie standardowe	Wartość minimalna	Wartość maksymalna
Dokładność	0,68	0,04	0,61	0,77
Precyzja	0,78	0,02	0,74	0,81
F1	0,60	0,07	0,47	0,74

- 0,95

	Średnia	Odchylenie standardowe	Wartość minimalna	Wartość maksymalna
Dokładność	0,68	0,06	0,51	0,80
Precyzja	0,77	0,08	0,47	0,85
F1	0,61	0,08	0,39	0,78

- Zbiór danych nr. 3 – studenci
  - 0,5

	Średnia	Odchylenie standardowe	Wartość minimalna	Wartość maksymalna
Dokładność	0,64	0,03	0,57	0,69
Precyzja	0,60	0,06	0,54	0,77
F1	0,57	0,03	0,47	0,62

- 0,8

	Średnia	Odchylenie standardowe	Wartość minimalna	Wartość maksymalna
Dokładność	0,65	0,03	0,58	0,70
Precyzja	0,61	0,07	0,53	0,76
F1	0,53	0,04	0,49	0,64

- 0,95

	Średnia	Odchylenie standardowe	Wartość minimalna	Wartość maksymalna
Dokładność	0,66	0,03	0,58	0,73
Precyzja	0,58	0,03	0,52	0,64
F1	0,58	0,04	0,49	0,65

Wnioski

Dla każdego z tych 3 zbiorów danych nie ma większego znaczenia jaki procent danych (0,5 i wyżej) wybierzemy do trenowania modelu o takich hiperparametrach.

Zmienne algorytmów - Wielkość zbioru do trenowania pojedynczego zbioru reguł

- Zbiór danych nr.1 - grzyby
    - 20 przykładów

	Średnia	Odchylenie standardowe	Wartość minimalna	Wartość maksymalna
Dokładność	0,93	0,02	0,89	0,97
Precyzja	0,93	0,01	0,91	0,97
F1	0,93	0,02	0,89	0,96

    - 50 przykładów

	Średnia	Odchylenie standardowe	Wartość minimalna	Wartość maksymalna
Dokładność	0,95	0,02	0,91	0,98
Precyzja	0,95	0,02	0,92	0,98
F1	0,95	0,02	0,91	0,98

    - 100 przykładów

	Średnia	Odchylenie standardowe	Wartość minimalna	Wartość maksymalna
Dokładność	0,95	0,02	0,91	0,98
Precyzja	0,96	0,02	0,92	0,98
F1	0,95	0,02	0,91	0,98

    - 200 przykładów

	Średnia	Odchylenie standardowe	Wartość minimalna	Wartość maksymalna
Dokładność	0,96	0,02	0,92	0,99
Precyzja	0,96	0,02	0,93	0,99
F1	0,95	0,02	0,92	0,99

    - 500 przykładów

	Średnia	Odchylenie standardowe	Wartość minimalna	Wartość maksymalna
Dokładność	0,96	0,02	0,92	0,99
Precyzja	0,96	0,02	0,93	0,99
F1	0,95	0,02	0,92	0,99

- Zbiór danych nr.2 - titanic
    - 20 przykładów

	Średnia	Odchylenie standardowe	Wartość minimalna	Wartość maksymalna
Dokładność	0,65	0,05	0,55	0,75
Precyzja	0,75	0,08	0,36	0,80

F1	0,55	0,09	0,39	0,73

- 50 przykładów

	Średnia	Odchylenie standardowe	Wartość minimalna	Wartość maksymalna
Dokładność	0,69	0,04	0,63	0,79
Precyzja	0,78	0,03	0,73	0,83
F1	0,62	0,07	0,52	0,77

- 100 przykładów

	Średnia	Odchylenie standardowe	Wartość minimalna	Wartość maksymalna
Dokładność	0,66	0,05	0,56	0,77
Precyzja	0,77	0,03	0,67	0,83
F1	0,58	0,08	0,42	0,74

- 200 przykładów

	Średnia	Odchylenie standardowe	Wartość minimalna	Wartość maksymalna
Dokładność	0,70	0,04	0,63	0,79
Precyzja	0,77	0,03	0,68	0,82
F1	0,63	0,06	0,53	0,76

- 500 przykładów

	Średnia	Odchylenie standardowe	Wartość minimalna	Wartość maksymalna
Dokładność	0,70	0,05	0,60	0,80
Precyzja	0,76	0,04	0,67	0,85
F1	0,65	0,07	0,51	0,79

- Zbiór danych nr.3 - studenci
  - 20 przykładów

	Średnia	Odchylenie standardowe	Wartość minimalna	Wartość maksymalna
Dokładność	0,62	0,04	0,54	0,67
Precyzja	0,56	0,02	0,51	0,59
F1	0,53	0,05	0,44	0,60

- 50 przykładów

	Średnia	Odchylenie standardowe	Wartość minimalna	Wartość maksymalna
Dokładność	0,65	0,02	0,58	0,69
Precyzja	0,59	0,06	0,50	0,76
F1	0,58	0,03	0,49	0,62

- 100 przykładów

	Średnia	Odchylenie standardowe	Wartość minimalna	Wartość maksymalna
Dokładność	0,64	0,02	0,59	0,68
Precyzja	0,57	0,02	0,51	0,62
F1	0,56	0,03	0,50	0,61

- 200 przykładów

	Średnia	Odchylenie standardowe	Wartość minimalna	Wartość maksymalna
Dokładność	0,64	0,03	0,55	0,68
Precyzja	0,60	0,07	0,52	0,76
F1	0,58	0,04	0,44	0,61

- 500 przykładów

	Średnia	Odchylenie standardowe	Wartość minimalna	Wartość maksymalna
Dokładność	0,64	0,03	0,58	0,69
Precyzja	0,57	0,03	0,52	0,68
F1	0,56	0,04	0,49	0,62

Wnioski

Trudno zauważyć trend związany z ilością przykładów na którym trenuje pojedynczy zbiór reguł. Najprawdopodobniej jest to związane z tym, że w przypadku większej ilości przykładów, szybko występują przykłady sprzeczne - co powoduje wybór najlepszej reguły w danej iteracji.

Klasyczny las

Zmieniamy maksymalną ilość wybieranych atrybutów do uczenia.

- Zbiór danych nr.1 - grzyby
  - 5

	Średnia	Odchylenie standardowe	Wartość minimalna	Wartość maksymalna
Dokładność	1,0	0,0	1,0	1,0
Precyzja	1,0	0,0	1,0	1,0
F1	1,0	0,0	1,0	1,0

- 10

	Średnia	Odchylenie standardowe	Wartość minimalna	Wartość maksymalna
Dokładność	1,0	0,0	1,0	1,0
Precyzja	1,0	0,0	1,0	1,0
F1	1,0	0,0	1,0	1,0

- Pierwiastek z liczby atrybutów

	Średnia	Odchylenie standardowe	Wartość minimalna	Wartość maksymalna
Dokładność	1,0	0,0	1,0	1,0
Precyzja	1,0	0,0	1,0	1,0
F1	1,0	0,0	1,0	1,0

- Log_2 z liczby atrybutów

	Średnia	Odchylenie standardowe	Wartość minimalna	Wartość maksymalna
Dokładność	1,0	0,0	1,0	1,0
Precyzja	1,0	0,0	1,0	1,0
F1	1,0	0,0	1,0	1,0

- Zbiór danych nr.2 - titanic
  - 5

	Średnia	Odchylenie standardowe	Wartość minimalna	Wartość maksymalna
Dokładność	0.82	0,03	0,76	0,87
Precyzja	0.83	0,02	0,78	0,88
F1	0.81	0,03	0,75	0,86

  - 10

	Średnia	Odchylenie standardowe	Wartość minimalna	Wartość maksymalna
Dokładność	0,81	0,03	0,76	0,89
Precyzja	0,81	0,03	0,75	0,89
F1	0,80	0,03	0,75	0,89

  - Pierwiastek z liczby atrybutów

	Średnia	Odchylenie standardowe	Wartość minimalna	Wartość maksymalna
Dokładność	0,82	0,03	0,75	0,88
Precyzja	0,82	0,03	0,76	0,88
F1	0,82	0,03	0,74	0,88

  - Log_2 z liczby atrybutów

	Średnia	Odchylenie standardowe	Wartość minimalna	Wartość maksymalna
Dokładność	0,81	0,03	0,74	0,87
Precyzja	0,81	0,04	0,74	0,88
F1	0,81	0,04	0,73	0,87

- Zbiór danych nr.3 - studenci
  - 5

	Średnia	Odchylenie standardowe	Wartość minimalna	Wartość maksymalna

Dokładność	0,78	0,01	0,76	0,81
Precyzja	0,77	0,01	0,74	0,80
F1	0,77	0,01	0,74	0,79

- 10

	Średnia	Odchylenie standardowe	Wartość minimalna	Wartość maksymalna
Dokładność	0,77	0,01	0,75	0,79
Precyzja	0,76	0,01	0,73	0,78
F1	0,76	0,01	0,74	0,78

- Pierwiastek z liczby atrybutów

	Średnia	Odchylenie standardowe	Wartość minimalna	Wartość maksymalna
Dokładność	0,78	0,01	0,75	0,80
Precyzja	0,76	0,02	0,74	0,80
F1	0,76	0,02	0,74	0,80

- Log_2 z liczby atrybutów

	Średnia	Odchylenie standardowe	Wartość minimalna	Wartość maksymalna
Dokładność	0,77	0,01	0,74	0,80
Precyzja	0,76	0,01	0,73	0,79
F1	0,76	0,01	0,73	0,79

Pojedynczy zbiór reguł

Zmieniamy maksymalną ilość wybieranych atrybutów do uczenia.

Hiperparametry:

- 30 reguł w zbiorze
- 0,9 procent danych dla testowania
- 5 reguł maksymalnie w trakcie indukcji reguł

Wartości statystyk

- Zbiór danych nr.1 - grzyby

	Średnia	Odchylenie standardowe	Wartość minimalna	Wartość maksymalna
Dokładność	0,99	0,00	0,98	1,00
Precyzja	0,99	0,00	0,98	1,00
F1	0,99	0,00	0,98	1,00

- Zbiór danych nr.2 - titanic

	Średnia	Odchylenie standardowe	Wartość minimalna	Wartość maksymalna
Dokładność	0.71	0,05	0,56	0,77

Precyzja	0.72	0,05	0,54	0,79
F1	0.71	0,05	0,55	0,77

- Zbiór danych nr.3 - studenci

	Średnia	Odchylenie standardowe	Wartość minimalna	Wartość maksymalna
Dokładność	0,59	0,02	0,54	0,62
Precyzja	0,58	0,02	0,54	0,63
F1	0,58	0,02	0,54	0,62

Wnioski

Dla sprawdzonych hiperparametrów zaimplementowany las zbiorów reguł daje gorsze wyniki niż klasyczny las i pojedynczy zbiór reguł. Przez tak dobry wynik pojedynczego zbioru reguł (w porównaniu do lasu reguł) podejrzewamy, że słaby wynik lasu ze zbiorem reguł jest spowodowany ograniczeniem ilości kolumn, które są ucinane dla każdego zbioru reguł. Dlatego wykonamy eksperyment testujący tą teorię.

Dodatkowy eksperyment

Zmienne algorytmów - Ilość kolumn w pojedynczym zbiorze reguł

- Zbiór danych nr.1 - grzyby
    - Wszystkie kolumny

	Średnia	Odchylenie standardowe	Wartość minimalna	Wartość maksymalna
Dokładność	0,99	0,00	0,98	1,00
Precyzja	0,99	0,00	0,98	1,00
F1	0,99	0,00	0,98	1,00

    - Pierwiastek kwadratowy wszystkich kolumn

	Średnia	Odchylenie standardowe	Wartość minimalna	Wartość maksymalna
Dokładność	0,95	0,02	0,91	0,99
Precyzja	0,96	0,02	0,92	0,99
F1	0,95	0,02	0,91	0,99

    - Logarytm o podstawie 2 ze wszystkich kolumn

	Średnia	Odchylenie standardowe	Wartość minimalna	Wartość maksymalna
Dokładność	0,96	0,02	0,92	0,99
Precyzja	0,96	0,02	0,93	0,99
F1	0,96	0,02	0,92	0,99

    - Połowa ze wszystkich kolumn

	Średnia	Odchylenie standardowe	Wartość minimalna	Wartość maksymalna
Dokładność	0,99	0,00	0,98	1,00
Precyzja	0,99	0,00	0,98	1,00
F1	0,99	0,00	0,98	1,00

- Zbiór danych nr.2 - titanic
    - Wszystkie kolumny

	Średnia	Odchylenie standardowe	Wartość minimalna	Wartość maksymalna
Dokładność	0,78	0,03	0,73	0,84
Precyzja	0,79	0,03	0,74	0,85
F1	0,78	0,03	0,72	0,84

    - Pierwiastek kwadratowy wszystkich kolumn

	Średnia	Odchylenie standardowe	Wartość minimalna	Wartość maksymalna
Dokładność	0,72	0,05	0,59	0,81
Precyzja	0,78	0,03	0,70	0,86
F1	0,68	0,07	0,52	0,79

    - Logarytm o podstawie 2 ze wszystkich kolumn

	Średnia	Odchylenie standardowe	Wartość minimalna	Wartość maksymalna
Dokładność	0,72	0,04	0,63	0,78
Precyzja	0,77	0,03	0,70	0,82
F1	0,68	0,06	0,59	0,76

    - Połowa ze wszystkich kolumn

	Średnia	Odchylenie standardowe	Wartość minimalna	Wartość maksymalna
Dokładność	0,78	0,03	0,73	0,86
Precyzja	0,80	0,03	0,72	0,87
F1	0,77	0,04	0,70	0,85

- Zbiór danych nr.3 - studenci
    - Wszystkie kolumny

	Średnia	Odchylenie standardowe	Wartość minimalna	Wartość maksymalna
Dokładność	0,71	0,01	0,68	0,73
Precyzja	0,67	0,04	0,60	0,78
F1	0,64	0,01	0,60	0,67

    - Pierwiastek kwadratowy wszystkich kolumn

	Średnia	Odchylenie standardowe	Wartość minimalna	Wartość maksymalna
Dokładność	0,65	0,02	0,61	0,69
Precyzja	0,58	0,05	0,51	0,75
F1	0,57	0,03	0,52	0,62

- Logarytm o podstawie 2 ze wszystkich kolumn

	Średnia	Odchylenie standardowe	Wartość minimalna	Wartość maksymalna
Dokładność	0,60	0,02	0,55	0,64
Precyzja	0,57	0,04	0,53	0,75
F1	0,52	0,03	0,45	0,57

- Połowa ze wszystkich kolumn

	Średnia	Odchylenie standardowe	Wartość minimalna	Wartość maksymalna
Dokładność	0,71	0,02	0,68	0,74
Precyzja	0,67	0,06	0,56	0,79
F1	0,64	0,02	0,61	0,67

Wnioski

Faktycznie, tak jak podejrzewaliśmy większa liczba kolumn przy specjalizacji AQ zwiększa jakość modelu. Połowa kolumn i wszystkie kolumny uzyskały podobne wyniki, które są dużo lepsze od wyników przy logarytmie czy pierwiastku. Porównując rezultat lasu zbiorów reguł ze zwiększoną ilością kolumn do pojedynczego zbioru reguł - który był testowany dla 10% wszystkich danych – widzimy, że las uzyskuje wynik lepszy – jest to rezultat, którego się spodziewaliśmy. Porównując zaś z klasycznym lasem, zbiory reguł wypadają zdecydowanie gorzej. Przyczyną może być to, że drzewa decyzyjne są prostsze, przez to mają mniejszą skłonność do przeuczenia. Kolejnym powodem może być jakiś błąd w naszej implementacji, ale mamy nadzieję, że tak nie jest.

Czego się nauczyliśmy?

Nauczyliśmy się niuansów związanych z specjalizacją AQ i różnych sytuacji brzegowych. Również nabyliśmy umiejętność znajdowania zbiorów danych testowych. Dowiedzieliśmy się również jak długo mogą trwać eksperymenty, szczególnie gdy korzystamy z własnej, słabo zoptymalizowanej implementacji. Została również dla nas (a raczej dla pozostałych prowadzących) postawiona poprzeczka, jeśli chodzi o precyzyjność wymagań projektów. Nauczyliśmy się jak można liczyć miary jakości modeli predykcyjnych.