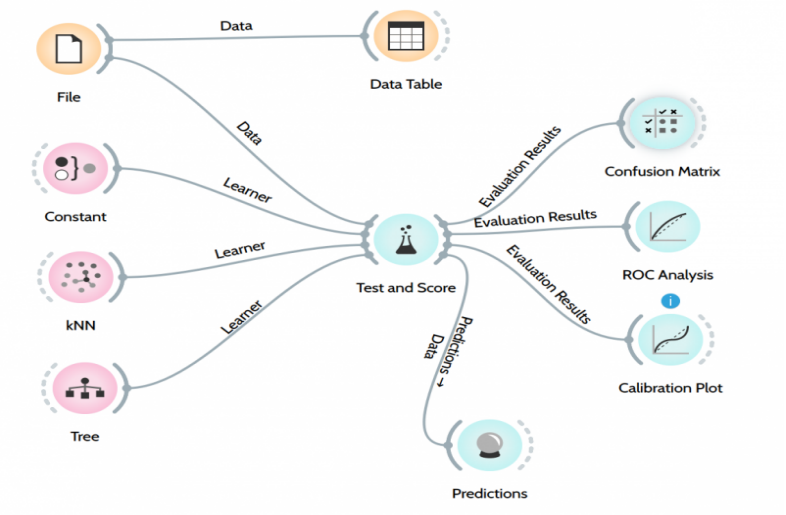
Wprowadzenie do klasyfikacji

Celem laboratorium jest przedstawienie podstawowych pojęć wykorzystywanych w zadaniach klasyfikacji, takich jak: *zbiór uczący*, *zbiór testujący*, *walidacja krzyżowa*, czy *macierz pomyłek*. W trakcie laboratorium sprawdzamy, jak wykonać najbardziej podstawowe algorytmy klasyfikacji w środowiskach Orange Data Mining i RapidMiner.

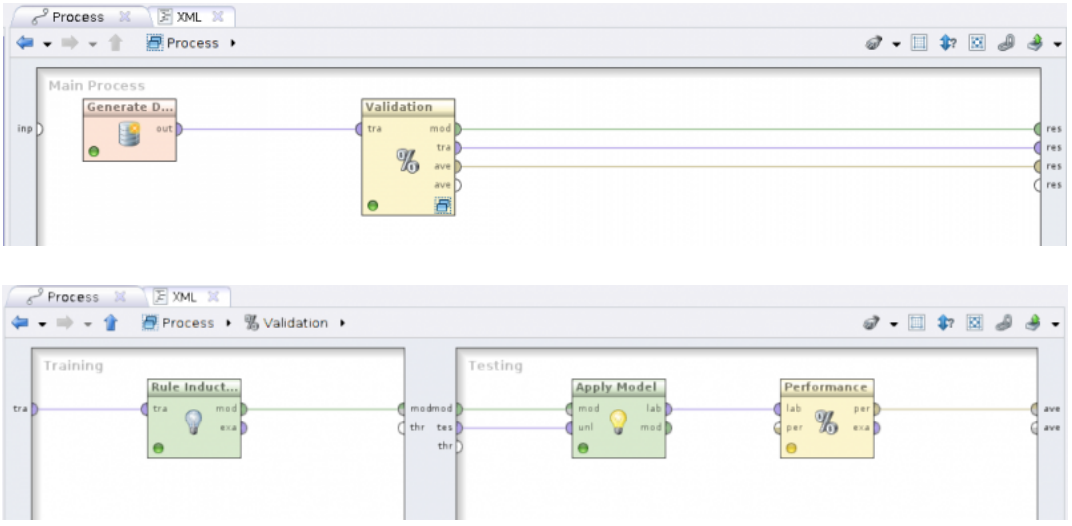
Orange Data Mining

- Uruchom narzędzie Orange Data Mining i korzystając z operatora `File` załaduj zbiór `titanic.tab`. Prześlij zbiór do operatora `Data Table` i zapoznaj się z jego charakterystyką. Wykorzystaj znane Ci narzędzia do wizualizacji aby lepiej poznać rozkłady poszczególnych zmiennych.
- Wyślij dane do operatora `Test and Score`.
- Dodaj do przepływu operator `Constant` i prześlij jego wynik do operatora `Test and Score`. Obejrzyj zawartość operatora `Test and Score`. Czy potrafisz powiedzieć, dlaczego dokładność klasyfikacji (CA) wynosi 67.7%?
- Dodaj operator `Confusion Matrix` i prześlij do niego wynik operatora `Test and Score`. Spróbuj samodzielnie zinterpretować uzyskaną macierz pomyłek.
- Dodaj operatory `Tree` i `k-NN` i prześlij je do operatora `Test and Score`. Porównaj główne metryki trzech modeli wewnątrz operatora `Test and Score`.
- Zmień sposób podziału danych na zbiór uczący i testujący. Sprawdź, czy zauważasz istotne różnice jeśli chodzi o trafność klasyfikacji. Co się dzieje, gdy testowanie odbywa się na zbiorze trenującym?
- Sprawdź w jaki sposób wybór liczby podziałów (k-folds) w walidacji krzyżowej wpływa na trafność klasyfikacji.
- Dodaj do przepływu operator `ROC Analysis` i porównaj ze sobą trzy analizowane modele klasyfikacji.
- Dodaj do przepływu operator `Calibration Plot` i sprawdź, w jakich zakresach modele są nadmiernie pesymistyczne/optymistyczne.
- Dodaj do przepływu operator `Predictions` i zaobserwuj, w jaki sposób poszczególne modele dokonują predykcji dla instancji.
- Twój ostateczny przepływ powinien wyglądać następująco:


Data
File
Data Table
Data
Constant
Learner
Learner
kNN
Learner
Tree
Test and Score
Evaluation Results
Confusion Matrix
Evaluation Results
ROC Analysis
Evaluation Results
Calibration Plot
Data
Predictions
Predictions

RapidMiner

- Uruchom narzędzie RapidMiner
- Znajdź operator `Generate Data`. Wskaż jako liczbę generowanych obiektów 1000, a jako funkcję zmiennej celu podaj *two gaussians classification*. Liczbę atrybutów ustaw na 2, ich zakres możesz zostawić z wartościami domyślnymi. Uruchom swój przepływ i obejrzyj wygenerowany zbiór danych. Zmień funkcję zmiennej celu na *gaussian mixture clusters* i jeszcze raz obejrzyj wynik.
- Wstaw do przepływu operator `Split Validation` i ustaw proporcje 60%-40%. Zauważ, że jest to operator dominujący, który wymaga sprecyzowania operatorów wewnętrznych.
- Kliknij dwukrotnie na operatorze `Split Validation`. W sekcji *Training* umieść operator `Rule Induction` i prześlij na wejście operatora zbiór trenujący, zaś wyjście operatora oznaczone mod (*model*) przekaż dalej.
- W sekcji *Testing* umieść sekwencję operatorów `Apply Model` i `Performance (Classification)`, przesyłając do operatora `Apply Model` przepływy mod i tes (*testing set*). Etykietowane dane z operatora `Apply Model` (port lab (*labeled data*)) przekaż do operatora `Performance (Classification)`. Port wyjściowy per (*performance vector*) przekaż jako wynik działania całego operatora złożonego.



- Uruchom przepływ i zaobserwuj uzyskane wyniki.
- Zamień operator `Split Validation` na `X-Validation` ustawiając 10-krotną walidację krzyżową. Zamień zbiór danych na Iris, a algorytm do klasyfikacji kolejno na `Tree to Rules` (to także operator dominujący, do środka możesz wstawić `Decision Tree` lub `Random Tree`) i `k-NN`. Za każdym razem sprawdź uzyskiwane wyniki.