An intelligent tool for expediting and automating data mining steps

Ourania Hatzi, **Nikolaos Zorbas**, Mara Nikolaidou and Dimosthenis Anagnostopoulos

Outline

- Data Mining,
  - current tools
- An intelligent tool for Managing Data Mining
  - Scope
  - Feasibility
- ADaMM
  - Architecture
  - Functionality
- Experiment Results
- Future Work

Data Mining

Tools and Uses

Data Mining

- The process of extracting useful knowledge, i.e. patterns, from data
  - A complicated process drawing from many fields
  - Artificial Intelligence, Machine Learning, Pattern Recognition, Statistics….
  - Used virtually everywhere from Business Analytics to Medicine

Tools for Data Mining

- A plethora of open source and commercial products
  - …ranging from complete graphical suites to
  - …specifically designed programming languages / frameworks
- KNIME, Orange, R, RapidMiner, SPSS , **Weka**…
  - Nearly all tools are aimed towards experts
  - or students that want to become experts
- **Powerful but complicated**

An intelligent tool for managing Data Mining

Scope

- Problem
  - Non-experts have seemingly no access to data mining
  - Experts have to spend time on simple repetitive tasks
- The solution
  - An intelligent tool for managing Data Mining

Feasibility

- Many steps of the process are standard
- …and are repeated a lot
- Generic workflows can be established and generalized
- **BUT**
  - Desired data sets and uses can be vastly different
  - It is nearly impossible to fully imitate an expert's insight on a problem

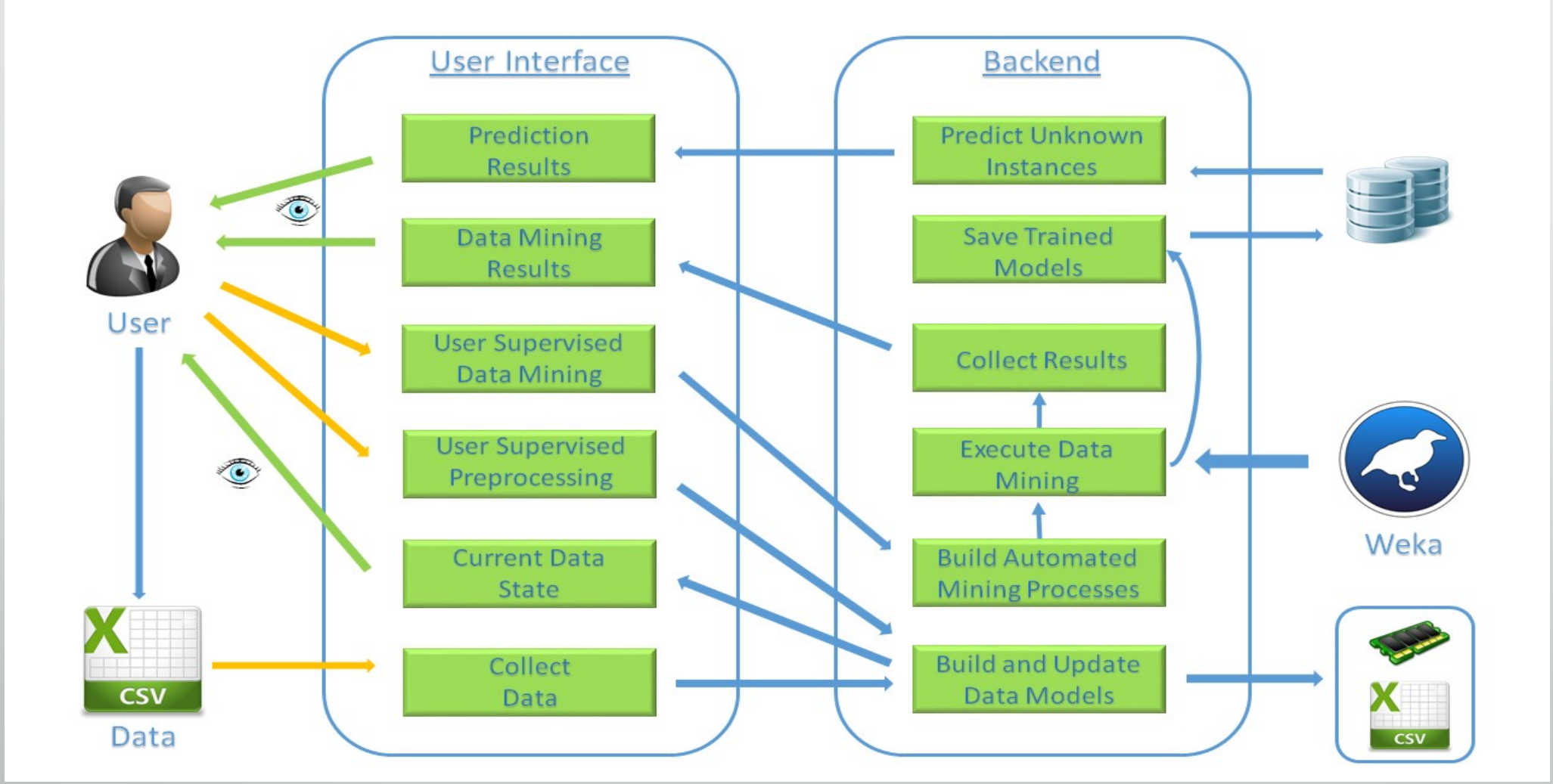
The Vision

- A step by step guided data mining process
  - Extendable and modular
  - Automated as much as possible
  - Generating optimized, comprehensive results
  - Requiring minimal input from the user

ADaMM

Automated Data Mining Manager

Architecture

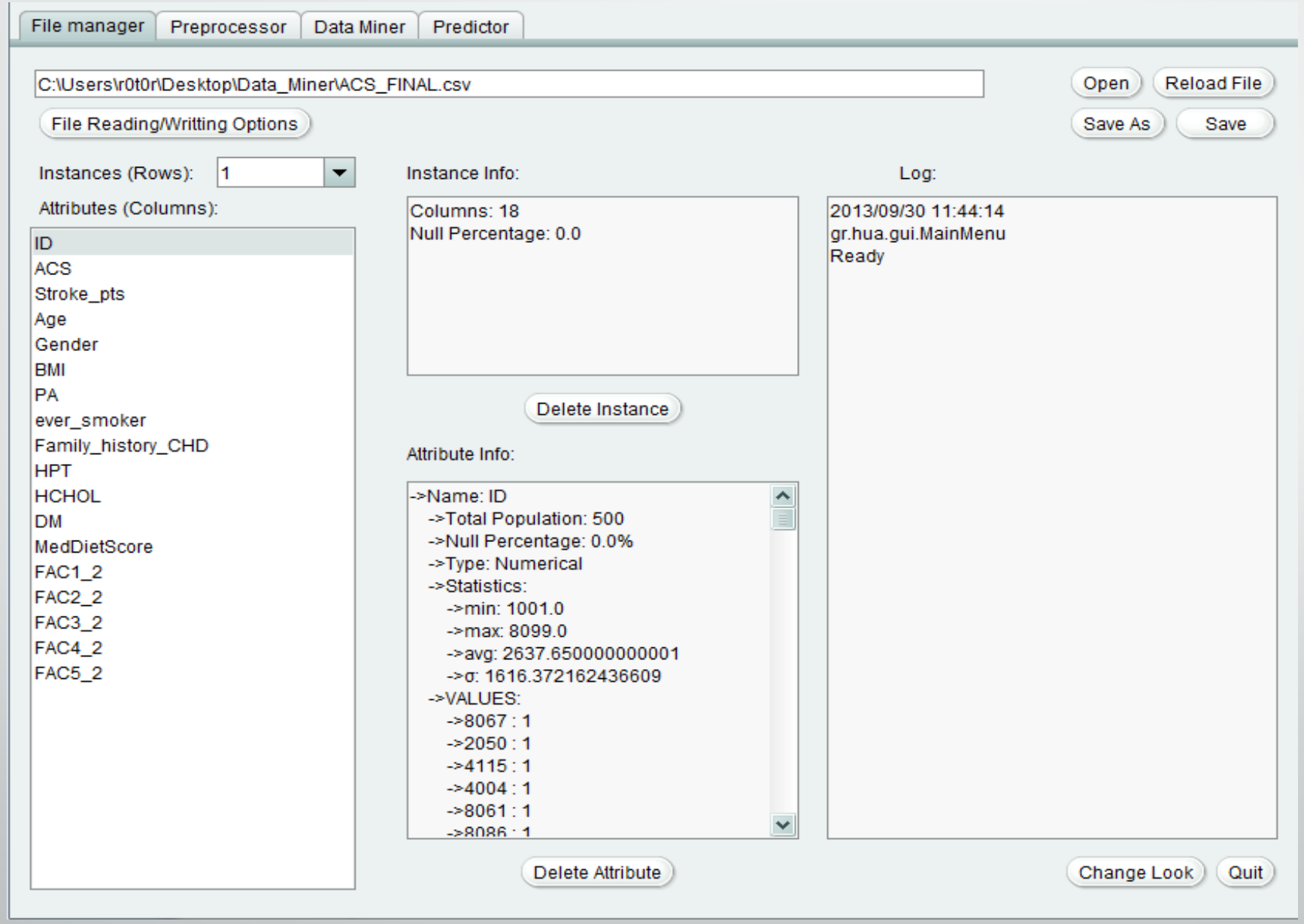


Functionality

- Data Handling
- Preprocessing
- Data Mining
- Prediction

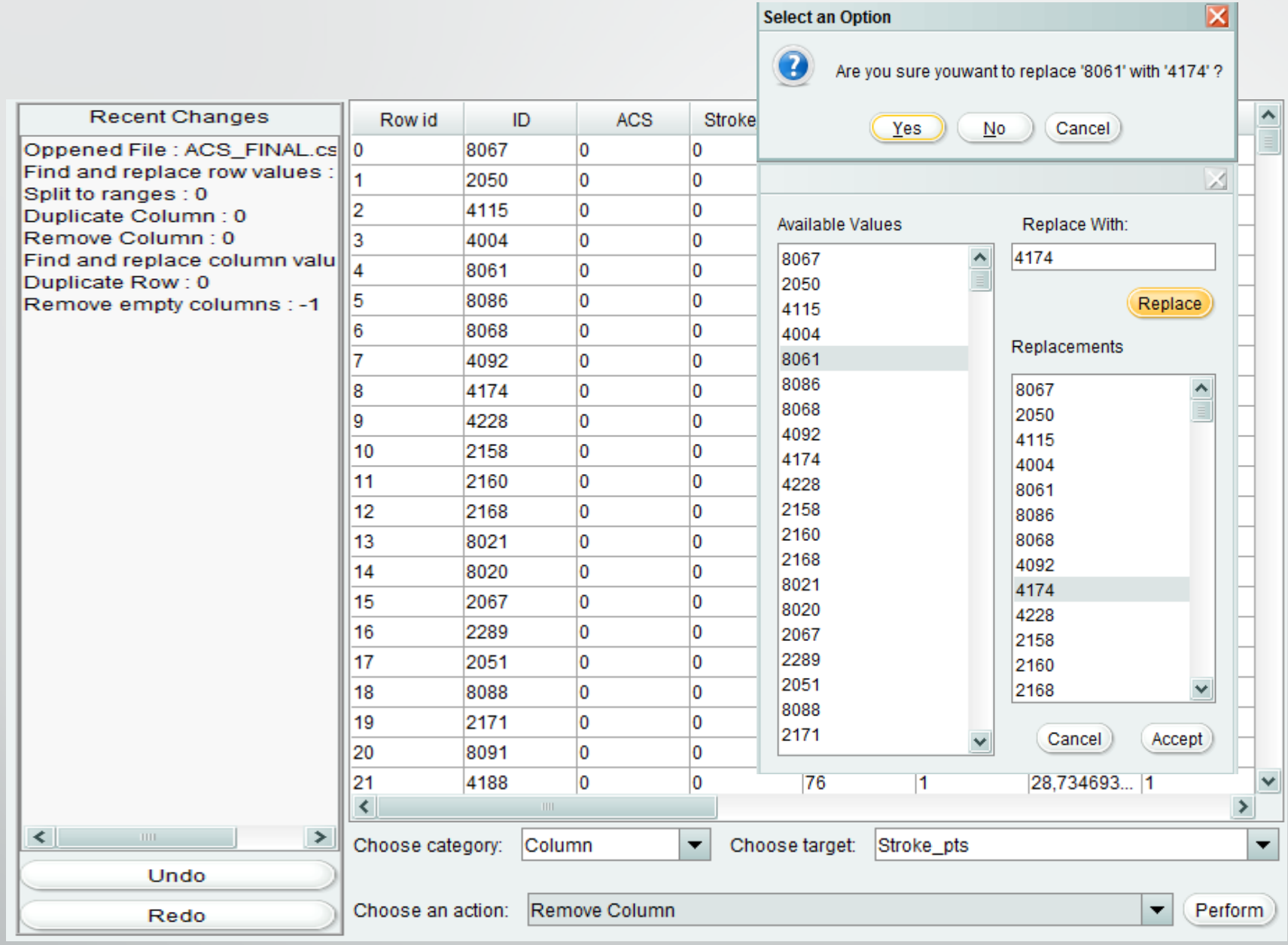
Data Handling

- Provides an easy way to import data
- … and useful information about the data


File manager
Preprocessor
Data Miner
Predictor
C:\Users\r0t0r\Desktop\Data_Miner\ACS_FINAL.csv
Open
Reload File
File Reading/Writting Options
Save As
Save
Instances (Rows):
1
Attributes (Columns):
ID
ACS
Stroke_pts
Age
Gender
BMI
PA
ever_smoker
Family_history_CHD
HPT
HCHOL
DM
MedDietScore
FAC1_2
FAC2_2
FAC3_2
FAC4_2
FAC5_2
Instance Info:
Columns: 18
Null Percentage: 0.0
Delete Instance
Attribute Info:
->Name: ID
->Total Population: 500
->Null Percentage: 0.0%
->Type: Numerical
->Statistics:
->min: 1001.0
->max: 8099.0
->avg: 2637.650000000001
->σ: 1616.372162436609
->VALUES:
->8067 : 1
->2050 : 1
->4115 : 1
->4004 : 1
->8061 : 1
->8086 : 1
Delete Attribute
Log:
2013/09/30 11:44:14
gr.hua.gui.MainMenu
Ready
Change Look
Quit

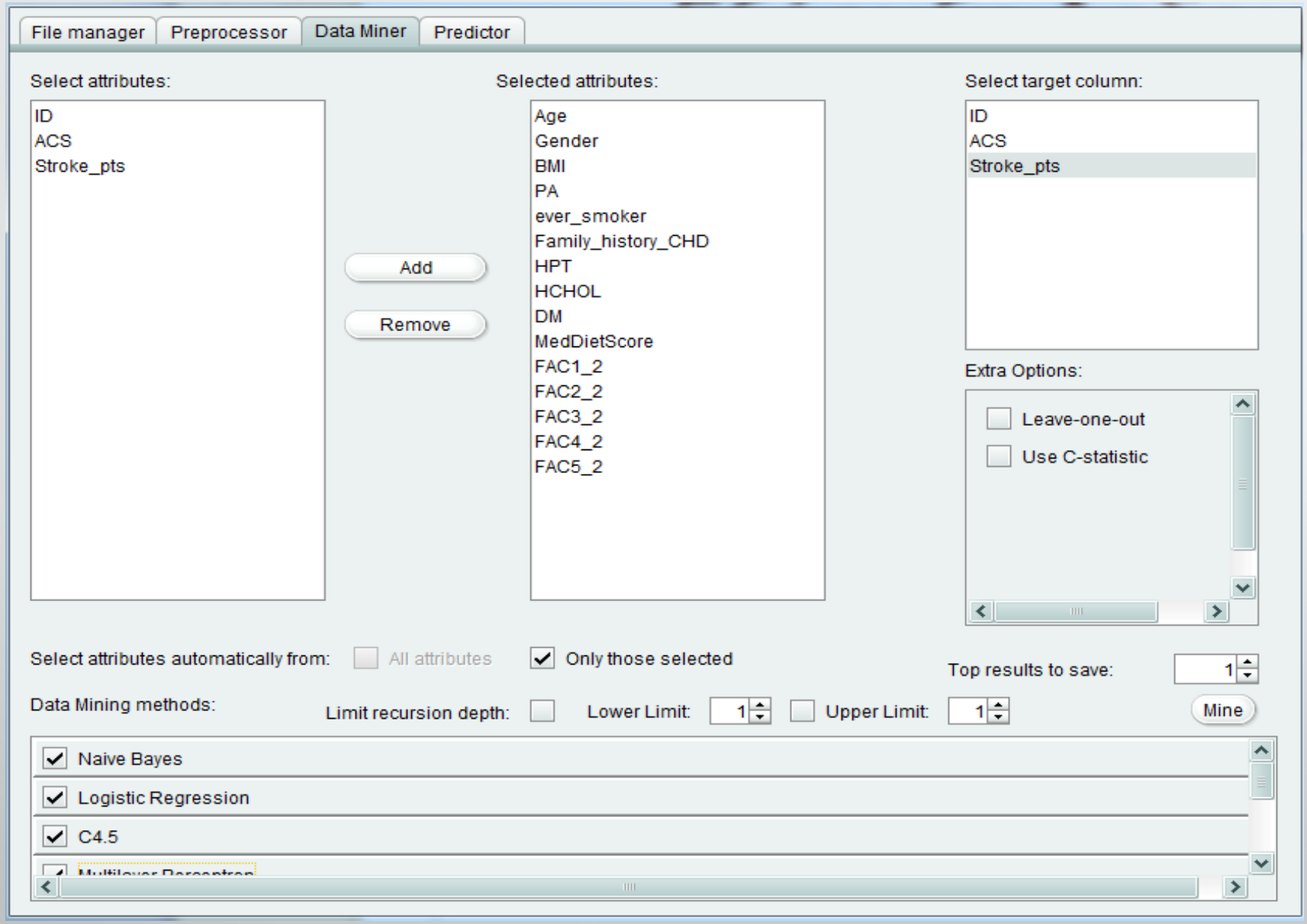
Data Preprocessing

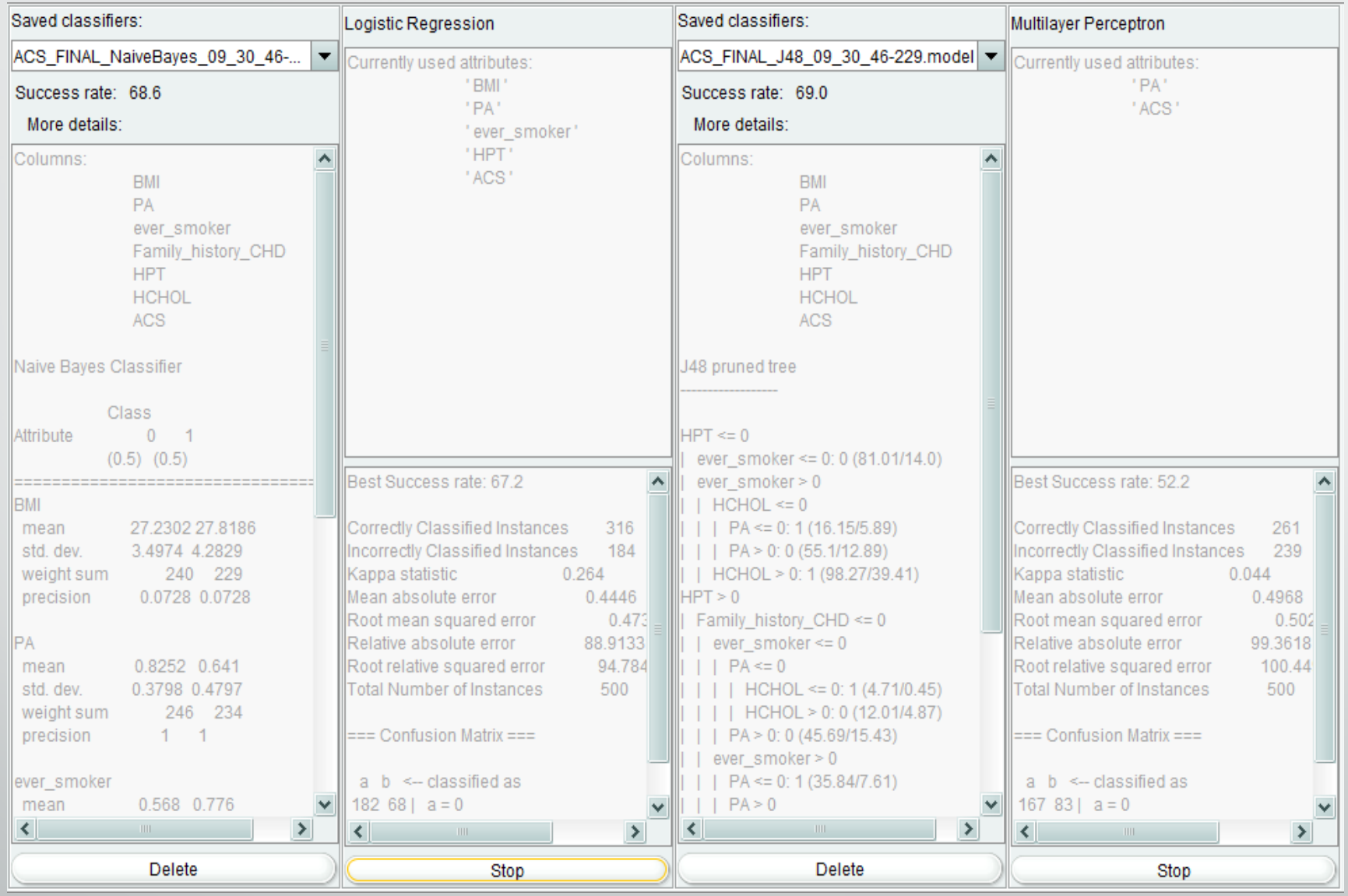
- Allows for many common preprocessing steps
- Complex steps are guided and accessible to non-experts


Recent Changes
Oppened File : ACS_FINAL.cs
Find and replace row values :
Split to ranges : 0
Duplicate Column : 0
Remove Column : 0
Find and replace column valu
Duplicate Row : 0
Remove empty columns : -1
Undo
Redo
Row id
ID
ACS
Stroke
0 8067 0 0
1 2050 0 0
2 4115 0 0
3 4004 0 0
4 8061 0 0
5 8086 0 0
6 8068 0 0
7 4092 0 0
8 4174 0 0
9 4228 0 0
10 2158 0 0
11 2160 0 0
12 2168 0 0
13 8021 0 0
14 8020 0 0
15 2067 0 0
16 2289 0 0
17 2051 0 0
18 8088 0 0
19 2171 0 0
20 8091 0 0
21 4188 0 0 76 1 28,734693... 1
Select an Option
Are you sure youwant to replace '8061' with '4174' ?
Yes
No
Cancel
Available Values
8067
2050
4115
4004
8061
8086
8068
4092
4174
4228
2158
2160
2168
8021
8020
2067
2289
2051
8088
2171
Replace With:
4174
Replace
Replacements
8067
2050
4115
4004
8061
8086
8068
4092
4174
4228
2158
2160
2168
Cancel
Accept
Choose category:
Column
Choose target:
Stroke_pts
Choose an action:
Remove Column
Perform

Data Mining

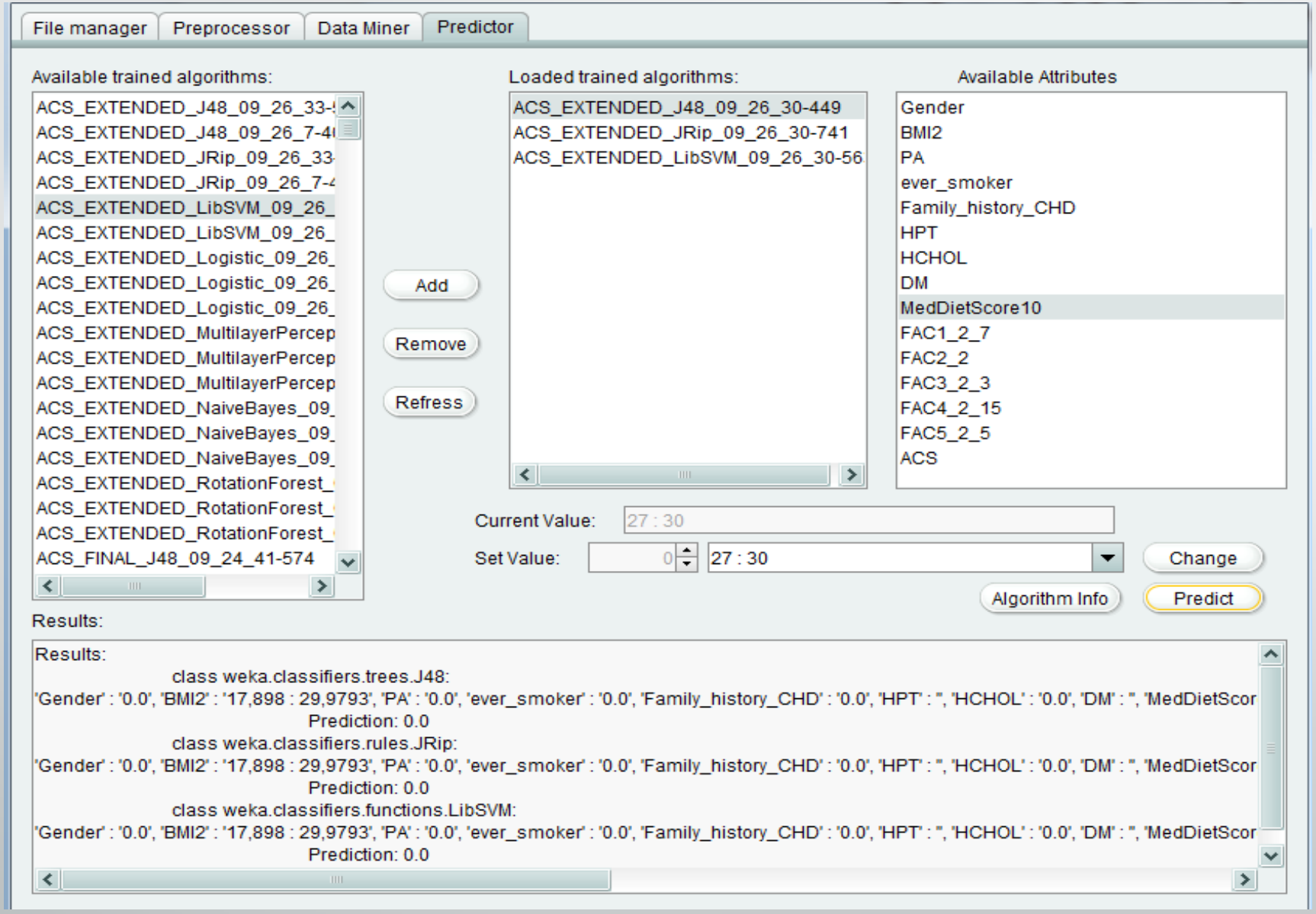
- Manages the data mining process
  - Automatically preforming specific optimizations
  - Checking, evaluating and storing results
  - Requiring minimal input from the user
- Allows users to monitor the progress


File manager
Preprocessor
Data Miner
Predictor
Select attributes:
ID
ACS
Stroke_pts
Add
Remove
Selected attributes:
Age
Gender
BMI
PA
ever_smoker
Family_history_CHD
HPT
HCHOL
DM
MedDietScore
FAC1_2
FAC2_2
FAC3_2
FAC4_2
FAC5_2
Select target column:
ID
ACS
Stroke_pts
Extra Options:
Leave-one-out
Use C-statistic
Select attributes automatically from:
All attributes
Only those selected
Top results to save:
1
Data Mining methods:
Limit recursion depth:
Lower Limit:
1
Upper Limit:
1
Mine
Naive Bayes
Logistic Regression
C4.5


Saved classifiers:
ACS_FINAL_NaiveBayes_09_30_46-...
Success rate: 68.6
More details:
Columns:
BMI
PA
ever_smoker
Family_history_CHD
HPT
HCHOL
ACS
Naive Bayes Classifier
Class
Attribute 0 1
(0.5) (0.5)
BMI
mean 27.2302 27.8186
std. dev. 3.4974 4.2829
weight sum 240 229
precision 0.0728 0.0728
PA
mean 0.8252 0.641
std. dev. 0.3798 0.4797
weight sum 246 234
precision 1 1
ever_smoker
mean 0.568 0.776
Delete
Logistic Regression
Currently used attributes:
'BMI'
'PA'
'ever_smoker'
'HPT'
'ACS'
Best Success rate: 67.2
Correctly Classified Instances 316
Incorrectly Classified Instances 184
Kappa statistic 0.264
Mean absolute error 0.4446
Root mean squared error 0.473
Relative absolute error 88.9133
Root relative squared error 94.784
Total Number of Instances 500
=== Confusion Matrix ===
a b <-- classified as
182 68 | a = 0
Stop
Saved classifiers:
ACS_FINAL_J48_09_30_46-229.model
Success rate: 69.0
More details:
Columns:
BMI
PA
ever_smoker
Family_history_CHD
HPT
HCHOL
ACS
J48 pruned tree
HPT <= 0
| ever_smoker <= 0: 0 (81.01/14.0)
| ever_smoker > 0
| | HCHOL <= 0
| | | PA <= 0: 1 (16.15/5.89)
| | | PA > 0: 0 (55.1/12.89)
| | HCHOL > 0: 1 (98.27/39.41)
HPT > 0
| Family_history_CHD <= 0
| | ever_smoker <= 0
| | | PA <= 0
| | | | HCHOL <= 0: 1 (4.71/0.45)
| | | | HCHOL > 0: 0 (12.01/4.87)
| | | PA > 0: 0 (45.69/15.43)
| | ever_smoker > 0
| | | PA <= 0: 1 (35.84/7.61)
| | | PA > 0
Delete
Multilayer Perceptron
Currently used attributes:
'PA'
'ACS'
Best Success rate: 52.2
Correctly Classified Instances 261
Incorrectly Classified Instances 239
Kappa statistic 0.044
Mean absolute error 0.4968
Root mean squared error 0.502
Relative absolute error 99.3618
Root relative squared error 100.44
Total Number of Instances 500
=== Confusion Matrix ===
a b <-- classified as
167 83 | a = 0
Stop

Prediction

- Guides users to make predictions on new cases
- Automatically using results from the data mining process


File manager
Preprocessor
Data Miner
Predictor
Available trained algorithms:
ACS_EXTENDED_J48_09_26_33-
ACS_EXTENDED_J48_09_26_7-4
ACS_EXTENDED_JRip_09_26_33
ACS_EXTENDED_JRip_09_26_7-
ACS_EXTENDED_LibSVM_09_26_
ACS_EXTENDED_LibSVM_09_26_
ACS_EXTENDED_Logistic_09_26_
ACS_EXTENDED_Logistic_09_26_
ACS_EXTENDED_Logistic_09_26_
ACS_EXTENDED_MultilayerPercep
ACS_EXTENDED_MultilayerPercep
ACS_EXTENDED_MultilayerPercep
ACS_EXTENDED_NaiveBayes_09_
ACS_EXTENDED_NaiveBayes_09_
ACS_EXTENDED_NaiveBayes_09_
ACS_EXTENDED_RotationForest_
ACS_EXTENDED_RotationForest_
ACS_EXTENDED_RotationForest_
ACS_FINAL_J48_09_24_41-574
Add
Remove
Refress
Loaded trained algorithms:
ACS_EXTENDED_J48_09_26_30-449
ACS_EXTENDED_JRip_09_26_30-741
ACS_EXTENDED_LibSVM_09_26_30-56
Available Attributes
Gender
BMI2
PA
ever_smoker
Family_history_CHD
HPT
HCHOL
DM
MedDietScore10
FAC1_2_7
FAC2_2
FAC3_2_3
FAC4_2_15
FAC5_2_5
ACS
Current Value:
27 : 30
Set Value:
0
27 : 30
Change
Algorithm Info
Predict
Results:
Results:
class weka.classifiers.trees.J48:
'Gender' : '0.0', 'BMI2' : '17,898 : 29,9793', 'PA' : '0.0', 'ever_smoker' : '0.0', 'Family_history_CHD' : '0.0', 'HPT' : '', 'HCHOL' : '0.0', 'DM' : '', 'MedDietScor
Prediction: 0.0
class weka.classifiers.rules.JRip:
'Gender' : '0.0', 'BMI2' : '17,898 : 29,9793', 'PA' : '0.0', 'ever_smoker' : '0.0', 'Family_history_CHD' : '0.0', 'HPT' : '', 'HCHOL' : '0.0', 'DM' : '', 'MedDietScor
Prediction: 0.0
class weka.classifiers.functions.LibSVM:
'Gender' : '0.0', 'BMI2' : '17,898 : 29,9793', 'PA' : '0.0', 'ever_smoker' : '0.0', 'Family_history_CHD' : '0.0', 'HPT' : '', 'HCHOL' : '0.0', 'DM' : '', 'MedDietScor
Prediction: 0.0

Experiment Results

Real Datasets

- Medical case-control study on dietary patterns (kindly provided by professor Panagiotakos*)
- Acute Coronary Syndrome (ACS)

and Stroke medical datasets

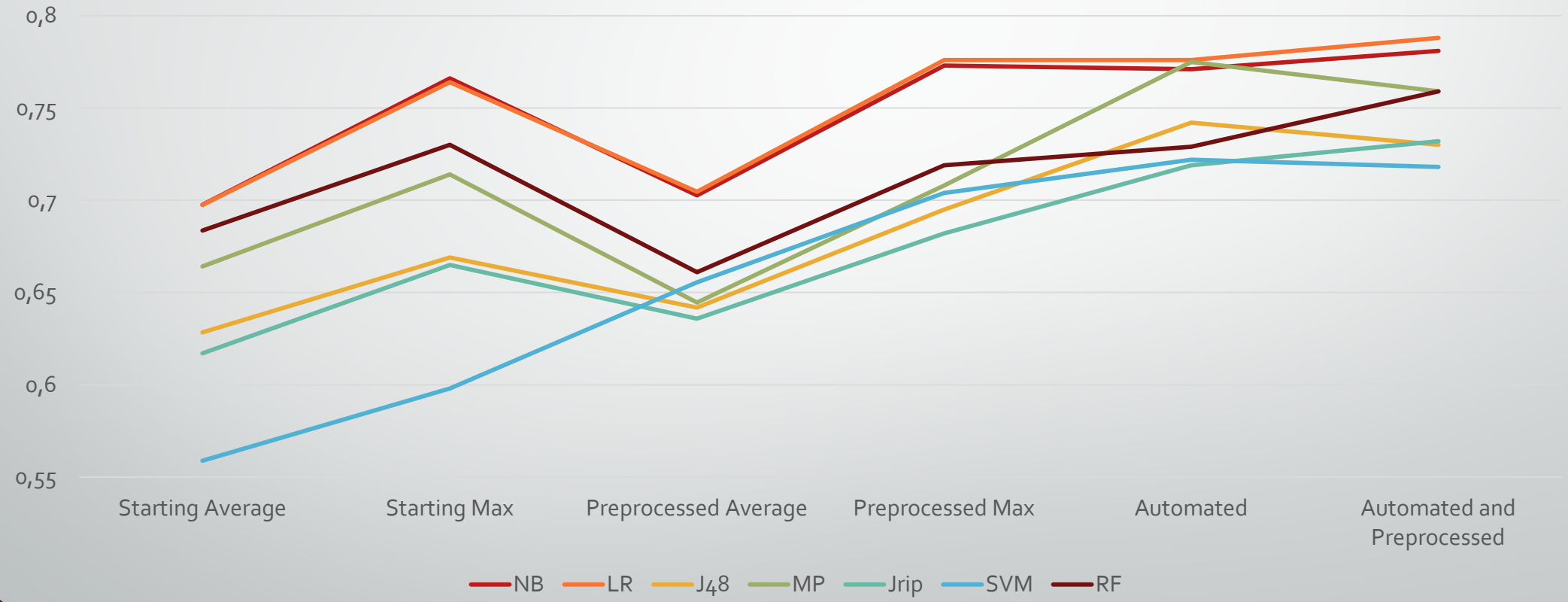
Instances:	500
Attributes:	16
Classes:	2

*Christina-Maria Kastorini, George Papadakis et al. Comparative analysis of a-priori and a-posteriori dietary patterns using state-of-the-art classification algorithms: A case/case-control study. Artificial Intelligence in Medicine 59 (2013) 175–183

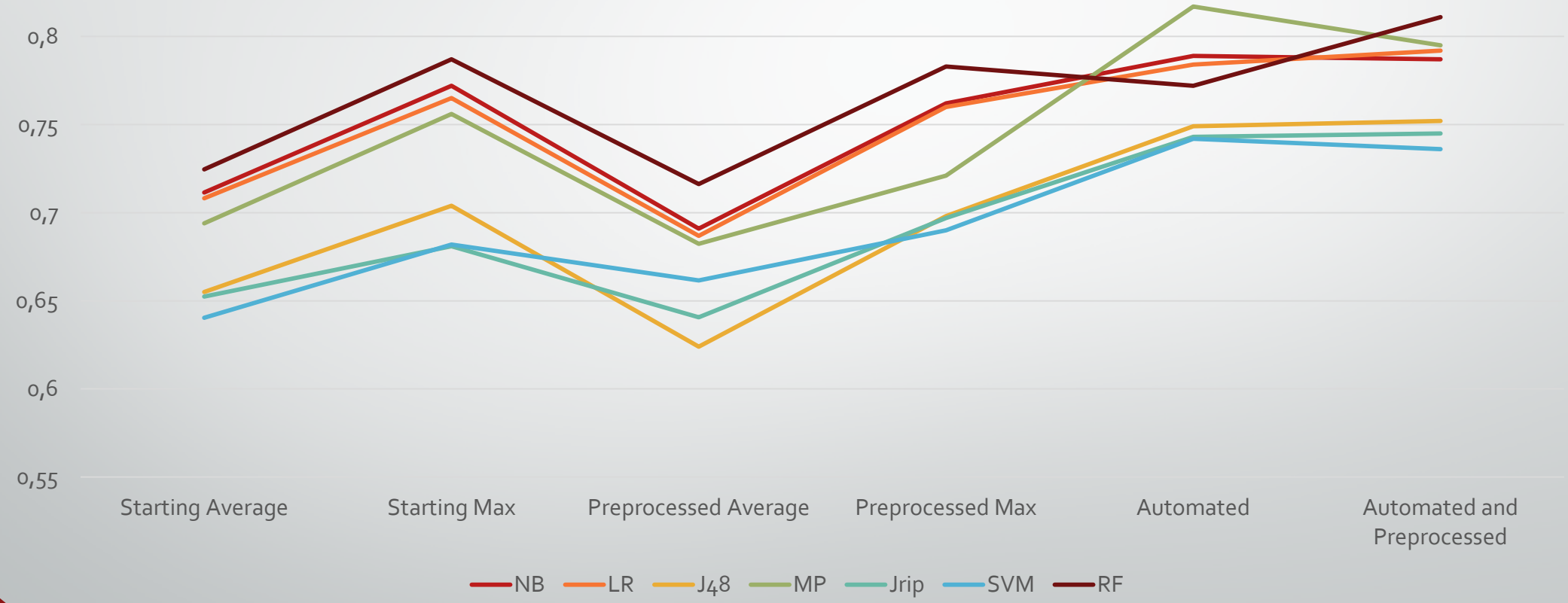
Case Study

- Using the previous study as a starting point
  - Are experiments easy and fast to set up?
  - Can the medical professional get equivalent results?
  - Is information, useful for data mining experts, produced regarding
    - performance improvements through preprocessing?
    - performance improvements through optimizations?
    - data mining algorithms' behavior?

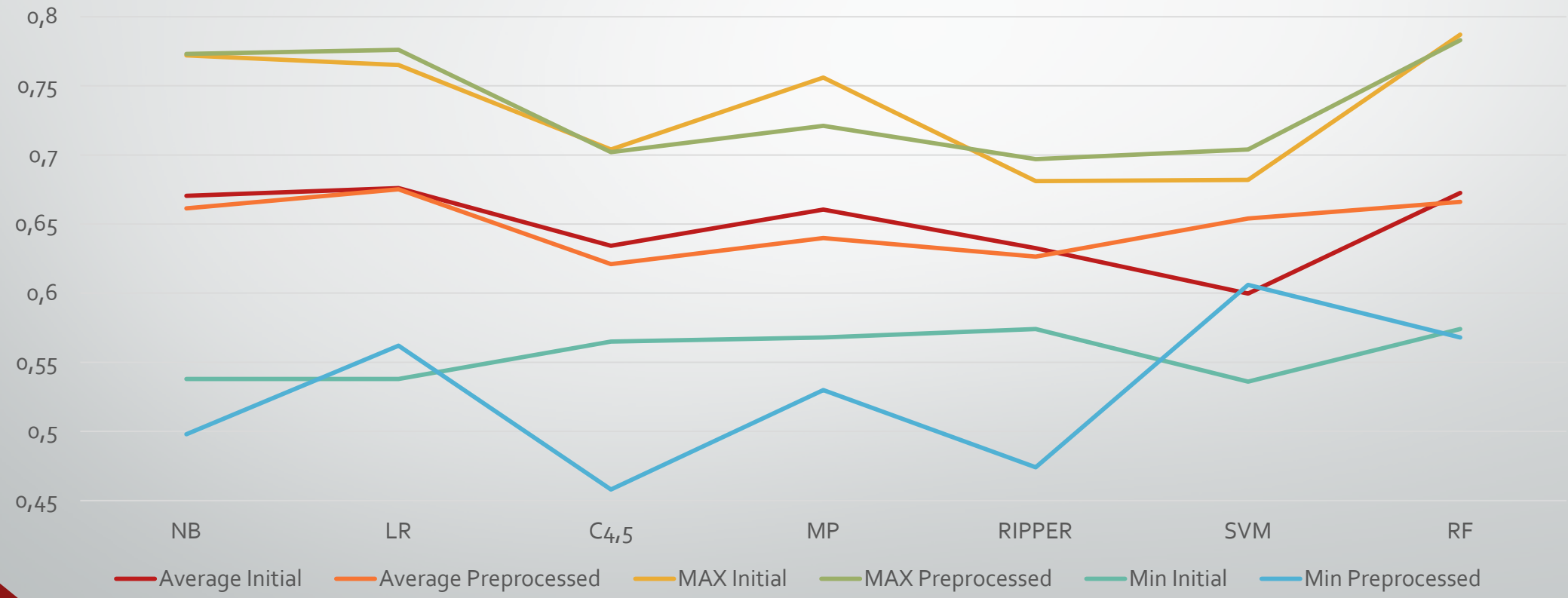
ACS


0,8
0,75
0,7
0,65
0,6
0,55
Starting Average
Starting Max
Preprocessed Average
Preprocessed Max
Automated
Automated and Preprocessed
NB
LR
J48
MP
Jrip
SVM
RF

Stroke


0,8
0,75
0,7
0,65
0,6
0,55
Starting Average
Starting Max
Preprocessed Average
Preprocessed Max
Automated
Automated and Preprocessed
NB
LR
J48
MP
Jrip
SVM
RF

Algorithm Robustness



Early Prototype Results

- Small time to set up an experiment
- Results support the usefulness of such automated procedures
- Relatively easy to understand and use
- An important first step towards the full vision

Future Work

Future Work
Short Term

- Better, step by step guided, graphical user interface
- Enhanced preprocessing capabilities
- Automated preprocessing procedures based on expert knowledge
- Preprocessing optimization
- Improved data mining optimization techniques for speed
- Better, easier to understand, presentation of the results

Future Work
Long Term

- Expand beyond simple classification
- Integration of more data sources
- Automated data collection from databases
- Support for big distributed processing
- Support for further optimizations (e.g. GPU acceleration)

Thank you!
Any questions?