Zadanie 2.1, 228884, Mateusz Guściora

2b

Próbujemy na podstawie informacji o procentowej wartości pierwiastków (zmienne Mg, Al, itd.) określić typ (Type).

Search Method: Attribute ranking.
Attribute Evaluator (supervised, Class (nominal): 10 Type): Information Gain Ranking Filter
Ranked attributes:
0.5628 6 Mg
0.5611 1 ï»¿Al
0.5307 5 K
0.4545 3 Ca
0.3334 7 Na
0.2986 8 RI
0.1884 2 Ba
0.0985 4 Fe
0 9 Si

Najbardziej wpływać (na Type) będą pierwiastki Mg,Al,K, itd. Widzimy to na powyższym rankingu. Przy klasyfikacji (domyślne) „ZeroR" przewidywaną Type jest „Window Float Processed" (ZeroR predicts class value: Window. Float Processed). Widzimy też(poniżej fragment), że błąd predykcji jest duży oraz dużo jest niezaklasyfikowanych wierszy/badań.

Correctly Classified Instances 87 40.6542 %
Incorrectly Classified Instances 127 59.3458 %
Kappa statistic 0
Mean absolute error 0.2752
Root mean squared error 0.3702
Relative absolute error 100 %
Root relative squared error 100 %
Total Number of Instances 214

2c

- obejrzyj otrzymane drzewa decyzyjne (Fit to Screen, Center on Top Node), zbiory reguł i ranking predyktorów; zapoznaj się z oceną jakości klasyfikacji; sformułuj wnioski

Algorytm klayfikacji (drzewo) Jrip najlepiej(największe TP) określił typ „window float procces"

Algorytm klasyfikacji oparty na regułach Part ma najwyższą wartość kappa oraz prawidłowo zakwalifikowanych instancji

Correctly Classified Instances 163 76.1682 %
Incorrectly Classified Instances 51 23.8318 %
Kappa statistic 0.6472

Algorytm drzewa J48 ma podobne wyniki (trochę gorsze). Poniżej wygląd drzewa

```
Mg <= 2.68
|   Na <= 13.78
|   |   ï»¿Al <= 1.38: Window. Non-float Processed (8.0/1.0)
|   |   ï»¿Al > 1.38
|   |   |   Fe <= 0.08: Container (10.0)
|   |   |   Fe > 0.08
|   |   |   |   Fe <= 0.22: Window. Non-float Processed (2.0)
|   |   |   |   Fe > 0.22
|   |   |   |   |   Ca <= 12.24: Container (2.0)
|   |   |   |   |   Ca > 12.24: Window. Non-float Processed (2.0)
|   Na > 13.78
|   |   ï»¿Al <= 1.76
|   |   |   Fe <= 0.03: Tableware (9.0/1.0)
|   |   |   Fe > 0.03: Window. Non-float Processed (2.0)
|   |   ï»¿Al > 1.76: Headlamp (26.0/2.0)
Mg > 2.68
|   ï»¿Al <= 1.41
|   |   Mg <= 3.86
|   |   |   Fe <= 0.11
|   |   |   |   RI <= 1.523: Window. Float Processed (65.0/5.0)
|   |   |   |   RI > 1.523
|   |   |   |   |   Ca <= 10.17: Headlamp (2.0/1.0)
|   |   |   |   |   Ca > 10.17: Window. Non-float Processed (2.0)
|   |   |   Fe > 0.11
|   |   |   |   K <= 0.23: Window. Float Processed (6.0)
|   |   |   |   K > 0.23
|   |   |   |   |   Mg <= 3.59
|   |   |   |   |   |   K <= 0.45: Window. Non-float Processed (2.0)
|   |   |   |   |   |   K > 0.45
|   |   |   |   |   |   |   Mg <= 3.26: Window. Non-float Processed (3.0/1.0)
|   |   |   |   |   |   |   Mg > 3.26: Window. Float Processed (7.0)
|   |   |   |   |   Mg > 3.59: Window. Non-float Processed (6.0)
|   |   Mg > 3.86
|   |   |   RI <= 1.51969: Window. Non-float Processed (6.0)
|   |   |   RI > 1.51969: Window. Float Processed (2.0)
|   ï»¿Al > 1.41
|   |   Ba <= 0
|   |   |   RI <= 1.51732: Window. Non-float Processed (40.0/4.0)
|   |   |   RI > 1.51732
|   |   |   |   RI <= 1.51797: Window. Float Processed (5.0)
|   |   |   |   RI > 1.51797: Window. Non-float Processed (5.0/1.0)
|   |   Ba > 0: Headlamp (2.0)
```

3a

Określenie problemu decyzyjnego: Określenie przewidywanej kwoty lub przedziału kwotowego na podstawie dostępnych zmiennych (oraz które zmienne będą wpływały i w jakim stopniu).

3c i 3d

Dla zmiennej nom-miesiac sprawdzamy ranking przydatności zmiennych. Widzimy, że zmienna nominalna miesiąc jest najbardziej zależna od daty rozmowy co jest oczywiste(bo w dacie też są miesiace) oraz trochę od numeru klienta, przedstawiciela, oddzialu, regionu.

Ranked attributes:
1.58451 4 data rozmowy
0.02868 1 numer klienta
0.00757 3 przedstawiciel
0.00472 5 Oddzia³
0.00268 6 Region
0 2 czas rozmowy
0 8 kwota zakupu

Dla naszego problemu decyzyjnego określenia przewidywanej kwoty lub przedziału kwoty zakupu stworzymy nominalną zmienną przedział (z kwoty zakupu która jest numeryczna).

3f

Powtórna ocena ważności zmiennych decyzyjnych(dla innej zmiennej konluzji)-tu „przedział" czyli przedziałów kwot zakupu.

Search Method: Attribute ranking.
Attribute Evaluator (supervised, Class (nominal): 9 przedzial): Information Gain Ranking Filter
Ranked attributes:
0.04694 1 numer klienta
0.01754 3 przedstawiciel
0.01615 5 Oddzia³
0.01304 6 Region
0.01086 8 przedzial_czas rozmowy
0.00569 7 miesi¹c
0 2 czas rozmowy
0 4 data rozmowy

3g

Pracując na zbiorze i stosując filtr InterquartileRange nie zaobserwowaliśmy zmiennych ekstremalnych (wszystkie „no") oraz wartości odstających (wszystkie „no"), więc usuwamy te cechy.

3h

Pogrupowałem czas rozmowy tworząc przedziały (6) używając filtra z grupy unsupervised/attributes. Dla sprawdzenia pogrupowałem wg filtra z grupy supervised, który stworzył jedną grupę.

//uwaga: Zmieniłem nazwy przedziałów czasu rozmowy w notatniku(plik: „228884_klienci_2.arff"). Nie zmieniłem nazw przedziału (kwot) z obawy możliwych niekompatybilności np. z plikami testowy, train. (Zmiany nazw robiłem na końcu pracy).

3l

Search Method: Attribute ranking.
Attribute Evaluator (supervised, Class (nominal): 10 przedzial): Information Gain Ranking Filter
Ranked attributes:
0.04694 1 numer klienta
0.0368 9 DzienTygodnia
0.01754 3 przedstawiciel
0.01615 5 Oddzia³
0.01304 6 Region
0.01086 8 przedzial_czas rozmowy
0.00569 7 miesi¹c
0 2 czas rozmowy
0 4 data rozmowy

Najważniejszymi zmiennymi (wobec przedziału kwotowego) będą numer klienta, dzień tygodnia. Trochę mniej zależne (ale wciąż zależne) będą przedstawiciel, oddzial, region, przedzial_czas rozmowy, miesiąc.

3ł.

	Jrip	PART
Incorrectly Classified Instances	81.5 %	81.9 %
Kappa	0.0051	0.0171
Średni błąd predykcji	0.277	0.2773

Bardzo niski wskaźnik kappa oraz wysoki procent „Incorrectly Classified Instances" świadczy o bardzo niskim poziomie przewidywanego przedziału //nie wiem w czym jest błąd(czy zmienne?)

Dla algorytmu Jrip i dla przedziału '(156.666667-190.833333]' TP rate wynosi 0,994 co jest najwyższym wynikiem dlatego jest to przewidywany przedział (kwoty).

3n

Przewidywana kwota to 83.6706, Średni błąd predykcji to 51.3947.

W modelu są warunki np. gdy DzienTygodnia="sob.","czw.","niedz.","?r." to dodajemy do wartości przewidywanej +7.9146

Linear Regression Model
kwota zakupu =
11.0463 * numer klienta=1,9,11,10,12,6,5,4,3,7,8,2,15 +
10.6515 * numer klienta=8,2,15 +
8.0234 * przedstawiciel=P03,P05,P04,P01 +
6.5877 * przedstawiciel=P04,P01 +
6.3248 * miesi¹c=10,8 +
6.4753 * przedzial_czas rozmowy=31-60,120-150 +
8.5855 * DzienTygodnia="pt.","wt.","sob.","czw.","niedz.","?r." +

7.9146 * DzienTygodnia="sob.","czw.","niedz.","?r." +
83.6706

Correlation coefficient	0.039
Mean absolute error	51.3947
Root mean squared error	59.5594
Relative absolute error	100.5667 %
Root relative squared error	100.9268 %