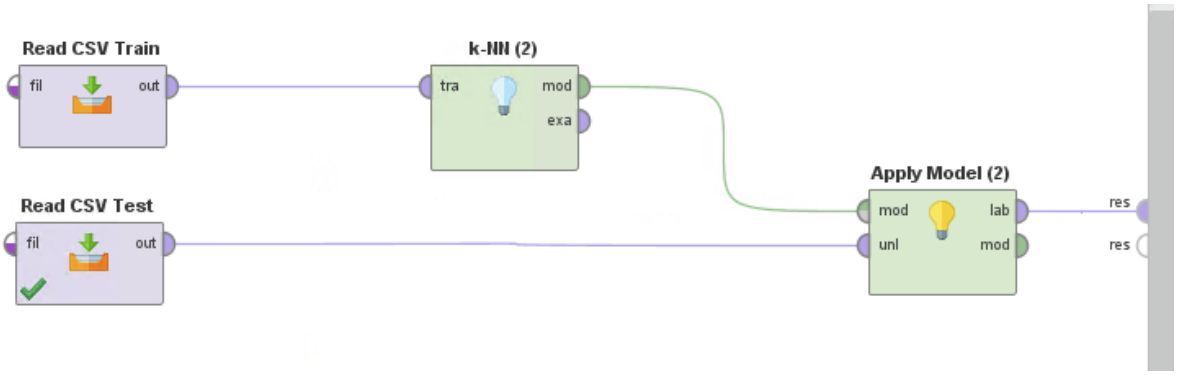
Klasyfikator kNN

Zadania

1. Co to jest wariancja modelu i obciążenie modelu
2. Badanie wpływu parametru *k* klasyfikatora k-NN na generalizację:
    a. Stwórz proces jak na obrazku



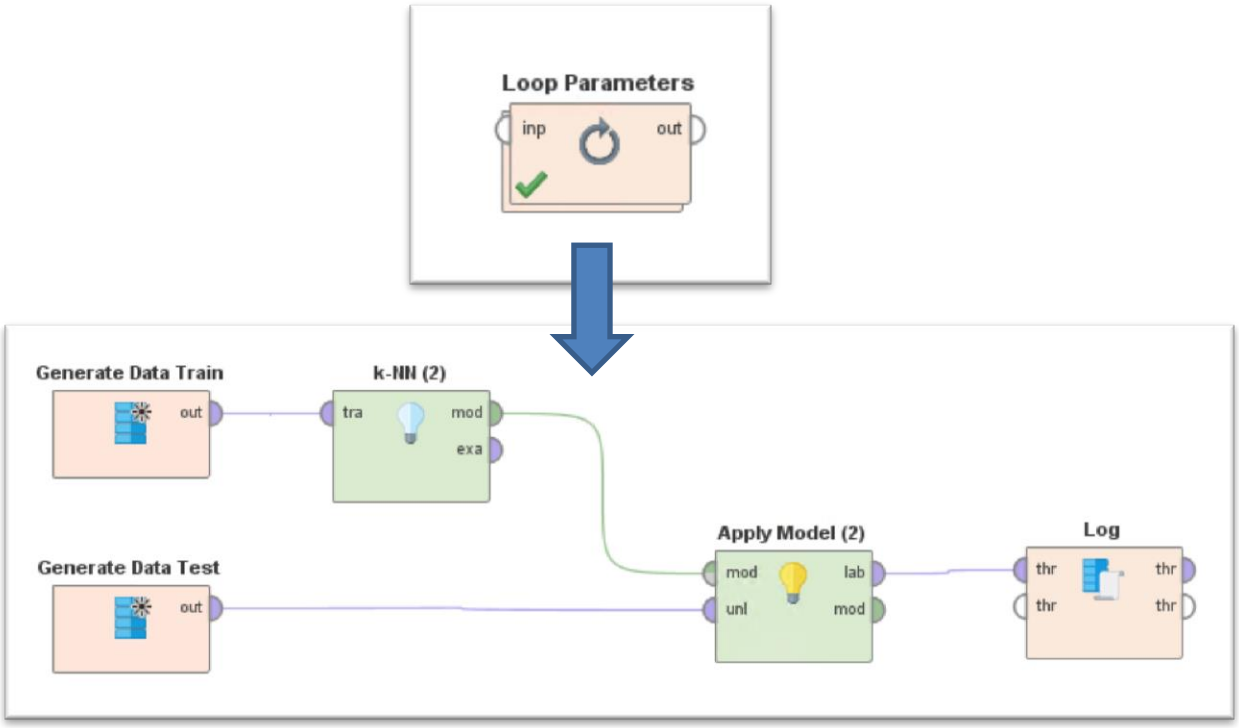
    b. Wczytaj zbiory danych *dataTrain_2G.csv* oraz *dataTest_2G.csv* – zbiory te reprezentują rozkłady problem klasyfikacji dwu klasowej, dla dwóch nakładających się rozkładów Gaussa o tej samej wariancji ale różnym położeniu środka rozkładów.
    c. Jaka jest w tym przypadku optymalna granica decyzji
    d. Dokonaj ewaluacji klasyfikatora kNN dla różnych parametrów *k* z zakresu od 1 do 300 (wartości k dobierz tak by zarejestrować zależność dokładność - k). Zarejestruj wykresy przedstawiające zależność x_1,x_2 dla wartości przewidywanych. (zastanów się)
    e. Zarejestruj dokładność predykcji uzyskana na zbiorze testowym z wykorzystaniem operatora *Performance*
    f. Narysuj zależność dokładności w funkcji wartości parametru *k*

    g. Wczytaj zbiór *dataTrain_GL.csv* i *dataTest_GL.csv* i powtórz podpunkty a,b,c,d,e,f.
    h. Czy oraz jak wpływa parametr k na dokładność predykcji.

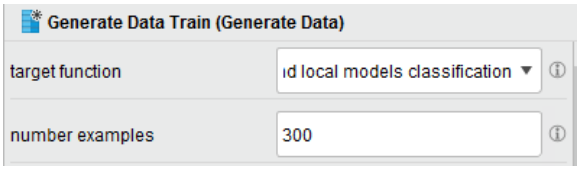
    i. Dla jakiej wartości parametru k uzyskujemy przeuczenie, a dla jakiej pojawia się obciążenie modelu. Uzasadnij swoją odpowiedź.

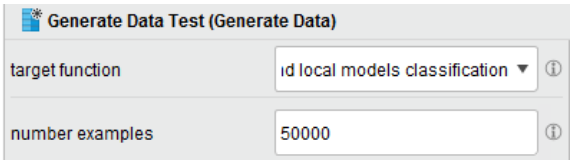
    j. Dla zbioru *dataTrain_GL.csv* wykonaj identyczne obliczenia z wykorzystaniem testu krzyżowego. Czy wyniki uzyskane na teście krzyżowym różnią się od wyników uzyskanych z wykorzystaniem zbioru testowego – narysuj wykres pokazujący dokładność predykcji uzyskaną z testu krzyżowego i dokładności uzyskanej na zbiorze testowym
    k. Czy są różnice? Skąd one wynikają?
3. Badanie złożoności obliczeniowej klasyfikatora kNN.
   W celu zbadania złożoności obliczeniowej dla zadanego (dużego) zbioru testowego zmieniaj rozmiar zbioru treningowego oraz zmieniaj wartość k i zarejestrować czas potrzebny na uczenie i predykcję. W tym celu

a. Stwórz proces jak na obrazku:

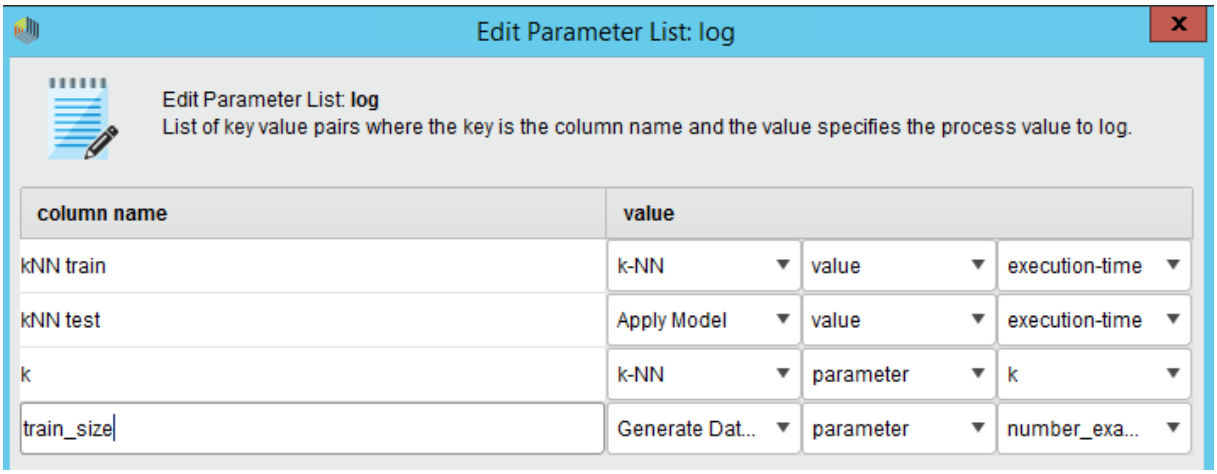


b. Ustaw operatory Generate Data odpowiednio jak na obrazku:





c. Ustaw operator log jak na obrazku:



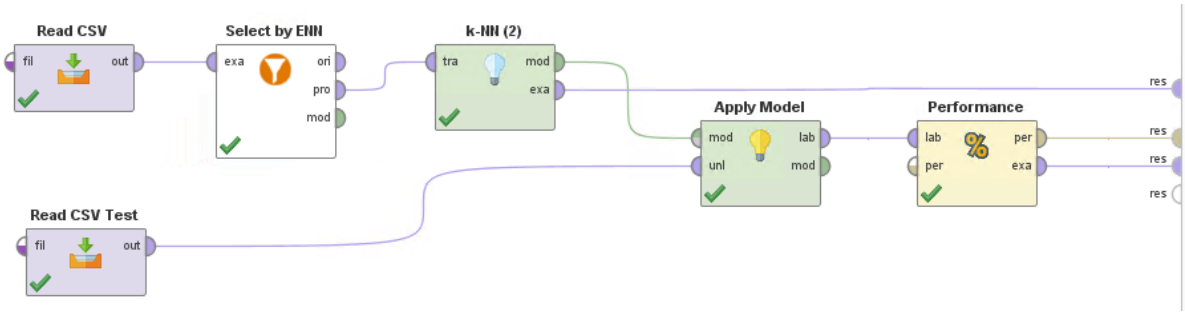
d. W operatorze loop parameters zmieniaj parametr k klasyfikatora kNN w zakresie 1-200 (10 różnych wartości) oraz parametr *number_of_examples* w opratorze *Generate Data Train* w zakresie 100-3000

e. Na podstawie uzyskanych wyników wykreśl zależności pokazujące jak poszczególne z parametrów wpływają na złożoność obliczeniową klasyfikatora kNN

f. Który z parametrów ma największy wpływ na złożoność?

g. Jakie kroki możemy podjąć by zapewnić niską złożoność obliczeniową klasyfikatora kNN
h. Poszukaj w internecie i zaproponuj rozwiązania

4. Dowiedz się z internetu co to jest Instance Selection w Machine Learning (używaj angielskiego nazewnictwa)
   a. Jeśli RapidMiner nie posiada zainstalowanego dodatku *Information Selection Extension* to zainstaluj go z *RapidMiner MarketPlace*
   b. Skorzystaj z procesu opisanego w zadaniu 2 ze zbiorami *dataTrain_GL.csv* i *dataTest_GL.csv*
   c. Dodaj operator *Select by ENN* tak aby powstał proces jak niżej



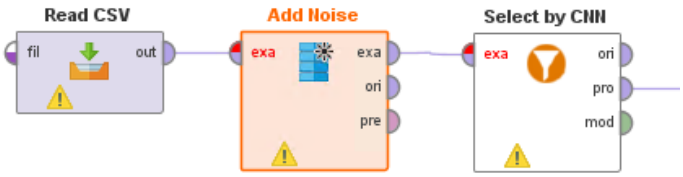
   d. Ustaw w klasyfikatorze kNN wartość k=1 i dobierz wartość parametru k w operatorze *Select by ENN* tak by zapewnić maksymalną dokładność
   e. Zaobserwuj jak parametr *k* w operatorze *Select by ENN* wpływa na zbiór uczący który trafia do klasyfikatora kNN
   f. Zarejestruj w procentach o ile zmniejszył się zbiór uczący dla klasyfikatora kNN względem zbioru całego zbioru danych. Parametr ten nazywany jest kompresją

$$cmp = \frac{|T| - |P|}{|T|}$$

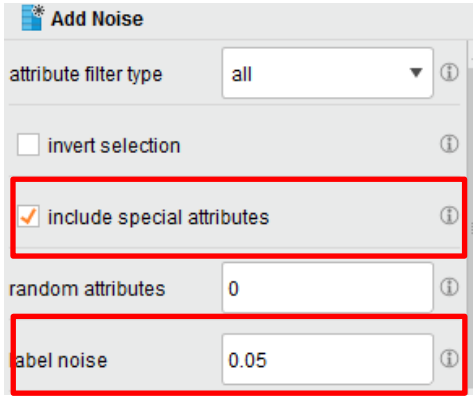
Gdzie T jest zbiorem danych przed kompresją, a P jest zbiorem danych po kompresji i || jest operatorem zwracającym rozmiar zbioru danych

Na obrazkach pokaż zbiór danych przed selekcją instancji i po dla najlepszej wartości parametru k w ENN. Jak scharakteryzowałbyś (opisał) punkty które zostały usunięte? – gdzie one sa położone?

   g. Zamień operator Select by ENN na operator Select by CNN i powtórz obliczenia. (operator CNN nie posiada żadnych parametrów)
   Zobacz jak operator Select by CNN wpływa na zbiór uczący klasyfikatora kNN, zmierz kompresję zbioru danych.
   Na obrazkach pokaż zbiór danych przed selekcją instancji i po dla algorytmu CNN. Jak scharakteryzowałbyś (opisał) punkty które zostały usunięte? – gdzie one sa położone?
   h. Dodaj operator Add Noise pomiędzy zbiór treningowy i algorytm selekcji instancji jak pokazano na obrazku



Następnie w konfiguracji operatora Add Noise dokonaj zmian w *Label Noise*



Zwiększając wartość parametru. Sprawdź jak poziom szumu w danych będzie wpływał na kompresję i dokładność predykcji klasyfikatora 1NN.
Zrób wykres pokazujący relację: poziom szumu – dokładność oraz poziom szumu-kompresja. Obliczenia powtórz dla obydwu algorytmów.

i. Który algorytm ma większą kompresją
j. Który algorytm ma większą dokładność
k. Czy algorytmy zachowują się podobnie?

Zaproponuj jak zintegrować korzyści poszczególnych rozwiązań. Dokonaj odpowiedniej modyfikacji procesu i przeprowadź badanie
Na obrazkach pokaż zbiór danych przed selekcją instancji i po selekcji

5. Badanie wpływu metod selekcji instancji na predyktory
   a. Zbuduj proces jak w zadaniu 3.
   b. Ustaw rozmiar zbioru treningowego na 20.000 wektorów, a testowego pozostaw jako (duży) i sprawdź jak wpływa zastosowanie każdej z metod selekcji instancji na czas predykcji klasyfikatora.
   c. Na podstawie wszystkich przeprowadzonych badań opisz jakie widzisz zastosowania metod selekcji instancji.