SPRAWOZDANIE 4

Podstawy Sztucznej Inteligencji

Klasyfikator Bayesa

Program WEKA 4

Natalia Gadocha 304165
Geoinformatyka III rok

Zadanie 1

P(A1) = 0,1

P(A2) = 0,9

P(T+/A1) = 0,5

P(T+/A2) = 0,02

P(A1/T+) = P(T+/A1)*P(A1)/P(T+) = 0,735

P(T+) = P(T+/A1) *P(A1) + P(T+/A2)*P(A2) = 0,068

Zadanie 2

P(A1) = 99,95% = 0,9995

P(A2) = 0,05% = 0,0005

P(T+/A1) = 3% = 0,03

P(T+/A2) = 98% = 0,98

P(A2/T+) = P(T+/A2)*P(A2)/P(T+) = 0,016078

P(T+) = P(T+/A1) *P(A1) + P(T+/A2)*P(A2) = 0,030475

Zadanie 3

Atrybut 1 K+	A2 K+	A3 K+	A4 K+	A1 K-	A2 K-	A3 K-	A4 K-
4/5	3/5	3/5	2/5	1/3	1/3	1/3	3/3

X(9,K+) = 4/5 * 3/5 * 3/5 * 2/5 = 72/625

72/625 * 5/8 = 0,048

X(9,K-) = 1/3 * 1/3 * 1/3 = 1/27

1/27 * 3/8 = 0,013883

Zadanie 4

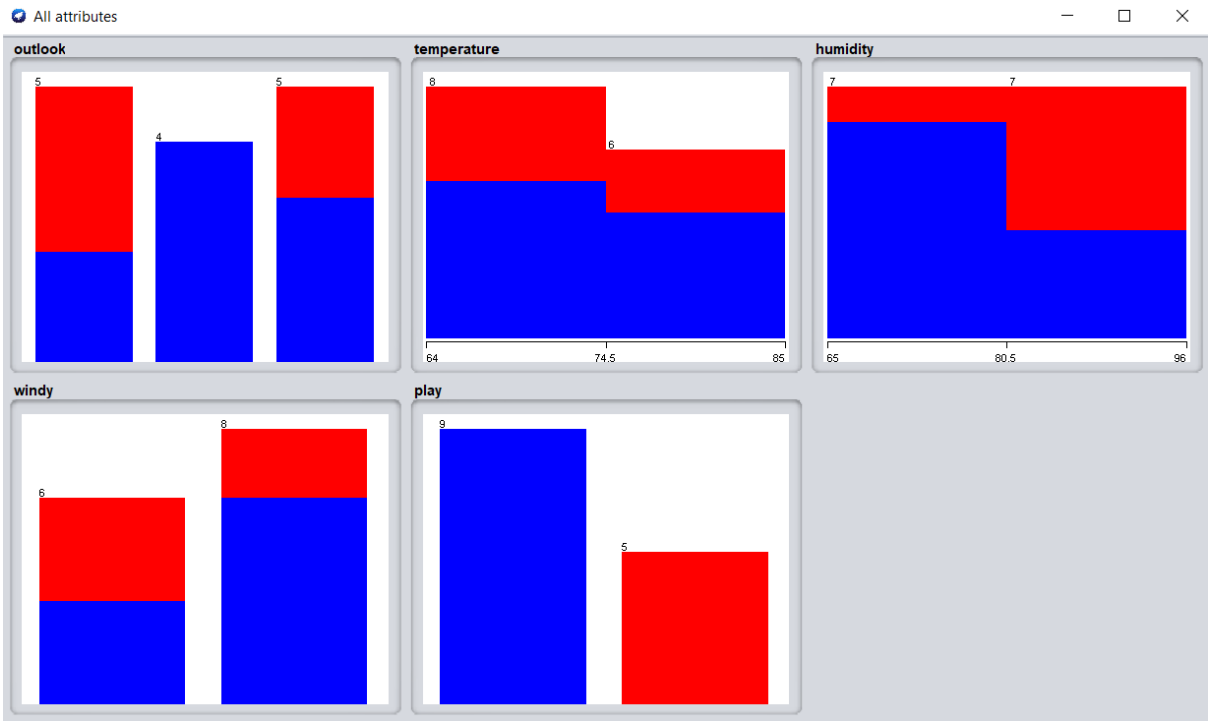
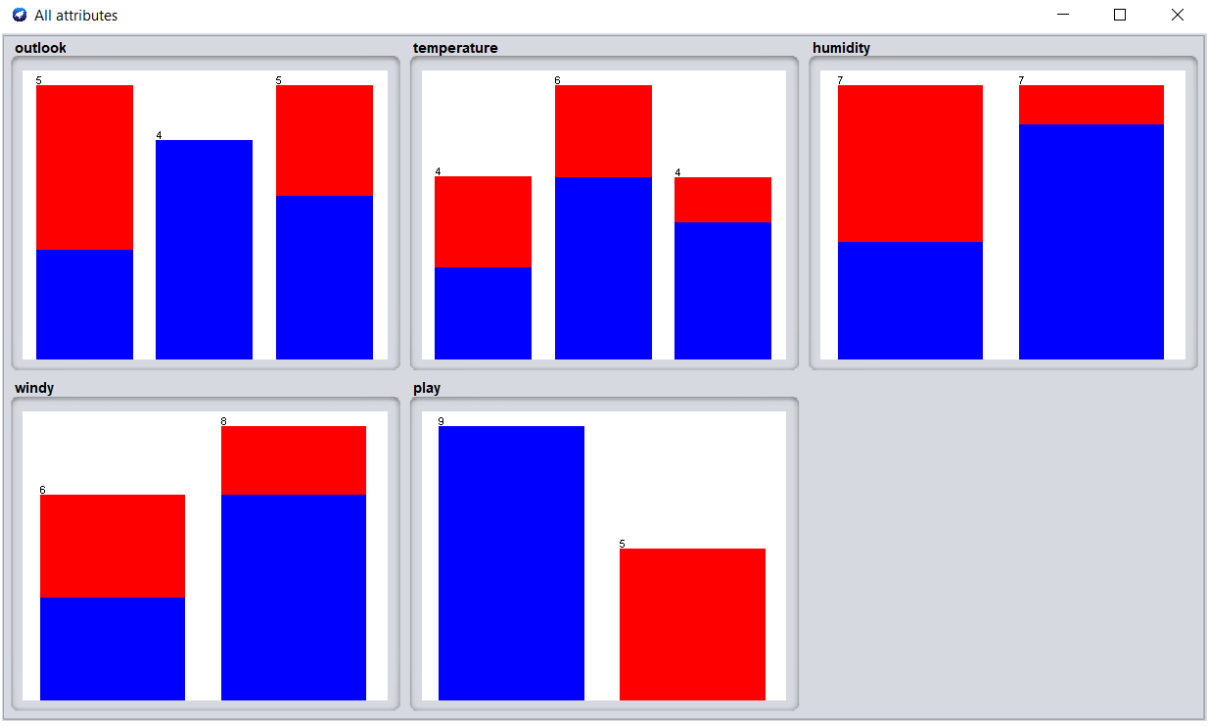
Naive Bayes

- Oparty jest na założeniu o wzajemnej niezależności zmiennych niezależnych.
- Przy jego pomocy klasyfikacja jest tak długo poprawna, jak długo poprawna klasa jest bardziej prawdopodobna od innych. Pozwala przewidzieć prawdopodobieństwo przynależności obiektu do klasy.
- Używany przy problemach o bardzo wielu wymiarach na wejściu.
- Prosty w użyciu i skuteczny
- Gdzie sobie poradzi? Z klasami binarnymi, brakującymi wartościami klasy, Atrybutami binarnymi; pustymi nominalnymi, nominalnymi, numerycznymi i jednoargumentowymi

Naive Bayes Updateable

- Używa domyślnej precyzji 0,1 dla atrybutów liczbowych przy wywołaniu z zerową liczbą instancji szkoleniowych
- Naive Bayes obsługuje uczenie się parametrów, natomiast – NB Updatable wskazuje tylko, czy implementacja obsługuje ową funkcję.

Analizę danych zaczniemy od wizualizalnego ich porównania. Jak możemy dostrzec, dwie zmienne przedstawione są w inny sposób. Są to bowiem temperature oraz humidity. Dla danych numeric zmienne te nie są rozdzielone, widać ich ciągłość.

weather numeric	
weather nominal	

Przejdźmy więc teraz już do wyników klasyfikacji Bayesa.

Weather Numeric

	NB	NBU
Total Number of Instances	14	14
Correctly Classified Instances	9 64.2857 %	9 64.2857 %
Incorrectly Classified Instances	5 35.7143 %	5 35.7143 %
Kappa statistic	0.1026	0.1026
Mean absolute error	0.4649	0.4649
Root mean squared error	0.543	0.543
Relative absolute error	97.6254 %	97.6254 %
Root relative squared error	110.051 %	110.051 %
Precision	0,607	0,607

Weather Nominal

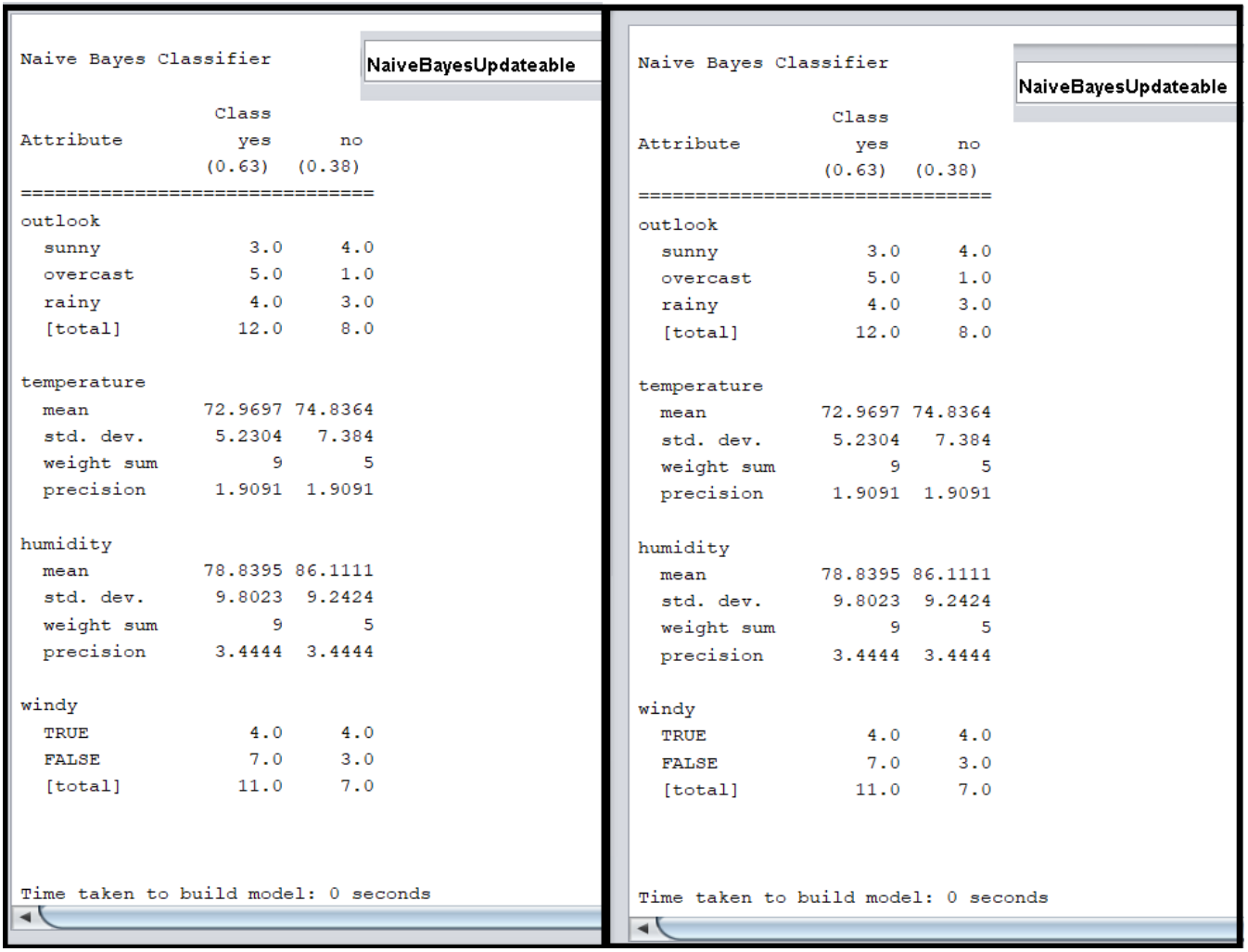
	NB	NBU
Total Number of Instances	14	14
Correctly Classified Instances	8 57.1429 %	8 57.1429 %
Incorrectly Classified Instances	6 42.8571 %	6 42.8571 %
Kappa statistic	-0.0244	-0.0244
Mean absolute error	0.4374	0.4374
Root mean squared error	0.4916	0.4916
Relative absolute error	91.8631 %	91.8631 %
Root relative squared error	99.6492 %	99.6492 %
Precision	0,528	0,528

Nominal

```
=== Confusion Matrix ===

 a b   <-- classified as
 8 1 | a = yes
 4 1 | b = no
```

Numeric

```
=== Confusion Matrix ===

 a b   <-- classified as
 7 2 | a = yes
 4 1 | b = no
```



```
Naive Bayes Classifier

                Class
Attribute         yes      no
                (0.63)   (0.38)
================================
outlook
  sunny            3.0      4.0
  overcast         5.0      1.0
  rainy            4.0      3.0
  [total]         12.0      8.0

temperature
  mean          72.9697  74.8364
  std. dev.      5.2304    7.384
  weight sum          9        5
  precision      1.9091   1.9091

humidity
  mean          78.8395  86.1111
  std. dev.      9.8023   9.2424
  weight sum          9        5
  precision      3.4444   3.4444

windy
  TRUE             4.0      4.0
  FALSE            7.0      3.0
  [total]         11.0      7.0


Time taken to build model: 0 seconds
```

NaiveBayesUpdateable

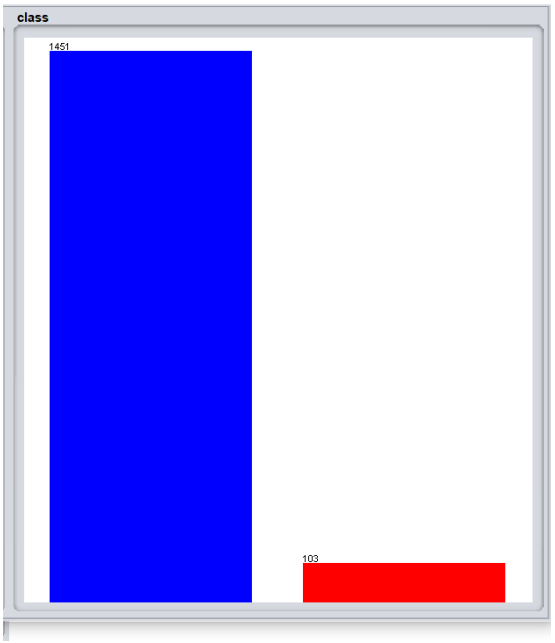
```
Naive Bayes Classifier

                Class
Attribute         yes      no
                (0.63)   (0.38)
================================
outlook
  sunny            3.0      4.0
  overcast         5.0      1.0
  rainy            4.0      3.0
  [total]         12.0      8.0

temperature
  mean          72.9697  74.8364
  std. dev.      5.2304    7.384
  weight sum          9        5
  precision      1.9091   1.9091

humidity
  mean          78.8395  86.1111
  std. dev.      9.8023   9.2424
  weight sum          9        5
  precision      3.4444   3.4444

windy
  TRUE             4.0      4.0
  FALSE            7.0      3.0
  [total]         11.0      7.0


Time taken to build model: 0 seconds
```

NaiveBayesUpdateable

Jeżeli chodzi o wartości Naive Bayes oraz Naive Bayes Updateable, dla obu danych nie zauważamy dużych różnic między poszczególnymi odpowiednimi wynikami. Dotyczy to obu plików. Można jednak dostrzec już różnice pomiędzy zbiorami weather_numeric i weather_nominals. Przede wszystkim poprawne dopasowanie dla danych _numeric jest większe. Różnica ta wynosi około 7%, czyli o wartość 1. Dla tego zbioru danych otrzymujemy również większą precyzję. Dla _nominals mamy jednak mniejszy błąd bezwzględny oraz średni i względny kwadratowy. Mocno wyróżniającym wynikiem dla niego jest ujemna wartość Kappa. Na powstające różnice ma wpływ szkielet i wygląd samych danych. Klasyfikator NB jest wysoce skalowalny oraz wymaga szeregu parametrów liniowych w stosunku do liczby zmiennych w problemie uczenia się.

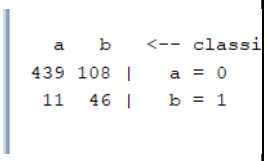
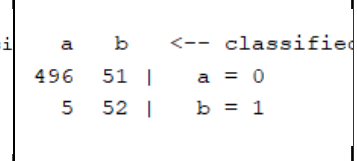
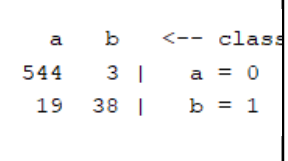
Multinominal Naive Bayes

Naszymi badanymi danymi jest ReuterGrain-train.arff. Widzimy że mamy dwa atrybuty: text (typ string) oraz class (typ nominal). Wizualnie przedstawiają się następująco:



Przejdźmy więc teraz do właściwej analizy. Otrzymane wyniki to:

	NB	NaiveBayesMultinomial	J48
Total Number of Instances	604	604	604
Correctly Classified Instances	485 80.298 %	548 90.7285 %	582 96.3576 %
Incorrectly Classified Instances	119 19.702 %	56 9.2715 %	22 3.6424 %
Kappa statistic	0.3459	0.6016	0.7563

Mean absolute error	0.1984	0.0946	0.043
Root mean squared error	0.4409	0.2944	0.1859
Relative absolute error	133.5501 %	63.6592 %	28.9093 %
Root relative squared error	150.1588 %	100.2715 %	63.3132 %
Precision	0,912	0,944	0,963
Confusion Matrix			

Na podstawie powyższych danych możemy zaobserwować, iż najlepsze wyniki otrzymał algorytm J48. Ma największą, wysoką precyzję oraz wysoką poprawną klasyfikację, a zarazem małe błędy.Wyniki dla Naive Bayes Multinomial są również wysokie oraz zbliżone do algorytmu J48. Najgorzej z tych trzech przypadków poradził sobie natomiast klasyfikator Naive Bayes. Ma najniższą precyzję oraz trafność. Multinominal jest również zdecydowanie szybszy od klasyfikatora Naive Bayes.

String To Word Vector

Filtr ten konwertuje atrybuty łańcuchów na zestaw atrybutów liczbowych reprezentujących informacje o występowaniu słów z tekstu zawartego w łańcuchach. Dane te pozyskiwane i zapamiętywane są głównie z pierwszej przefiltrowanej partii danych, zazwyczaj uczących.

- outputWordCounts - wyjściowe słowa są wyświetlane jako wartości logiczne: 0 (brak wystąpienia) albo 1 (słowo wystąpiło).
- lowerCaseTokens - wszystkie tokeny słów są konwertowane na małe litery przed dodaniem do słownika.
- useStoplist - ignoruje wszystkie słowa, które znajdują się na stopliście.

Bez dodatkowych ustawień:

Correctly Classified Instances	548	90.7285 %
Root mean squared error	0.2944	
Relative absolute error	63.6592 %	
Precision	0,944	
Time	0.16 seconds	

Z ustawnionym OutputWordCounts

Poprawa poprawnie sklasyfikowanych obiektów na 550 91.0596%
Względny błąd również się zmniejsza na 60.6365%
Podobnie jest z błędem średniokwadratowym - 0.2908
Precyzja również wzrasta 0,949
Krótszy czas potrzebny do wykonania o 0,03 s

Dodatkowe ustawienie loweCaseTokens

Ponowna poprawa poprawnie sklasyfikowanych obiektów na 554 91,7218%
Zmniejszenie błędów o około 4%, oraz jednoczesne zwiększenie precyzji
Skrócenie czasu wykonywania o 0,2 s

Kolejne ustawienie useStoplist

Correctly Classified Instances	560	92.7152 %
Root mean squared error	0.2638	
Relative absolute error	50.1639 %	
Precision	0,957	
Time	0.10 seconds	

Kolejne ustawienia więc powodowały poprawę wyników i czasu generacji. Skuteczność algorytmu wzrosła. Opcje te są zatem bardzo efektywne, potrafią zoptymalizować wyniki oraz szybkość ich otrzymywania.