

Базисни софтуерни продукти за работа с данни

Ангел Марчев, мл.

Калоян Харалампиев

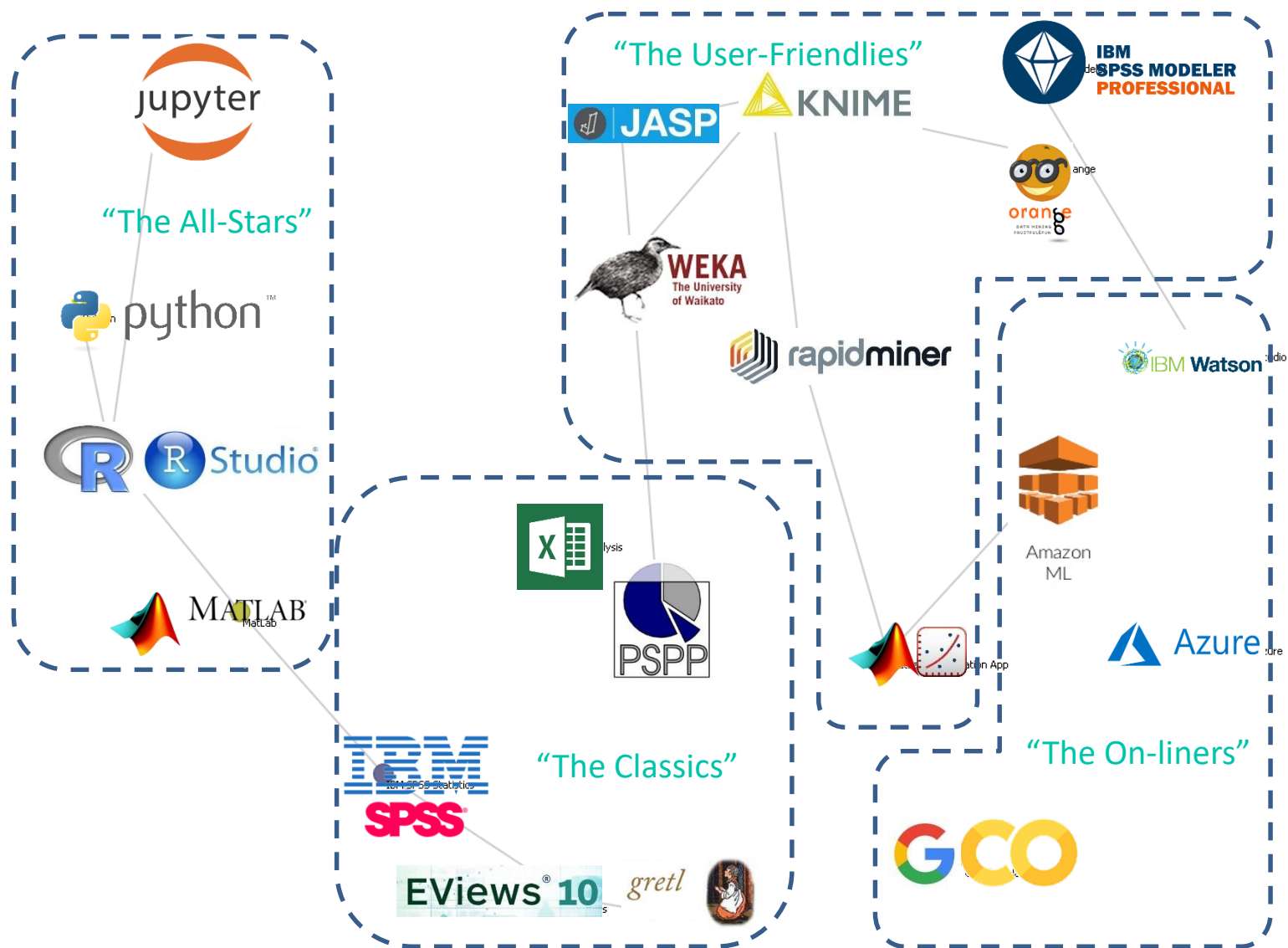
Въведение

- Невъзможно е да се покрият всички съществуващи продукти
- Затова ние редуцираме броя на продуктите до тези, които самите ние използваме
- Но задачата все още е трудна, защото:
 - Има различни типове продукти
 - Всеки продукт може да се класифицира по много критерии
 - Всеки продукт може да се използва за много цели

Критерии за класификация

- Приложение
 - Статистика
 - Иконометрия
 - Data mining
- Начин на работа
 - Писане на програмен код
 - Работа с менюта и прозорци
 - Работа с графични обекти
 - Работа онлайн
- Цена
 - Безплатни
 - Платени
- Свързаност (приемственост)
- Популярност
- Интерактивност

Една възможна класификация



The Classics

Excel Data Analysis

- Приложение: Статистика
- Начин на работа: Менюта и прозорци
- Цена: Платен
- Предимства: Достъпност (почти всеки има Excel)
- Недостатъци: Работи се чрез избор на клетки, а не чрез имената на променливите

Data Analysis

Analysis Tools

Anova: Single Factor
Anova: Two-Factor With Replication
Anova: Two-Factor Without Replication
Correlation
Covariance
Descriptive Statistics
Exponential Smoothing

OK

Cancel

Help

Correlation

Input

Input Range:

\$AF\$1:\$AH\$74

Grouped By:

☒ Columns

☐ Rows

☒ Labels in First Row

OK

Cancel

Help

Output options

☐ Output Range:

☒ New Worksheet Ply:

☐ New Workbook

	A	B	C	D	E
1		Mean Plausible Value in Mathematics	Mean Plausible Value in Reading	Mean Plausible Value in Science	
2	Mean Plausible Value in Mathematics	1			
3	Mean Plausible Value in Reading	0,936674148	1		
4	Mean Plausible Value in Science	0,972625261	0,963215102	1	
5					

IBM SPSS Statistics



- Приложение: Статистика, иконометрия
- Начин на работа: Менюта и прозорци; писане на програмен код
- Цена: Платен
- Предимства: Огромен набор от възможни анализи
- Недостатъци: Ниска степен на интерактивност

Correlations				
		Mean Plausible Value in Mathematics	Mean Plausible Value in Reading	Mean Plausible Value in Science
Mean Plausible Value in Mathematics	Pearson Correlation	1	,937**	,973**
	Sig. (2-tailed)		,000	,000
	N	73	73	73
Mean Plausible Value in Reading	Pearson Correlation	,937**	1	,963**
	Sig. (2-tailed)	,000		,000
	N	73	73	73
Mean Plausible Value in Science	Pearson Correlation	,973**	,963**	1
	Sig. (2-tailed)	,000	,000	
	N	73	73	73

** . Correlation is significant at the 0.01 level (2-tailed).

CORRELATIONS

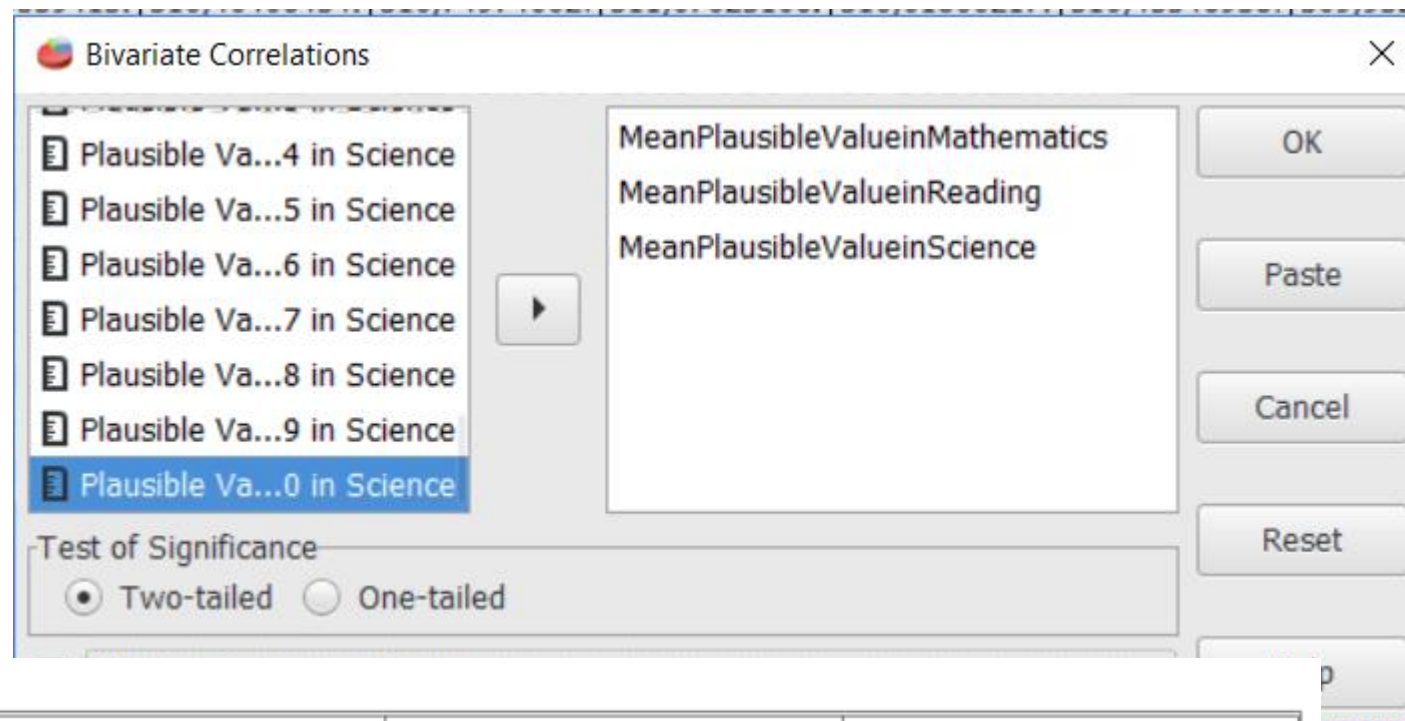
```

/VARIABLES=MeanPlausibleValueinMathematics MeanPlausibleValueinReading MeanPlausibleValueinScience
/PRINT=TWOTAIL NOSIG
/MISSING=PAIRWISE.
    
```

PSPP



- Приложение: Статистика
- Начин на работа: Менюта и прозорци
- Цена: Безплатен
- Предимства: Наподобява IBM SPSS Statistics
- Недостатъци: Относително малък набор от възможни анализи; не е интерактивен



Correlations

		Mean Plausible Value in Mathematics	Mean Plausible Value in Reading	Mean Plausible Value in Science
Mean Plausible Value in Mathematics	Pearson Correlation	1,00	,94	,97
	Sig. (2-tailed)		,000	,000
	N	73	73	73
Mean Plausible Value in Reading	Pearson Correlation	,94	1,00	,96
	Sig. (2-tailed)	,000		,000
	N	73	73	73
Mean Plausible Value in Science	Pearson Correlation	,97	,96	1,00
	Sig. (2-tailed)	,000	,000	
	N	73	73	73

EViews

EViews® 10

- Приложение:
Иконометрия
- Начин на работа:
Менюта и прозорци;
писане на програмен
код
- Цена: Платен
- Предимства: Ефективни
изчисления
- Недостатъци: Проблеми
с използването на данни
от други продукти

VAR Specification

Basics Cointegration VEC Restrictions

VAR type

☐ Standard VAR

☒ Vector Error Correction

☐ Bayesian VAR

Endogenous variables

women jewel

Estimation sample

1989m01 1998m12

Lag Intervals for D(Endogenous):

1 11

Exogenous variables

Do NOT include C or Trend in VEC's

OK Cancel

View	Proc	Object	Print	Name	Freeze	Estimate	Forecast	Stats	Impulse	Resids
------	------	--------	-------	------	--------	----------	----------	-------	---------	--------

Vector Error Correction Estimates
Date: 03/12/18 Time: 22:52
Sample (adjusted): 1990M01 1998M12
Included observations: 108 after adjustments
Standard errors in () & t-statistics in []

Cointegrating Eq:	CointEq1
WOMEN(-1)	1.000000
JEWEL(-1)	-33.57107 (10.2931) [-3.26153]
C	520459.3

Error Correction:	D(WOMEN)	D(JEWEL)
CointEq1	0.068665 (0.02768) [2.48089]	0.053523 (0.01673) [3.19907]
D(WOMEN(-1))	-0.961166 (0.15470) [-6.21327]	-0.080627 (0.09351) [-0.86220]
D(WOMEN(-2))	-0.802443 (0.19463) [-4.12295]	0.037585 (0.11765) [0.31946]
D(WOMEN(-3))	-0.805171 (0.21540) [-3.73799]	-0.104838 (0.13021) [-0.80515]
D(WOMEN(-4))	-0.628510 (0.22110) [-2.84259]	-0.008221 (0.13366) [-0.06151]
D(WOMEN(-5))	-0.298444 (0.22719) [-1.31365]	0.147138 (0.13733) [1.07139]

Gretl

gretl



- Приложение: Иконометрия
- Начин на работа: Менюта и прозорци; писане на програмен код
- Цена: Безплатен
- Предимства: Наподобява EViews
- Недостатъци: Ограничения за обема на данните

gretl: векторен модел за коригиране н...

Векторен модел за коригиране на грешката (VECM)

date
men
women
jewel
mail
page
phone

лагов порядък: 11
ранг: 1

Ендогенни променливи
women
jewel

ни променливи

VECM система, лагов порядък 11
Максимално правдоподобие оценки, наблюдения 1989:12-1998:12 (T = 109)
Ранг на коинтеграция = 1
Случай 3: Неограничена константа

beta (коинтеграционни вектори, стандартна грешка в скобите)

women 1,0000
(0,00000)
jewel -425,50
(86,834)

alpha (коригиращи вектори)

women 0,0086199
jewel 0,0062820

Log-likelihood = -2127,6344
Детерминанта на ковариационната матрица = 3,087056e+014
AIC-Критерий на Акайке = 39,8832
BIC - Критерий на Шварц = 41,0190
HQC - Крит. на Ханан-Куин = 40,3438

Уравнение 1: d_women

	коэффициент	стандартна грешка	t-критерий	p-value	
const	61519,8	15925,9	3,863	0,0002	***
d_women_1	-0,816250	0,177479	-4,599	1,43e-05	***
d_women_2	-0,343692	0,211974	-1,621	0,1086	
d_women_3	-0,406954	0,238931	-1,703	0,0921	*
d_women_4	-0,252055	0,248397	-1,015	0,3130	
d_women_5	0,106608	0,252535	0,4222	0,6740	
d_women_6	0,194529	0,251673	0,7729	0,4417	
d_women_7	-0,101355	0,242738	-0,4175	0,6773	
d_women_8	0,177352	0,222018	0,7988	0,4266	

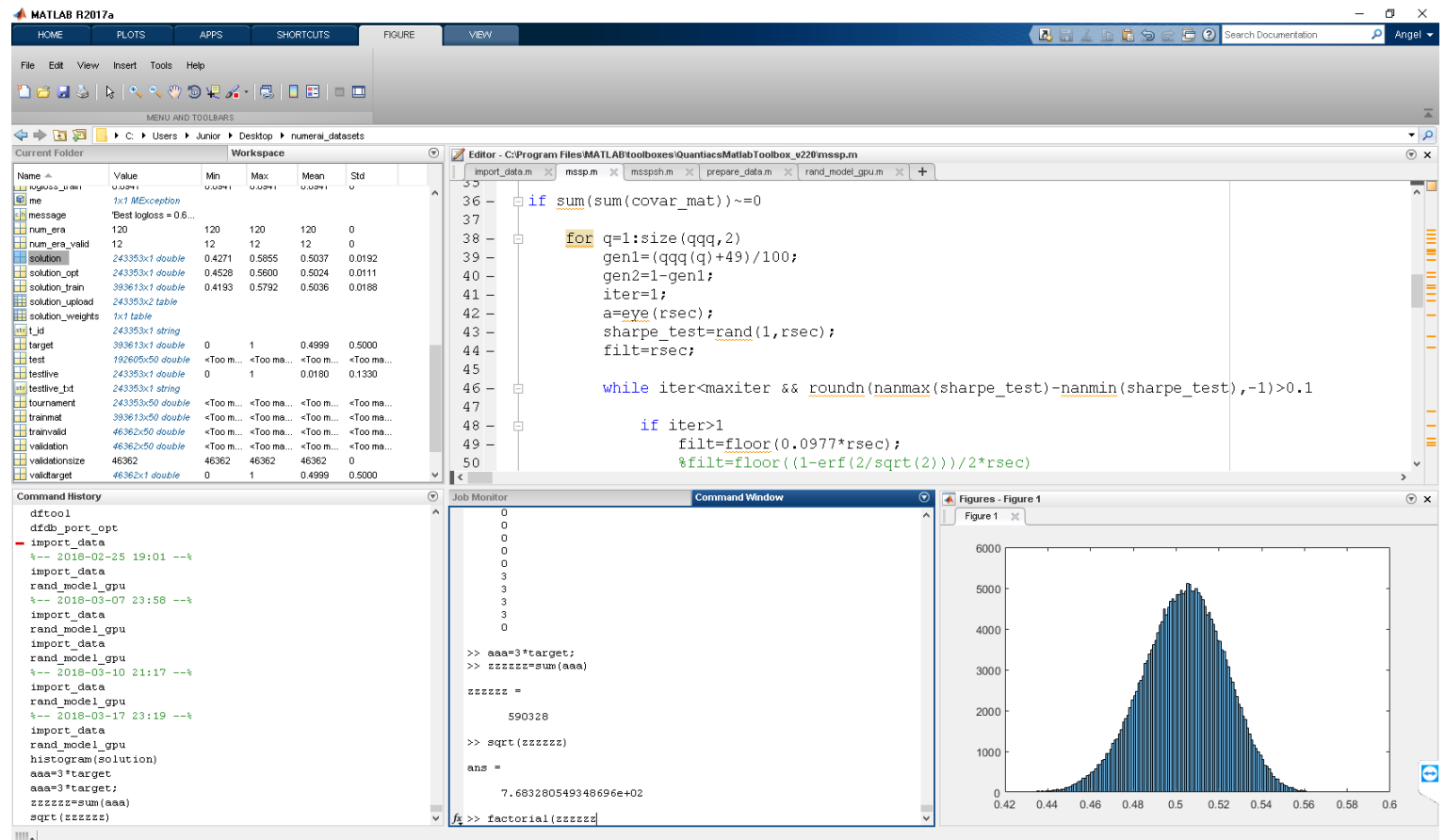
Добре

The All-Stars

MatLab



- Приложение: Статистика; иконометрия
- Начин на работа: Писане на програмен код
- Цена: Платен
- Предимства: Много добра документация; паралелни изчисления
- Недостатъци: Много висока цена



Python



- Приложение: Статистика; иконометрия; data mining
- Начин на работа: Писане на програмен код
- Цена: Безплатен
- Предимства: Глобална общност от разработчици
- Недостатъци:

The screenshot shows two windows from a Python 3.4.3 environment. The left window is a 'Python 3.4.3 Shell' displaying a list of installed modules in a three-column format. The right window is an editor for 'example_model.py' showing a script that imports pandas, numpy, and sklearn, loads training and prediction data from CSV files, and uses a LogisticRegression model for prediction.

```
*Python 3.4.3 Shell*
File Edit Shell Debug Options Window Help
_funcools      cythonmagic    pickle
_hashlib       datetime       pip
_heapq         dateutil       pipes
_imp           dbm            pkg_re
_io            decimal        pkgutil
_json          decorator      platform
_locale        difflib        plist
_lsprof        dill           poplib
_lzma          dis            posix
_markerlib     distutils      pprint
_markupbase    doctest        profile
_md5           docutils       pstats
_msi           dummy_threading
_multibytecodec
_multiprocessing
_opcode        email          py_compile
_operator      encodings      pydoc
_osx_support   ensurepip      pyexpat
_overlapped    enum           pygments
_pickle        errno          pylab
_pyio          faulthandler   pyparser
_random        filecmp        pyqtgraph
_shal          fileinput      pyreadline
_sha256        fnmatch        pytz
_sha512        formatter      queue
_sitebuiltins  fractions      quopri

Enter any module name to get more help. Or, type 'h'
for modules whose name or summary contain the
help>
```

```
example_model.py - C:\Users\acer\Desktop\numerai_codes\example_model.py (3.4.3)
File Edit Format Run Options Window Help

import pandas as pd
import numpy as np
from sklearn import metrics, preprocessing, linear_model

def main():
    # Set seed for reproducibility
    np.random.seed(0)

    print("Loading data...")
    # Load the data from the CSV files
    training_data = pd.read_csv('numerai_training_data.csv', header=0)
    prediction_data = pd.read_csv('numerai_tournament_data.csv', header=0)

    # Transform the loaded CSV data into numpy arrays
    features = [f for f in list(training_data) if "feature" in f]
    X = training_data[features]
    Y = training_data["target"]
    x_prediction = prediction_data[features]
    ids = prediction_data["id"]

    # This is your model that will learn to predict
    model = linear_model.LogisticRegression(n_jobs=-1)

    print("Training...")
    # Your model is trained on the training_data
    model.fit(X, Y)

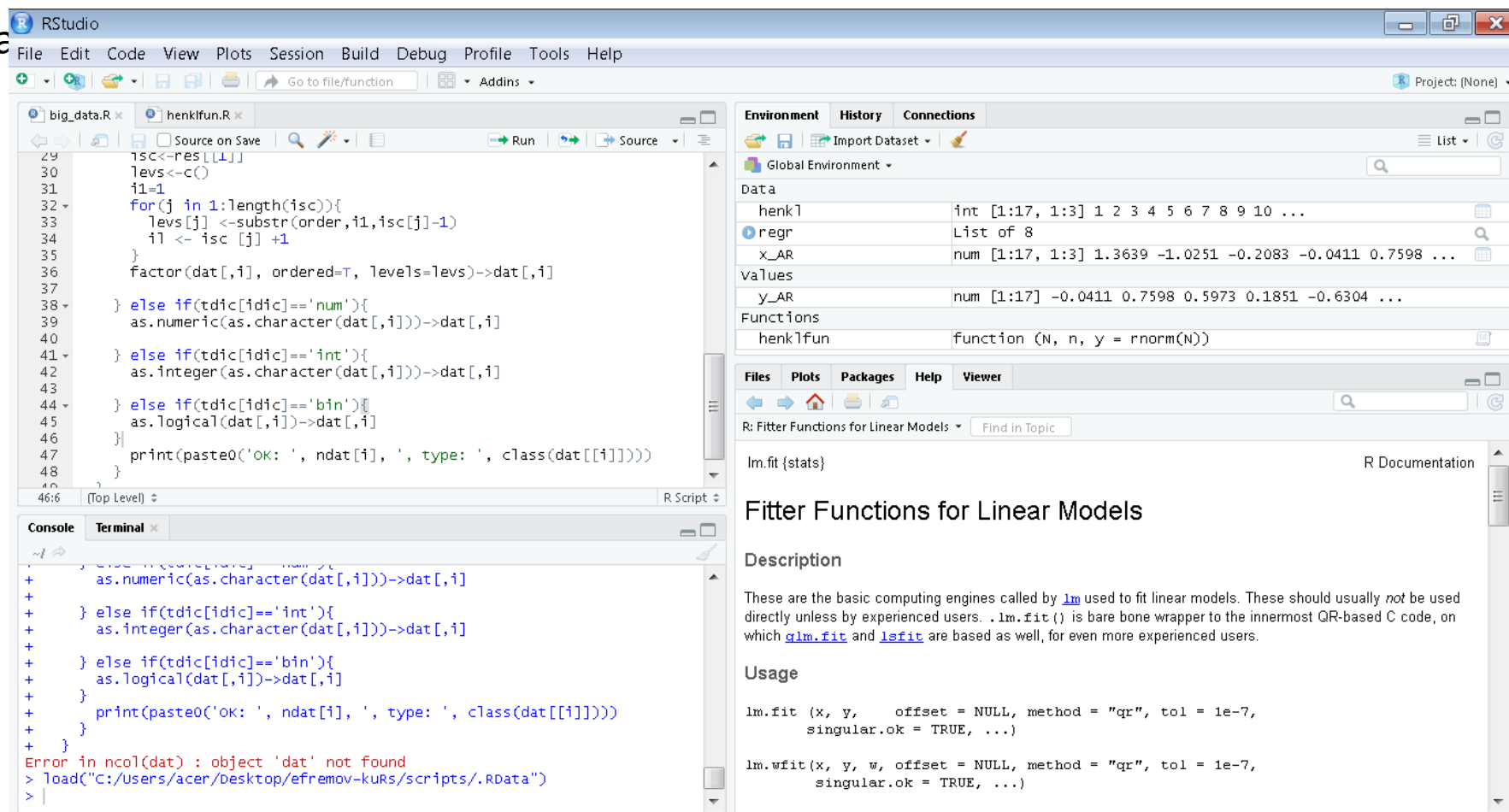
    print("Predicting...")
    # Your trained model is now used to make predictions on the numerai tournament data
```

Ln: 1 Col: 0

R (+R studio)



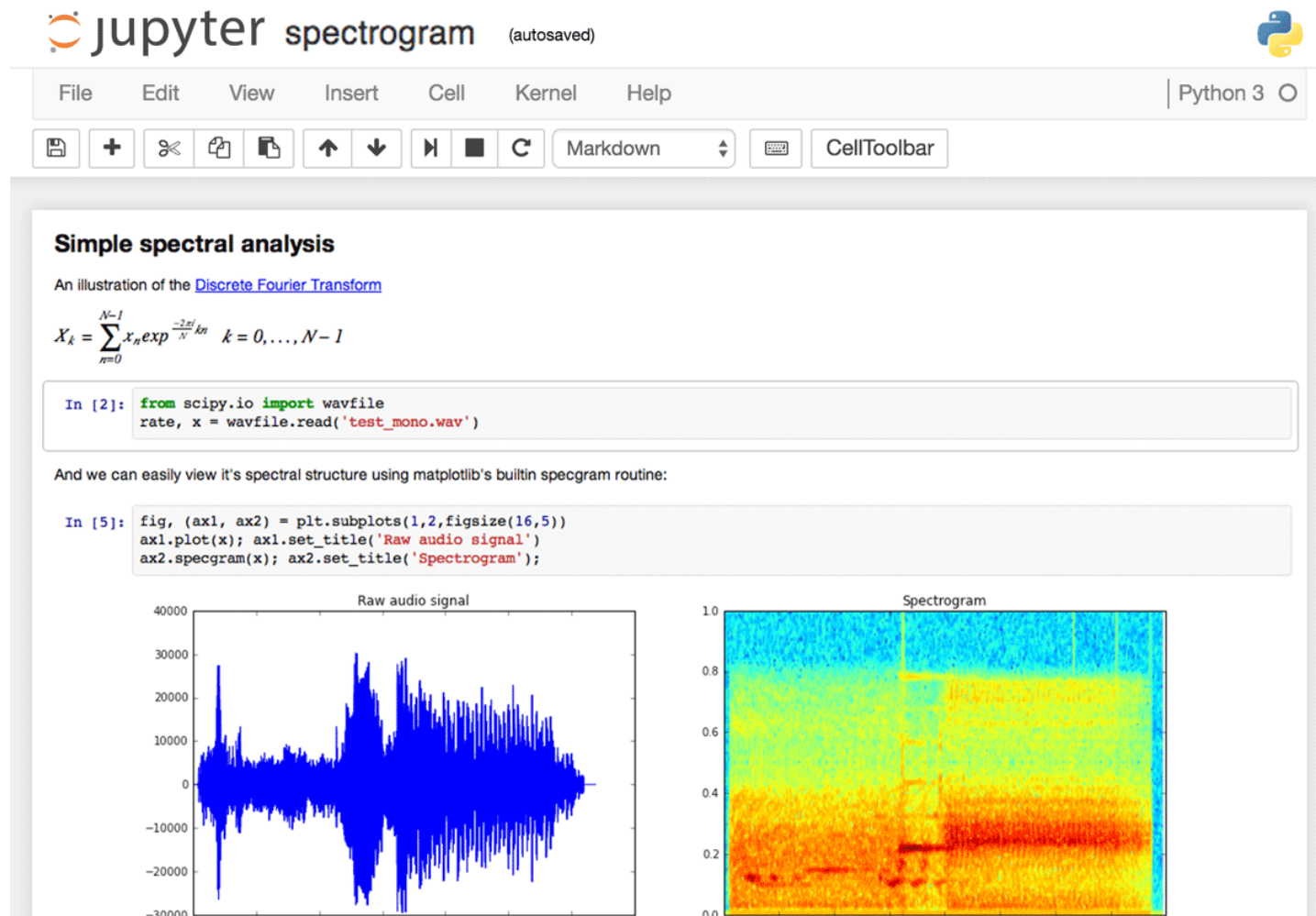
- Приложение: Статистика; иконометрия; data mining
- Начин на работа: Писане на програмен код
- Цена: Безплатен
- Предимства: Глобална общност от разработчици
- Недостатъци: Малко странен език



Jupyter Notebook



- Приложение: Data mining
- Начин на работа: Онлайн
- Цена: Безплатен
- Предимства: Индустриален стандарт за Data Science
- Недостатъци:



The User-Friendlies

JASP



- Приложение: Статистика
- Начин на работа: Менюта и прозорци
- Цена: Безплатен
- Предимства: Висока степен на интерактивност
- Недостатъци: Относително малък набор от възможни анализи

The screenshot displays the JASP software interface. The top menu bar includes 'File', 'Common', and a '+' icon. Below the menu is a toolbar with icons for Descriptives, T-Tests, ANOVA, Regression, Frequencies, and Factor. The main window is divided into two panes. The left pane contains a list of variables on the left, a list of selected variables on the right, and a list of correlation coefficients (Pearson, Spearman, Kendall's tau-b) with checkboxes for various options like 'Display pairwise table', 'Report significance', 'Flag significant correlations', 'Confidence intervals', and 'Vovk-Sellke maximum p-ratio'. The right pane shows the 'Results' section, specifically the 'Correlation Matrix' for Pearson Correlations. The matrix shows the correlation between 'Mean Plausible Value in Science', 'Mean Plausible Value in Mathematics', and 'Mean Plausible Value in Reading'. The correlation between Science and Mathematics is 0.973***, and between Science and Reading is 0.963***. The correlation between Mathematics and Reading is 0.937***. A legend at the bottom indicates that * p < .05, ** p < .01, and *** p < .001.

Results

Correlation Matrix

Pearson Correlations

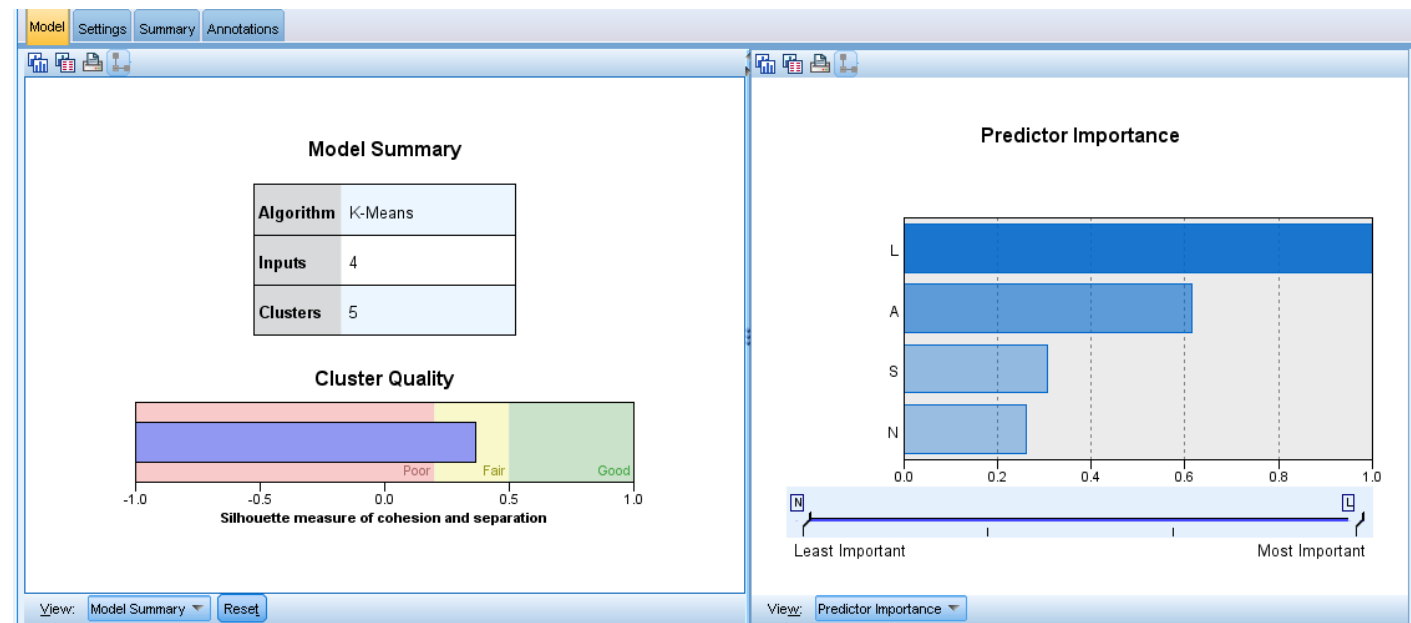
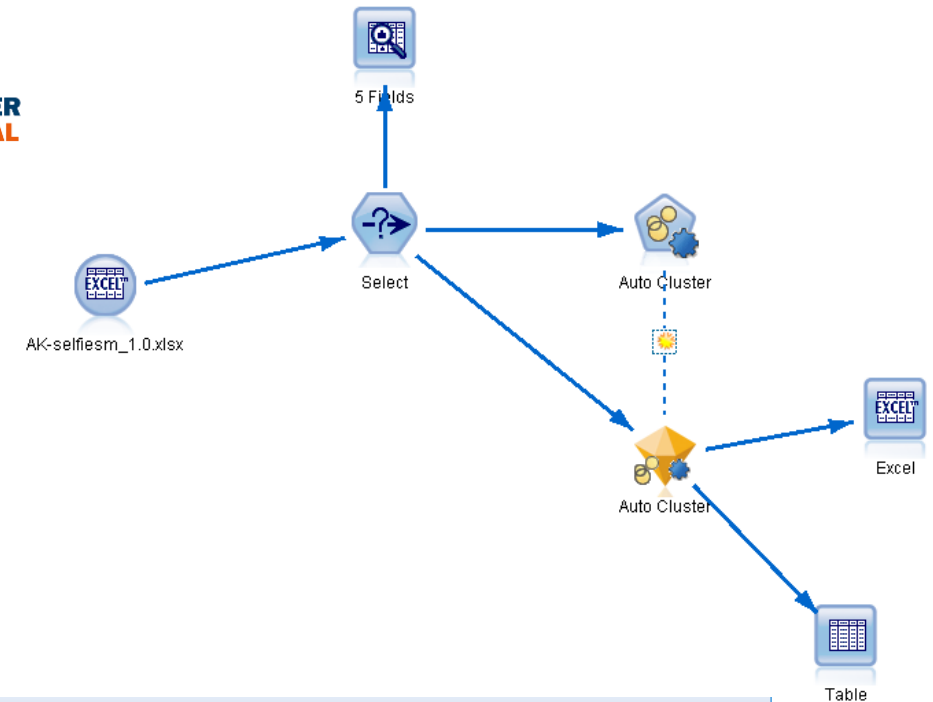
	Mean Plausible Value in Science	Mean Plausible Value in Mathematics	Mean Plausible Value in Reading
Mean Plausible Value in Science	—		
Mean Plausible Value in Mathematics	0.973***	—	
Mean Plausible Value in Reading	0.963***	0.937***	—

* p < .05, ** p < .01, *** p < .001

IBM SPSS Modeler



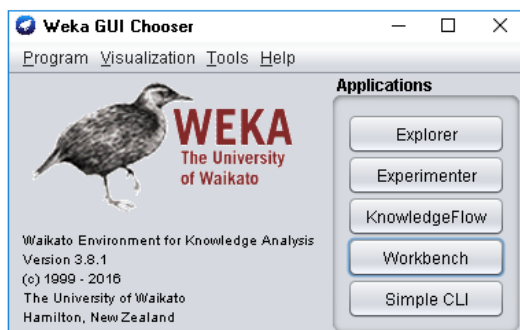
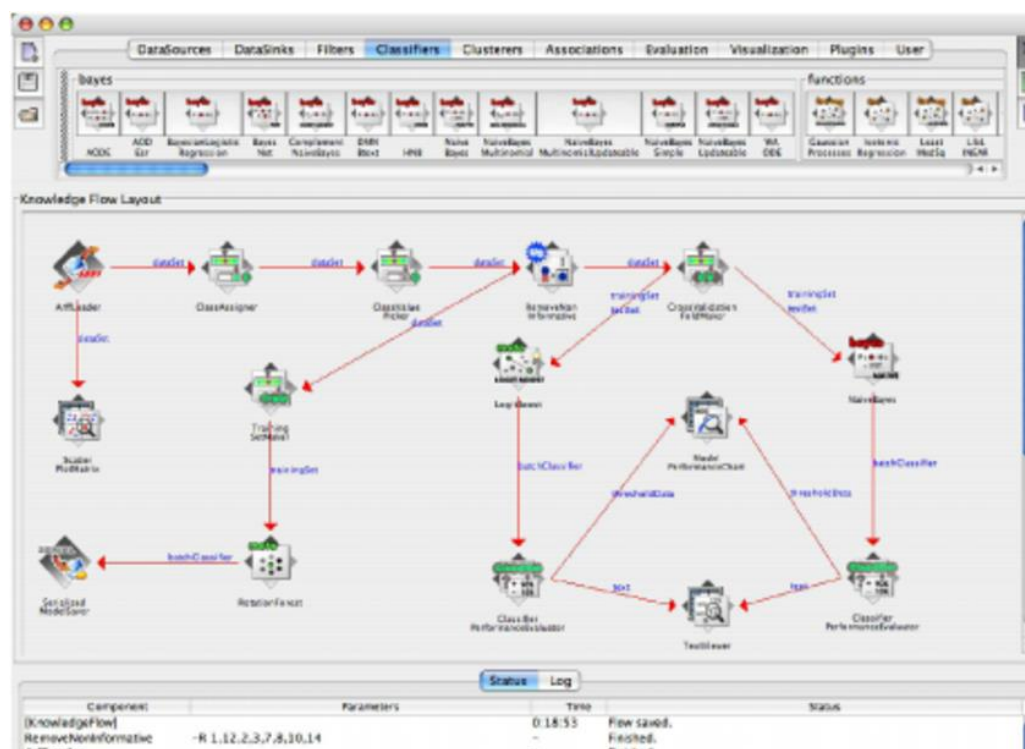
- Приложение: Иконометрия; data mining
- Начин на работа: Графични обекти
- Цена: Платен
- Предимства: Добро използване на ресурсите
- Недостатъци: Не е много user friendly, когато се работи с много променливи



Weka



- Приложение: Статистика; data mining
- Начин на работа: Графични обекти
- Цена: Безплатен
- Предимства: Един от пионерите в областта
- Недостатъци: Демодирани и тромави



The Weka Explorer window shows the following classification results for the Iris dataset:

Test options	Classifier
Use training set	IBk -K 1 -W 0 -A "weka.core.neighboursearch.LinearNNSearch -A "weka.core.EuclideanDistance -R first-last"
Supplied test set	
Cross-validation Folds: 10	
Percentage split %: 66	

Classifier output:

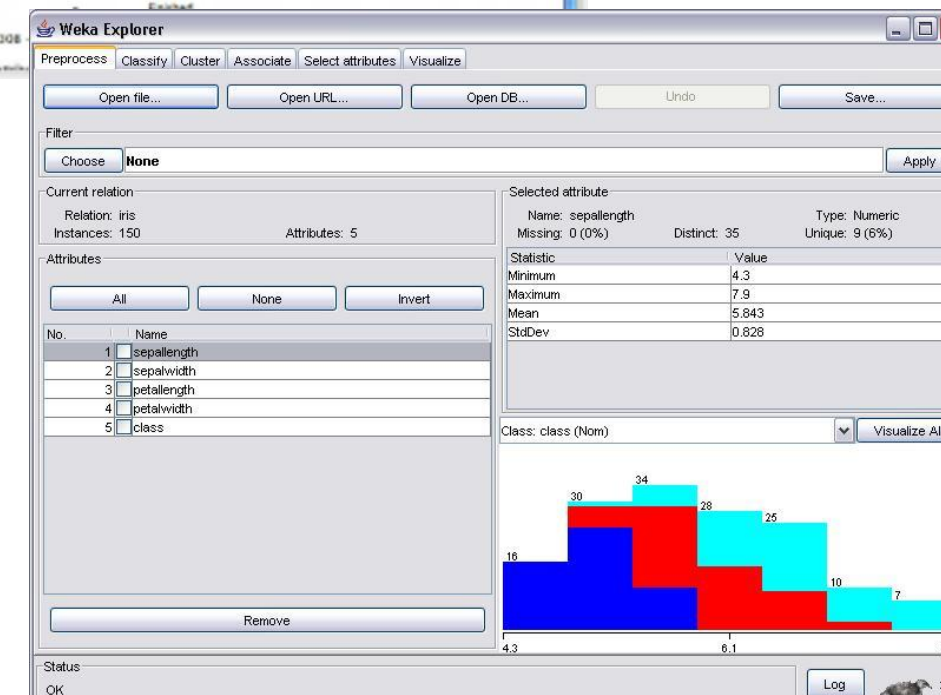
Correctly Classified Instances	2663	88.7667 %
Incorrectly Classified Instances	337	11.2333 %
Kappa statistic	0.7748	
Mean absolute error	0.1326	
Root mean squared error	0.2573	
Relative absolute error	26.522 %	
Root relative squared error	51.462 %	
Total Number of Instances	3000	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area
0	0.95	0.177	0.847	0.95	0.896	0.972
1	0.823	0.05	0.941	0.823	0.878	0.972
Weighted Avg.	0.888	0.114	0.893	0.888	0.887	0.972

=== Confusion Matrix ===

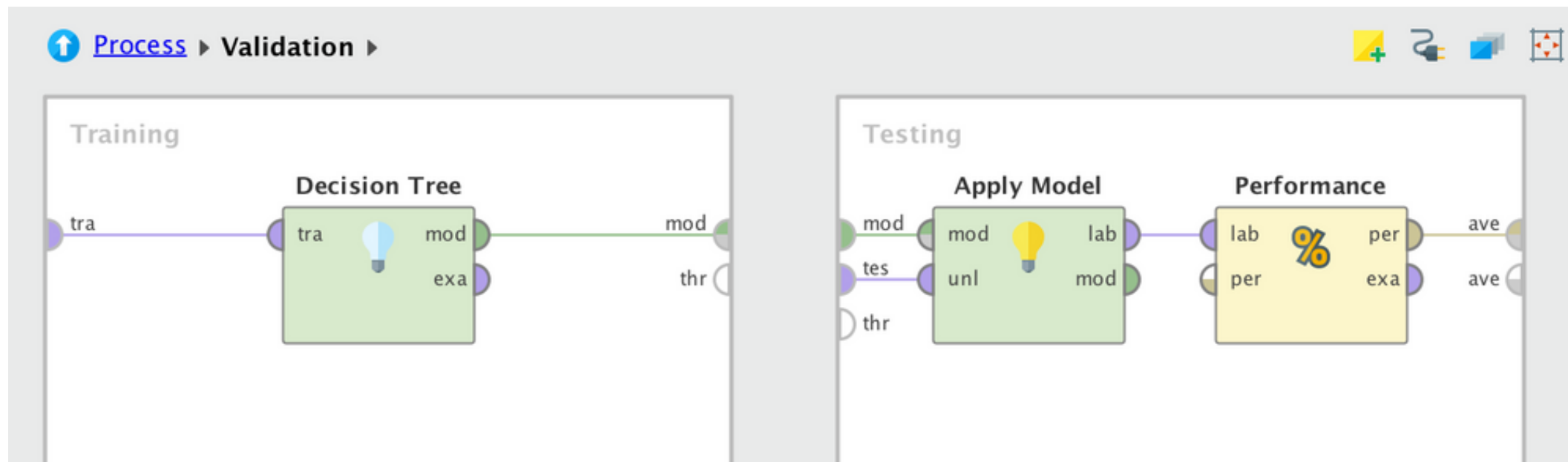
a	b	classified as	
1449	76	1	a = 1
261	1214	1	b = 0



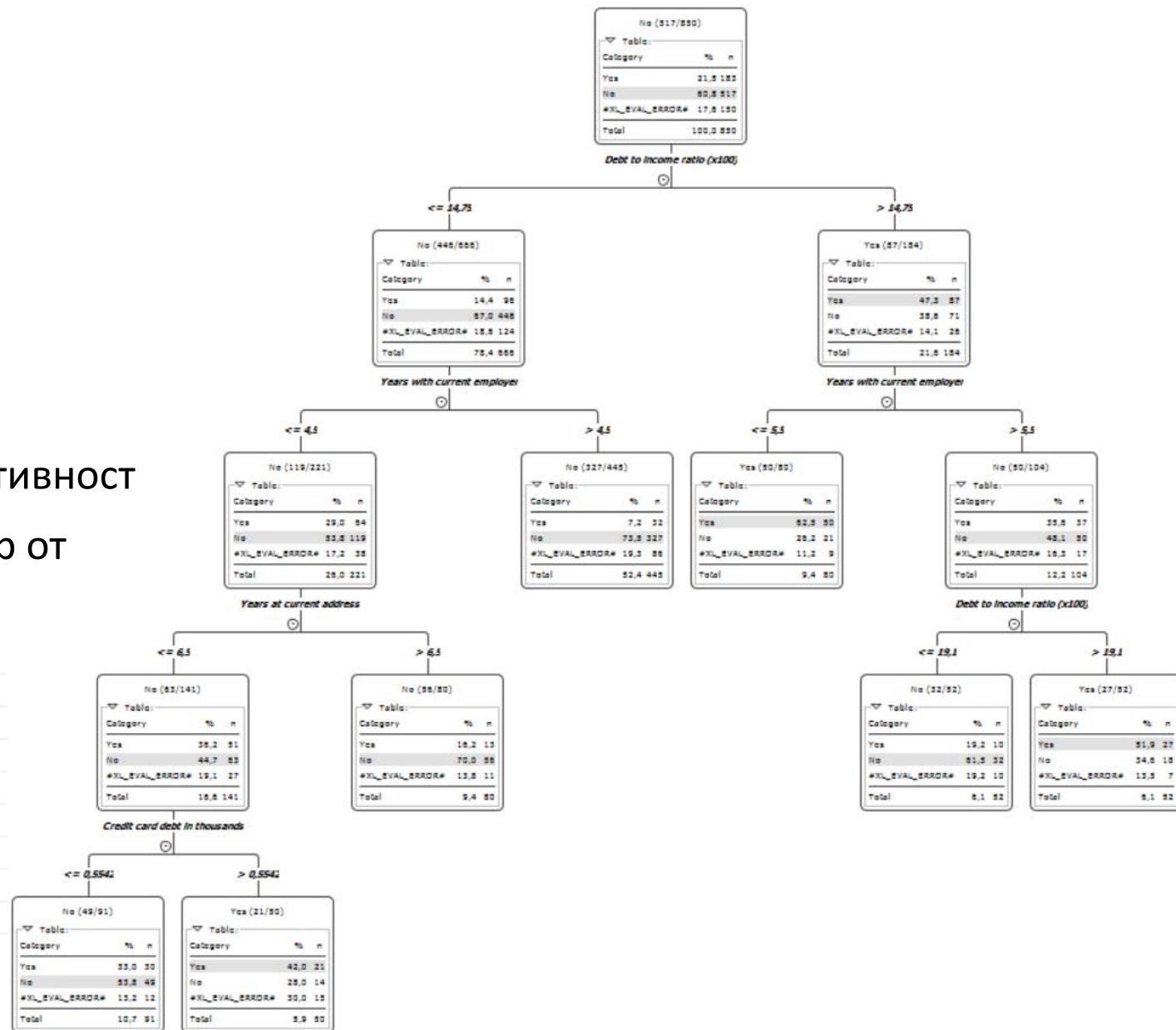
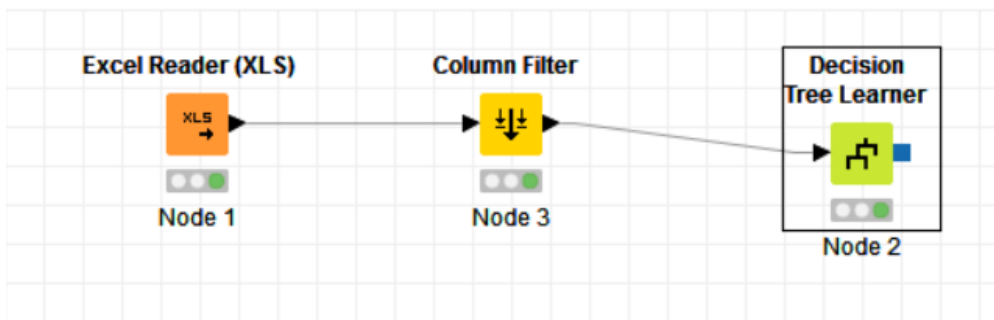
Rapid Miner



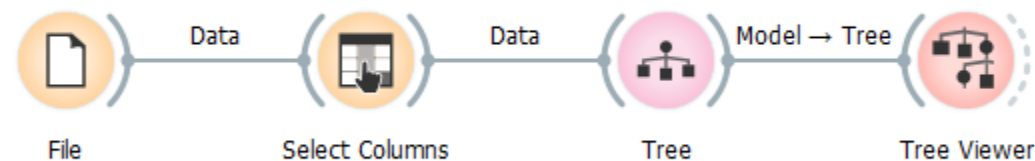
- Приложение: Статистика; data mining
- Начин на работа: Графични обекти
- Цена: Платен
- Предимства: Вероятно най-интуитивният интерфейс
- Недостатъци:



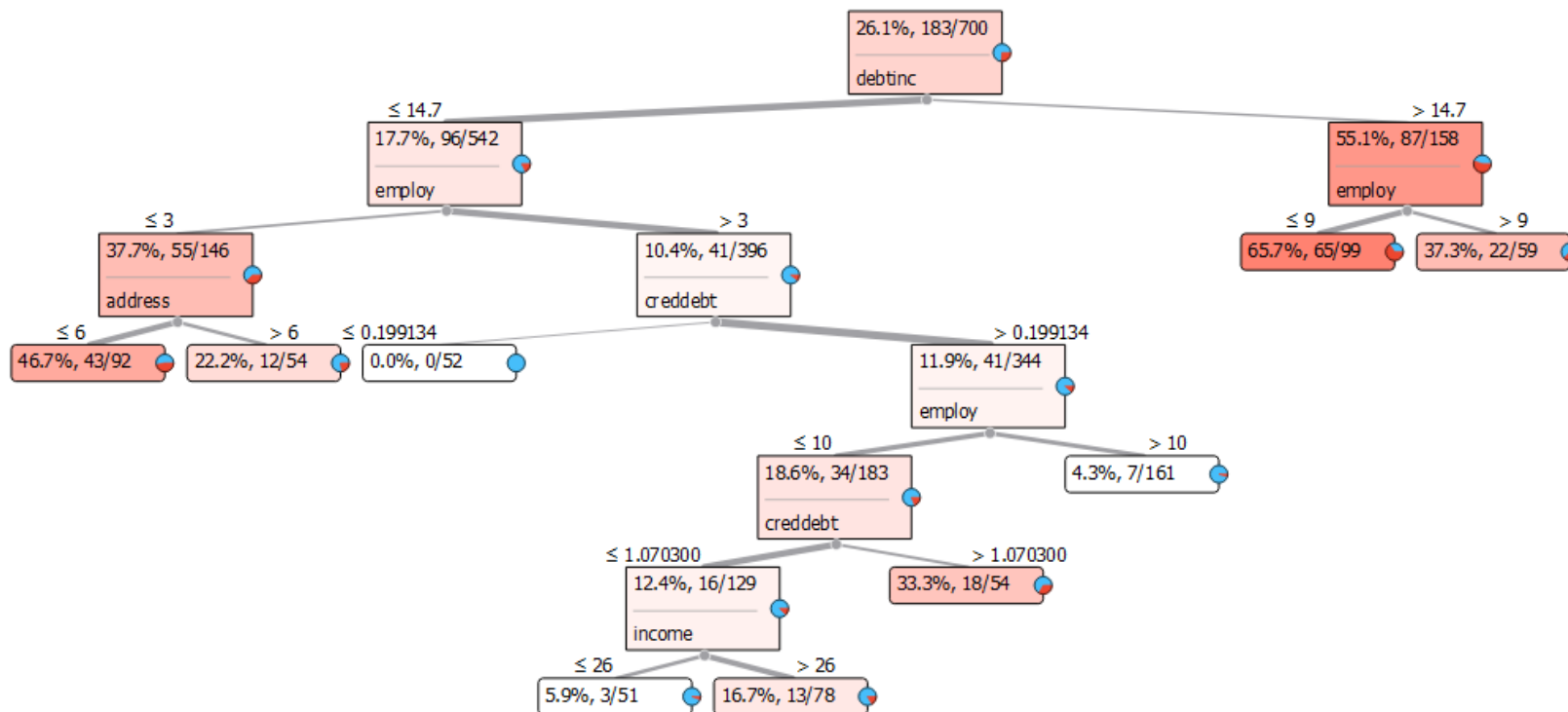
- Приложение: Статистика; data mining
- Начин на работа: Графични обекти
- Цена: Безплатен
- Предимства: Висока степен на интерактивност
- Недостатъци: Относително малък набор от възможни анализи



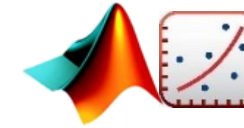
Orange



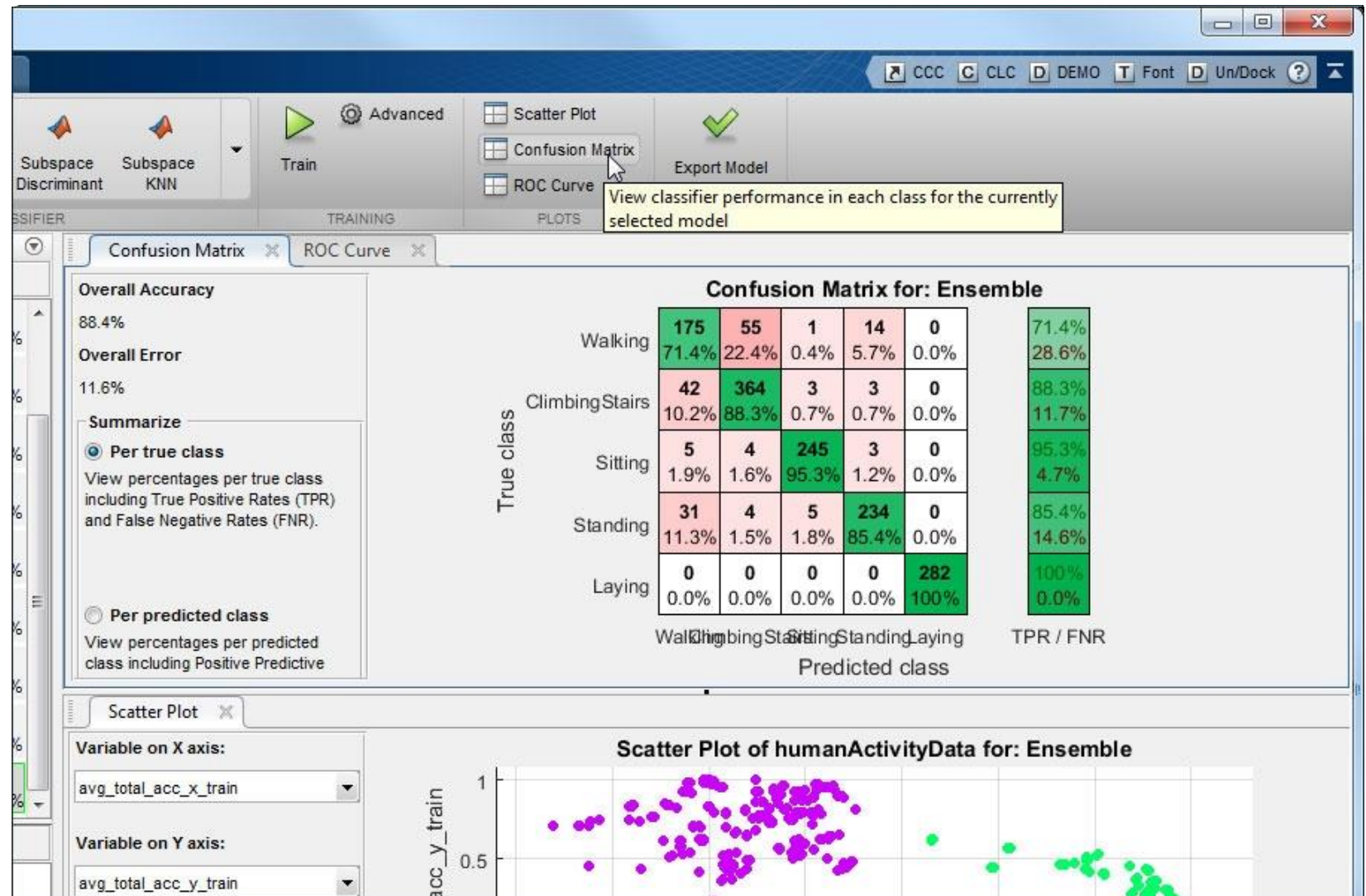
- Приложение: data mining
- Начин на работа: Графични обекти
- Цена: Безплатен
- Предимства: Висока степен на интерактивност
- Недостатъци: Относително малък набор от възможни анализи



MatLab Classification Learner



- Приложение: data mining
- Начин на работа:
Графични обекти
- Цена: Платен
- Предимства: Част от средата на Matlab
- Недостатъци: Все още се разработва, за да включи още модели

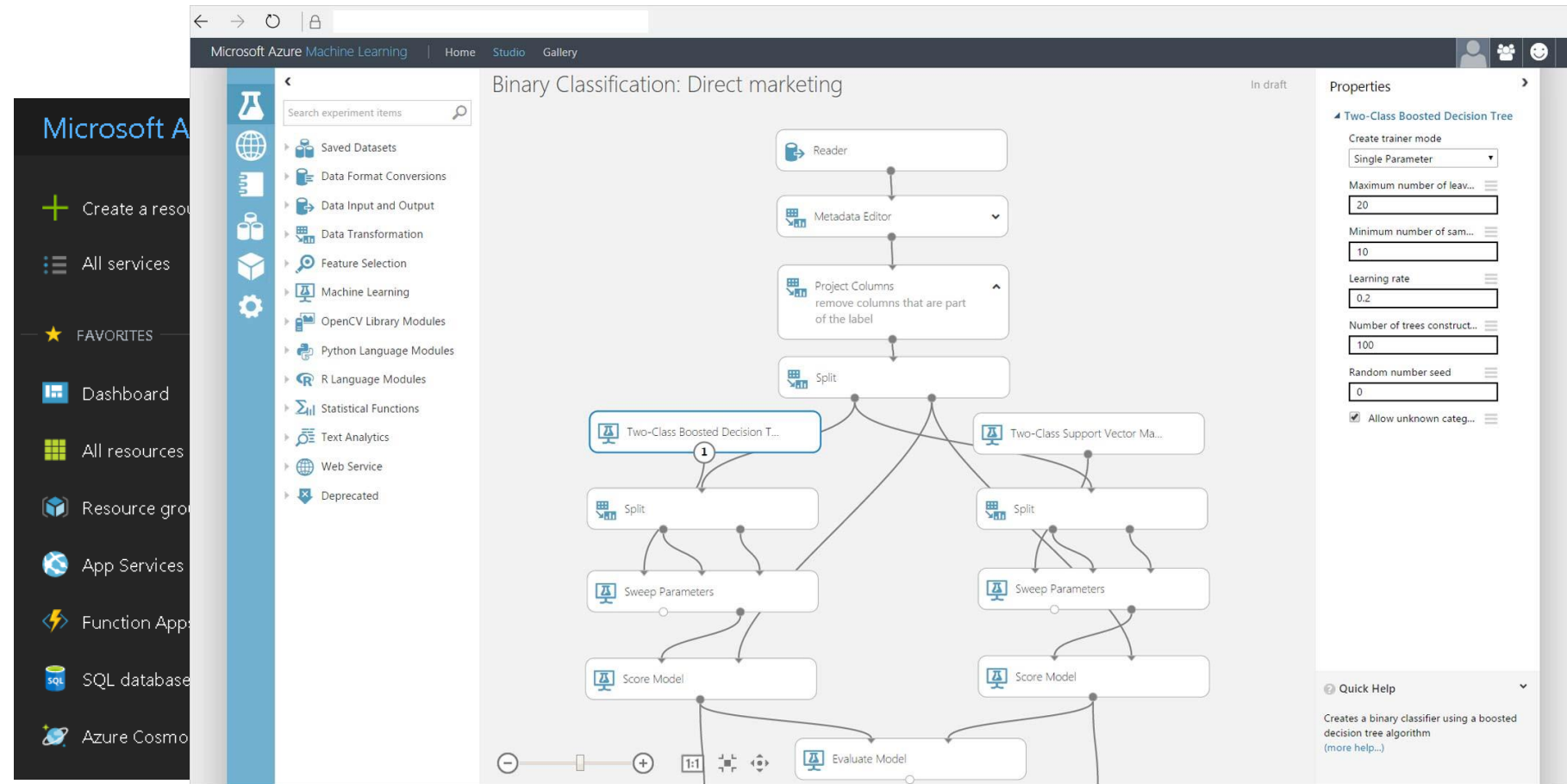


The On-liners

Microsoft Azure



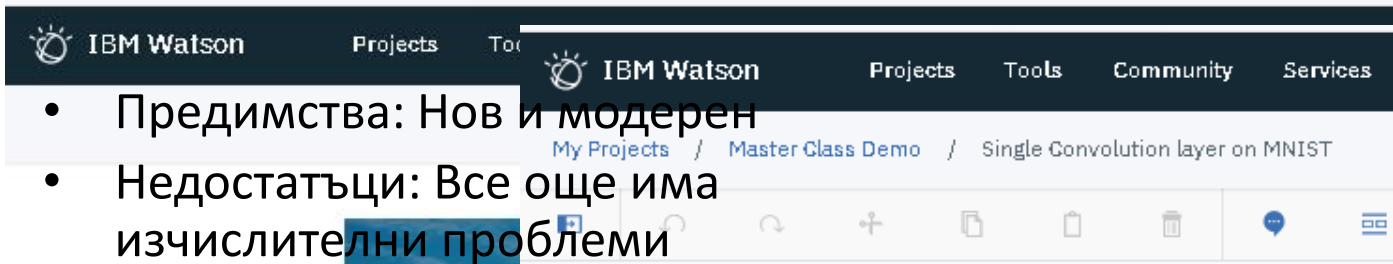
- Приложение: data mining
- Начин на работа: онлайн
- Цена: Платен
- Предимства: Вече са достъпни много инструменти
- Недостатъци: Може би е малко труден за инсталиране



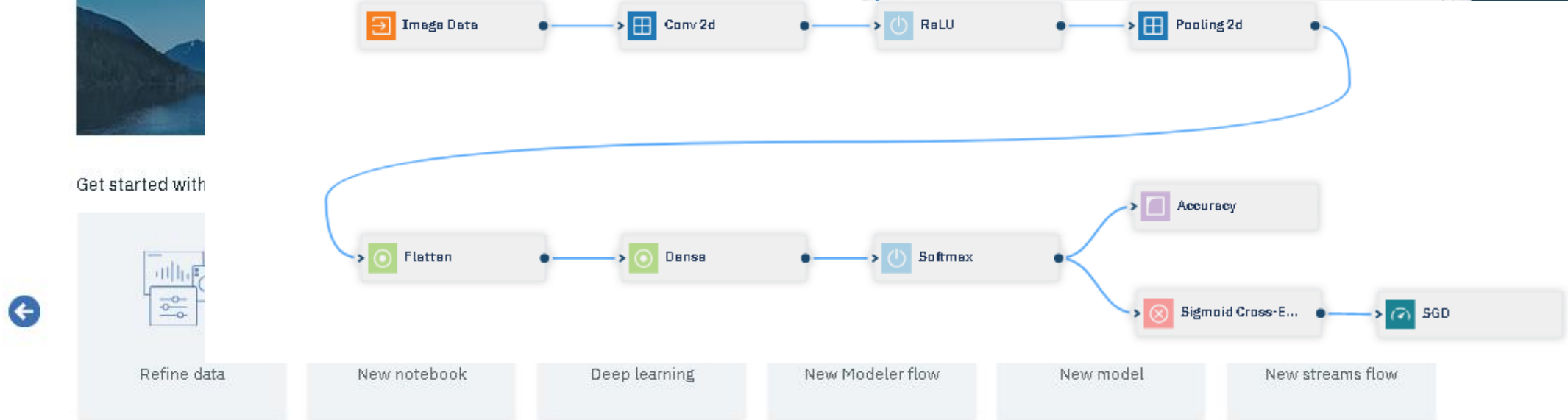
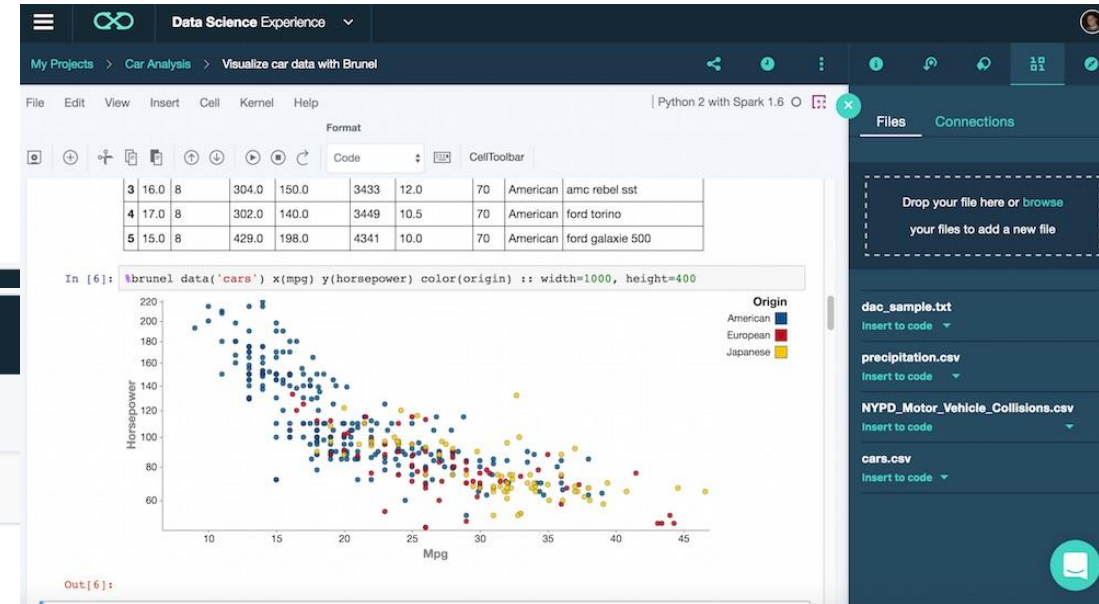
IBM Watson Studio



- Приложение: data mining
- Начин на работа: онлайн
- Цена: Платен



- Предимства: Нов и модерен
- Недостатъци: Все още има изчислителни проблеми





Amazon ML

- Приложение: data mining
- Начин на работа: онлайн
- Цена: Платен

- Предимства: Интегриран с AWS S3 и може да работи в реално време
- Недостатъци: Все още се разработва, за да включи още модели

ML Model Name : ML Model: Banking Data

Enable real time predictions

If your application requires predictions to be generated on demand, you can enable your model for real-time predictions at low latency.
[Read API documentation.](#)

- ML Model:
- ☒ Enable Real Time Predictions
 - ☐ Disable Real Time Predictions

Cancel

Confirm

ML

This whe

Adju num

E

of Records

AWS

Services

Edit

Amazon Machine Learning

Batch Predictions

Create batch prediction

1. ML model for batch prediction

2. Data for batch prediction

3. Batch prediction results

4. Review

Data for batch prediction

Locate the input data to use for the batch prediction. [Learn more about S3 permissions.](#)

Locate the input data

☐ I already created a datasource pointing to my S3 data

☒ My data is in S3, and I need to create a datasource

Datasource name

S3 location

Does the first line in your CSV contain the column names? ☒ Yes ☐ No

[Cancel](#) [Previous](#) [Verify](#)

Google Colab GCO

- Приложение: data mining
- Начин на работа: онлайн
- Цена: Безплатен
- Предимства: GPU изчисления чрез TensorFlow
- Недостатъци: 12-часови сесии

Colaboratory

Colaboratory is a research project created to help disseminate machine learning education and research. It's a Jupyter notebook environment that requires no setup to use. For more information, see our [FAQ](#).

[Sign in to request access to Colaboratory](#)

[Explore now](#)

My Drive > app ▾

The screenshot displays the Google Colaboratory web interface. At the top, there's a navigation bar with the 'co' logo, a 'Hello, Colaboratory' greeting with a star icon, and a menu with 'File', 'Edit', 'View', 'Insert', 'Runtime', 'Tools', and 'Help'. On the right, there are 'COMMENT' and 'SHARE' buttons. Below the navigation bar, a status bar shows 'TEXT', 'CODE', 'CELL', and 'CELL' tabs, along with a 'CONNECTED' status and an 'EDITING' mode indicator. The main content area features a 'Welcome to Colaboratory!' message, explaining that it's a data analysis tool combining text, code, and outputs. It shows a simple code cell with `print 'Hello, Colaboratory!'` and its output. Below this, it explains how to execute TensorFlow code and provides an example of adding two matrices, showing the mathematical representation and the corresponding Python code using TensorFlow and NumPy. The code cell is highlighted, and its output is shown at the bottom. At the very bottom, it mentions that Colaboratory includes libraries like [matplotlib](#) for visualization.

co Hello, Colaboratory ☆
File Edit View Insert Runtime Tools Help

TEXT CODE CELL CELL CONNECTED EDITING

Welcome to Colaboratory!

Colaboratory is a data analysis tool that combines text, code, and code outputs into a single collaborative document.

```
[ ] print 'Hello, Colaboratory!'
```

Hello, Colaboratory!

Colaboratory allows you to execute TensorFlow code in your browser with a single click. The example below adds two matrices.

$$\begin{bmatrix} 1. & 1. & 1. \\ 1. & 1. & 1. \end{bmatrix} + \begin{bmatrix} 1. & 2. & 3. \\ 4. & 5. & 6. \end{bmatrix} = \begin{bmatrix} 2. & 3. & 4. \\ 5. & 6. & 7. \end{bmatrix}$$

```
import tensorflow as tf
import numpy as np

with tf.Session():
    input1 = tf.constant(1.0, shape=[2, 3])
    input2 = tf.constant(np.reshape(np.arange(1.0, 7.0, dtype=np.float32), (2, 3)))
    output = tf.add(input1, input2)
    result = output.eval()
    print result
```

```
[[ 2.  3.  4.]
 [ 5.  6.  7.]]
```

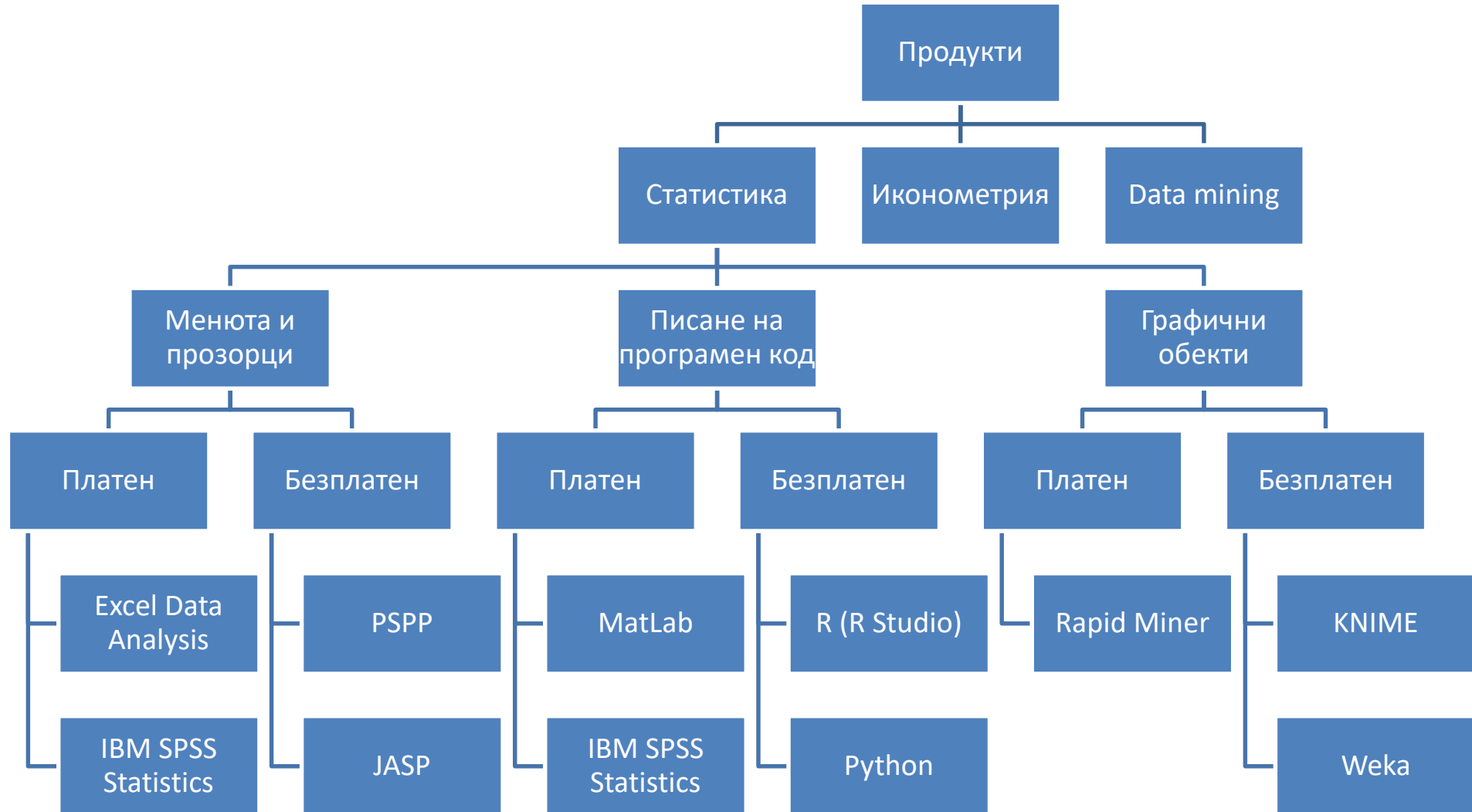
Colaboratory includes widely used libraries like [matplotlib](#), simplifying visualization

Критерии за избор на точния продукт

- Какъв проблем решавате? (Приложение)
- Как искате да работите? (Начин на работа)
- Колко сте склонни да платите? (Цена)

Приложение	Начин на работа	Цена	Продукт		
Статистика	Менюта и прозорци	Платен	Excel Data Analysis	IBM SPSS Statistics	
		Безплатен	PSPP	JASP	
	Писане на програмен код	Платен	MatLab	IBM SPSS Statistics	
		Безплатен	R (+ R Studio)	Python	
	Графични обекти	Платен	Rapid Miner		
		Безплатен	KNIME	Weka	
Иконометрия	Менюта и прозорци	Платен	eViews	IBM SPSS Statistics	
		Безплатен	Gretl		
	Писане на програмен код	Платен	eViews	IBM SPSS Statistics	MatLab
		Безплатен	Gretl	R (+ R Studio)	Python
	Графични обекти	Платен	IBM SPSS Modeler		
Data mining	Писане на програмен код	Платен	Matlab		
		Безплатен	R (R Studio)	Python	
	Графични обекти	Платен	IBM SPSS Modeler	Rapid Miner	Matlab Classification App
		Безплатен	Orange	KNIME	Weka
	Онлайн	Платен	IBM Watson Studio	Microsoft Azure	Amazon ML
		Безплатен	Google Colab	Jupyter Notebook	

Дърво на решенията



Дърво на решенията

