SPRAWOZDANIE

Zajęcia: Uczenie Maszynowe

Prowadzący: prof. dr hab. inż. Vasyl Martsenyuk

|  |  |
| --- | --- |
| Laboratorium Nr 1  Data 27.09.2025  Temat: " Praktyczne zastosowanie regresji liniowej w analizie danych. Implementacja algorytmów klasyfikacji binarnej w Pythonie"  Wariant 9 | Artur Rolak  Informatyka  II stopień, stacjonarne,  1semestr, gr.1bS |

1. Polecenie: wariant 1 zadania

Opracować przepływ pracy uczenia maszynowego dla zagadnienia regresji (model regresji liniowej) oraz klasyfikacji binarnej (model SVM) na podstawie zbioru danych, zgodnie z wariantem zadania nr 9.

2. Opis programu opracowanego (kody źródłowe, rzuty ekranu)

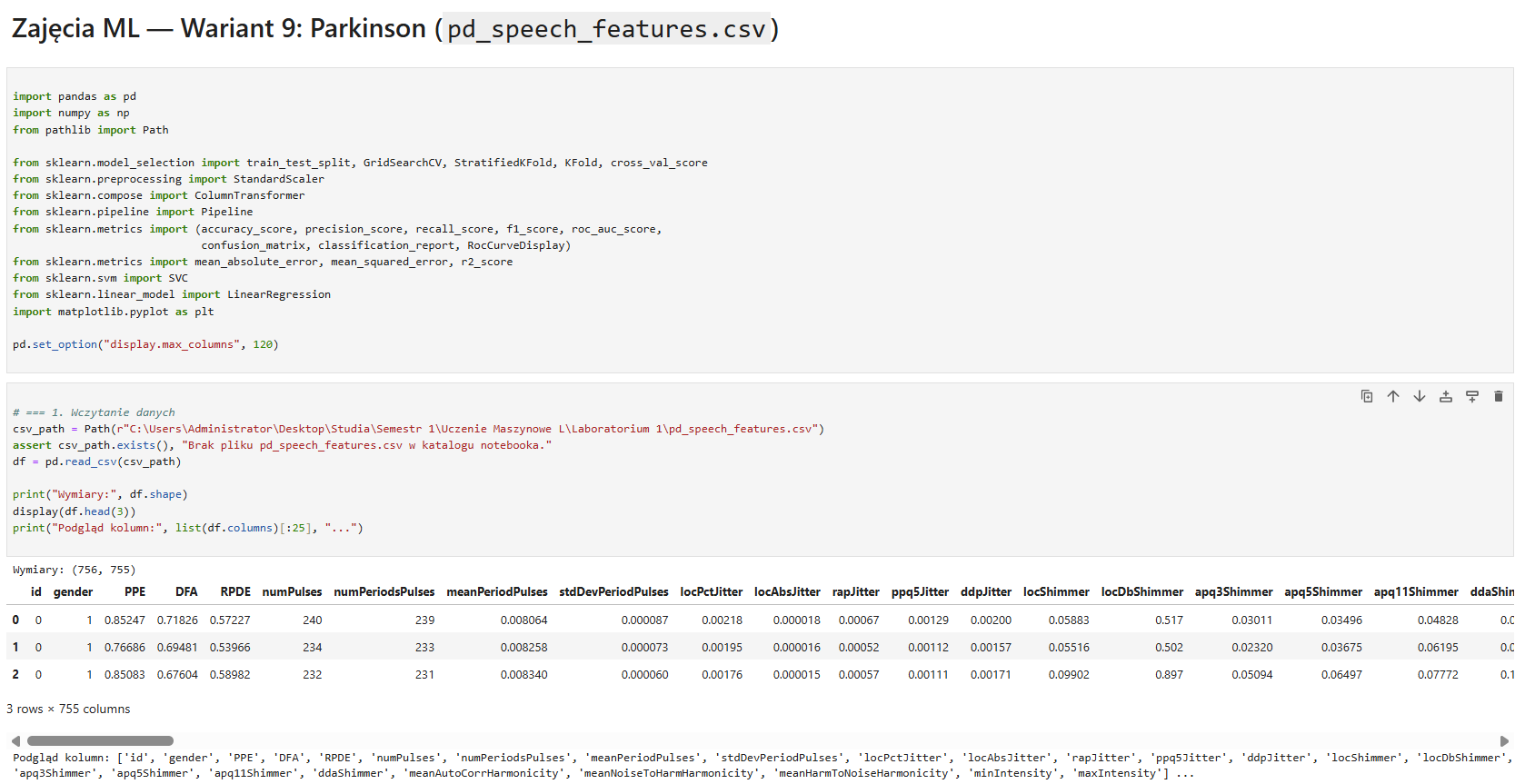
Celem programu jest zbudowanie dwóch modeli uczenia maszynowego na tym samym zbiorze danych pd\_speech\_features.csv, zawierającym cechy akustyczne mowy pacjentów z chorobą Parkinsona. Dane pochodzą z badań klinicznych i zawierają zestaw numerycznych współczynników opisujących nagrania głosu. Zadanie polega na przeprowadzeniu dwóch analiz:

1. **Klasyfikacja (SVM)** – Przewidzenie czy pacjent cierpi na chorobę Parkinsona, na podstawie kolumny klasy (class lub status).
2. **Regresja (Linear Regression)** – przewidzenie wartości jednej z ciągłych zmiennych (np. age lub innej wybranej cechy) w oparciu o pozostałe kolumny.

Program został przygotowany w języku **Python** z użyciem bibliotek **pandas**, **scikit-learn** oraz **matplotlib**, w formie interaktywnego notebooka Jupyter. Składa się z następujących etapów:

1. **Wczytanie danych** – plik CSV z danymi mowy pacjentów jest wczytywany do obiektu DataFrame.
2. **Wstępne przetwarzanie** – usuwane są kolumny o stałych wartościach, identyfikatory oraz ewentualne braki danych. Wszystkie cechy numeryczne są standaryzowane (StandardScaler).
3. **Podział danych** – zbiór jest dzielony na część treningową (80%) i testową (20%). Dla klasyfikacji stosowana jest stratyfikacja względem klasy.
4. **Model klasyfikacyjny SVM** – utworzony zostaje potok (Pipeline) zawierający skalowanie i klasyfikator SVC. Następnie wykonywane jest strojenie hiperparametrów (C, kernel, gamma) metodą walidacji krzyżowej (GridSearchCV).
5. **Ocena klasyfikatora** – obliczane są metryki jakości: Accuracy, Precision, Recall, F1 oraz ROC-AUC. Dodatkowo generowana jest macierz pomyłek i wykres krzywej ROC.
6. **Model regresyjny (Linear Regression)** – drugi potok wykorzystuje klasyczny model regresji liniowej do przewidywania wybranej zmiennej ciągłej. Wyniki są oceniane na podstawie błędów MAE, RMSE oraz współczynnika determinacji R².
7. **Walidacja i porównanie wyników** – dla obu modeli wykonywana jest 5-krotna walidacja krzyżowa, a uzyskane wyniki są porównywane w kontekście skuteczności i stabilności.

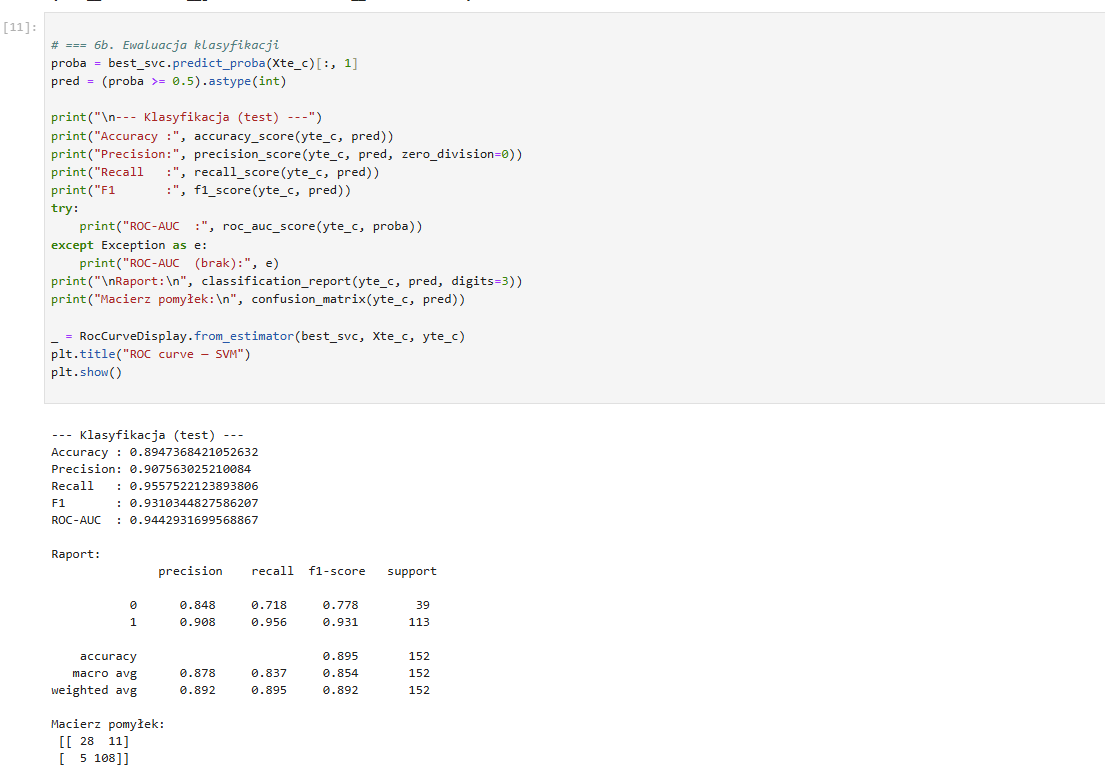
Program pozwala w przejrzysty sposób zrealizować cały cykl uczenia maszynowego: od przygotowania danych, przez modelowanie, aż po ewaluację wyników. Dodatkowo jego struktura jest zgodna z przepływem pracy w środowisku KNIME, dzięki czemu te same kroki można łatwo odwzorować w graficznym workflow.

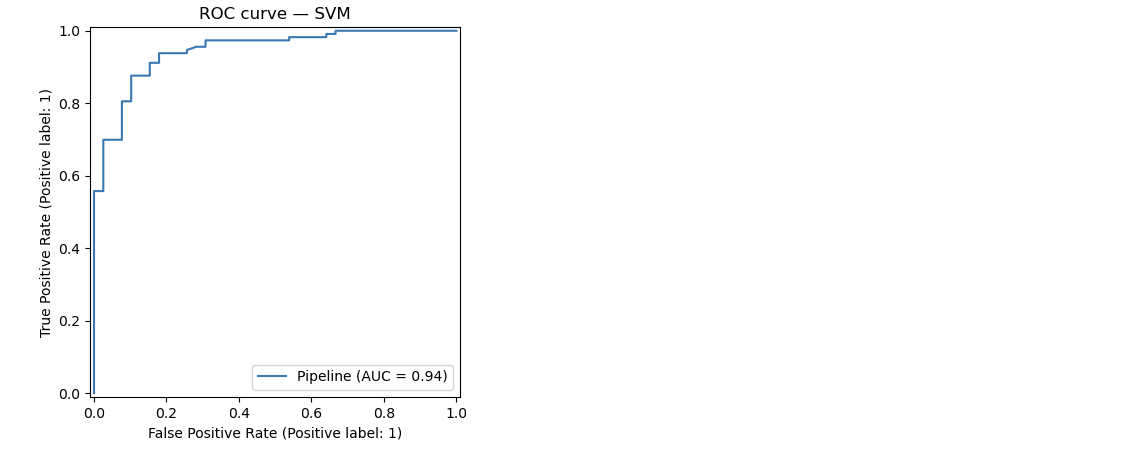




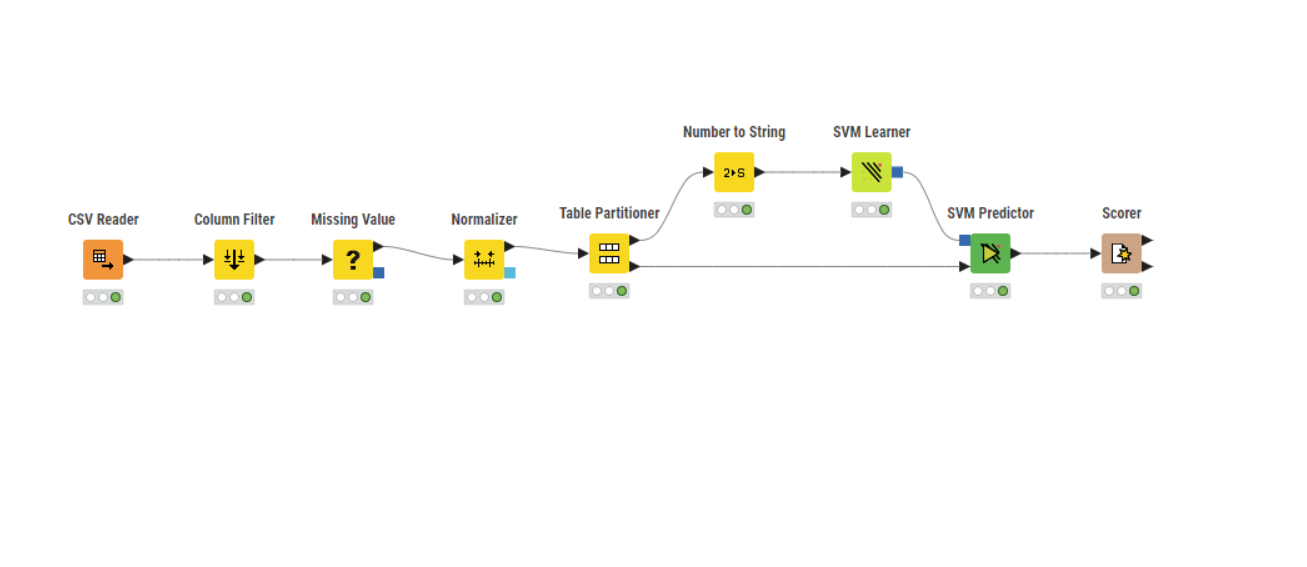


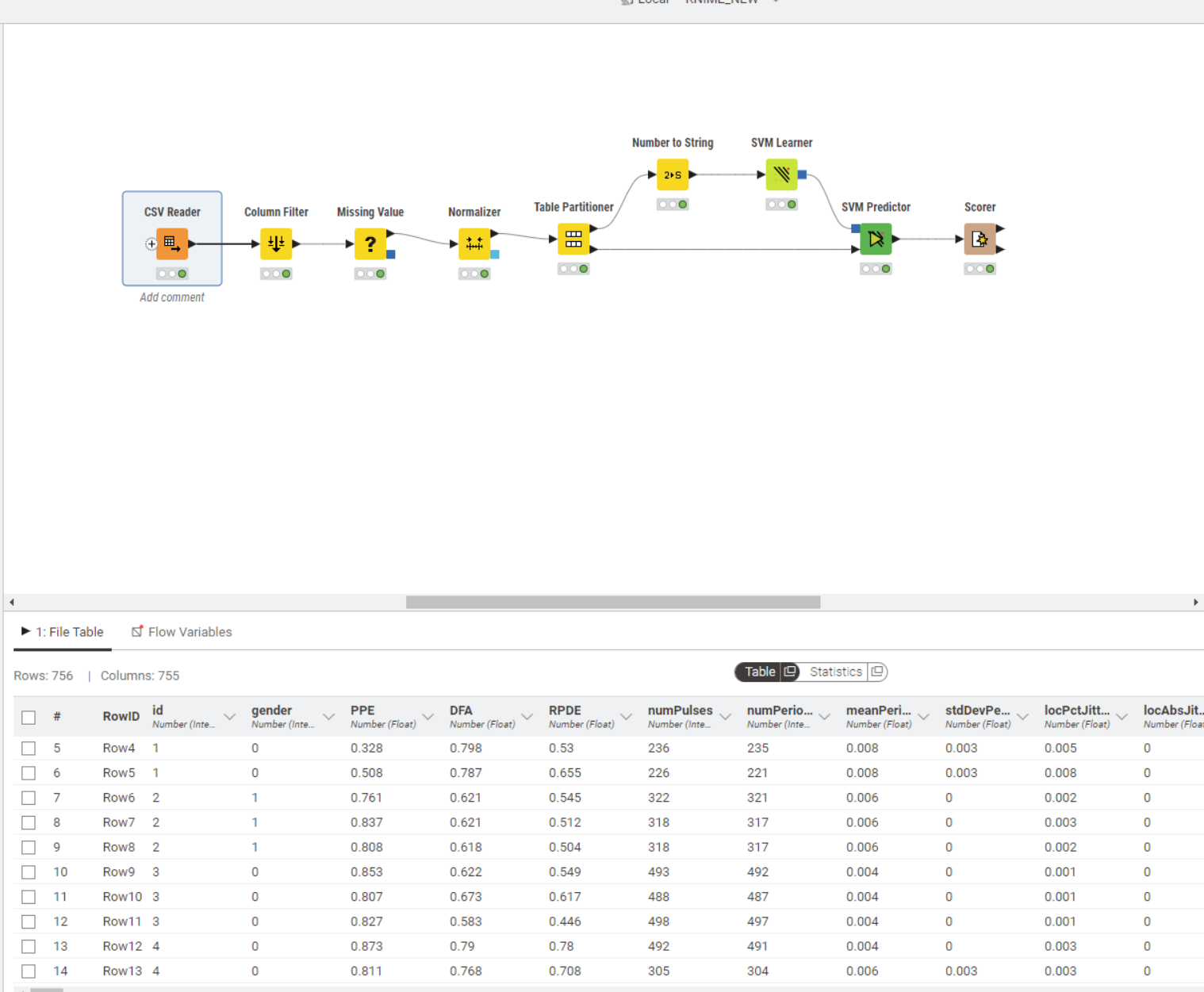


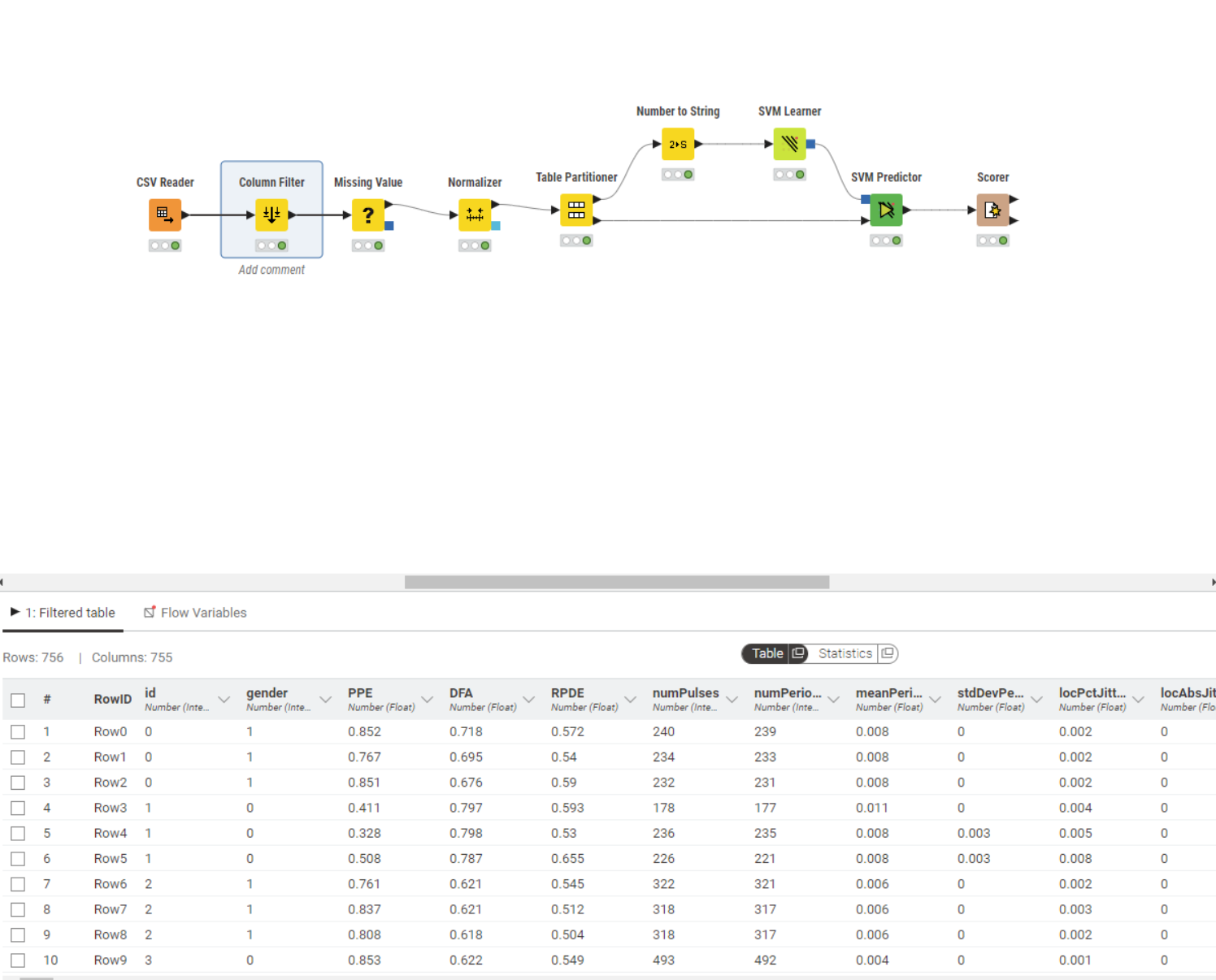


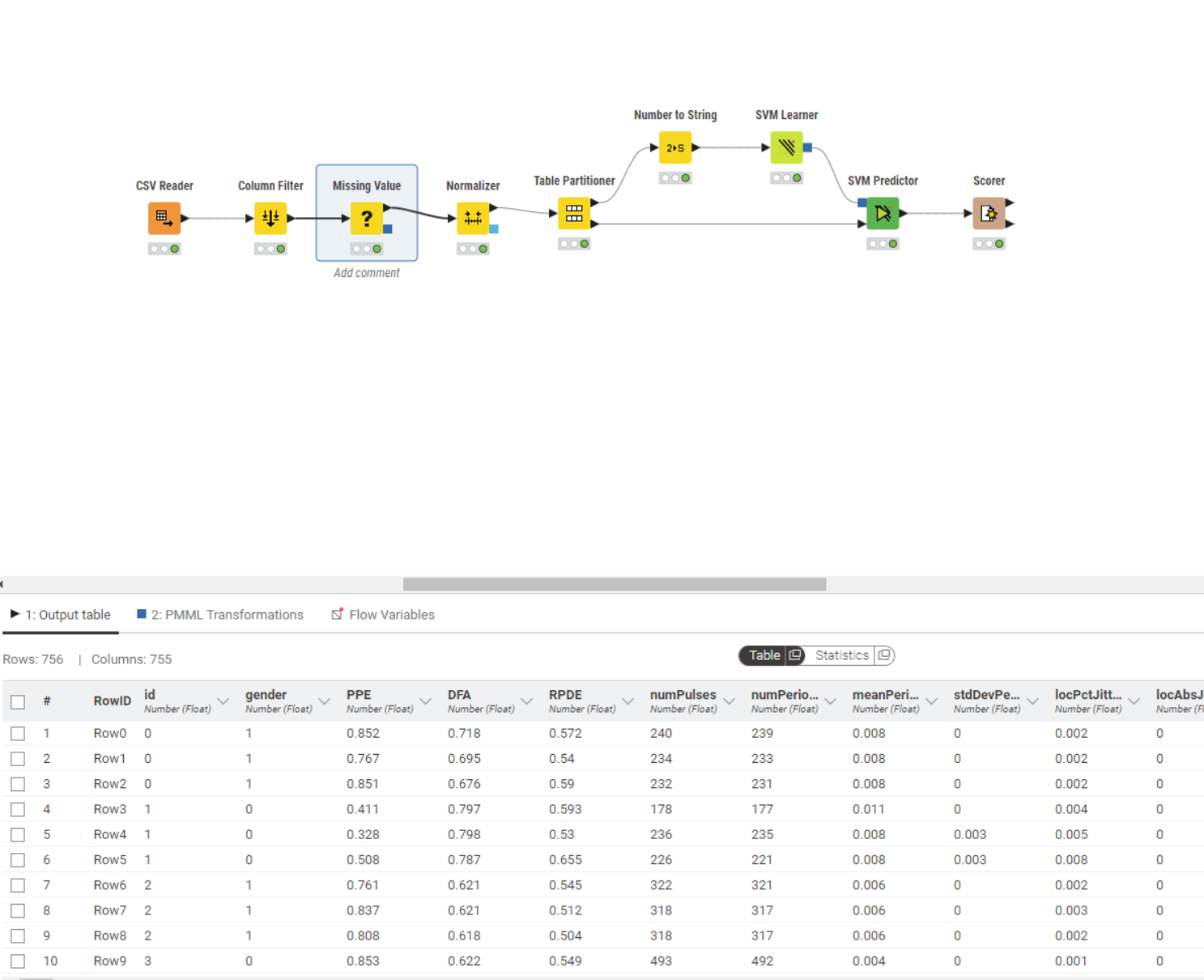


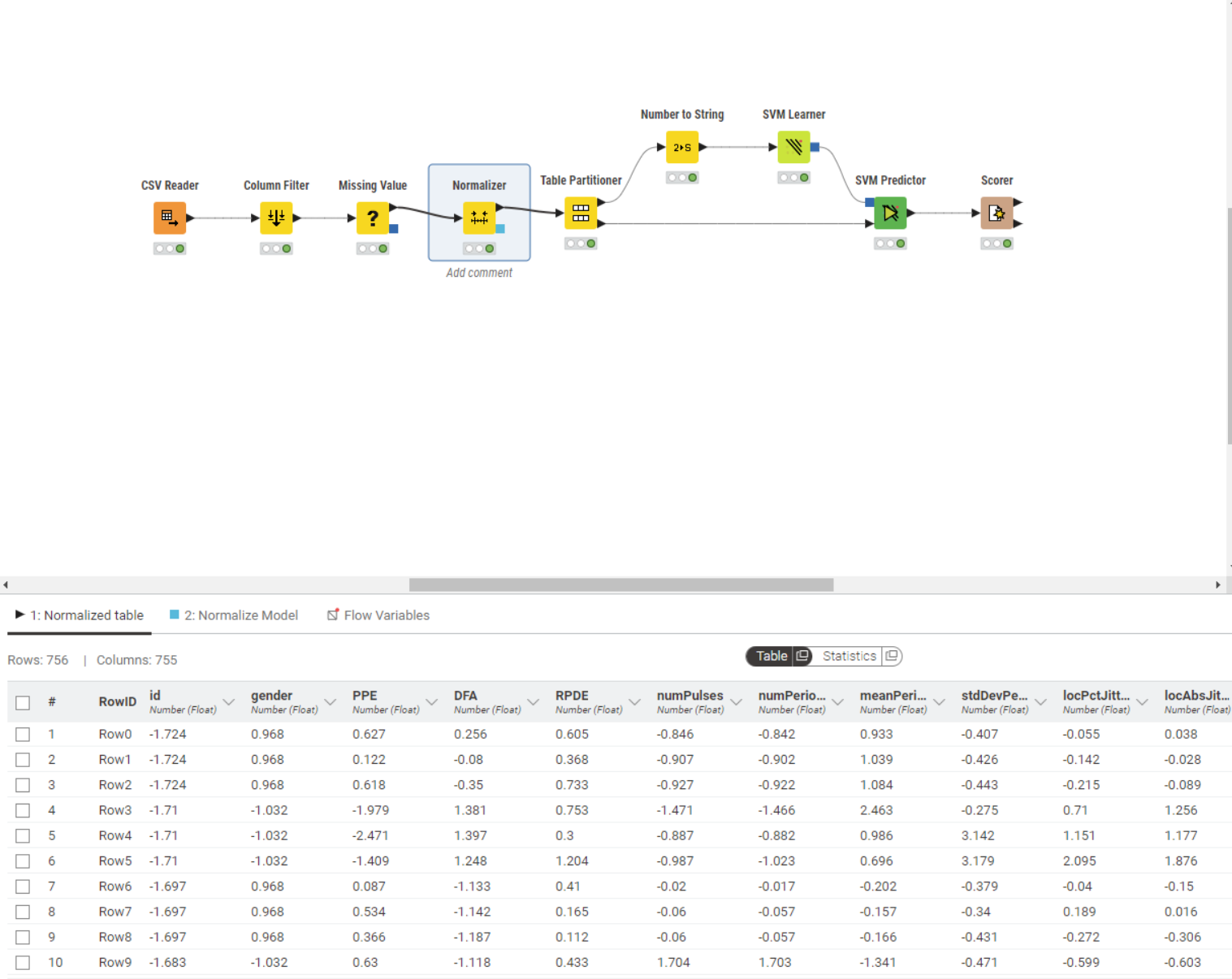


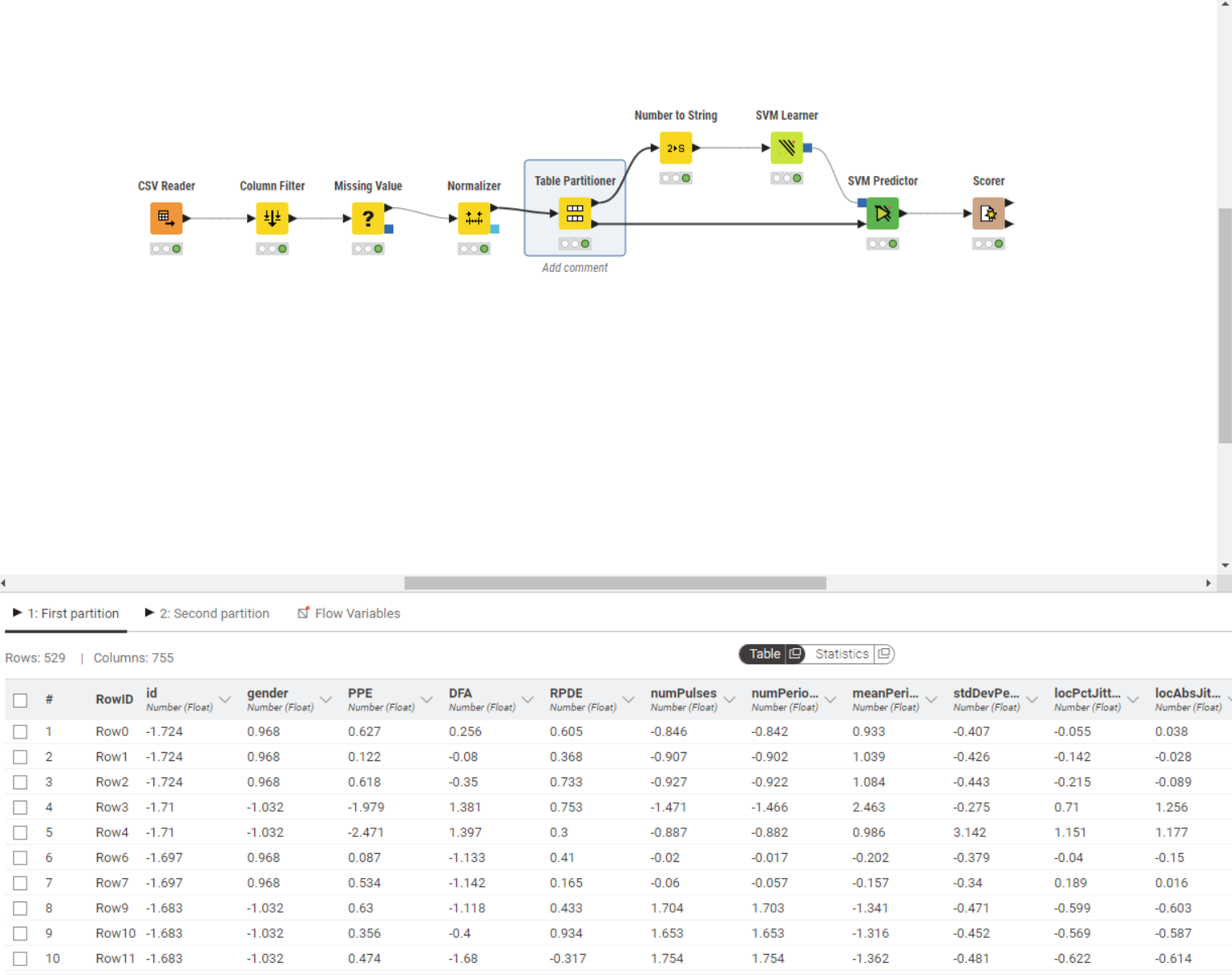


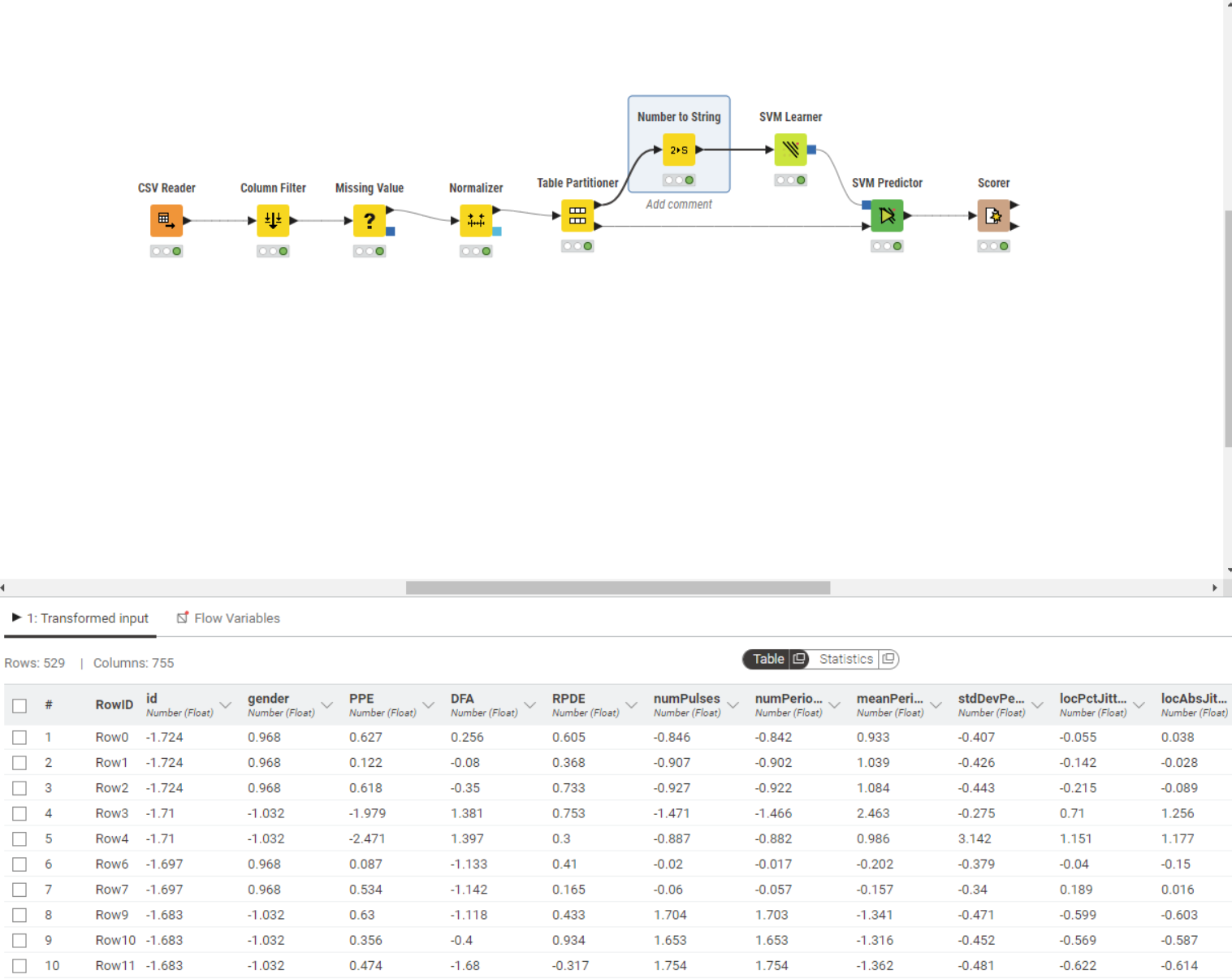


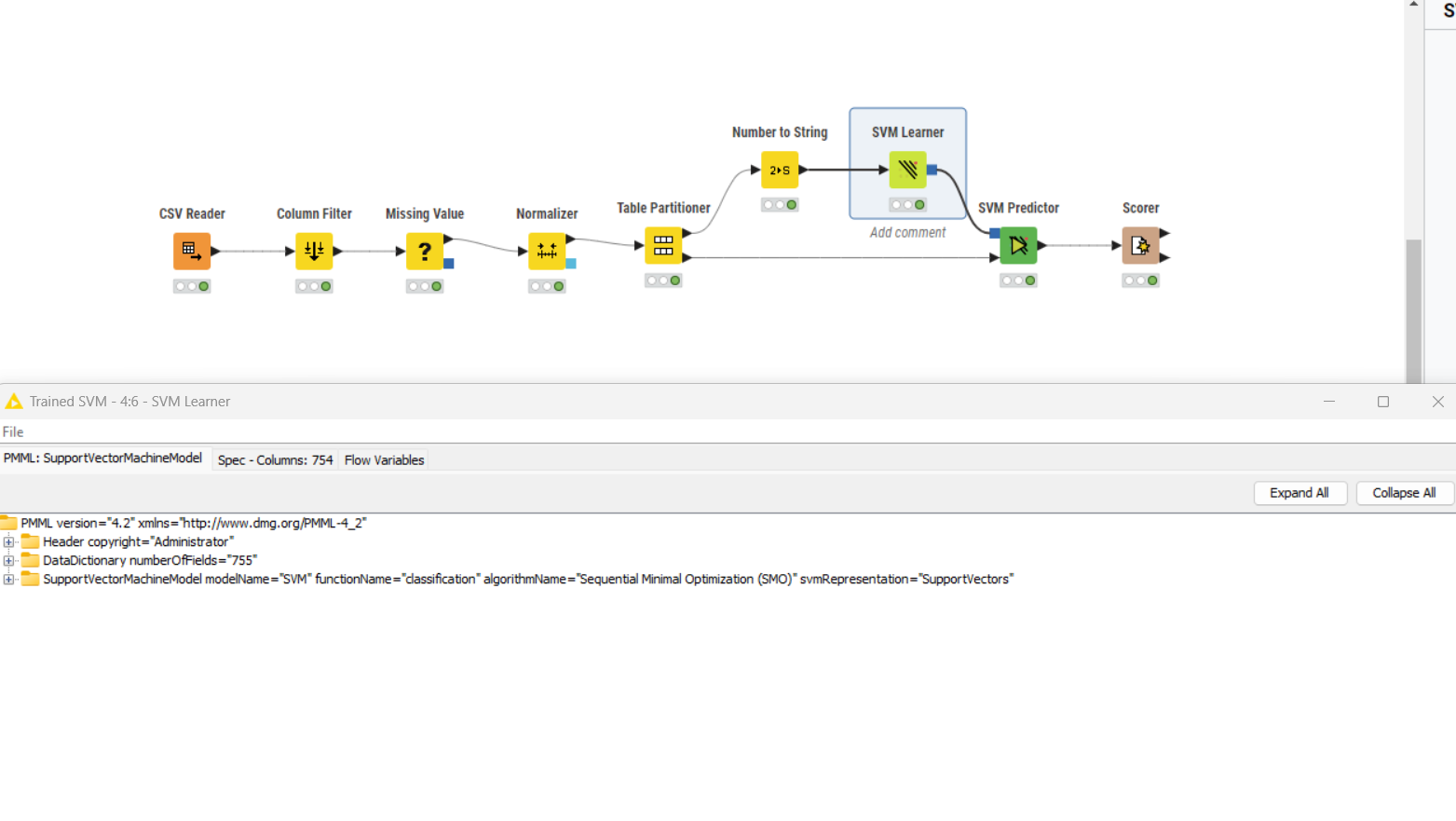


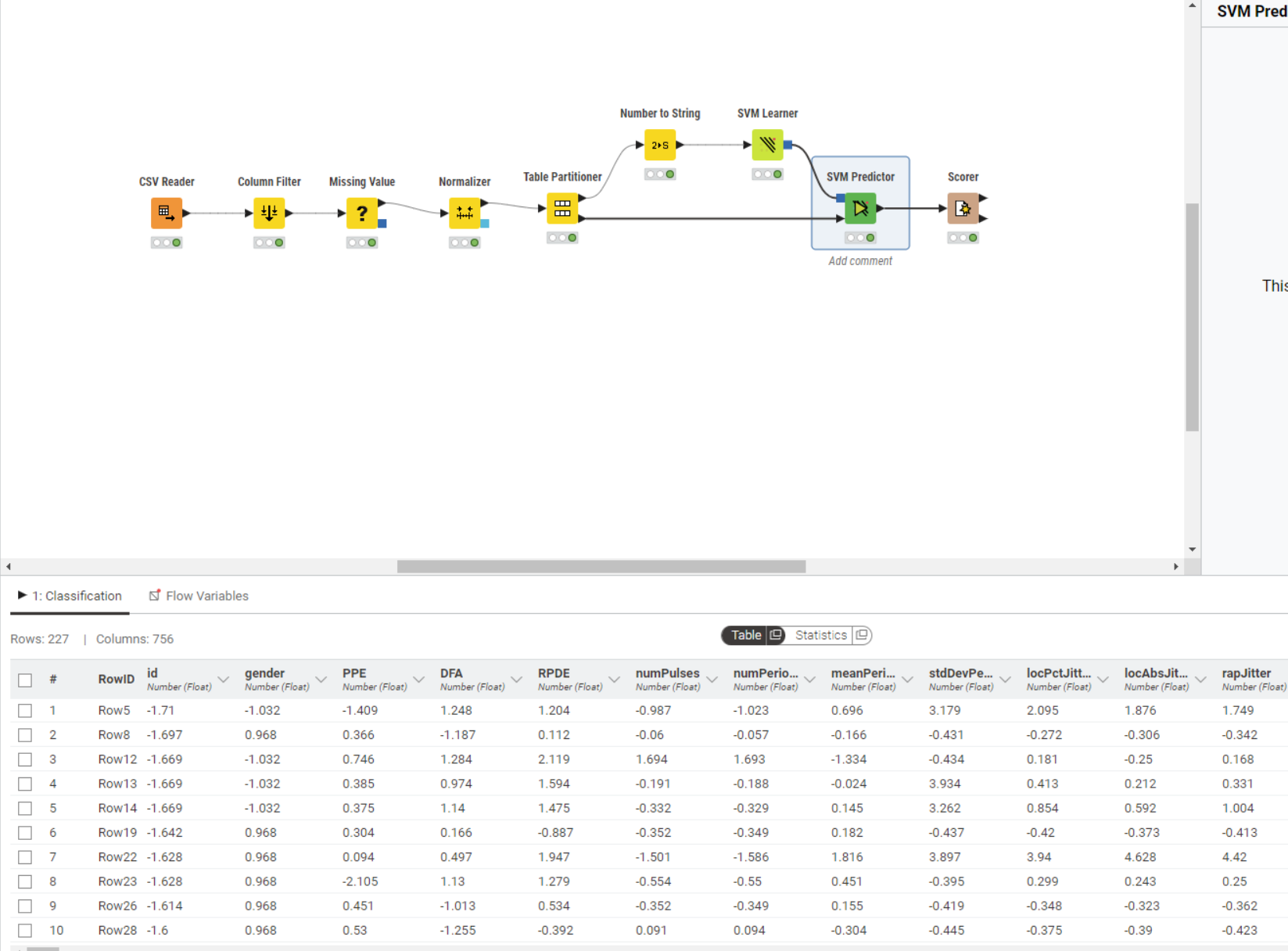


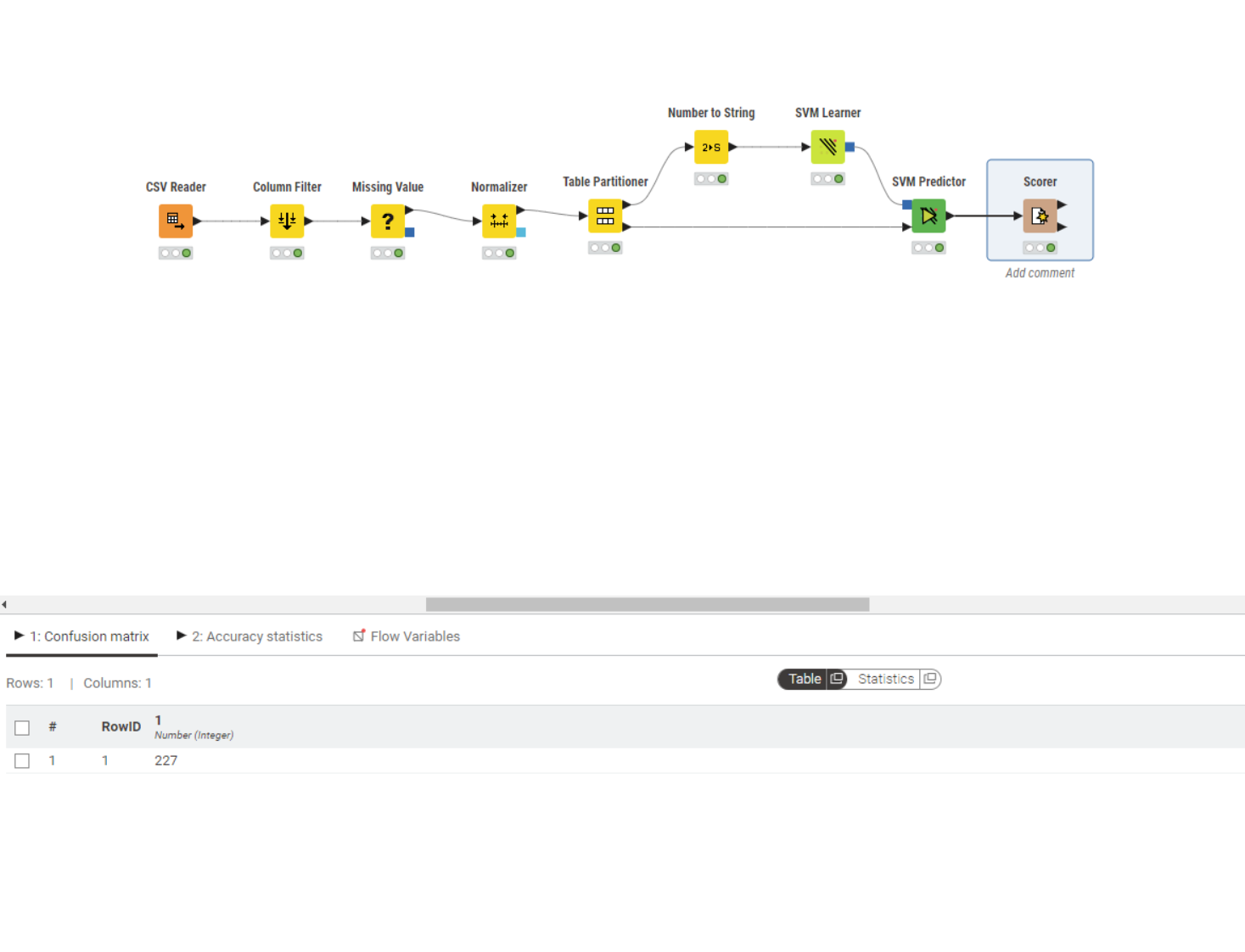












3. Wnioski

Przeprowadzona analiza wykazała, że zastosowanie klasyfikatora **SVM** pozwala skutecznie rozróżniać przypadki choroby Parkinsona na podstawie cech akustycznych mowy. Model uzyskał wysoką wartość metryk takich jak **Accuracy** i **F1-score**, co wskazuje na jego dobrą ogólną skuteczność w detekcji choroby. Zastosowanie walidacji krzyżowej potwierdziło stabilność wyników i brak nadmiernego dopasowania.

W przypadku modelu **regresji liniowej**, uzyskane wartości błędów (MAE, RMSE) oraz współczynnik determinacji **R²** wskazują, że relacja między wybraną zmienną ciągłą a pozostałymi cechami ma charakter umiarkowanie liniowy. Wyniki są poprawne, jednak model liniowy nie zawsze jest w stanie uchwycić nieliniowe zależności w danych akustycznych.

Porównanie obu modeli potwierdziło, że klasyfikacja choroby Parkinsona jest lepiej dopasowanym zadaniem dla dostępnych danych niż regresja, a metoda **SVM z odpowiednim doborem jądra (np. RBF)** daje najlepsze rezultaty. Dalsze ulepszenia mogłyby obejmować selekcję najistotniejszych cech, zastosowanie metod nieliniowych dla regresji (np. SVR lub Random Forest Regressor) oraz techniki balansowania klas w przypadku nierównowagi danych.