Monika Szymańska 238359

Zadanie 3.1

Uruchomienie przykładowego procesu: Direct Marketing

RapidMiner jest programem, który pozwala nam dokonywać różnego rodzaju analiz na danych, korzystając z różnego rodzaju algorytmów. Pierwszy kontakt z tym środowiskiem odbywał się na przykładzie Direct Marketing.

W bloku CrossValidation został zastąpiony dotychczasowy algorytm NaiveBayes, algorytmem DecisionTree i w jego wyniku otrzymaliśmy drzewo klasyfikacyjne o bardzo złożonej strukturze. Aby uzyskać bardziej skondensowany wynik, poszczególne parametry zostały zmienione (Tabela1).

Zmienne	Pierwotne wartości	Wartości po zmianie
Minimal gain	0,01	0,18
Minimal leaf size	2	7
Minimal size for split	4	3
Numer of prepruning alternatives	3	4

Tabela 1 - Zmiana parametrów algorytmu DecisionTree na zbiorze Direct Marketing

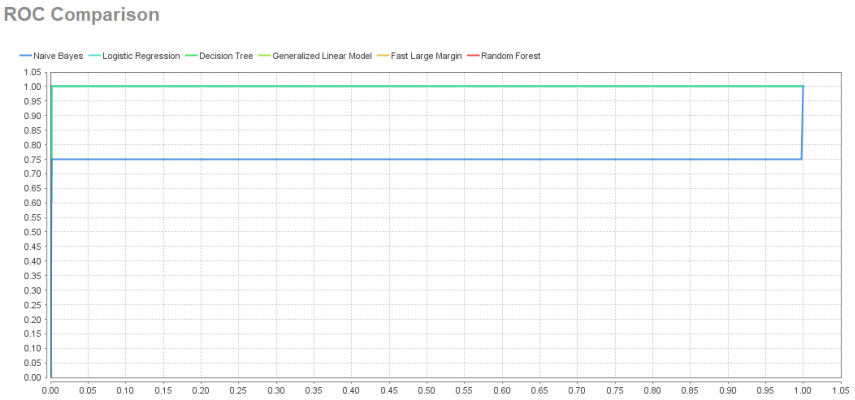
Otrzymane drzewo zostało zapisane do pliku DecisionTree_238359.png. Z przeprowadzonych prób można dodać, że najbardziej na drzewo wpływa zmiana pierwszego oraz drugiego parametru wymienionego w Tabeli1, przy trzecim widzimy zmianę w przypadku dużej zmiany, natomiast ostatni nie wpłynął na wygląd drzewa.

W trybie Auto Model istnieje możliwość dokonania predykcji zmiennych. W tym przypadku jako zmienna wyjściowa zostało wybrane pole „Sales total". Wyniki dla poszczególnych algorytmów zostały przedstawione w Tabeli2.

Algorytm	Wynik predykcji
Naive Bayes	0,9930
Generalized Linear Model	0,9966
Logistic Regression	0,9754
Fast Large Margin	0,9966
Decision Tree	0,9930
Random Forest	0,9966

Tabela 2 - Wyniki predykcji dla poszczególnych algorytmów przy zastosowaniu opcji Auto Model

Jak można zauważyć powyżej, wszystkie algorytmy charakteryzują się bardzo wysokim wynikiem predykcji. Dodatkowo na Rysunku 1 została zaprezentowana krzywa ROC.



Rysunek 1 - Krzywa ROC przy zastosowaniu poszczególnych algorytmów dla danych Direct Marketing

Budowa nowego procesu dla danych klienci6.arff

Dla zbioru danych kienci6.arff utworzony został model procesu klasyfikacyjnego z wykorzystaniem algorytmów Decision Tree oraz Rule Induction. W ich wyniku otrzymaliśmy obraz drzewa klasyfikacyjnego, którego fragment został zapisane jako drzewo_klienci6_1.png oraz zbiór reguł klasyfikacyjnych, w wyniku którego poprawie przyporządkowanych zostało 591 z 995 argumentów.

W kolejnym kroku zbiór danych został podzielony na dwie części za pomocą bloku Split Data i ponownie dokonano klasyfikacji przy pomocy Decision Tree oraz Rule Induction. Dla obu zbiorów danych otrzymaliśmy bardzo rozległe drzewa klasyfikacyjne, których fragmenty zostały zapisane kolejno jako drzewo_klienci6_2.png oraz drzewo_klienci6_3.png. Reguły klasyfikacyjne pozwoliły natomiast dla pierwszego zbioru danych przyporządkować poprawnie 339 argumentów z 494, a dla drugiego 306 z 495.

Do programu Rapid Miner został doinstalowany pakiet Weka Extension, dzięki czemu można było zastosować algorytm W-JRip. W jego wyniku otrzymaliśmy zbiór 6 reguł klasyfikacyjnych, a zatem był on znacznie krótszy od tego wygenerowanego przez algorytm Rule Induction.