

PENAMBANGAN DATA FORMAT TEXT EXCEL DENGAN SOFTWARE TANAGRA

Oleh :
Julsam¹⁾, Handryawan Adnan Mooduto¹⁾, Alexyusandria²⁾

¹⁾Staf pengajar jurusan Elektro Politeknik Negeri Padang

²⁾Staf pengajar Institut Teknologi Padang

ABSTRACT

This article describes about Tanagra software application on data mining. Tanagra is data mining software which can be used to access some existing data mining method. Data is built using excel with text data type. This application use the dataset of women with benign or malignant. The result inform from 699 sample for Univariate Continuous Statistic have begin 65,52% and malignant 34,48%. The other information is the benign has less mitos (1,6 : 1,59), the malignant has more ucellshape (6,56 : 3,21)

Keyword: *tanagra software, data mining, benign, malignant*

PENDAHULUAN

Dalam melaksanakan penelitian, peneliti muda ataupun peneliti yang akan melakukan pengolahan data sering terbentur pada proses pengolahan data secara statistik. Banyak software yang dapat digunakan, masing-masing dengan kelebihan dan kekurangannya memberikan kemudahan pada peneliti untuk melaksanakan pengolahan data. Yang terpenting adalah bagaimana melakukan penambangan informasi dari data yang diperoleh (*data mining*).

Tanagra adalah sebuah *software data mining* yang dibangun dengan tujuan untuk memberikan akses kepada sejumlah metode *data mining* yang ada. Tanagra tidak mempersoalkan format bilangan dari file data set yang diimport, dan juga tidak berurusan dengan perbaikan data yang hilang. Dalam tulisan ini akan dijelaskan bagaimana langkah Tanagra dalam mengimport data yang dibangun dalam format spreadsheet

Excel kedalam Tanagra untuk diproses. Dengan Excel pembangunan data dapat dilakukan dengan mudah namun disimpan dengan format text.

Rumusan Masalah

Dari penjelasan diatas dapat dirumuskan permasalahan sebagai berikut :

1. Bagaimana memanfaatkan software Tanagra dalam penambangan data.
2. Bagaimana memproses data yang dibangun dengan Excel.

Tujuan

Adapun yang menjadi tujuan dari tulisan ini adalah sebagai berikut :

1. Menjelaskan penggunaan Tanagra dalam penambangan data
2. Menjelaskan kerja Tanagra dalam memproses data yang dibuat dengan Excel

Membangun Data dengan Excel

Pembangunan data dengan Excel dilakukan tidak beda dengan membangun data yang selama ini dilakukan oleh pengguna Excel. Seperti diperlihatkan pada gambar 1, data yang dibangun dalam Excel adalah data tentang cuaca. Setelah data terbentuk

langkah berikutnya adalah melakukan penyimpanan data. Penyimpanan data dilakukan seperti halnya menyimpan data Excel lainnya, namun dalam tipe text (*Tab delimited*) dengan meng-klik *File/Save As*, selanjutnya set tipe data seperti dicontohkan pada gambar 2,

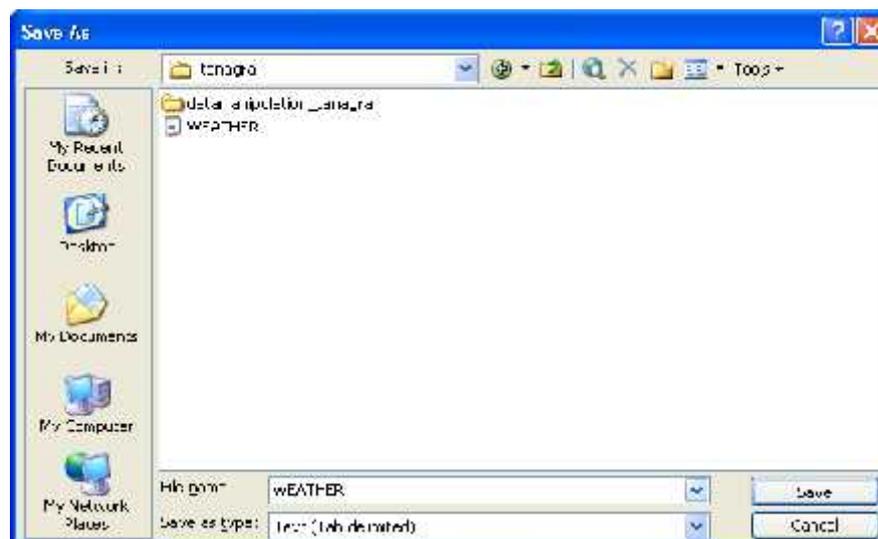


	A	B	C	D	E
1	Outlook	temperature	Humidity	Windy	Class
2	sunny	75	70	yes	Play
3	sunny	70	90	yes	Don't Play
4	sunny	85	85	no	Don't Play
5	sunny	72	85	no	Don't Play
6	sunny	69	70	no	Play
7	overcast	72	90	yes	Play
8	overcast	83	78	no	Play
9	overcast	64	85	yes	Play
10	overcast	81	75	no	Play
11	rain	71	70	yes	Don't Play
12	rain	65	70	yes	Don't Play
13	rain	76	80	no	Play
14	rain	80	70	no	Play
15	rain	70	90	no	Play

Gambar 1. Contoh Data Dibangun Dengan Excel

Setelah memasukkan nama file (contoh : Weather) dan memilih tipe filenya, diteruskan dengan mengeksekusinya dengan menekan tombol *save*, setelah itu data siap untuk

di import kedalam Tanagra. File yang dibuat ini dapat dilihat kembali isinya dengan editor lain seperti Windows notepad dan editor lainnya.



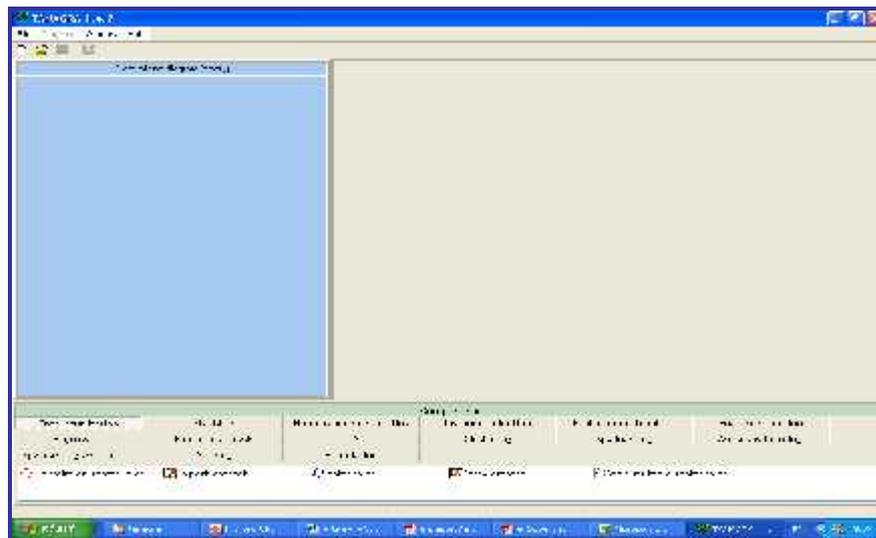
Gambar 2. Penyimpanan File Text

Penggunaan Tanagra

1. Meng-Import Data dari Text File Excel

Setelah file data set dibentuk dalam format text Excel, untuk

selanjutnya proses penambangan data dapat dilakukan menggunakan software Tanagra. Untuk itu software Tanagra di aktifkan sehingga User Interface Tanagra seperti gambar 3 ditampilkan

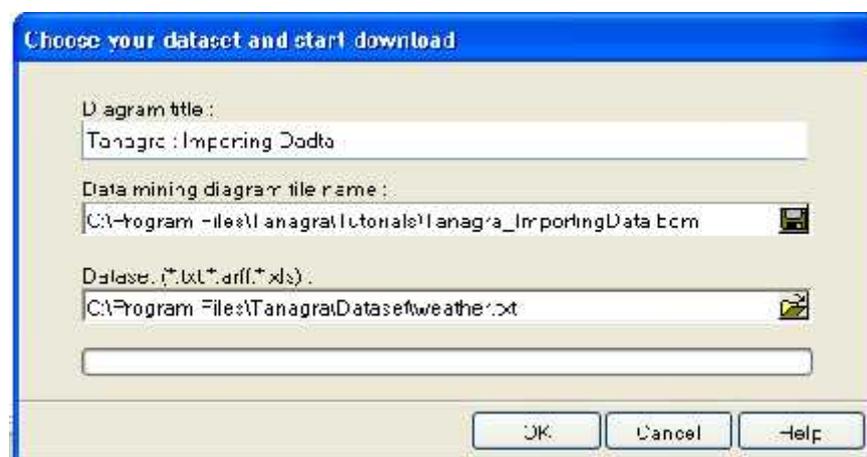


Gambar 3. User Interface TANAGRA

Pada bagian bawah UI Tanagra disediakan komponen-komponen yang dapat digunakan untuk melakukan proses penambangan data, sebagian diantaranya akan dijelaskan pada tulisan ini.

Untuk meng-import data set text Excel buka lembaran baru diagram data mining dengan meng-klik menu

File/New pada menu utama selanjutnya dialog box pemilihan dataset seperti gambar 4 akan ditampilkan. Ada 3 (tiga) informasi yang harus diisikan kedalam dialig box ini yaitu: Diagram Title, Data mining diagram file name dan Dataset. Dalam gambar 4 diberikan contoh pengisian informasi ini.



Gambar 4. Setting Dokumen

Setelah setting selesai, tekan OK untuk memastikan setting yang dibuat dan selanjutnya pada *User Interface* Tanagra ditampilkan sebuah diagram data mining baru berdasarkan pada file < weather.txt >. Pada bagian kanan *User Interface* tanagra diperlihatkan informasi tentang download file yang dilakukan dan data set description, seperti gambar 5. Ketika meng-create diagram baru, disini kita menggunakan extension

(.BDM) untuk nama file diagram (< tanagra_ImportingData.BDM >). BDM (Binary Diagram) adalah extension untuk file Tanagra yang mengandung deskripsi diagram, juga data set yang di import (dalam hal ini data dari file < weather.txt >). Dengan demikian kita akan kehilangan asosiasi (hubungan) ke filetext data set. Modifikasi ataupun penghapusan data filetext data set tidak mempengaruhi kerja dalam Tanagra.



Gambar 5. Diagram yang di create berdasarkan file weather.txt

Format file diagram Tanagra lainnya adalah dengan extension TDM (Text Diagram). Dengan TDM, selain menyimpan deskripsi diagram kita hanya menyimpan referensi file data set. Sehingga jika dilakukan perubahan pada file data set (< weather.txt >) akan terjadi perubahan hasil saat dilakukan eksekusi pada urutan operasi Tanagra berikutnya. Jadi jangan menghapus file data asal jika ingin membuka kembali file diagram Tanagra, karena akan diperoleh hasil yang berbeda.

2. Pengolahan untuk *Basic Statistic*

Pada bagian ini akan dicontohkan karakteristik dasar dari *User Interface* Tanagra melalui analisis file data set *breast.txt*. Data set ini berasal dari *medical domain* yang terdiri dari karakteristik sel yang disampel dari sel yang mengandung tumor ganas atau tidak. Contoh ini menjelaskan penggunaan dari beberapa komponen seperti pada table 1. Bagian data yang diolah dengan Tanagra diperlihatkan pada tabel 2.

Tabel 1. Komponen yang digunakan

Tab	Operator (Component)	Function
Data visualization	View dataset	Memperlihatkan isi data file dalag grid
Feature selection	Define status	Spesifikasi penggunaan atribut
Descriptive statistic	Univariate continuous statistic	Descriptive statistic untuk continuous attribute
Descriptive statistic	Univariate discrete statistic	Descriptive statistic untuk discrete attribute
Descriptive statistic	Group characterization	Statistik untuk sub-populasi

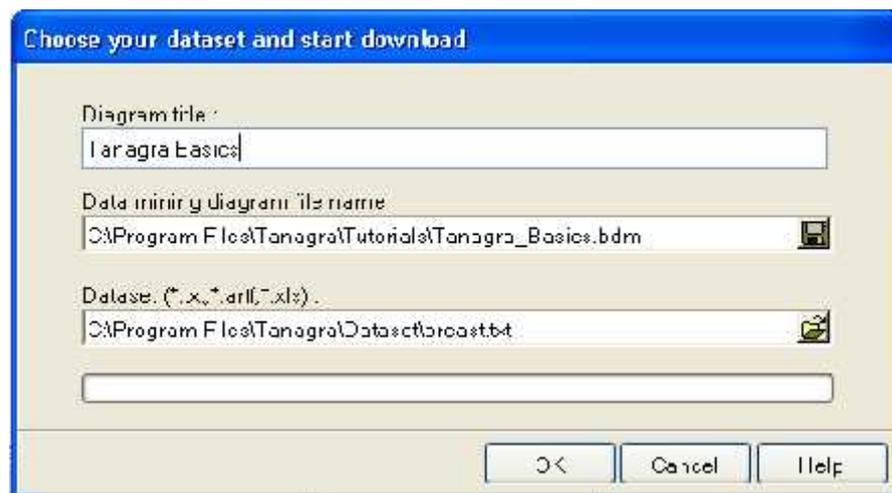
Tabel 2. Data dari file *breast.txt*

Cellshape	ucellsize		ucellshape		mgadhesion		sepics	b nuclei	bchromatin
4	2	2	1	2	1	2	1	1	begin
1	1	1	1	2	1	2	1	1	begin
2	1	1	1	2	1	2	1	1	malignant
10	6	6	2	4	10	9	7	1	begin
4	1	1	1	2	1	2	1	1	begin
1	1	1	1	2	1	1	1	1	begin
1	1	1	1	2	1	2	1	1	begin
5	1	1	1	2	1	2	1	1	begin
3	1	1	1	2	1	2	1	1	begin
1	1	1	1	2	4	2	1	1	begin
5	3	3	2	3	1	3	1	1	begin
4	2	2	1	2	1	2	1	1	begin
1	1	1	1	2	1	2	1	1	begin
.
.
.

Dan seterusnya ada sebanyak 699 sampel data

Analisis ini diawali dengan meng-import file data dengan cara seperti yang telah dijelaskan pada bagian diatas dengan setting document seperti

diperlihatkan pada gambar 5, dilanjutkan dengan menekan tombol OK untuk penetapan.



Gambar 6. Setting dokumen untuk analisis file breast.txt

Selanjutnya dimulai melakukan analisis, dengan tahapan sebagai berikut:

- Tambahkan komponen *View Dataset* (pada *component palette* “*DATA VISUALIZATION*”, ke diagram dibawah *node Dataset(Breast.txt)*.
- Klik *View Dataset node* untuk memilihnya, selanjutnya klik kanan *View Dataset* tersebut untuk menampilkan popup menu: pilih *View command*, data akan ditampilkan dibagian kanan *frame*, seperti pada gambar 6.

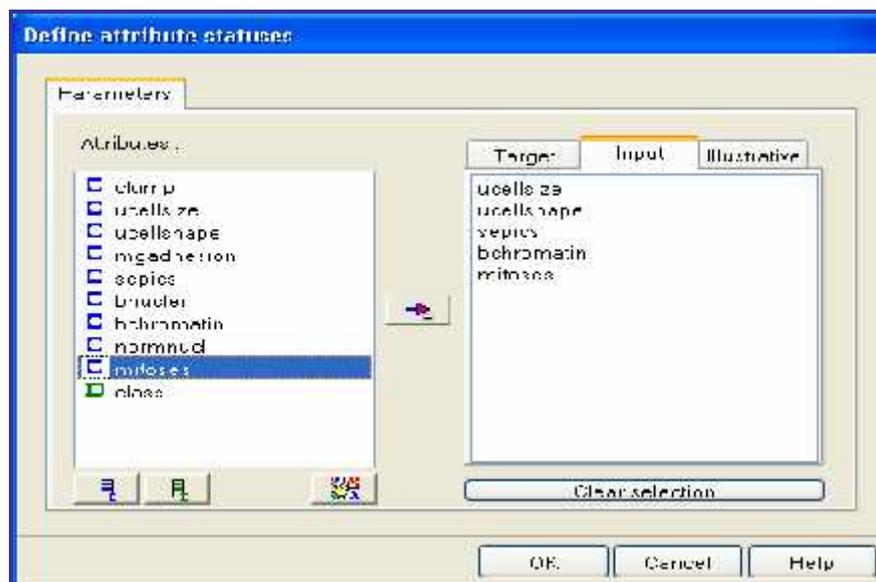
Dalam Tanagra dapat dibangun daftar urutan operasi, namun hampir semua operator memerlukan pen-definisian

kegunaan atribut dan bagaimana penggunaannya sebelum dieksekusi. Komponen Define status ada didalam *component palette FEATURE SELECTION* tab. Langkah pendefinisian adalah sebagai berikut:

- Tempatkan komponen Define status dibawah *node Dataset(Breast.txt)*, click kanan, pilih parameter ...Command.
- Didalam dialog box terlihat sejumlah pilihan continuous variables, klik pilihan yang disebelah kiri kemudian klik tombol panah. List yang diklik akan berfungsi sebagai input, karena tab input sedang aktif (gambar 6). Klik OK

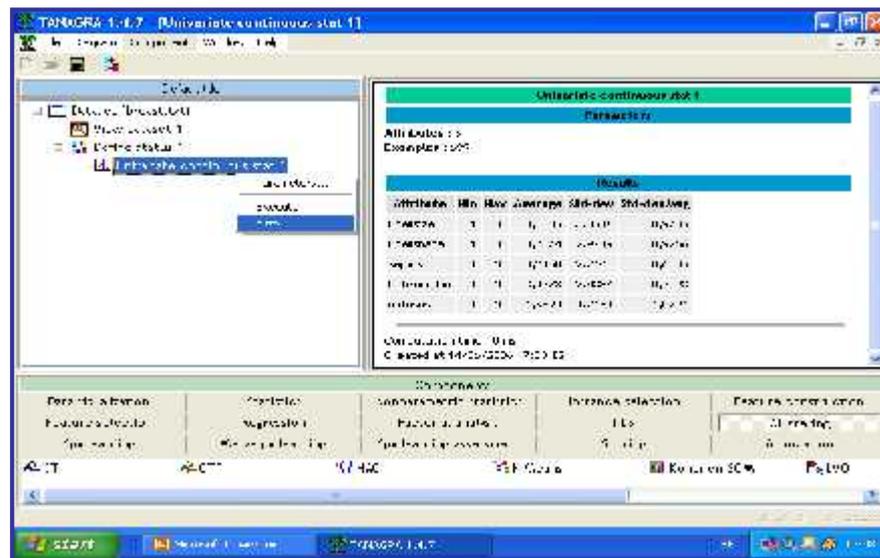


Gambar 7. Tampilan data yang diimport dari file breast.txt



Gambar 8. Dialog box penetapan parameter

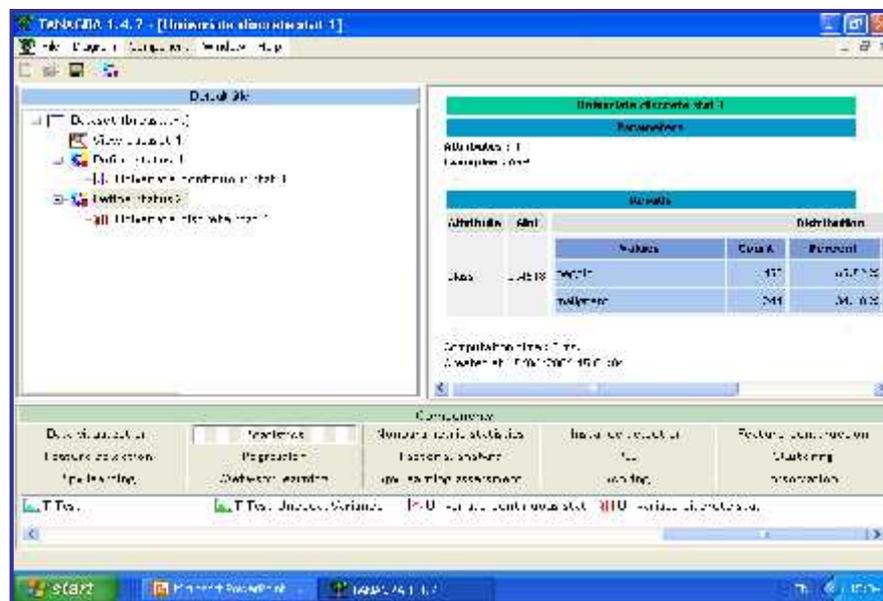
- Tambahkan komponen Univariate continuous statistic (DESCRIPTIVE STATISTIC tab) ke diagram dibawah Define status 1. Pada popup menu pilih View command. Descriptive statistic untuk atribut yang dipilih akan ditampilkan di frame kanan seperti pada gambar 8.



Gambar 9. Hasil Univariate Continuous Statistic

- Tambahkan Define status 2, seperti sebelumnya. Pilih discrete attribute (class). Tambahkan Univariate discrete statistic dibawah define status 2. Pilih View command seperti sebelumnya. Hasilnya seperti gambar

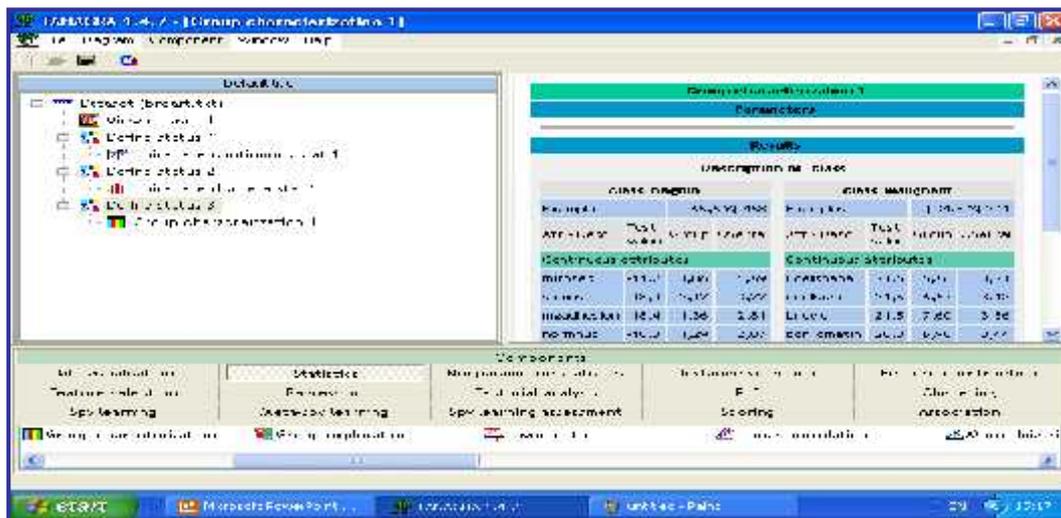
9. Diperlihatkan jumlah dan persentase sampel yang menderita tumor jinak (benign) atau ganas (malignant) sebesar 458 dan 241 orang (65,52% dan 34,48%).



Gambar 10. Hasil Univariate Discrete Statistic

Untuk statistic sub-populasi (karakteristik wanita mengandung tumor ganas atau tidak), tambahkan define status 3 seperti sebelumnya.

Tempatkan Group characterization kebawahnya. Pilih View pada popup menu. Hasilnya seperti gambar 9.



Gambar 9. Hasil Group Characterization Hasil ini memperlihatkan bahwa rata-rata wanita dengan tumor jinak (benign) mempersembahkan nilai mitosis yang lebih kecil (1,6 : 1,59 untuk populasi lengkap). Disisi lain, nilai ucellshape attribute masih pada rata-rata, lebih tinggi untuk wanita dengan tumor ganas (6,56 : 3,21).

KESIMPULAN

Dari uraian yang diberikan diatas dapat diambil kesimpulan sebagai berikut :

1. Tanagra merupakan software data mining yang dapat mengolah data
2. Contoh aplikasi Tanagra ini menggunakan data pasien yang tidak diperlihatkan informasi-informasi yang dapat diperoleh.
3. *Univariate Continuous Statistic* menunjukkan hasil benign 65,52% dan malignant 34,48%
4. Informasi lain yang diperoleh adalah karakteristik grup dari penderita tumor jinak memberikan nilai motosis yang lebih kecil (1,6 : 1,59), serta nilai ucellshape tumor ganas lebih tinggi (6,56 : 3,21)

Saran

1. Untuk selanjutnya data yang digunakan dapat disesuaikan dengan kebutuhan peneliti.
2. Data yang ditambah merupakan data-data dalam jumlah yang besar.

dalam format text yang dibentuk dengan Excel.
mengandung penyakit tumor ganas atau

Daftar Pustaka

1. <http://en.wikipedia.org/wiki/tanagra>, *Tanagra*
2. <http://eric.univ-Iyon2.fr/-ricco/tanagra/en/tanagra.html>, *Tanagra*, Tutorial Data Manipulation
3. <http://data-mining-tutorials.blogspot.com/Tanagra-Data-Mining-Tutorials>
4. <http://eric.univ-Iyon2.fr/-ricco/tanagra/fichiers/enBasics.pdf>, *Tutorial overview*

