SPRAWOZDANIE

Zajęcia: Uczenie Maszynowe

Prowadzący: prof. dr hab. Vasyl Martsenyuk

Laboratorium Nr 2 Data 09.11.2024 Temat: "Praktyczne Zastosowanie Drzew Decyzyjnych i Metod Ensemble w Analizie Danych" Wariant 10	Anna Więzik Informatyka II stopień, niestacjonarne, 1 semestr, gr.1b

1. Polecenie:

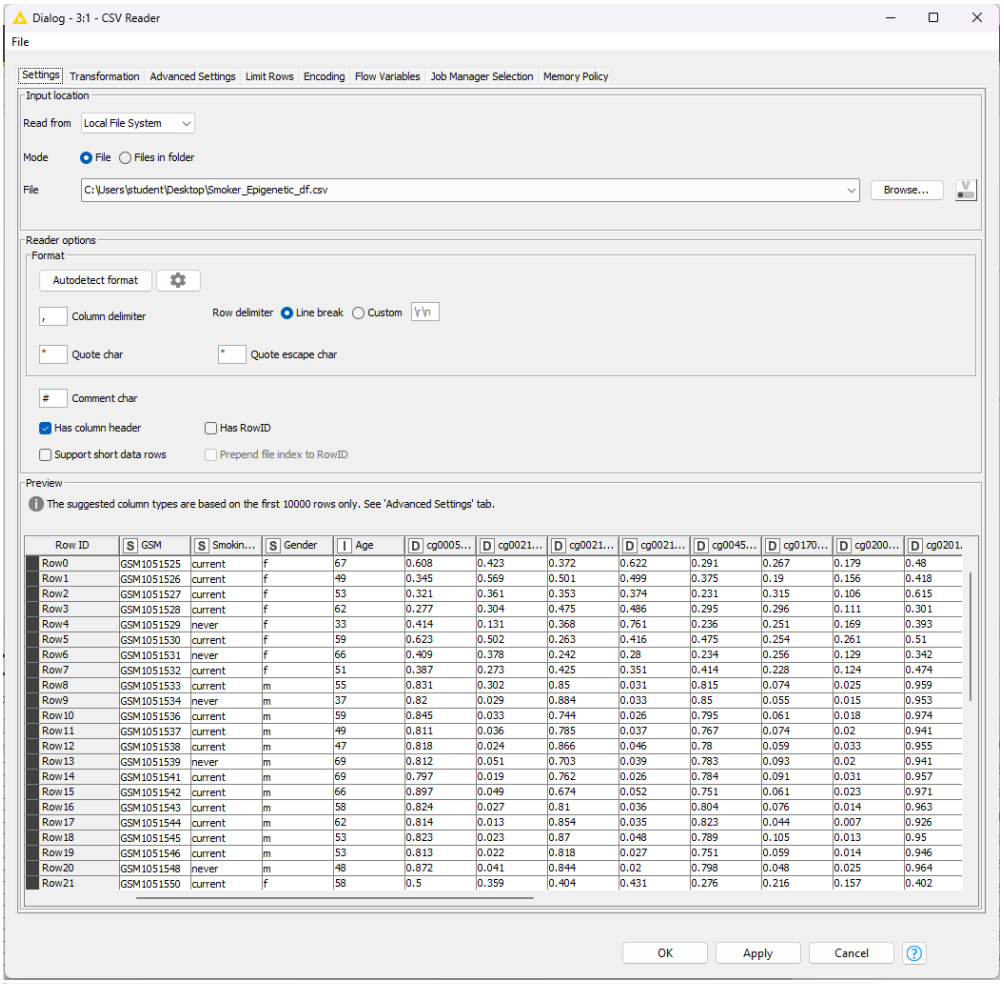
   Smoking patient: https://www.kaggle.com/datasets/thomaskonstantin/cpg-values-of-smoking-and-non-smoking-patients

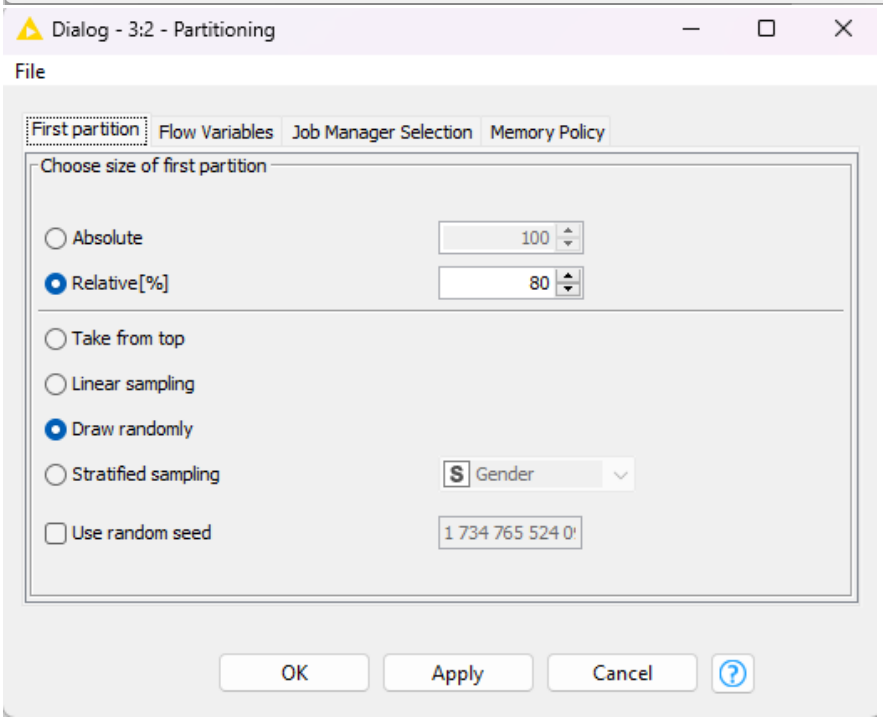
2. Link do repozytorium:

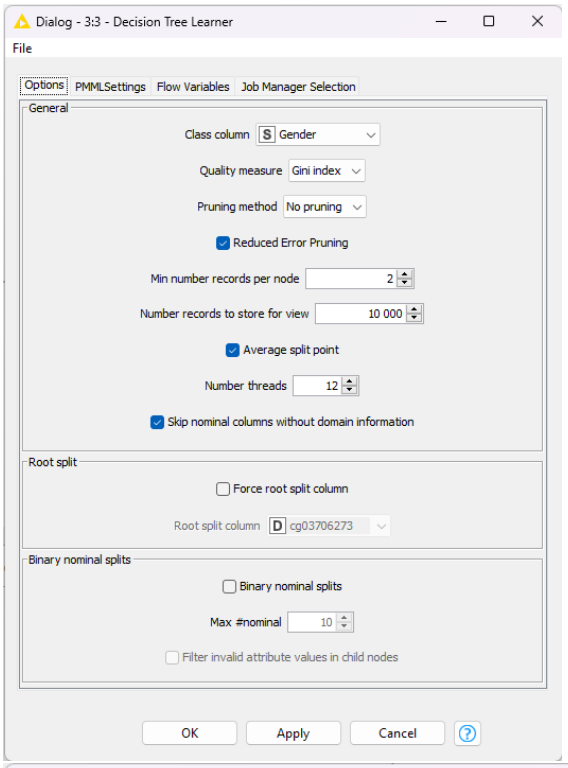
   Link: https://github.com/AnaShiro/UM_2024

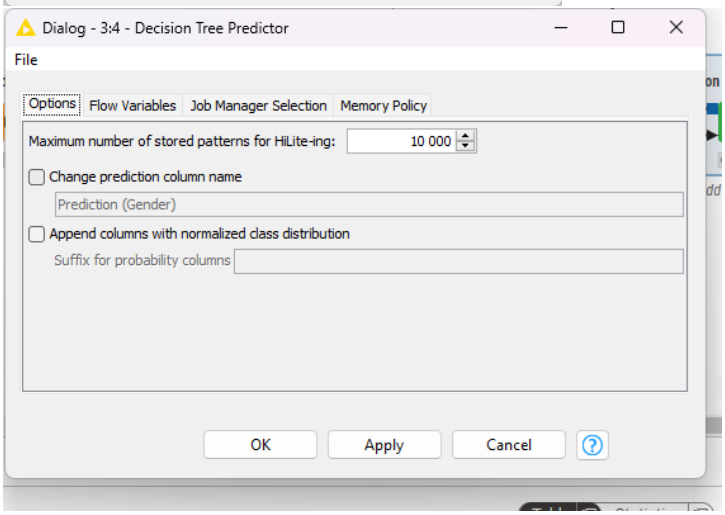
3. Opis programu opracowanego

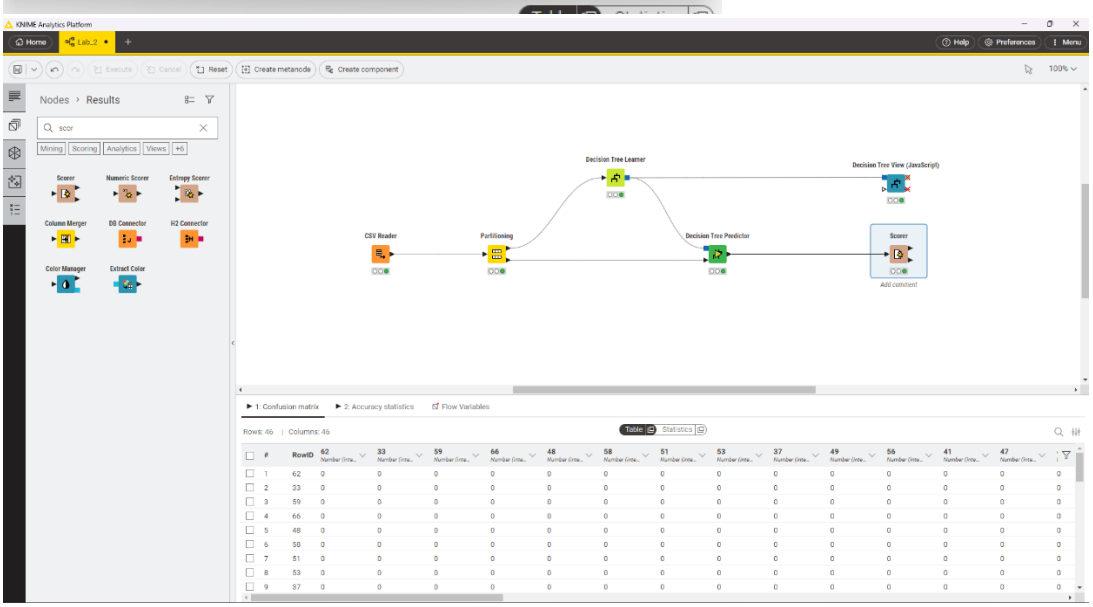
- Decyzjonalne drzewo przepływów

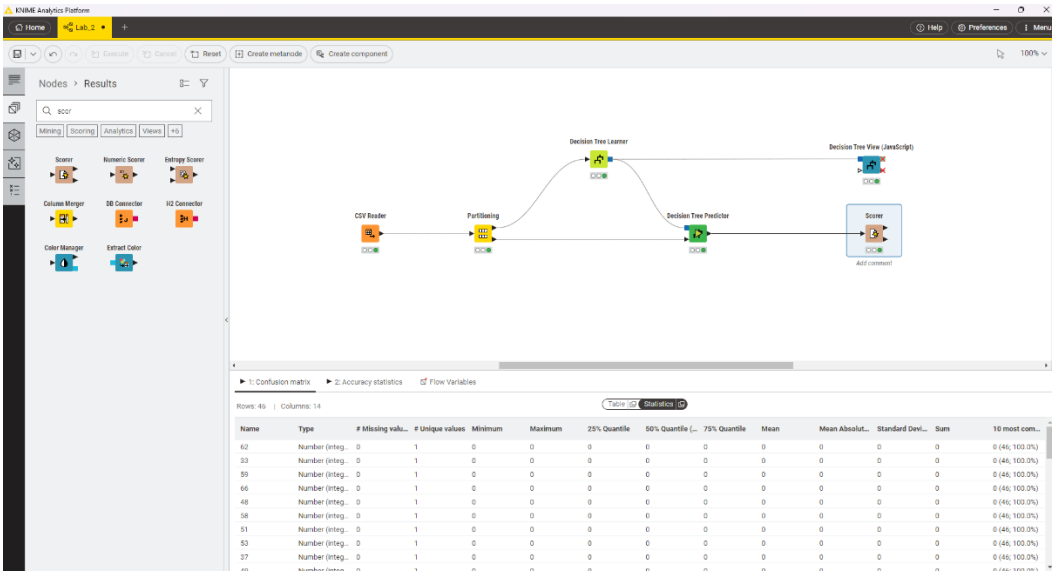


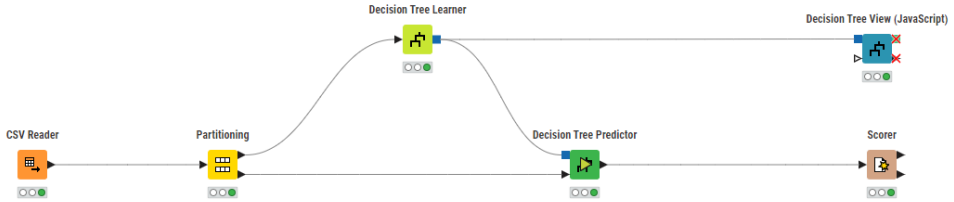



Dialog - 3:3 - Decision Tree Learner
File
Options
PMMLSettings
Flow Variables
Job Manager Selection
General
Class column
Gender
Quality measure
Gini index
Pruning method
No pruning
Reduced Error Pruning
Min number records per node
2
Number records to store for view
10 000
Average split point
Number threads
12
Skip nominal columns without domain information
Root split
Force root split column
Root split column
cg03706273
Binary nominal splits
Binary nominal splits
Max #nominal
10
Filter invalid attribute values in child nodes
OK
Apply
Cancel

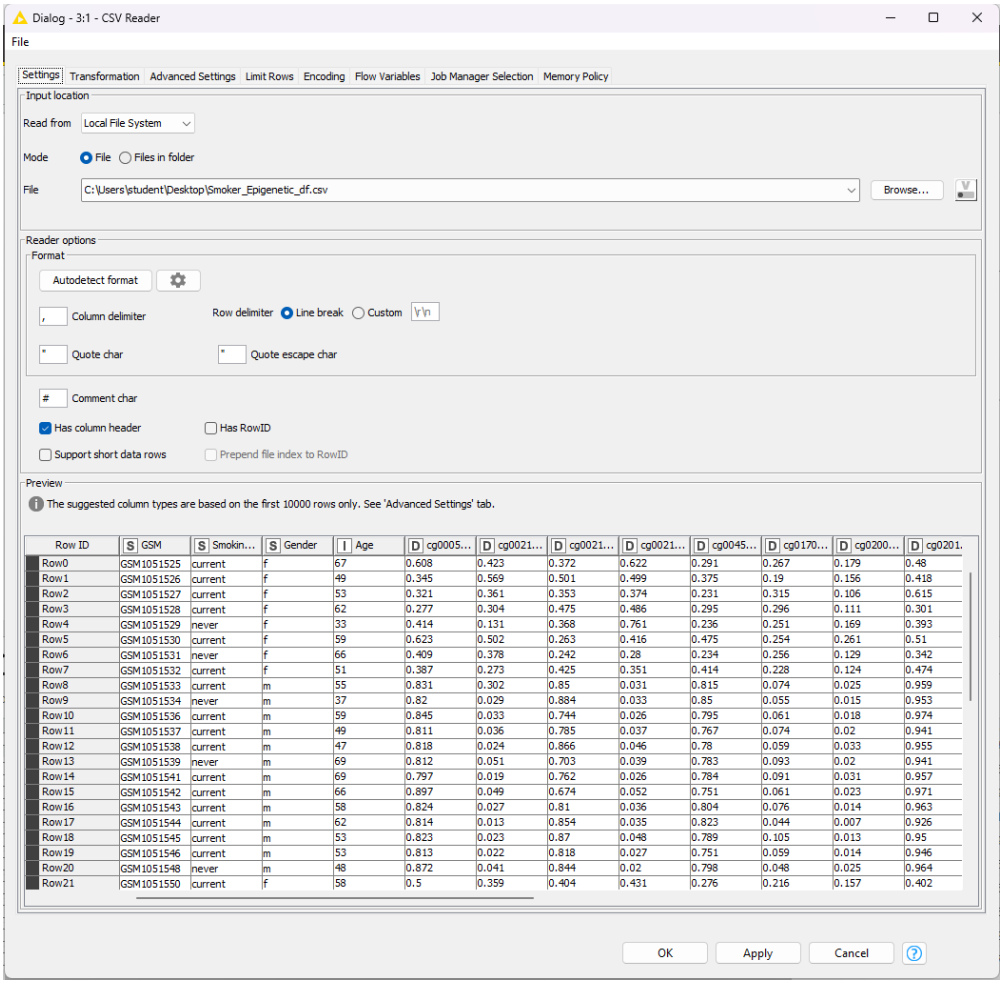

Dialog - 3:4 - Decision Tree Predictor
File
Options
Flow Variables
Job Manager Selection
Memory Policy
Maximum number of stored patterns for HiLite-ing:
10 000
Change prediction column name
Prediction (Gender)
Append columns with normalized class distribution
Suffix for probability columns
OK
Apply
Cancel

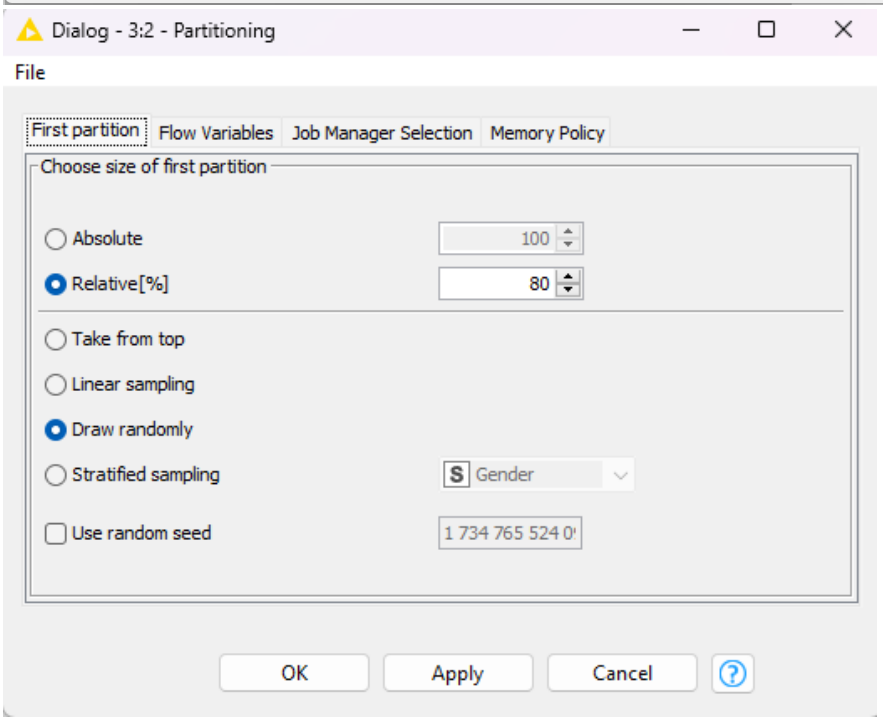

KNIME Analytics Platform
Nodes
Results
CSV Reader
Partitioning
Decision Tree Learner
Decision Tree Predictor
Decision Tree View (JavaScript)
Scorer

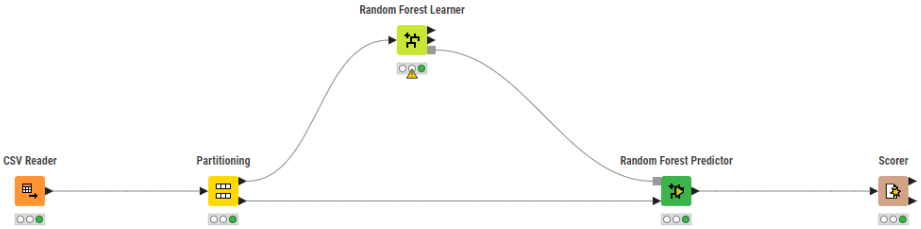

KNIME Analytics Platform
Home
Lab 2
Help
Preferences
Reset
Create metanode
Create component
Nodes › Results
scor
Mining
Scoring
Analytics
Views
Scorer
Numeric Scorer
Entropy Scorer
Column Merger
DB Connector
H2 Connector
Color Manager
Extract Color
Decision Tree Learner
Decision Tree View (JavaScript)
CSV Reader
Partitioning
Decision Tree Predictor
Scorer
Add comment
1: Confusion matrix
2: Accuracy statistics
Flow Variables
Rows: 46 | Columns: 14
Table
Statistics
Name
Type
Missing valu...
Unique values
Minimum
Maximum
25% Quantile
50% Quantile (...
75% Quantile
Mean
Mean Absolut...
Standard Devi...
Sum
10 most com...

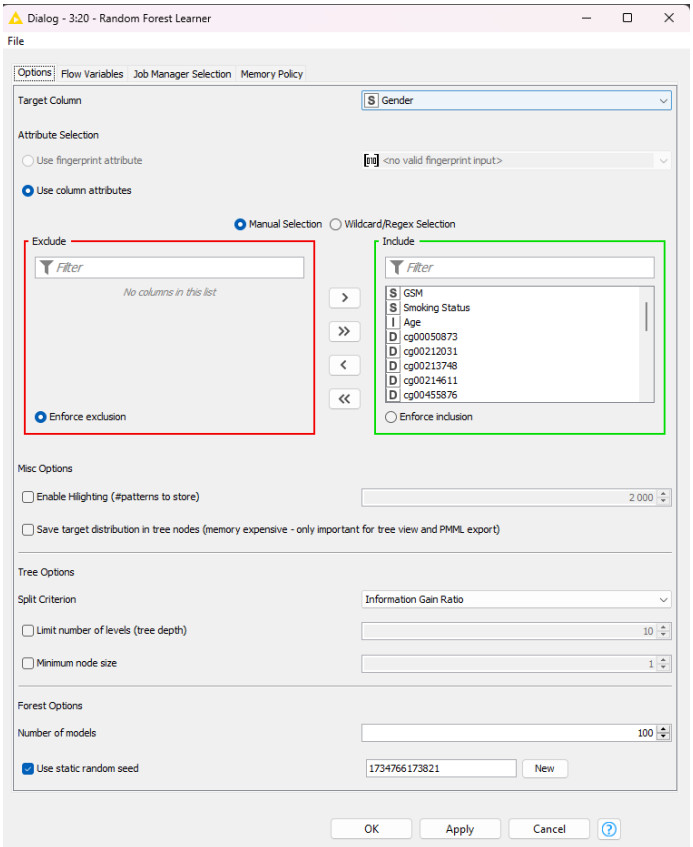

Decision Tree Learner
Decision Tree View (JavaScript)
CSV Reader
Partitioning
Decision Tree Predictor
Scorer

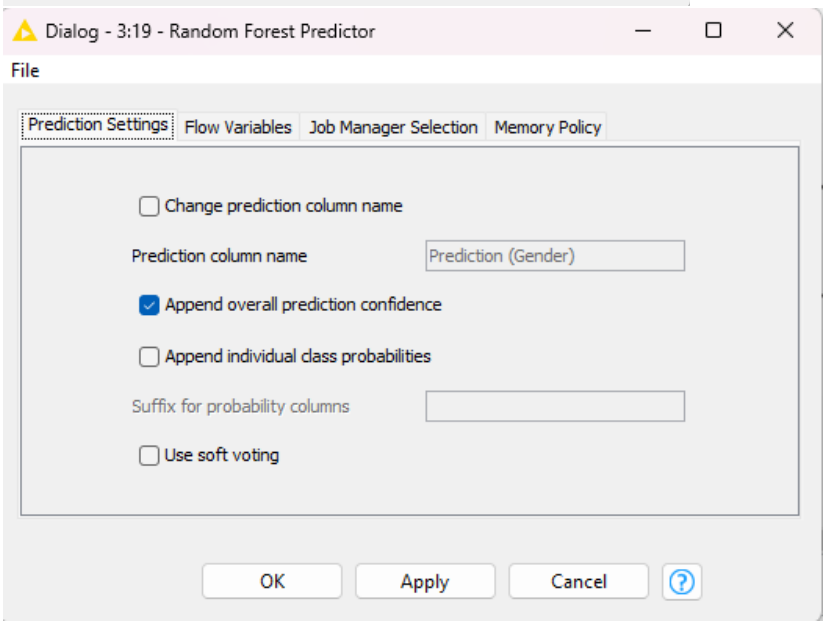
- Las losowy

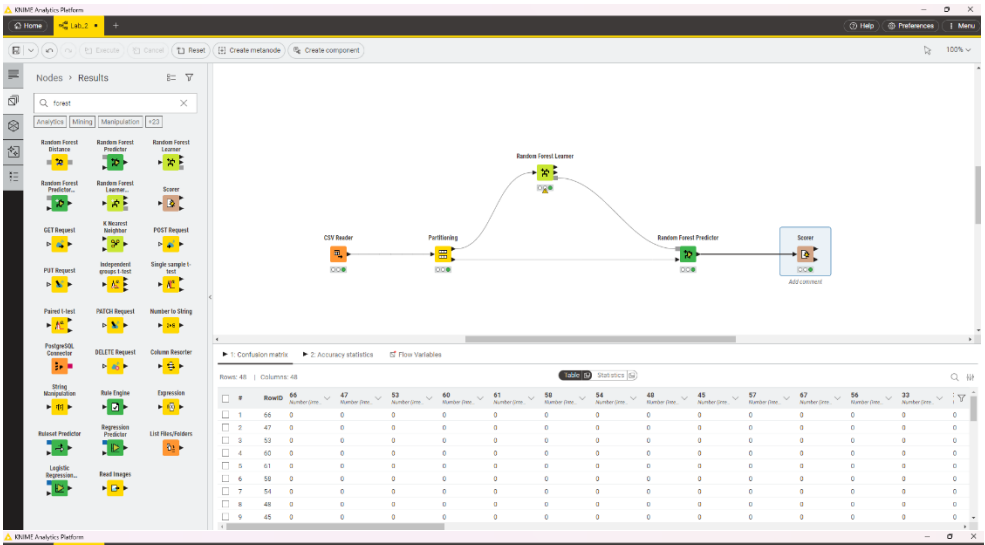

Dialog - 3:1 - CSV Reader
File
Settings
Transformation
Advanced Settings
Limit Rows
Encoding
Flow Variables
Job Manager Selection
Memory Policy
Input location
Read from
Local File System
Mode
File
Files in folder
File
C:\Users\student\Desktop\Smoker_Epigenetic_df.csv
Browse...
Reader options
Format
Autodetect format
Column delimiter
Row delimiter
Line break
Custom
Quote char
Quote escape char
Comment char
Has column header
Has RowID
Support short data rows
Prepend file index to RowID
Preview
The suggested column types are based on the first 10000 rows only. See 'Advanced Settings' tab.
Row ID
GSM
Smokin...
Gender
Age
cg0005...
cg0021...
cg0021...
cg0021...
cg0045...
cg0170...
cg0200...
cg0201.
Row0 GSM1051525 current f 67 0.608 0.423 0.372 0.622 0.291 0.267 0.179 0.48
Row1 GSM1051526 current f 49 0.345 0.569 0.501 0.499 0.375 0.19 0.156 0.418
Row2 GSM1051527 current f 53 0.321 0.361 0.353 0.374 0.231 0.315 0.106 0.615
Row3 GSM1051528 current f 62 0.277 0.304 0.475 0.486 0.295 0.296 0.111 0.301
Row4 GSM1051529 never f 33 0.414 0.131 0.368 0.761 0.236 0.251 0.169 0.393
Row5 GSM1051530 current f 59 0.623 0.502 0.263 0.416 0.475 0.254 0.261 0.51
Row6 GSM1051531 never f 66 0.409 0.378 0.242 0.28 0.234 0.256 0.129 0.342
Row7 GSM1051532 current f 51 0.387 0.273 0.425 0.351 0.414 0.228 0.124 0.474
Row8 GSM1051533 current m 55 0.831 0.302 0.85 0.031 0.815 0.074 0.025 0.959
Row9 GSM1051534 never m 37 0.82 0.029 0.884 0.033 0.85 0.055 0.015 0.953
Row10 GSM1051536 current m 59 0.845 0.033 0.744 0.026 0.795 0.061 0.018 0.974
Row11 GSM1051537 current m 49 0.811 0.036 0.785 0.037 0.767 0.074 0.02 0.941
Row12 GSM1051538 current m 47 0.818 0.024 0.866 0.046 0.78 0.059 0.033 0.955
Row13 GSM1051539 never m 69 0.812 0.051 0.703 0.039 0.783 0.093 0.02 0.941
Row14 GSM1051541 current m 69 0.797 0.019 0.762 0.026 0.784 0.091 0.031 0.957
Row15 GSM1051542 current m 66 0.897 0.049 0.674 0.052 0.751 0.061 0.023 0.971
Row16 GSM1051543 current m 58 0.824 0.027 0.81 0.036 0.804 0.076 0.014 0.963
Row17 GSM1051544 current m 62 0.814 0.013 0.854 0.035 0.823 0.044 0.007 0.926
Row18 GSM1051545 current m 53 0.823 0.023 0.87 0.048 0.789 0.105 0.013 0.95
Row19 GSM1051546 current m 53 0.813 0.022 0.818 0.027 0.751 0.059 0.014 0.946
Row20 GSM1051548 never m 48 0.872 0.041 0.844 0.02 0.798 0.048 0.025 0.964
Row21 GSM1051550 current f 58 0.5 0.359 0.404 0.431 0.276 0.216 0.157 0.402
OK
Apply
Cancel

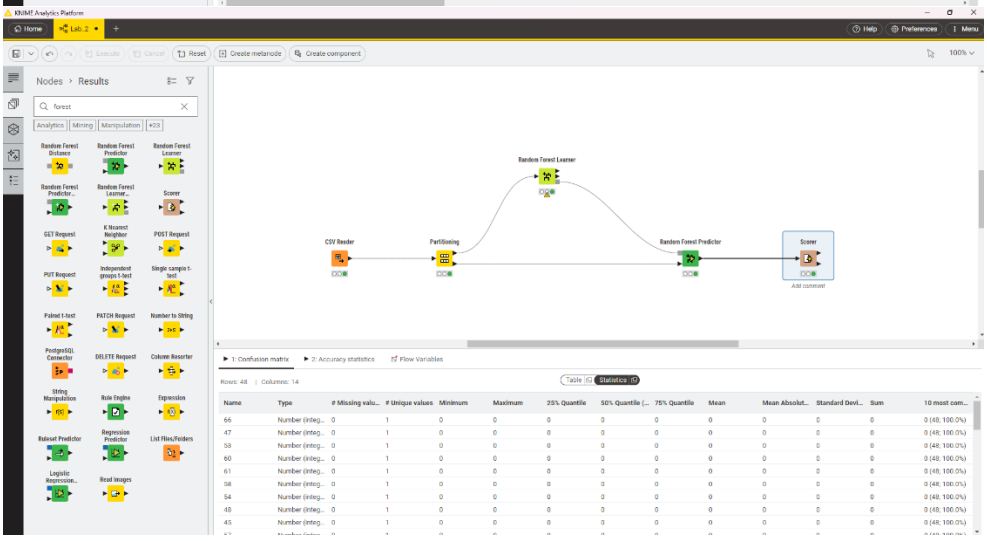

Dialog - 3:2 - Partitioning
File
First partition
Flow Variables
Job Manager Selection
Memory Policy
Choose size of first partition
Absolute
100
Relative[%]
80
Take from top
Linear sampling
Draw randomly
Stratified sampling
Gender
Use random seed
1 734 765 524 0
OK
Apply
Cancel


Random Forest Learner
CSV Reader
Partitioning
Random Forest Predictor
Scorer

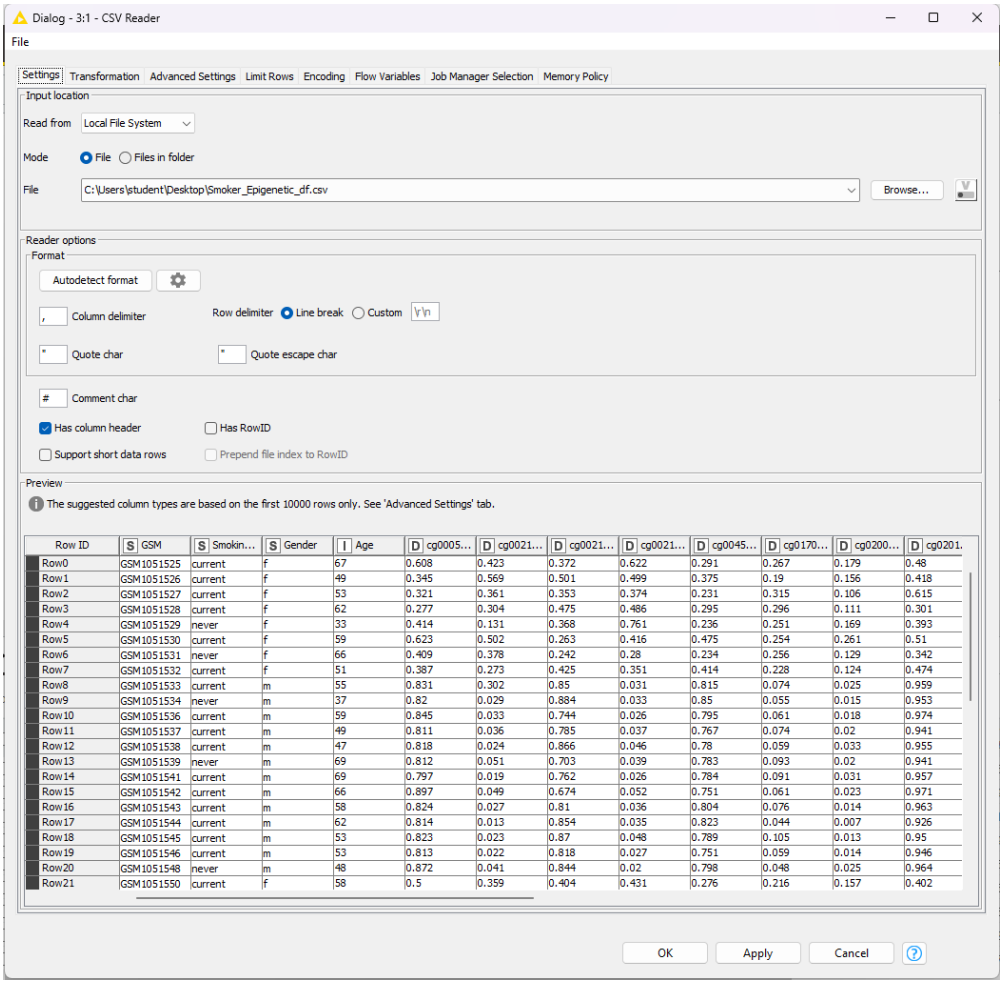

Dialog - 3:20 - Random Forest Learner
File
Options
Flow Variables
Job Manager Selection
Memory Policy
Target Column
Gender
Attribute Selection
Use fingerprint attribute
<no valid fingerprint input>
Use column attributes
Manual Selection
Wildcard/Regex Selection
Exclude
Filter
No columns in this list
Enforce exclusion
Include
Filter
GSM
Smoking Status
Age
cg00050873
cg00212031
cg00213748
cg00214611
cg00455876
Enforce inclusion
Misc Options
Enable Hilighting (#patterns to store)
2 000
Save target distribution in tree nodes (memory expensive - only important for tree view and PMML export)
Tree Options
Split Criterion
Information Gain Ratio
Limit number of levels (tree depth)
10
Minimum node size
1
Forest Options
Number of models
100
Use static random seed
1734766173821
New
OK
Apply
Cancel

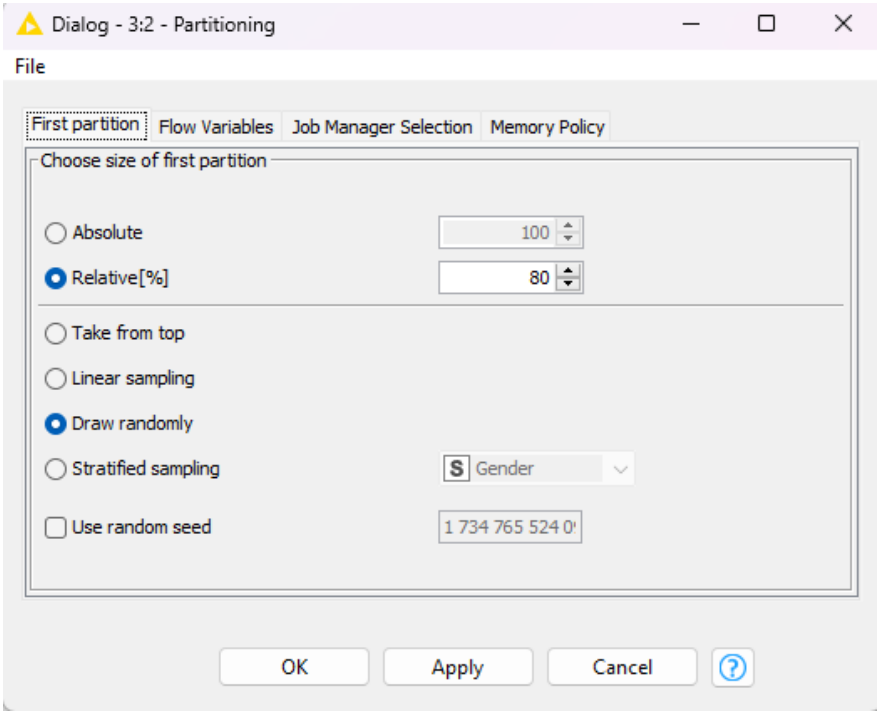

Dialog - 3:19 - Random Forest Predictor
File
Prediction Settings
Flow Variables
Job Manager Selection
Memory Policy
Change prediction column name
Prediction column name
Prediction (Gender)
Append overall prediction confidence
Append individual class probabilities
Suffix for probability columns
Use soft voting
OK
Apply
Cancel

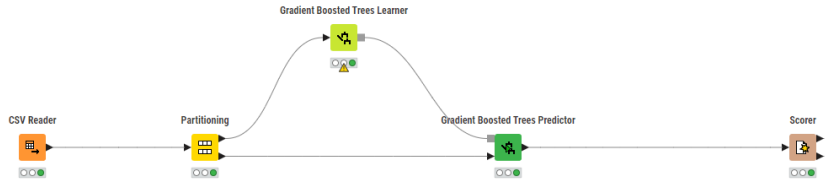

KNIME Analytics Platform
Home
Lab 2
Help
Preferences
Menu
Reset
Create metanode
Create component
Nodes › Results
forest
Analytics
Mining
Manipulation
Random Forest Learner
CSV Reader
Partitioning
Random Forest Predictor
Scorer
1: Confusion matrix
2: Accuracy statistics
Flow Variables

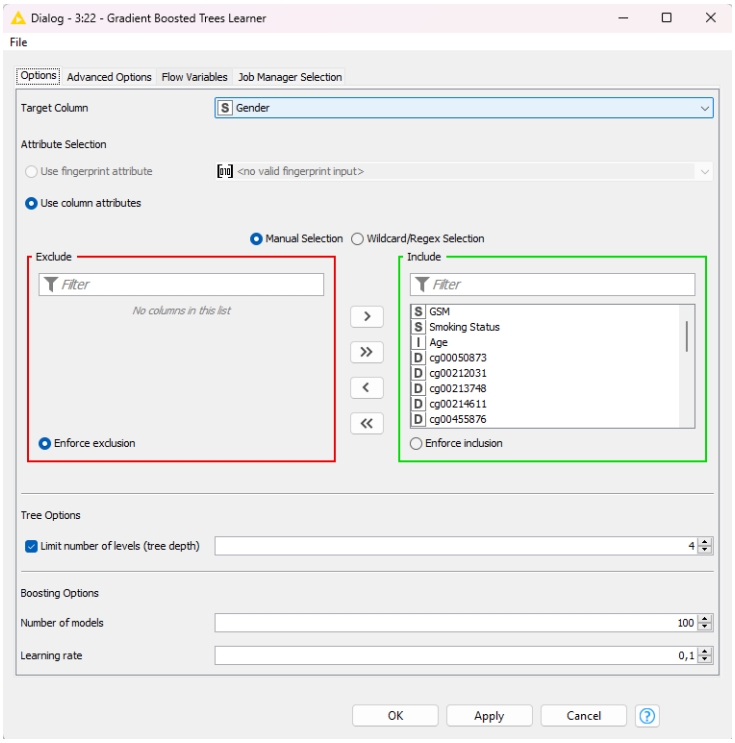

KNIME Analytics Platform
Home
Lab 2
Help
Preferences
Menu
Reset
Create metanode
Create component
Nodes › Results
forest
Analytics
Mining
Manipulation
Random Forest Learner
CSV Reader
Partitioning
Random Forest Predictor
Scorer
1: Confusion matrix
2: Accuracy statistics
Flow Variables
Table
Statistics

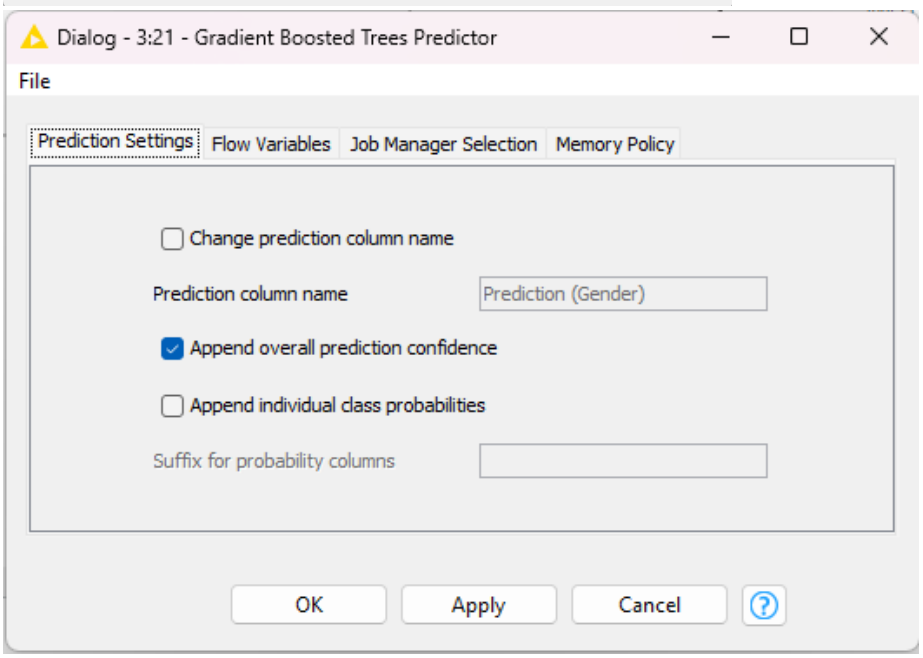
- Boosted trees

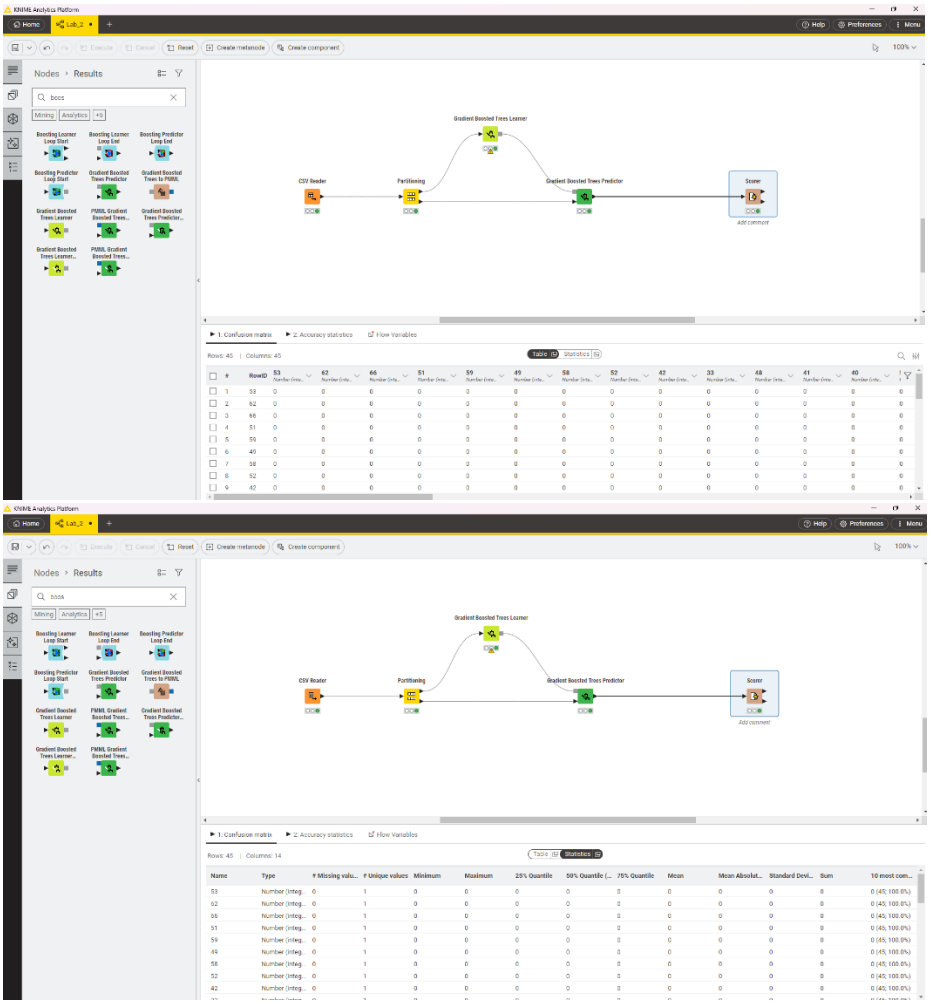





Gradient Boosted Trees Learner
CSV Reader
Partitioning
Gradient Boosted Trees Predictor
Scorer


Dialog - 3:22 - Gradient Boosted Trees Learner
File
Options
Advanced Options
Flow Variables
Job Manager Selection
Target Column
Gender
Attribute Selection
Use fingerprint attribute
<no valid fingerprint input>
Use column attributes
Manual Selection
Wildcard/Regex Selection
Exclude
Filter
No columns in this list
Enforce exclusion
Include
Filter
GSM
Smoking Status
Age
cg00050873
cg00212031
cg00213748
cg00214611
cg00455876
Enforce inclusion
Tree Options
Limit number of levels (tree depth)
4
Boosting Options
Number of models
100
Learning rate
0,1
OK
Apply
Cancel


Dialog - 3:21 - Gradient Boosted Trees Predictor
File
Prediction Settings
Flow Variables
Job Manager Selection
Memory Policy
Change prediction column name
Prediction column name
Prediction (Gender)
Append overall prediction confidence
Append individual class probabilities
Suffix for probability columns
OK
Apply
Cancel



4. Wnioski

KNIME umożliwia intuicyjne i efektywne tworzenie modeli klasyfikacyjnych za pomocą graficznych przepływów pracy. Modele, takie jak drzewa decyzyjne, Random Forest i boosting, mogą być łatwo wdrażane za pomocą odpowiednich węzłów KNIME, co pozwala na analizę danych bez potrzeby pisania kodu. Drzewa decyzyjne są prostymi, ale skutecznymi modelami uczenia maszynowego, szczególnie w kontekście analizy danych. Metody z grupy ensemble, takie jak bagging, Random Forest oraz boosting, poprawiają dokładność i stabilność modeli poprzez łączenie wielu słabszych klasyfikatorów. Random Forest dodatkowo wprowadza element losowości przy wyborze cech.