Vasil Kusmartsev, Tsimafei Kotski

---

1. Wybrane atrybuty:

`1. Failures`

`2. Studytime`

`3. Absences`

`4. Medu`

`5. Goout`

`6. Dalc`

`7. Higher`

`8. Famrel`

Wybór ośmiu atrybutów został dokonany w oparciu o analizę dziedzinową oraz czynniki mające potencjalnie największy wpływ na wyniki akademickie uczniów. Atrybuty `failures` i `studytime` bezpośrednio odzwierciedlają zaangażowanie akademickie i historię sukcesów. `Absences` wskazuje na frekwencję i dyscyplinę. `Medu` reprezentuje kapitał kulturowy rodziny. `Goout` i `Dalc` odzwierciedlają czynniki społeczno-behawioralne mogące dystraktować od nauki. `Higher` pokazuje motywację długoterminową, a `famrel` - stabilność środowiska domowego. Taka kombinacja zapewnia holistyczne ujęcie czynników wpływających na sukces edukacyjny.

3. Wybrane metryki:

- F1-score

- Accuracy

- AUC-ROC

Do oceny modeli klasyfikacji binarnej wybrano trzy kluczowe metryki: F1-Score, AUC-ROC i Accuracy. F1-Score zapewnia zbalansowaną ocenę między precyzją a czułością, co jest krytyczne przy potencjalnie niezbalansowanych klasach. AUC-ROC mierzy fundamentalną zdolność modelu do rozróżniania między klasami, niezależnie od progu klasyfikacji. Accuracy dostarcza intuicyjnej miary ogólnej skuteczności. Ta kombinacja pozwala na kompleksową ocenę modelu z różnych perspektyw, zapewniając zarówno techniczną rygorystyczność, jak i praktyczną interpretowalność wyników.

4.

№ testu	confidenceFactor	minNumObj	binarySplits	Accuracy (%)	F1-Score	AUC	Confusion Matrix [a=fail, b=pass]	
1	0.25	2	False	55.38%	0.446	0.566	73/45	42/35
2	0.1	2	False	55.90%	0.449	0.567	74/44	42/35
3	0.5	2	False	54.36%	0.411	0.542	75/43	46/31
4	0.25	5	False	55.38%	0.416	0.539	77/41	46/31
5	0.25	**10**	False	**57.95%**	0.414	0.535	**84/34	48/29**
6	0.25	2	True	55.38%	0.446	0.566	73/45	

Po przeprowadzeniu serii eksperymentów z różnymi parametrami algorytmu J48 na zbiorze student-mat z wykorzystaniem wybranych 8 atrybutów, najlepszy wynik na zbiorze testowym uzyskano dla następującego zestawu parametrów:

- confidenceFactor (współczynnik ufności): 0.25
- minNumObj (min. liczba obiektów w liściu): 10
- binarySplits (podziały binarne): False

5.

№	confidenceFactor	minNumObj	binarySplits	Accuracy (%)	F1-Score	AUC	Confusion Matrix	
1	0.25	2	False	50.77 %	0.377	0.486	70/48	48/29
2	0.1	2	False	50.77 %	0.377	0.488	70/48	48/29
3	0.5	2	False	49.74 %	0.364	0.481	69/49	49/28
4	0.25	5	False	50.77 %	0.200	0.495	87/31	65/12
5	0.25	**10**	False	**63.08 %**	0.163	0.558	**116/2	70/7** *
6	0.25	2	**True**	**54.87 %**	**0.457**	**0.505**	**70/48	40/37 **

Porównanie z Zadaniem 4 (8 atrybutów):

- Najlepszy model z 8 atrybutów: Accuracy=57.95%, F1-Score=0.414
- Najlepszy model ze wszystkimi atrybutami: Albo wysoki Accuracy z F1=0.163, albo F1=0.457 z niższym Accuracy

Wniosek końcowy:
Utilizacja wszystkich atrybutów nie przyniosła jakościowej poprawy modelu w porównaniu z naszym intuicyjnym wyborem 8 atrybutów. Nasz wybór okazał się trafny i efektywny - pozwolił na uzyskanie modelu o zrównoważonych i generalizujących mocach, usuwając zbędny szum i zapobiegając przeuczeniu.

6.

Parametry algorytmu: confidenceFactor=0.25, minNumObj=10, binarySplits=False (najlepsze parametry z Zadania 4)

Dokładność (Accuracy)	70.416%
Miary F1 (dla klasy '(11.5-inf)')	0.747
Pole pod krzywą ROC (AUC)	0.739
Macierz pomyłek	173 (TN) / 128 (FP) | 64 (FN) / 284 (TP)

Struktura drzewa decyzyjnego (fragment):

```
failures <= 0
|   higher = yes
|   |   school = GP
|   |   |   Walc <= 3: '(11.5-inf)' (295.0/69.0)
|   |   |   Walc > 3: ...
|   |   school = MS: ...
|   higher = no: '(-inf-11.5]' (36.0/5.0)
failures > 0: '(-inf-11.5]' (100.0/7.0)
```

Model dla języka portugalskiego osiągnął znacząco lepsze wyniki niż model dla matematyki, uzyskując wysoką, zrównoważoną skuteczność w przewidywaniu obu klas. Lepsza wydajność może wynikać z większego rozmiaru zbioru danych (649 instancji vs 200 dla matematyki), co umożliwiło skuteczniejsze uczenie się modelu.

7.
Podobieństwa:

- Atrybut korzenia: W obu drzewach failures (liczba poprzednich niepowodzeń) jest atrybutem korzeniowym, co wskazuje na jego kluczowe znaczenie dla prognozowania wyniku egzaminu w obu przedmiotach.
- Kolejny ważny atrybut: higher (zamiar podjęcia studiów wyższych) pojawia się na drugim poziomie drzewa dla portugalskiego i jest istotnym atrybutem w obu modelach, co podkreśla rolę motywacji długoterminowej.

Różnice:

- Złożoność drzewa: Drzewo dla języka portugalskiego jest bardziej złożone (29 węzłów, 17 liści) w porównaniu z drzewem dla matematyki (17 węzłów, 9 liści), pomimo użycia tych samych parametrów upraszczających (minNumObj=10).
- Specyficzne atrybuty przedmiotowe:

    W drzewie dla portugalskiego pojawiają się atrybuty Walc (spożycie alkoholu w weekendy) i school (typ szkoły), które nie były tak istotne w drzewie dla matematyki.

    Drzewo dla portugalskiego wykorzystuje więcej atrybutów społeczno-demograficznych (np. Mjob, guardian).

Atrybuty "uważnego studenta":

Na podstawie struktury obu drzew za najważniejsze atrybuty charakteryzujące studenta z pozytywnymi wynikami można uznać:

- failures (0 poprzednich niepowodzeń) - UNIWERSALNY I NAJWAŻNIEJSZY WSKAŹNIK
- higher (chęć kontynuowania nauki) - ISTOTNY CZYNNIK MOTYWACYJNY
- studytime (czas na naukę) - widoczny w gałęziach obu drzew

Wnioski końcowe:

Pomimo różnic w złożoności i doborze atrybutów szczegółowych, rdzenne czynniki sukcesu są spójne dla obu przedmiotów. Brak wcześniejszych niepowodzeń akademickich oraz wewnętrzna motywacja do dalszej edukacji okazują się uniwersalnymi prognostykami sukcesu szkolnego, niezależnie od przedmiotu. Różnice w drugorzędnych atrybutach sugerują, że czynniki społeczno-środowiskowe mogą mieć różne znaczenie w zależności od kontekstu przedmiotowego.

8.

Tabela porównawcza wyników:

Algorytm	Dokładność (Accuracy)	F1-Score	AUC	Macierz pomyłek
J48	57,95%	41,4%	0,535	-
Naive Bayes	53,33%	53,1%	0,625	50-68 / 23-54

Top 5 atrybutów według Naive Bayes:

failures (liczna niepowodzeń) - najważniejszy atrybut

Medu (wykształcenie matki)

absences (nieobecności)

goout (czas z przyjaciółmi)

higher (plany dotyczące studiów wyższych)

Odpowiedzi na pytania:

1. Który algorytm osiągnął lepsze wyniki?
Algorytm J48 osiągnął wyższą dokładność (57,95% vs 53,33%), natomiast Naive Bayes wykazał lepszą zbalansowaną jakość predykcji (wyższy F1-Score i AUC).

2. Porównanie atrybutów wpływowych:
5 z 8 atrybutów wybranych intuicyjnie w Zadaniu 1 pokrywa się z najbardziej wpływowymi atrybutami identyfikowanymi przez Naive Bayes. Potwierdza to trafność naszej wstępnej selekcji atrybutów.

3. Atrybuty wskazujące na uwagę studenta:
Na podstawie analizy obu algorytmów, atrybutami najbardziej wskazującymi na uwagę studenta są:

failures (historia niepowodzeń)

absences (frekwencja)

studytime (czas nauki)

higher (motywacja długoterminowa)

4. Jaki algorytm zapewnia lepszą interpretowalność?
J48 oferuje lepszą interpretowalność poprzez czytelne drzewo decyzyjne, podczas gdy Naive Bayes dostarcza statystycznej analizy ważności atrybutów.

Wnioski końcowe:

Wybór algorytmu zależy od celu - J48 dla maksymalnej dokładności, Naive Bayes dla zbalansowanej klasyfikacji obu klas. Nasza intuicyjna selekcja 8 atrybutów została potwierdzona analitycznie.