Sprawozdanie Drzewa Decyzyjne cz. 2
Jakub Kłosiński 145959 Krzysztof Rybczyński 148321

Zad1. Na podstawie opisu atrybutów wybrano 6, które mogą najlepiej przewidzieć, czy student zaliczy przedmiot, biorąc pod uwagę potencjalne znaczenie edukacyjne, społeczne oraz wpływ związany z zaangażowaniem:

- **studytime** - Tygodniowy czas nauki, ponieważ może wpływać na wyniki akademickie.
- **failures** - Liczba wcześniejszych niepowodzeń, jako wskaźnik wcześniejszych trudności.
- **schoolsup** - Dodatkowe wsparcie edukacyjne, które może poprawiać wyniki uczniów.
- **famsup** - Wsparcie rodzinne w nauce, mogące mieć pozytywny wpływ na osiągnięcia.
- **higher** - Aspiracje edukacyjne, ponieważ motywacja do nauki wyższej może wpływać na wyniki.
- **absences** - Liczba nieobecności, która może wskazywać na zaangażowanie i obecność na zajęciach.

Zad 2/3. Uruchomiono klasyfikację przy użyciu algorytmu J48. Początkowo ustawione parametry wynosiły np. confidenceFactor = 0.25, minNumObj = 2. Użyto pliku student-mat-train.arff oraz student-train-test.arff jako supplied test set.

Na podstawie raportu można wyróżnić trzy najważniejsze metryki:

- **Accuracy** – uzyskała wartość 50.77%, co jest wskaźnikiem ogólnej skuteczności modelu. Na ten moment jest onabliska połowie, a więc dość niska.
- **F-Measure** – dla każdej klasy jest stosunkowo niska (0.593 i 0.377). F-Measure wskazuje, że model ma trudności z równoczesnym osiąganiem wysokiej precyzji i czułości, zwłaszcza dla klasy '11.5-inf'.
- **ROC Area** – wartość 0.486 dla obu klas sugeruje, że model ma problemy z rozróżnianiem klas, ponieważ wartość ROC dla dobrze działającego modelu powinna być jak najbliższa 1.

Następnie podjęto próbę dostrojenia parametrów confidenceFactor i minNumObj oraz binarySplits. Ostatecznie po przyjęciu confidenceFactor = 0.8 i minNumObj = 10 udało się poprawnie sklasyfikować 123 na 195 przykładów. Teraz accuracy = 63.08%, liczba liści zmniejszyła się do 8, a rozmiar drzewa do 15. Drzewo więc zostało uproszczone.

- **Klasa '(-inf-11.5]'**:
  - **TP Rate (Recall)**: 0.983, co oznacza, że model bardzo dobrze rozpoznaje przypadki z tej klasy.
  - **Precision**: 0.624, co wskazuje, że model ma pewne trudności z precyzyjnym klasyfikowaniem przypadków jako '(-inf-11.5]'.
  - **F-Measure**: 0.763, co jest stosunkowo wysoką wartością dla tej klasy.

- **Klasa '(11.5-inf)'**:
  - **TP Rate (Recall)**: 0.091, co jest bardzo niskie i wskazuje, że model ma problem z rozpoznawaniem przypadków z tej klasy.
  - **Precision**: 0.778, co sugeruje, że kiedy model klasyfikuje przypadek jako '(11.5-inf)', jest to zazwyczaj trafne, ale zdarza się to rzadko.
  - **F-Measure**: 0.163, co wskazuje na niską skuteczność modelu dla tej klasy.

Zad 4. Dla wyżej wymienionych atrybutów, przy uruchomieniu algorytmu J48 ustawiano parametry:

- confidenceFactor: 0.25 , minNumObj = 2, binarySplits = false.

Macierz pomyłek

	Przewidywana klasa '(-inf-11.5]'	**Przewidywana klasa '(11.5-inf)'**
Rzeczywista klasa '(-inf-11.5]'	**52**	**66**
Rzeczywista klasa '(11.5-inf)'	**33**	**44**

- confidenceFactor: 0.5 , minNumObj = 5, binarySplits = false.

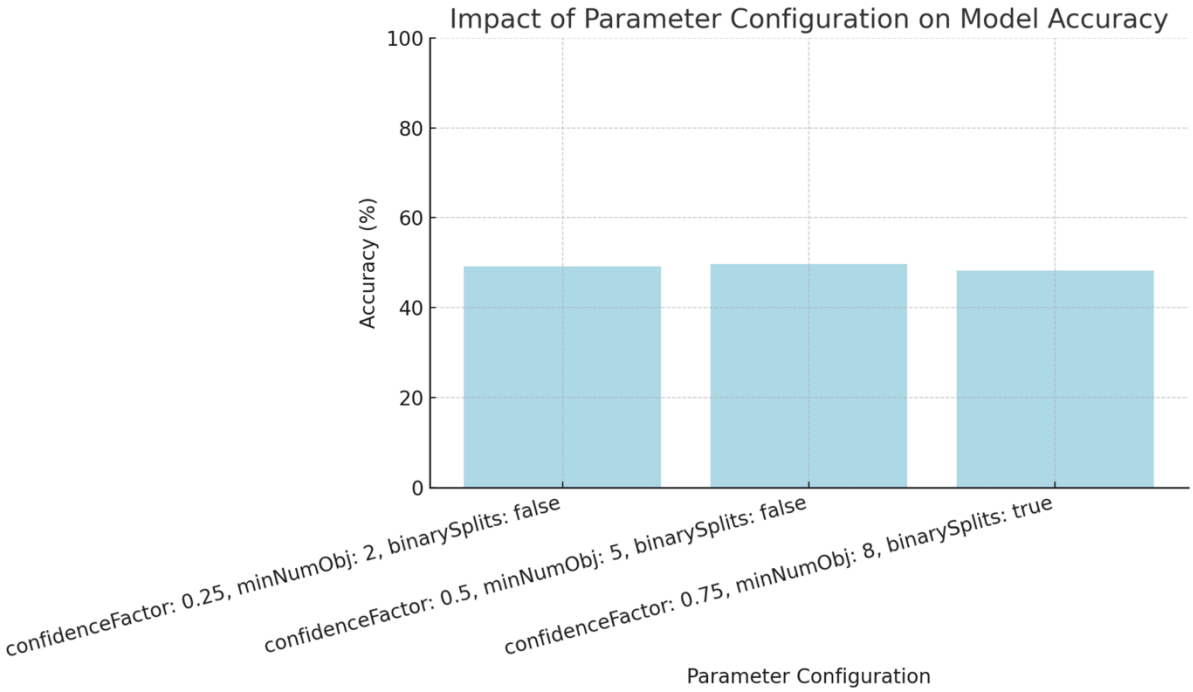
Macierz pomyłek

	Przewidywana klasa '(-inf-11.5]'	**Przewidywana klasa '(11.5-inf)'**
Rzeczywista klasa '(-inf-11.5]'	**52**	**66**
Rzeczywista klasa '(11.5-inf)'	**32**	**45**

- confidenceFactor: 0.75 , minNumObj = 8, binarySplits = true.

Macierz pomyłek

	Przewidywana klasa '(-inf-11.5]'	**Przewidywana klasa '(11.5-inf)'**
Rzeczywista klasa '(-inf-11.5]'	**40**	**78**
Rzeczywista klasa '(11.5-inf)'	**23**	**54**



Wnioski

- Wartości dokładności nie różnią się znacząco pomiędzy konfiguracjami, co może wskazywać na ograniczoną skuteczność tych parametrów w poprawie klasyfikacji dla tego konkretnego zbioru danych.
- Dokładność modelu waha się w przedziale 45-50%, co sugeruje, że model J48 ma trudności z dokładnym rozróżnianiem klas przy wybranych atrybutach i konfiguracjach.

Zad 5. Dla wszystkich atrybutów

- confidenceFactor: 0.25, minNumObj = 2, binarySplits = false.

Poprawnie sklasyfikowane: 50,77%

Macierz pomyłek

	Przewidywana klasa '(-inf-11.5]'	**Przewidywana klasa '(11.5-inf)'**
Rzeczywista klasa '(-inf-11.5]'	**72**	**48**
Rzeczywista klasa '(11.5-inf)'	**48**	**29**

- confidenceFactor: 0.5, minNumObj = 5, binarySplits = false.

Poprawnie sklasyfikowane: 51,28%

Macierz pomyłek

	Przewidywana klasa '(-inf-11.5]'	**Przewidywana klasa '(11.5-inf)'**
Rzeczywista klasa '(-inf-11.5]'	**87**	**31**
Rzeczywista klasa '(11.5-inf)'	**64**	**13**

- confidenceFactor: 0.75, minNumObj = 8, binarySplits = true.

Poprawnie sklasyfikowane: 55,9%

Macierz pomyłek

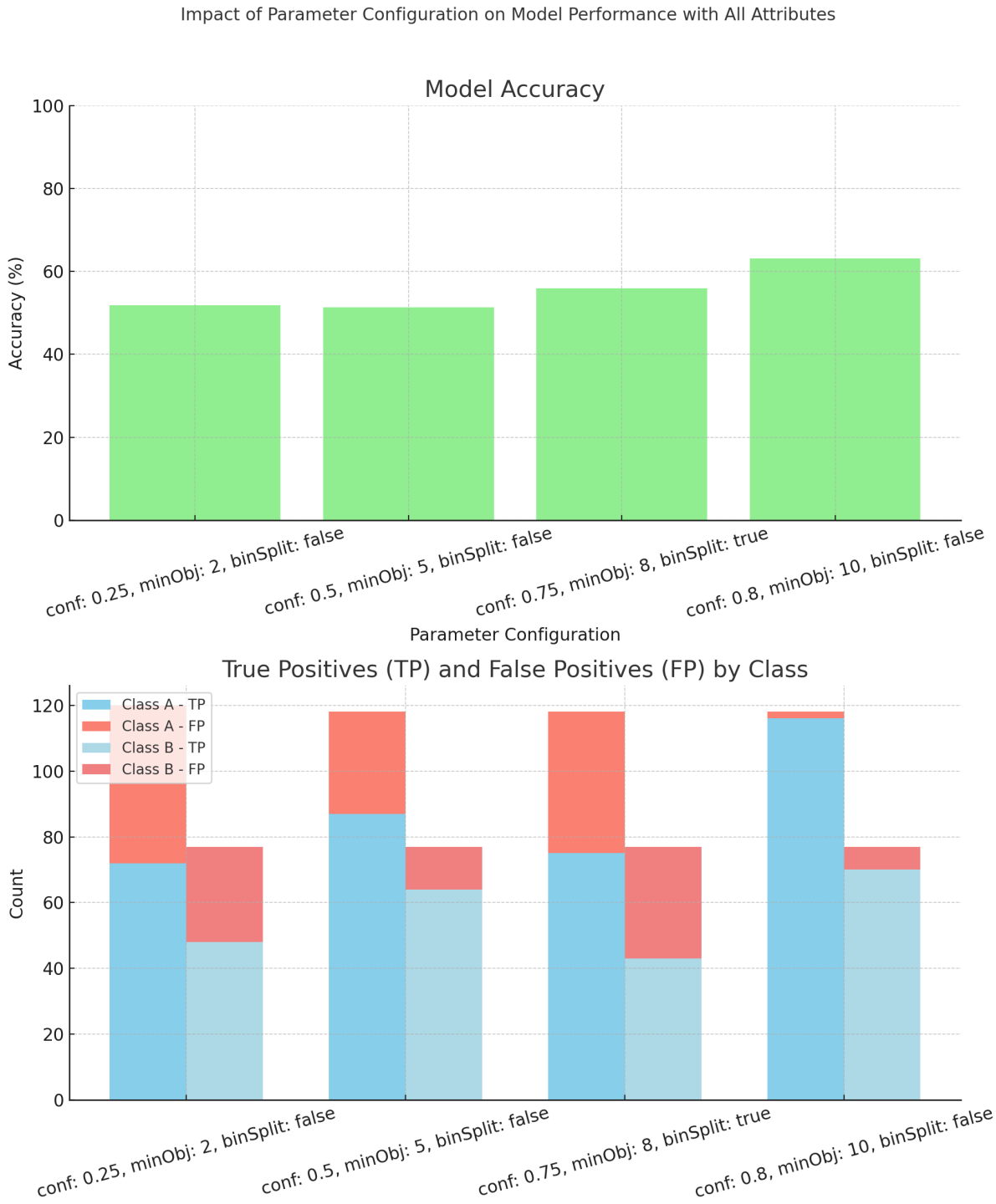
	Przewidywana klasa '(-inf-11.5]'	**Przewidywana klasa '(11.5-inf)'**
Rzeczywista klasa '(-inf-11.5]'	**75**	**43**
Rzeczywista klasa '(11.5-inf)'	**43**	**34**

- confidenceFactor: 0.8, minNumObj = 10, binarySplits = false.

Poprawnie sklasyfikowane: 63,08%

Macierz pomyłek

	Przewidywana klasa '(-inf-11.5]'	**Przewidywana klasa '(11.5-inf)'**
Rzeczywista klasa '(-inf-11.5]'	**116**	**2**
Rzeczywista klasa '(11.5-inf)'	**70**	**7**



Zad 6.

Uruchomiono plik student-por.arff z wykorzystanie algorytmu J48 dla cross-validation k = 10.
Po raz kolejny wykorzystano parametry, które poprzednio okazały się najbardziej trafne tzn. confidenceeFactor = 0.8, binarySplits = false, minNumObj = 10.
Poprawnie skalsyfikowano : 69, 49 %
Liczba liści drzewa: 45
Rozmiar drzewa: 72
Czas budowania modelu: 0.15 s

Macierz pomyłek

	Przewidywana klasa '(-inf-11.5]'	**Przewidywana klasa '(11.5-inf)'**
Rzeczywista klasa '(-inf-11.5]'	**186**	**115**
Rzeczywista klasa '(11.5-inf)'	**83**	**265**

Zad 7.

Uruchomiono plik student-por.arff oraz student-mat-train.arff z wykorzystanie algorytmu J48 dla cross-walidacji k = 10. ConfidenceeFactor = 0.8, binarySplits = false, minNumObj = 10.

1. Porównanie

Portugalski:

- Rozmiar drzewa: 72 węzłów i 45 liści. Jest to rozbudowane drzewo, które uwzględnia różnorodne atrybuty i ich wartości.
- Kluczowe atrybuty: **failures, higher, schoolsup, Walc** (spożycie alkoholu w weekendy), **internet**, **Fedu** (wykształcenie ojca), **guardian**, **health** i **studytime**.
- Drzewo jest głębsze i bardziej szczegółowe, co może sugerować, że ocena wyników studentów języka portugalskiego wymaga więcej zmiennych socjalnych oraz związanych z czasem wolnym, takich jak Walc czy goout.

Matematyka:

- Rozmiar drzewa: 15 węzłów i 8 liści, co wskazuje na prostszą strukturę drzewa w porównaniu do drzewa dla student-por.
- Kluczowe atrybuty: **failures**, **schoolsup**, **age**, **sex**, **health**, **Medu** (wykształcenie matki). Drzewo jest mniej szczegółowe, a atrybuty związane są głównie z podstawowymi danymi demograficznymi i poziomem wsparcia edukacyjnego.
- Drzewo koncentruje się bardziej na atrybutach związanych bezpośrednio z edukacją i zdrowiem, a także na liczbie nieobecności, co może sugerować, że dla przedmiotu matematyka te czynniki są kluczowe.

2. Podobieństwa między drzewami

- Oba drzewa używają atrybutów **failures** (liczba niezdanych przedmiotów) oraz **schoolsup** (dodatkowe wsparcie szkolne) jako kluczowych węzłów decyzyjnych. Wskazuje to, że zarówno w matematyce, jak i języku portugalskim wcześniejsze niepowodzenia edukacyjne i wsparcie szkolne mają istotny wpływ na wyniki uczniów.
- Atrybut **health** występuje w obu drzewach, co może sugerować, że stan zdrowia wpływa na ogólne wyniki uczniów.

3. Różnice między drzewami

- Więcej atrybutów społecznych w drzewie dla student-por: W drzewie języka portugalskiego uwzględniono takie atrybuty jak Walc, Dalc, internet, romantic i freetime, które nie pojawiają się w drzewie dla matematyki. Może to wynikać z tego, że w przedmiotach związanych z językiem, relacje społeczne i zachowania mogą bardziej wpływać na wyniki.
- Prostota drzewa dla student-mat: Drzewo matematyczne jest bardziej uproszczone i skoncentrowane na atrybutach związanych bezpośrednio z edukacją oraz wsparciem szkolnym, co może sugerować, że matematyka jest przedmiotem, gdzie wynik zależy bardziej od konkretnych zdolności i regularnej obecności.

4. Wnioski o uważności uczniów

Na podstawie obu drzew możemy wskazać atrybuty, które mogą sugerować, czy uczeń jest uważny i zaangażowany w naukę:

- **failures**: Niska liczba niepowodzeń wskazuje na to, że uczeń mógł być uważny i pilnować swojej nauki wcześniej.
- **schoolsup**: Dodatkowe wsparcie szkolne może wskazywać na to, że uczeń wymaga dodatkowej pomocy, co może być oznaką zaangażowania, jeśli uczeń aktywnie z niego korzysta.
- **absences**: Niska liczba nieobecności wskazuje na zaangażowanie i systematyczność ucznia, co jest oznaką jego uważności.
- **health**: Lepszy stan zdrowia może wpływać pozytywnie na wyniki, ponieważ zdrowi uczniowie są bardziej skłonni do regularnego uczęszczania na zajęcia i angażowania się w naukę.

Zad 8. Porównanie algorytmów J48 i Naive Bayes na zbiorze student-mat

Wyniki dla Naive Bayes

1. **Dokładność (Accuracy)**: Naive Bayes osiągnął dokładność na poziomie 62%, co jest nieco niższą wartością w porównaniu do najlepszego wyniku uzyskanego przez algorytm J48 (63.08%).
2. **Kluczowe metryki**:
    - **TP Rate** (Recall) dla klasy '(-inf-11.5]' wynosi 0.513, a dla klasy '(11.5-inf)' wynosi 0.765, co sugeruje, że Naive Bayes lepiej klasyfikuje przypadki z klasy '(11.5-inf)'.
    - **Precision**: Naive Bayes osiąga wyższą precyzję dla klasy '(-inf-11.5]' (0.747), podczas gdy dla klasy '(11.5-inf)' jest niższa (0.537).
    - **Kappa Statistic**: 0.2632 – wartość ta wskazuje na umiarkowaną zgodność modelu z danymi, ale jest niższa niż dla J48, co sugeruje, że model J48 jest bardziej stabilny.

Porównanie wyników algorytmów

- **Dokładność**: Model J48 osiągnął nieznacznie wyższą dokładność (63.08%) w porównaniu do Naive Bayes (62%).
- **Stabilność wyników**: Algorytm J48 lepiej radził sobie z rozróżnianiem klas w sposób bardziej zrównoważony, co sugerują wyższe wartości TP Rate dla obu klas.
- **Precision i Recall**: Naive Bayes osiąga wyższą precyzję dla klasy '(-inf-11.5]', co oznacza, że model ma mniejsze błędy fałszywie pozytywne, ale ma niższą precyzję dla klasy '(11.5-inf)', co oznacza większą ilość błędnych przypisań.

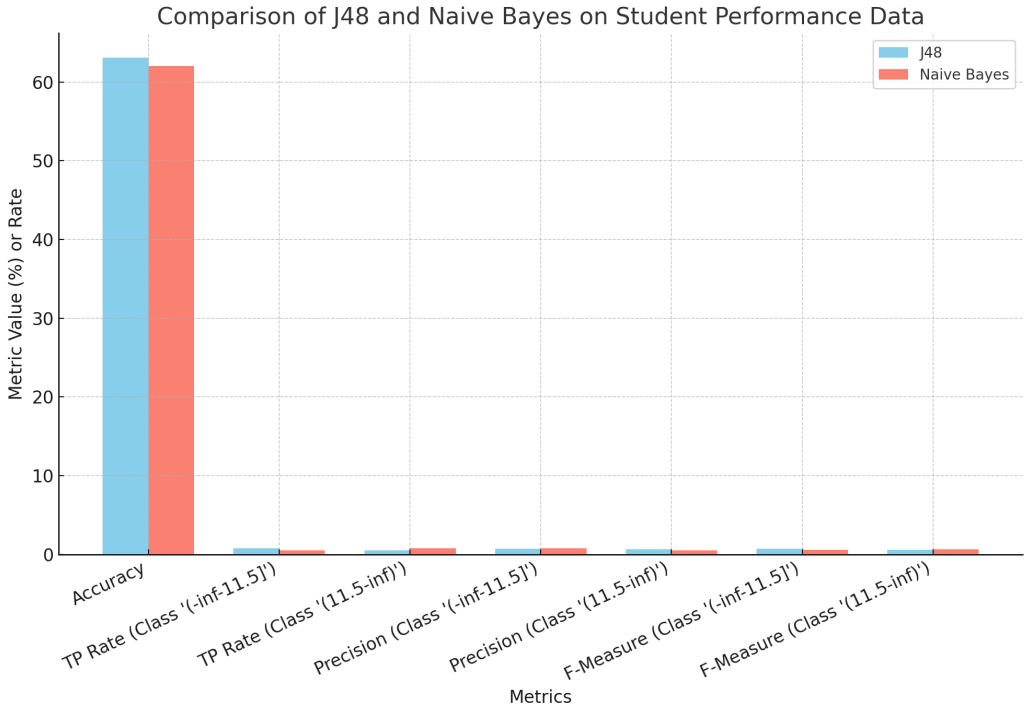
Analiza wpływowych atrybutów

Na podstawie wyników dla Naive Bayes można określić, które atrybuty miały największy wpływ na wynik:

- **failures**: Średnia liczba niepowodzeń była wyższa dla klasy '(-inf-11.5]', co sugeruje, że wcześniejsze porażki edukacyjne znacząco wpływają na wynik.
- **schoolsup**: Większa liczba studentów, którzy mieli wsparcie szkolne, znalazła się w klasie '(-inf-11.5]', co potwierdza, że dodatkowe wsparcie szkolne jest istotnym czynnikiem.
- **Walc i goout**: Średnie spożycie alkoholu w weekendy i czas spędzany na wyjściach to atrybuty, które miały większe wartości dla klasy '(-inf-11.5]', co może sugerować mniejsze zaangażowanie w naukę.
- **Medu i Fedu**: Wyższe poziomy wykształcenia rodziców są widoczne w klasie '(11.5-inf)', co może sugerować, że wyższy poziom edukacji rodziców wpływa pozytywnie na wyniki ucznia.

Wnioski końcowe

- **Podobieństwa z intuicyjnie wybranymi atrybutami**: Wybrane atrybuty, takie jak failures, schoolsup, goout, i absences, potwierdzają swoją istotność w wynikach obu algorytmów. Te atrybuty są kluczowe w przewidywaniu wyników ucznia i były już uwzględnione w początkowych intuicyjnych wyborach.
- **Lepszy algorytm**: Chociaż różnice w dokładności nie są znaczące, model J48 wydaje się lepszy, ponieważ jest bardziej stabilny i dokładniejszy w rozróżnianiu klas. Naive Bayes natomiast ma swoje zalety w precyzji dla klasy '(-inf-11.5]', co czyni go dobrym wyborem, jeśli priorytetem jest minimalizacja błędów fałszywie pozytywnych.


Comparison of J48 and Naive Bayes on Student Performance Data
J48
Naive Bayes
Metric Value (%) or Rate
0
10
20
30
40
50
60
Accuracy
TP Rate (Class '(-inf-11.5]')
TP Rate (Class '(11.5-inf)')
Precision (Class '(-inf-11.5]')
Precision (Class '(11.5-inf)')
F-Measure (Class '(-inf-11.5]')
F-Measure (Class '(11.5-inf)')
Metrics