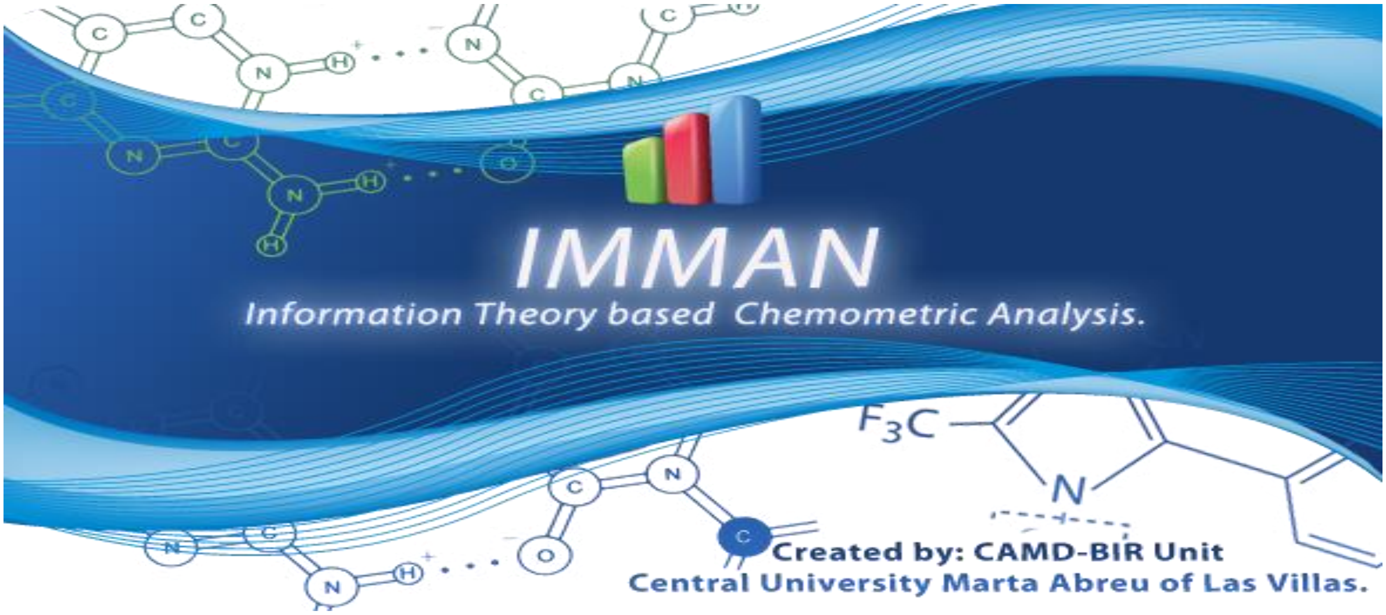
IMMAN

Information Theory based Chemometric Analysis

IMMAN v1.0


IMMAN
Information Theory based Chemometric Analysis.
Created by: CAMD-BIR Unit
Central University Marta Abreu of Las Villas.

Software User Manual

IMMAN is a user-friendly program designed to perform unsupervised and supervised feature selection tasks using information theory. In addition, this program allows the graphic analysis of the information theoretic parameters, computed for different datasets, using diverse statistical representations.

USER'S MANUAL

IMMAN 1.0

IMMAN 1.0 is a feature selection program that uses information theoretic methods.

CEI and CAMD-BIR Unit
Department of Computer Sciences and Faculty of Chemistry-Pharmacy
Universidad Central "Marta Abreu" de Las Villas (UCLV)
Santa Clara, Cuba

October, **2014**

USER'S MANUAL

TABLE OF CONTENTS

Page #

INDEX OF FIGURES

1.0 GENERAL INFORMATION

1.0 GENERAL INFORMATION

1.1 System Overview

IMMAN is an interactive and user-friendly *free* multi-platform program designed to perform supervised and unsupervised feature selection tasks and to compare datasets, using information theoretic measures. The program constitutes seven information theoretic parameters for supervised feature selection, namely: Differential Shannon's Entropy (DSE), Standardized Differential Shannon's Entropy (sDSE), Jeffrey's Information (JI), Mutual Information for Differential Shannon's Entropy (MI-DSE), Information Gain (IG), Gain Ratio (GR) and Symmetrical Uncertainty (SU). On the other hand unsupervised feature selection relies on 10 measures, divided in two main groups: 1) discretization-based measures, viz.: Shannon's Entropy (SE), Standardized Shannon's Entropy (sDSE), Brillouin Redundancy Index (rSE), Gini Index (gSE) and Information Energy Content (iSE) 2) Non discretization-based measures, namely: Singular Value Decomposition (SVDE), Degenerative Entropy Raid (DGSE), Degenerated Value (DV) and Distance based Entropy (EDSE). These measures could be used independently or following a multi-criteria approach based fused numeric scores or positions of variables, using different statistical operators such as the average, sum, product or the geometric mean. In addition, this program allows the graphic analysis of the information theoretic parameters, computed for different datasets, using different representations such as the correlation, distribution and importance graphs.

1.2 System requirements

IMMAN Software runs on a wide variety of operating systems and computers including multi-processor clusters, multi-processor or multi-core desktops (PC and MAC), high-performance scientific workstations, and laptops. This release can run either interactively or in batch mode, which permits sequential execution to be distributed across multiple processors (and/or cores) workstations, even in a heterogeneous computing environment. In general terms the minimal and recommended system requirements are:

Hardware:

Processor: All processors developed hereafter by Intel Corp. are supported on the assembly level optimization. All AMD current processors work as old Pentium with higher clock frequency (no special optimization).

Processor Clock Speed: minimum Intel(R) Celeron(R) M processor 1.40GHz or equivalent. Recommended Intel(R) Core2Quad processor 2.5GHz or above.

Memory: 256MB minimum, 512MB default tuning. We recommend 4096 MB or above in order to improve performance.

Software:

Operation system: **IMMAN** is designed to run on any UNIX/LINUX or MAC platforms, as well as on microcomputers running Windows 95, 98, ME, 2000 or XP, Vista, 7 and above. **IMMAN** is platform-independent software.

Operation system extensions: **IMMAN** requires Java(TM) 7 Runtime Environment on the target system. It runs under any host operating system, which supports Java(TM) 7 Runtime Environment and also works on X86 and X64 based architecture.

1.3 Points of Contact

1.3.1 Information

For all comments, suggestions, information, and inquiries about **ToMoCoMD-CARDD QuBiLS-MAS** Software please contact:

Prof. Yovani Marrero Ponce, PhD
Enviromental and Computational Chemistry Group,
Facultad de Química Farmacéutica,
Universidad de Cartagena,
Cartagena de Indias, Bolivar,
Colombia.

CAMD-BIR Unit Group
Faculty of Chemistry-Pharmacy
Central University of Las Villas
Santa Clara, Villa Clara
Cuba.
E-mail: ymarrero77@yahoo.es;
ymponce@gmail.com

Prof. Stephen J. Barigye, PhD
Faculty of Chemistry
Federal University of Lavras
37200 Lavras
Minas Gerais
Brazil
Faculty of Chemistry-Pharmacy
Universidad Central "Marta Abreu" de Las Villas
Santa Clara, Villa Clara
Cuba.
E-mail: stvnjns.barigye@gmail.com

1.3.2 Technical Support

Prof. Ricardo Wilfredo Pino Urias, Eng.
CAMD-BIR Unit Group
Researcher at Artificial Intelligence Laboratory, Center of Informatics Studies
Department of Computer Sciences,
Faculty of Mathematics Physics and Computation
Universidad Central "Marta Abreu" de Las Villas, Santa Clara, Villa Clara
Cuba
Work Phone: +53 42281063
E-mail: rwpino@uclv.cu
rwpu1989@gmail.com

2.0 SYSTEM SUMMARY

2.0 SYSTEM SUMMARY

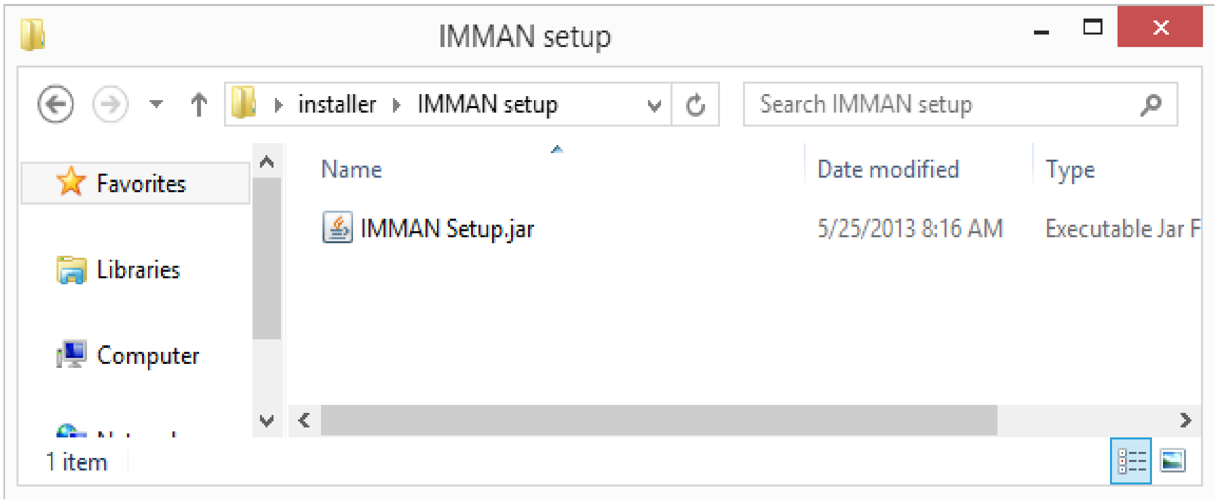
2.1 System Configuration

The system is prepared to maintain its default configuration regardless of the platform on which is executed. It does not require any parameters or initial configuration file, so it fits natively over the Java™ virtual machine.

2.2 Installation of the program

The Java™ RE version 7 are required on the target operating system, in a Microsoft Windows platform (i.e. XP, Vista, Windows 7) follows these steps:

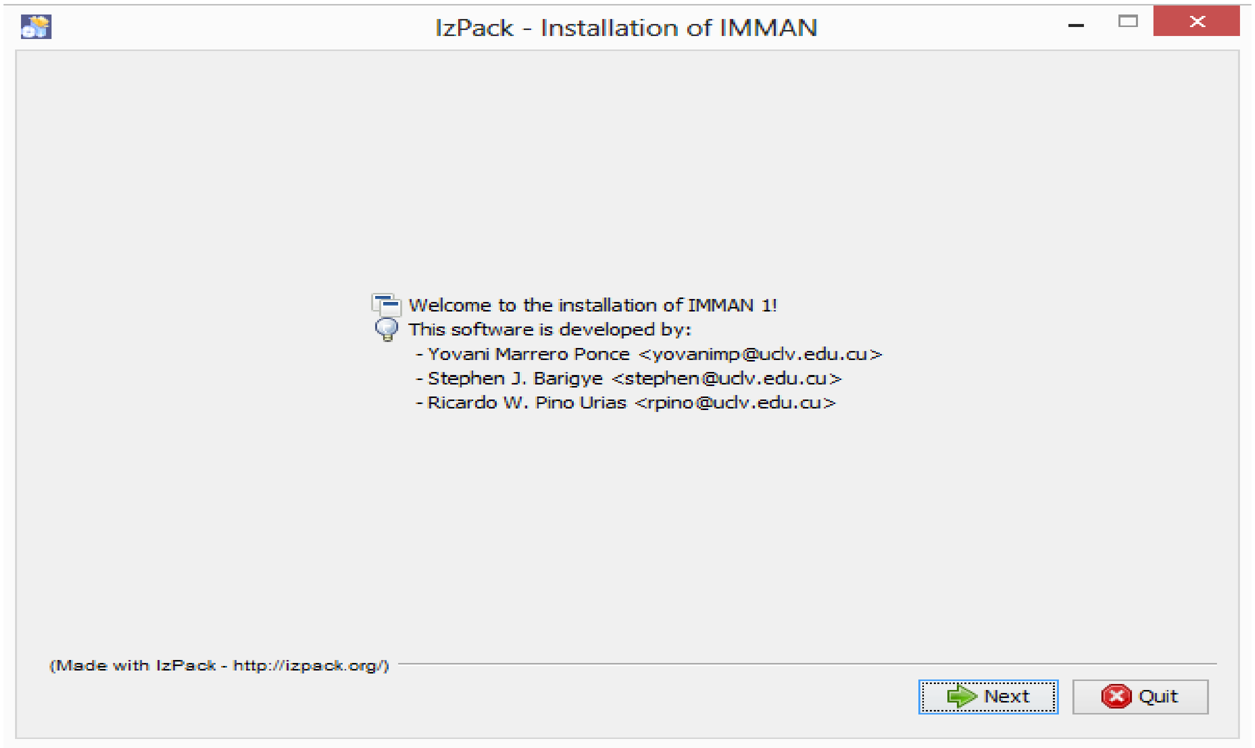
1. Insert the **IMMAN** CD in the CD/DVD-ROM drive of your computer.
2. If the Windows *Autorun* feature is turned on, the installation options will be displayed automatically. Skip next instruction.
3. If you have downloaded **IMMAN** suite from our website on internet or obtained from a colleague recommendation, locate the Java™ based installer file and execute it, see figure bellow.



Name	Date modified	Type
IMMAN Setup.jar	5/25/2013 8:16 AM	Executable Jar F

Figure 1 IMMAN installer file

4. Follow the on screen instructions for installation. See below guided sequence of screenshot for the step by step installation process on a Windows 7 workstation.

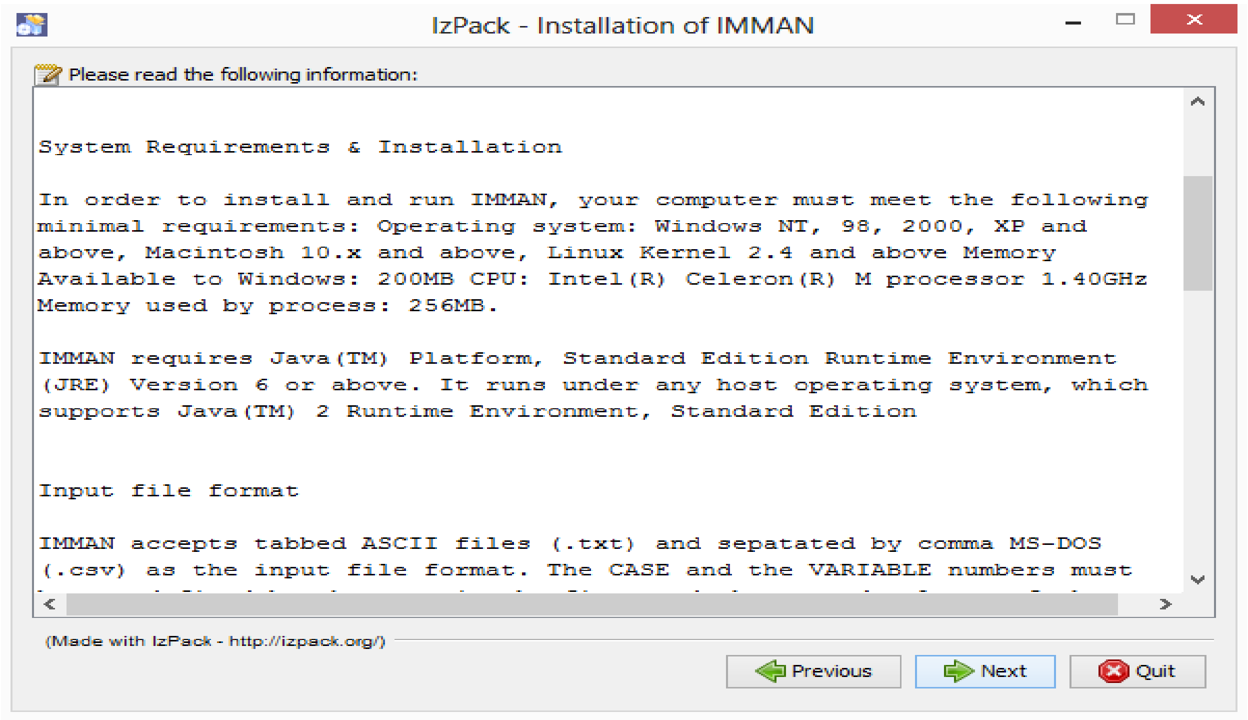


IzPack - Installation of IMMAN

Welcome to the installation of IMMAN 1!
This software is developed by:
- Yovani Marrero Ponce <yovanimp@uclv.edu.cu>
- Stephen J. Barigye <stephen@uclv.edu.cu>
- Ricardo W. Pino Urias <rpino@uclv.edu.cu>

(Made with IzPack - http://izpack.org/)

Figure 2 Welcome screen



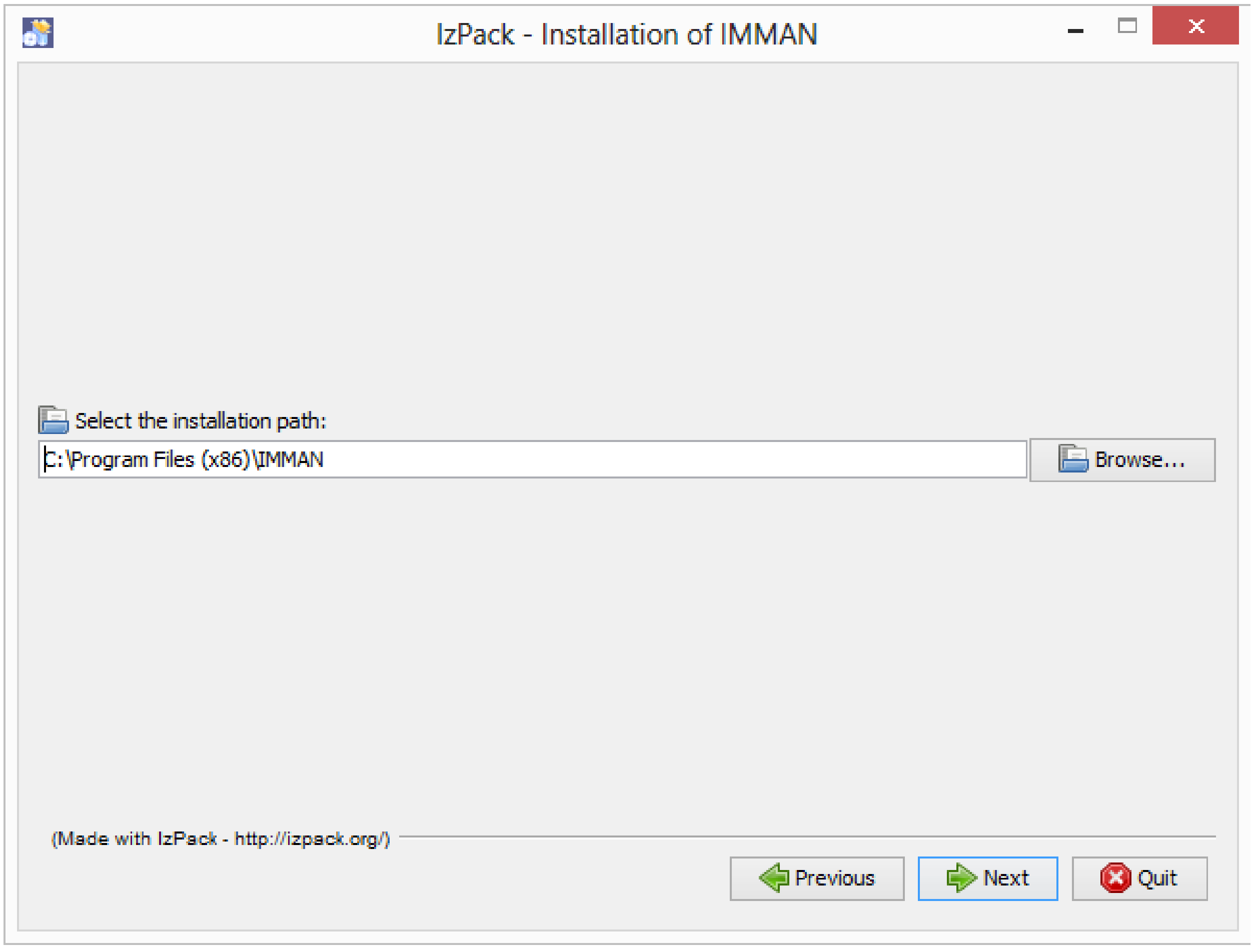
IzPack - Installation of IMMAN

Please read the following information:

```
System Requirements & Installation

In order to install and run IMMAN, your computer must meet the following
minimal requirements: Operating system: Windows NT, 98, 2000, XP and
above, Macintosh 10.x and above, Linux Kernel 2.4 and above Memory
Available to Windows: 200MB CPU: Intel(R) Celeron(R) M processor 1.40GHz
Memory used by process: 256MB.

IMMAN requires Java(TM) Platform, Standard Edition Runtime Environment
(JRE) Version 6 or above. It runs under any host operating system, which
supports Java(TM) 2 Runtime Environment, Standard Edition


Input file format

IMMAN accepts tabbed ASCII files (.txt) and sepatated by comma MS-DOS
(.csv) as the input file format. The CASE and the VARIABLE numbers must
```

(Made with IzPack - http://izpack.org/)

Figure 3 Summary information and notes

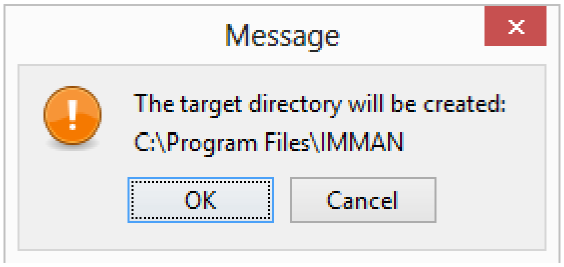


IzPack - Installation of IMMAN

Select the installation path:

C:\Program Files (x86)\IMMAN

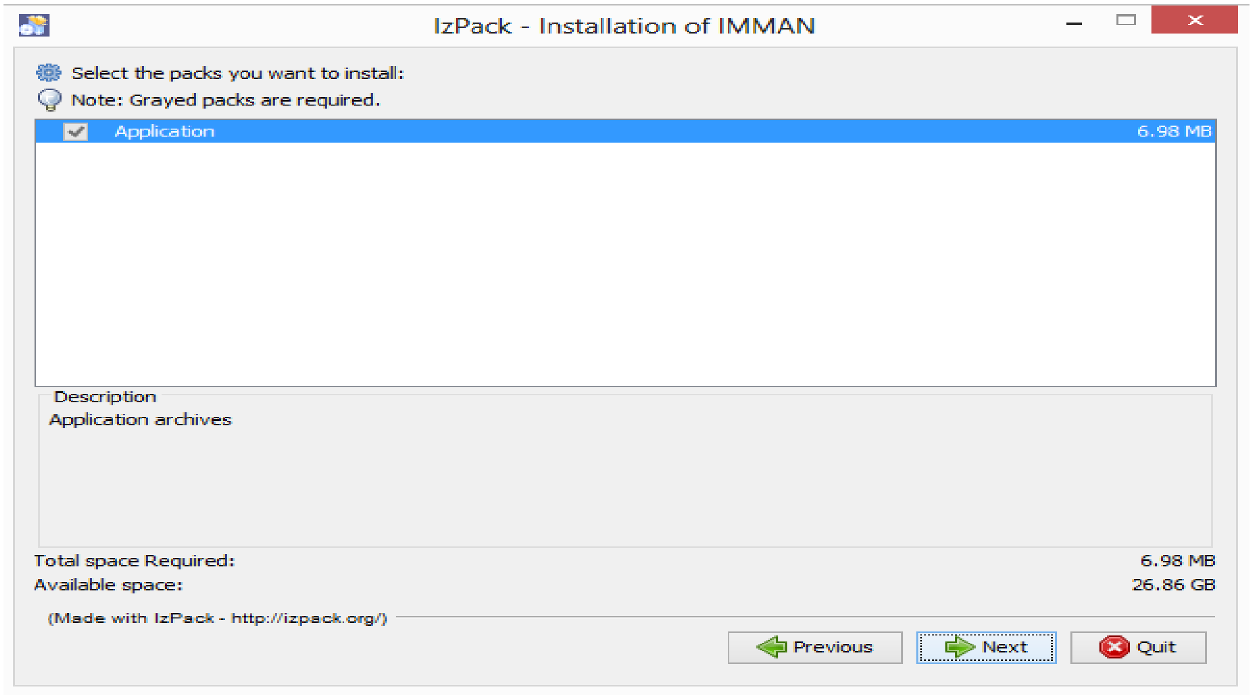
Figure 4 Browse installation path



Message

The target directory will be created:
C:\Program Files\IMMAN

Figure 5 Confirmation for new folder creation


IzPack - Installation of IMMAN

Select the packs you want to install:

Note: Grayed packs are required.

☑	Application	6.98 MB

Description

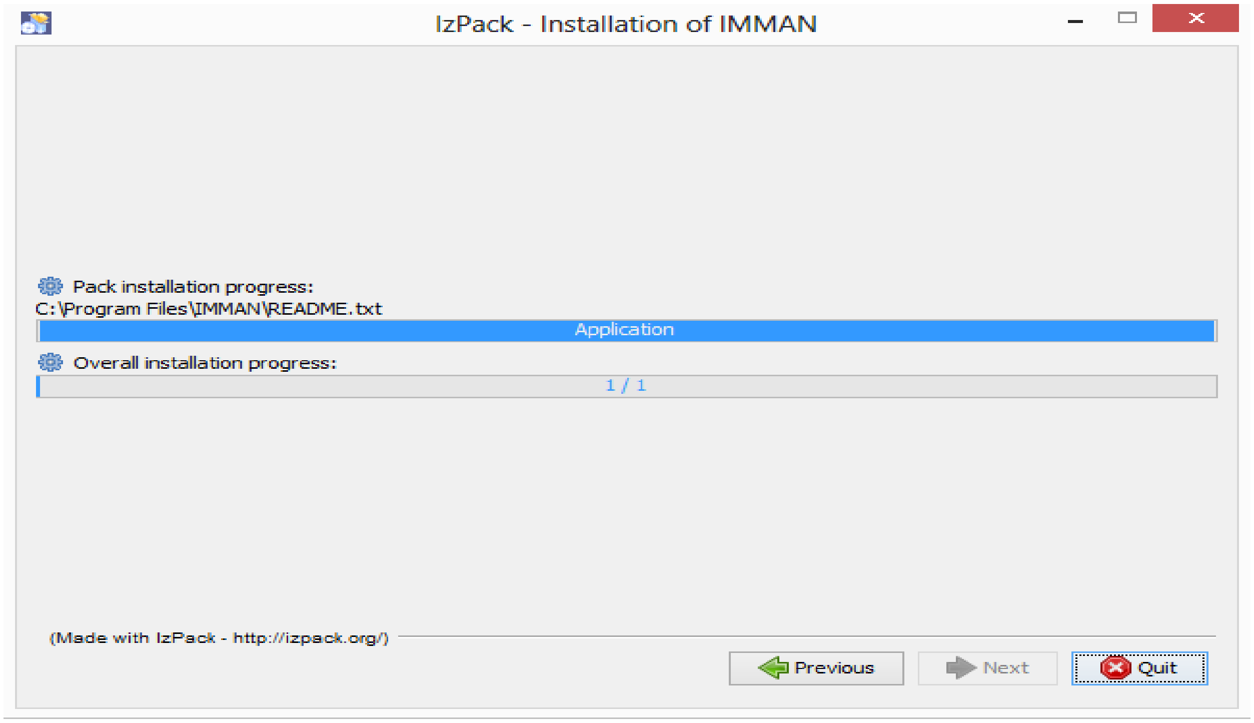
Application archives

Total space Required: 6.98 MB

Available space: 26.86 GB

(Made with IzPack - http://izpack.org/)

Figure 6 An optional installation


IzPack - Installation of IMMAN

Pack installation progress:

C:\Program Files\IMMAN\README.txt

Application

Overall installation progress:

1 / 1

(Made with IzPack - http://izpack.org/)

Figure 7 Installation process

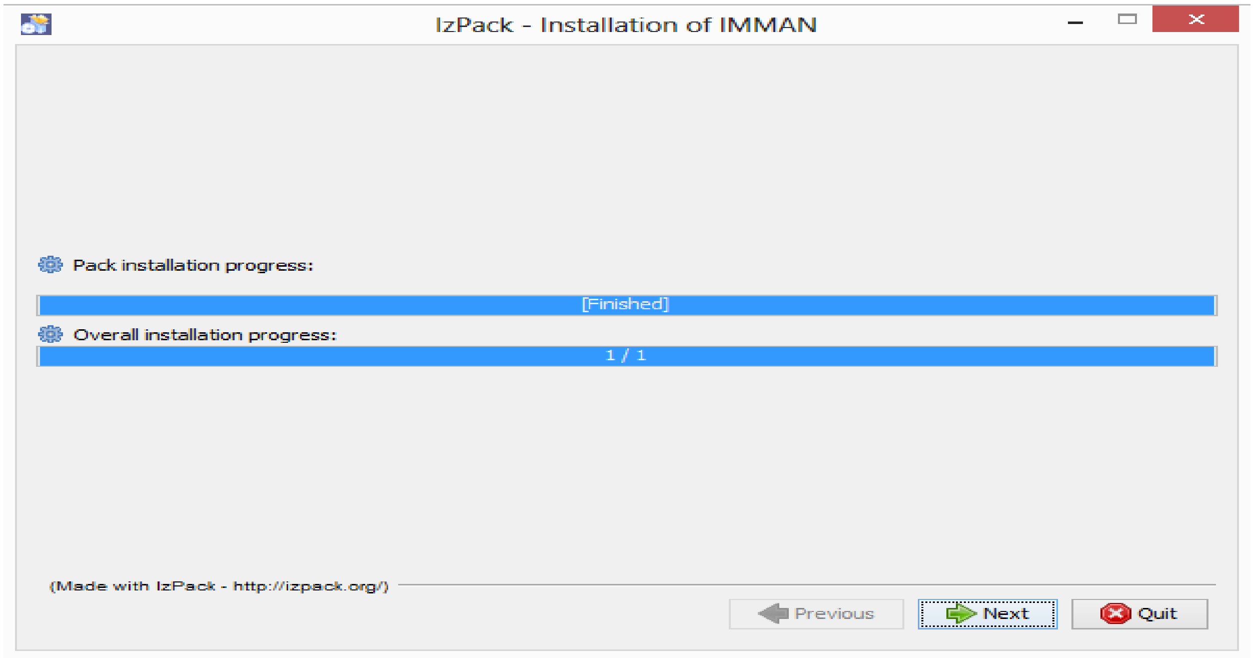
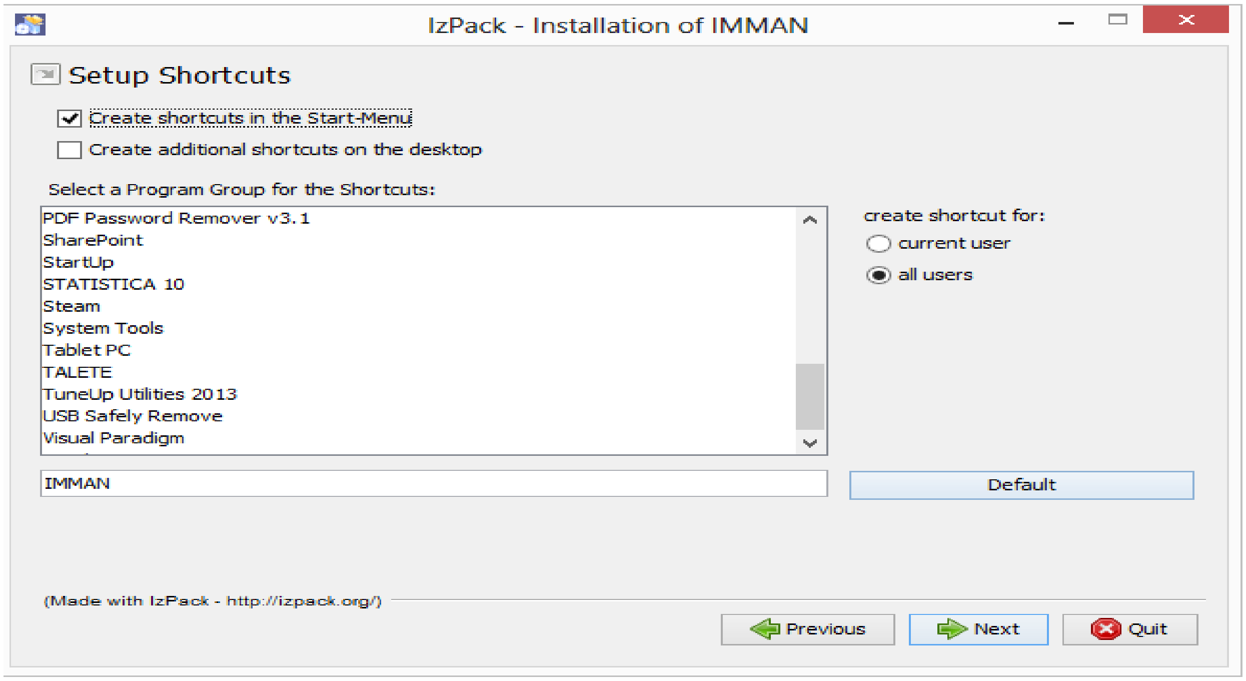

IzPack - Installation of IMMAN
Pack installation progress:
[Finished]
Overall installation progress:
1 / 1
(Made with IzPack - http://izpack.org/)
Previous
Next
Quit

Figure 8 Finish copying files


IzPack - Installation of IMMAN
Previous
Next
Quit
(Made with IzPack - http://izpack.org/)
Default

Setup Shortcuts

☑ Create shortcuts in the Start-Menu

□ Create additional shortcuts on the desktop

Select a Program Group for the Shortcuts:

PDF Password Remover v3.1
SharePoint
StartUp
STATISTICA 10
Steam
System Tools
Tablet PC
TALETE
TuneUp Utilities 2013
USB Safely Remove
Visual Paradigm

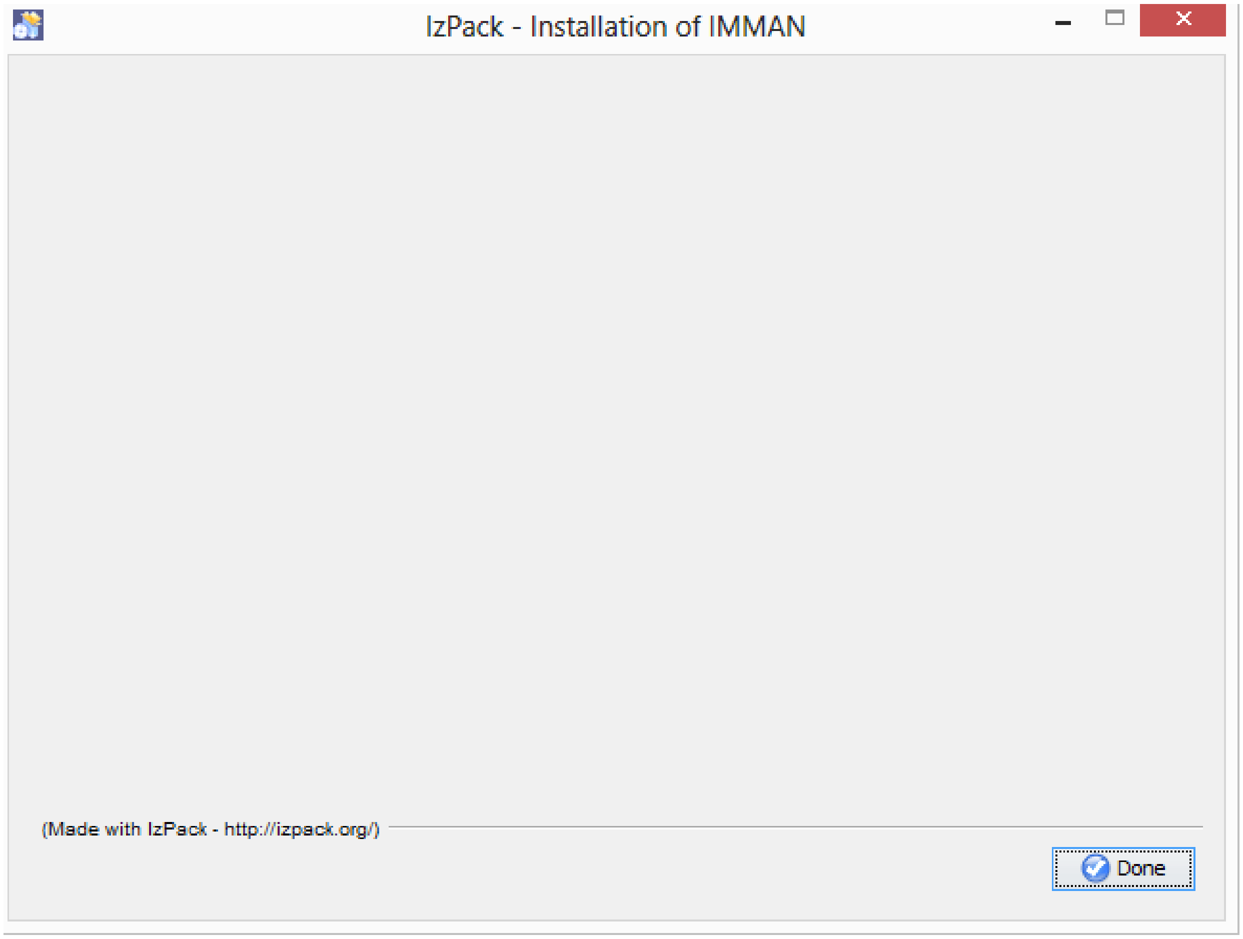
IMMAN

create shortcut for:

○ current user

● all users

Figure 9 Configure desktop Icon and Start Menu items



IzPack - Installation of IMMAN
(Made with IzPack - http://izpack.org/)
Done

Figure 10 Installation and configuration has been completed successfully

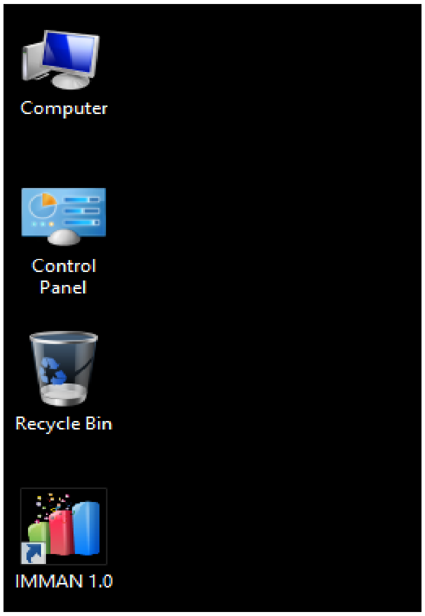


Figure 11 Desktop Icon

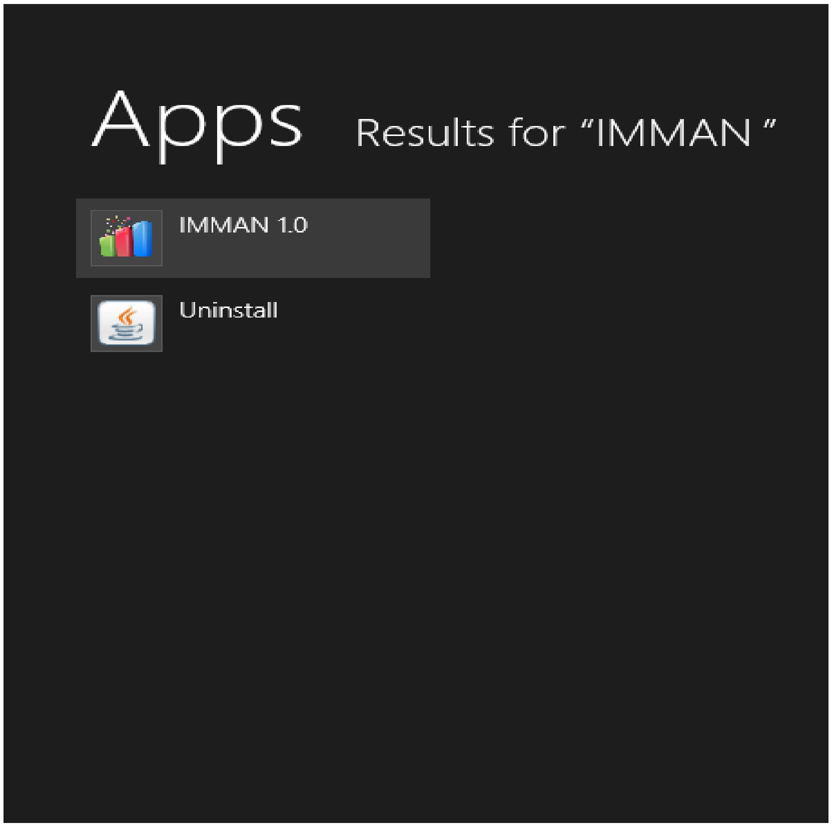


Figure 12 Start Menu files

The IMMAN setup program will detect an existing previous version in your computer and offers the choice of uninstalling or keeping the older version of the program.

Each version of IMMAN could be uninstalled either by using the option "*Control Panel -> Add or Remove Programs*" or by clicking the uninstall shortcut in *"Uninstall" link located in "Start -> All Programs -> IMMAN -> Uninstall*".

If the target operating system is any UNIX/LINUX or MAC platforms, or if the Windows PATH environment variable does not recognize the executable *.jar* file, just execute the corresponding script for launch the installer application, or run directly the Java executable program from the command prompt: (assuming E: is your CD-ROM drive letter):

```
java – jar e:\installer\IMMAN_Setup.jar
```

IMMAN module is available in two different forms, these are:

a) A guided installation program: *IMMAN Setup.jar* *, which allows standard and common chemo-informatics users to deploy the whole program through a friendly step by step GUI, creating easy access shortcut icon in the user's profile Desktop and Start Menu (Windows platform).
b) A Java™ portable application: *IMMAN Portable.jar*, for users that frequently use different workstations. No matter the operating system or workstation hardware configuration, IMMAN users always will have the IMMAN software one click away.

**NOTICE:* Installation program also requires Java™ Runtime Environment on the target system and it is platform-independent. It runs under any host operating system, which supports Java(TM) 7 Runtime Environment, it's also works on X86 and X64 architecture.

3.0   GETTING STARTED

3.0 GETTING STARTED

This section provides a general walkthrough of the system from initiation through exit.

Loading application

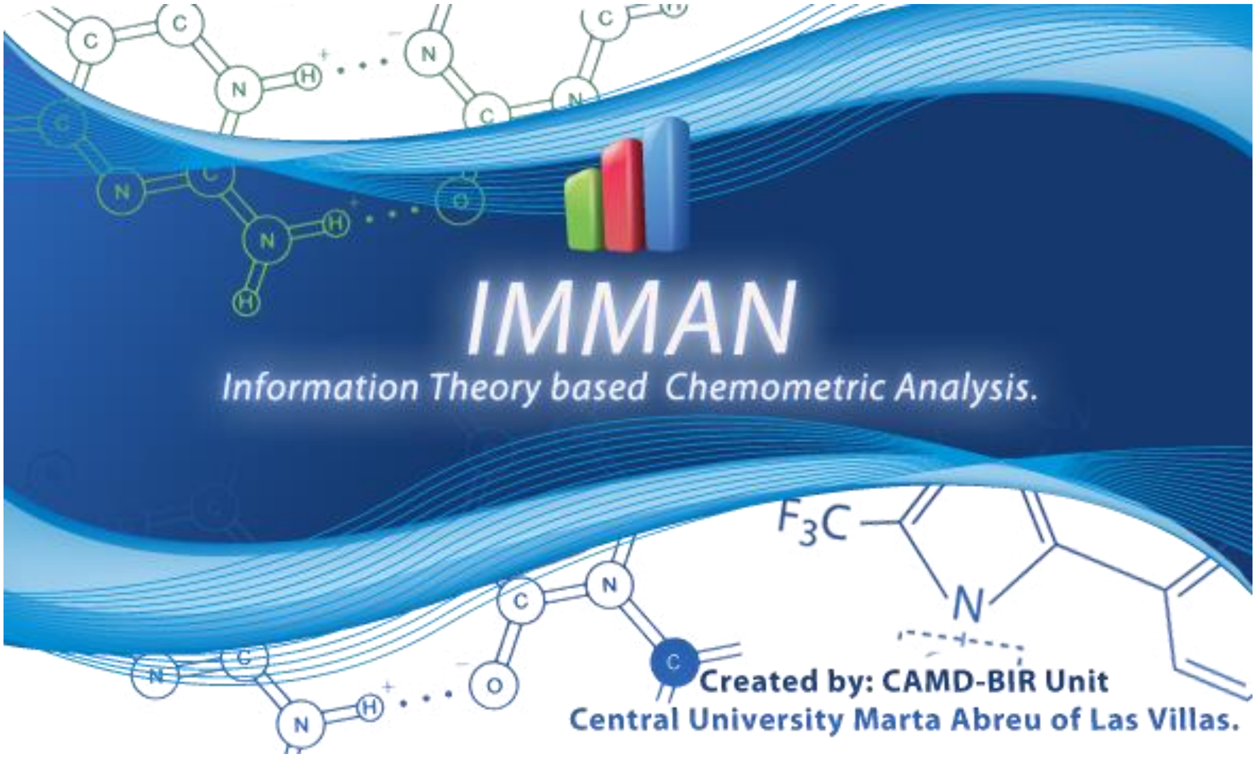

IMMAN
Information Theory based Chemometric Analysis.
Created by: CAMD-BIR Unit
Central University Marta Abreu of Las Villas.

Figure 13 Loading Splash Screen for IMMAN

The software does not require any additional information to login or warm up, as soon as you execute the main program the Splash Screen is launched instantly.

IMMAN Graphical Visual Interface (GUI)

*The **IMMAN** GUI has the following screen areas* (see Figure 14):

Title Bar: Bears the title of program
Menu Bar: Menus related with the different tasks performed by IMMAN.

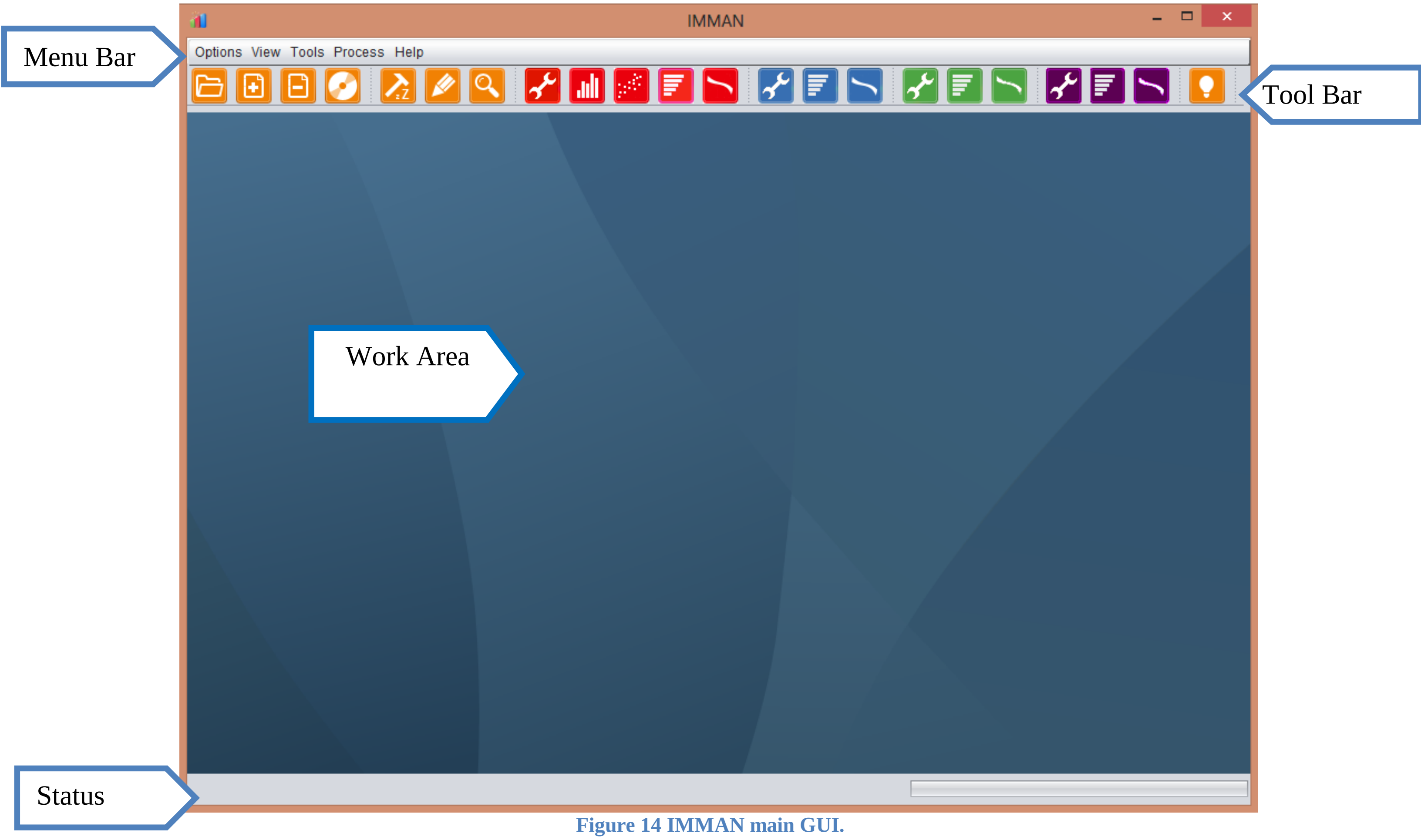


Figure 14 IMMAN main GUI.

- **Tool Bar:** Quick access shortcuts to commonly performed tasks, displayed as graphical icons instead of the classical menu items.
- **Status Bar:** Shows the current process running.
- **Working Area:** This is the *main client area*

System Menu Bar

This section describes in general terms the system menu first encountered by the user, as well as the navigation paths to functions noted on the screen. Each system function should be under a separate section header.

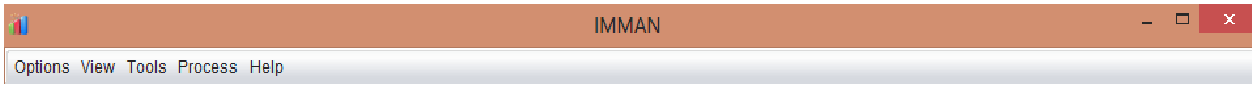


Figure 15 System Menu Bar

Options menu commands

Commands of the *Options menu* allow the user to open several datasets and save results.

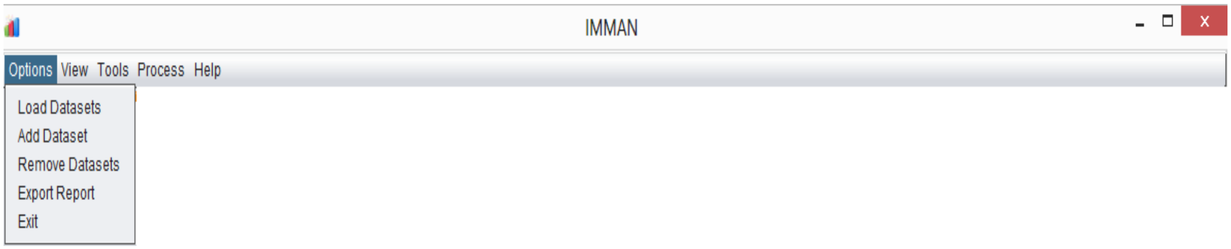


Figure 16 The Options menu

Load Dataset

Launch the load Datasets windows. Configure the dataset's missing values, load multiple Datasets and reset the existing configuration options of the all panels.

Add Dataset

Import one Dataset to the current project without resetting the current configuration.

Remove Datasets

Remove several datasets from the current project

Export Report

Save in *.txt* file format the output report.

Exit

Safely close application (terminates the current **IMMAN** session)

View menu commands

Commands of the *View menu* enable the user to visualize the contents of datasets and the memory usage window.

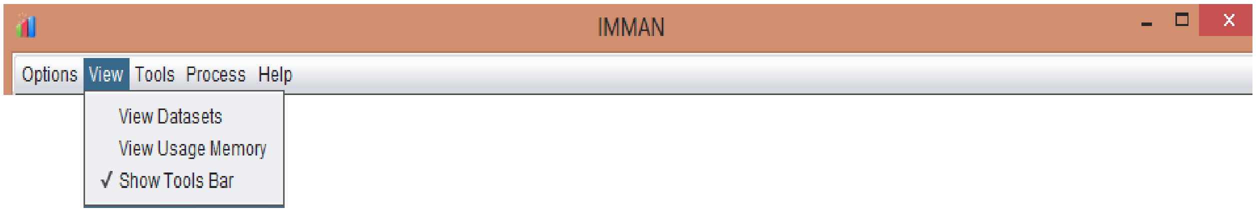


Figure 17 The View menu

Show Tools Bar

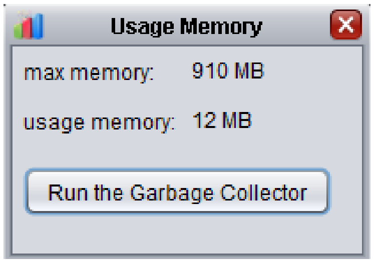
Enable or disable the Tool Bar.

View Datasets

Launch the Datasets Manager windows.

Memory manager

Launch the Memory Manager windows, an example screenshot are available in Figure 18b.



max memory: 910 MB

usage memory: 12 MB

Figure 18 The Memory Manager Tool

Tools menu commands

Commands of the *Tools menu* permit the client to reduce the number of dataset's features

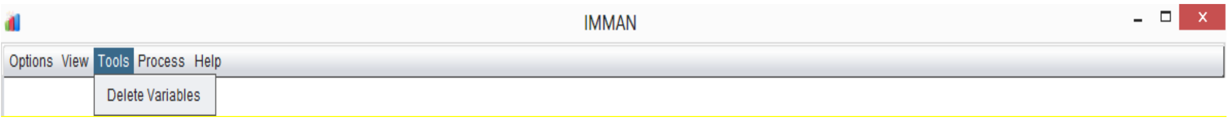


Figure 19 The Tools menu

Delete Variables

The Delete Variables option launches the Delete variables Window. From here the user can reduce the number of variables and create a reduced dataset.

Process menu commands

The *Process* Menu provides access to the feature selection methods and graphic analysis available in the IMMAN software.

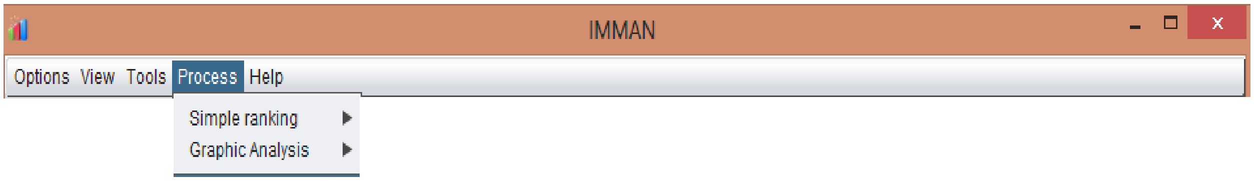


Figure 20 The Process menu

Single Ranking Menu commands

The *Single Ranking* menu provides access to the ranking types or options available in IMMAN program.

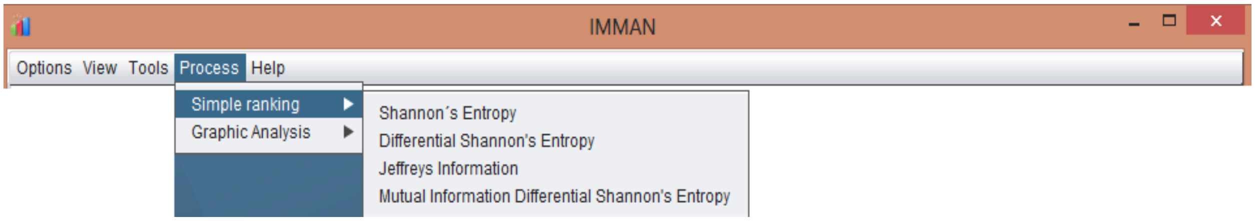

IMMAN
Options View Tools Process Help
Simple ranking
Graphic Analysis
Shannon´s Entropy
Differential Shannon's Entropy
Jeffreys Information
Mutual Information Differential Shannon's Entropy

Figure 21 The Simple ranking menu

Shannon's Entropy

The Shannon's Entropy option launches the window for the unsupervised feature selection methods.

Differential Shannon's Entropy

The Differential Shannon's Entropy option launches the window for the supervised methods based on this entropy type and related measures.

Jeffreys Information

The Jeffreys Information option launches the window for the supervised methods based on this entropy type.

Mutual Information Differential Shannon's Entropy

The Mutual Information Differential Shannon's Entropy option launches the window for the supervised methods based on this entropy type.

Graphic Analysis menu commands

The *Graphic Analysis* menu provides access to the different graphic analyses based on the information theoretic measures implemented in the IMMAN program.

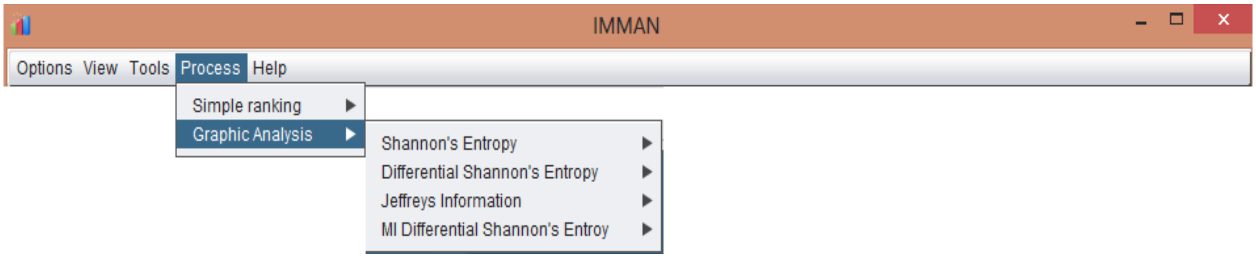

IMMAN
Options View Tools Process Help
Simple ranking
Graphic Analysis
Shannon's Entropy
Differential Shannon's Entropy
Jeffreys Information
MI Differential Shannon's Entroy

Figure 22 The Graphic Analysis menu

Shannon’s Entropy Graphic Analysis menu commands

The Shannon’s Entropy submenu provides access to the different graphic analyses according to unsupervised methods.

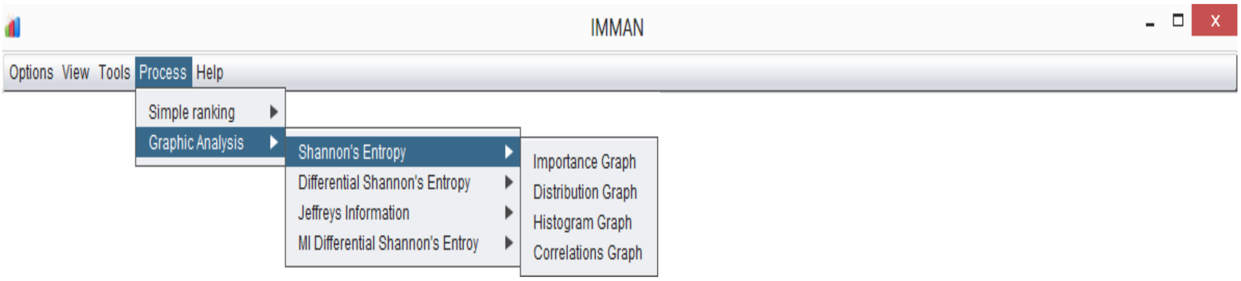


Figure 23 The Shannon’s Entropy Graphic Analysis menu

Shannon’s Distribution Graph

Launches the window for configuring the variable graphic analysis using the cumulative distribution scheme according to the selected entropy measures.

Histogram Graph

Launches the window for the histogram-based distribution analysis according to the Shannon’s entropy of the variables.

Importance Graph

Launches the window for configuring the variable graphic analysis using the importance distribution scheme according to the selected entropy measures.

Correlation Graph

Launches the window for configuring the graphic analysis for the correlation between datasets or classes according to the selected entropy measures.

Differentials Shannon’s Entropy Graphic Analysis menu commands

The Differential Shannon’s Entropy submenu provides access to the different graphic analyses according to supervised methods.

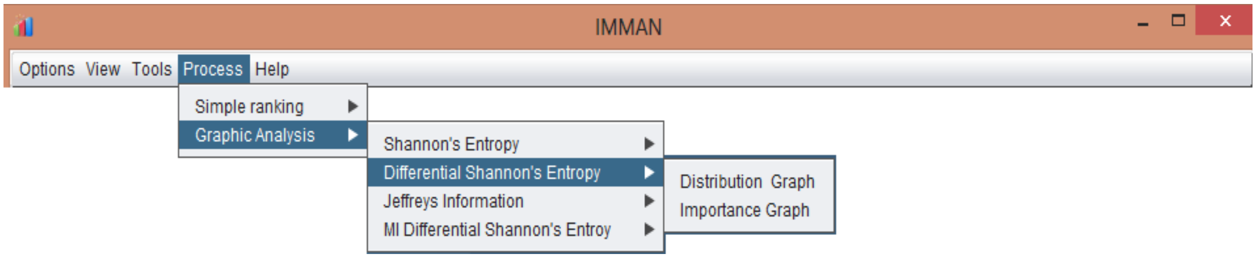


Figure 24 The Differentials Shannon's Entropy Graphic Analysis menu

Shannon's Distribution Graph

Launches the window for configuring the variable graphic analysis using the cumulative distribution scheme according to the Differential Shannon's Entropy.

Importance Graph

Launches the window for configuring the variable graphic analysis using the importance distribution scheme according to the Differential Shannon's Entropy.

Jeffreys Information Graphic Analysis menu commands

The Jeffreys Information submenu provides access to the different graphic analyses according to supervised methods.

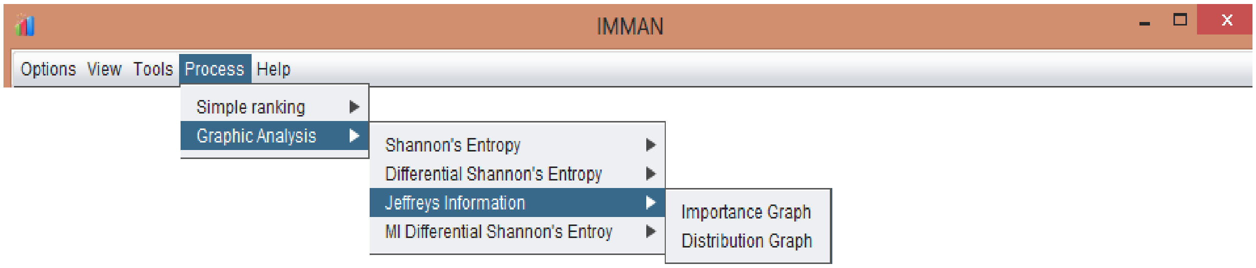


Figure 25 The Jeffreys Information Graphic Analysis menu

Shannon's Distribution Graph

Launches the window for configuring the variable graphic analysis using the cumulative distribution scheme according to Jeffreys Information.

Importance Graph

Launches the window for configuring the variable graphic analysis using the importance distribution scheme according to Jeffreys Information.

Mutual Information Differentials Shannon's Entropy Graphic Analysis menu commands

The Mutual Information Differentials Shannon's Entropy submenu provides access to the different graphic analyses according to supervised methods.

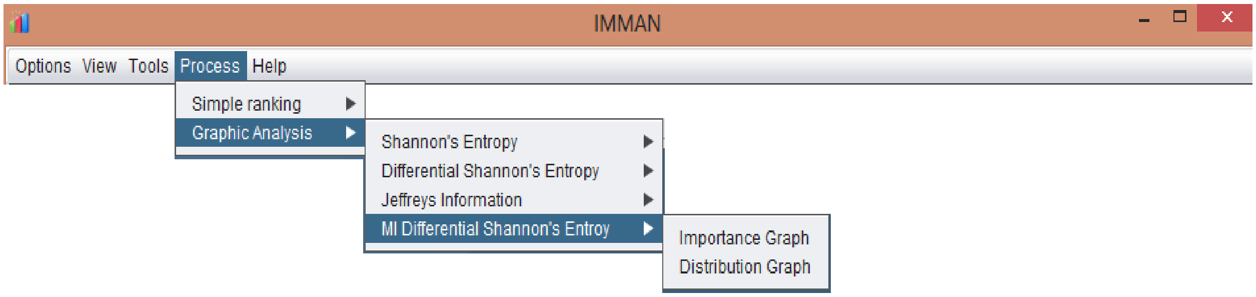


Figure 26 Mutual Information The Differentials Shannon's Entropy Graphic Analysis menu

Shannon's Distribution Graph

Launches the window for configuring the variable graphic analysis using the cumulative distribution scheme according to Mutual Information Differentials Shannon's Entropy.

Importance Graph

Launches the window for configuring the variable graphic analysis using the importance distribution scheme according to Mutual Information Differentials Shannon's Entropy.

Help menu commands

Moreover, the **IMMAN** main window contains some icons which may be clicked in order to obtain specific information. The *Help menu* contains the following commands:

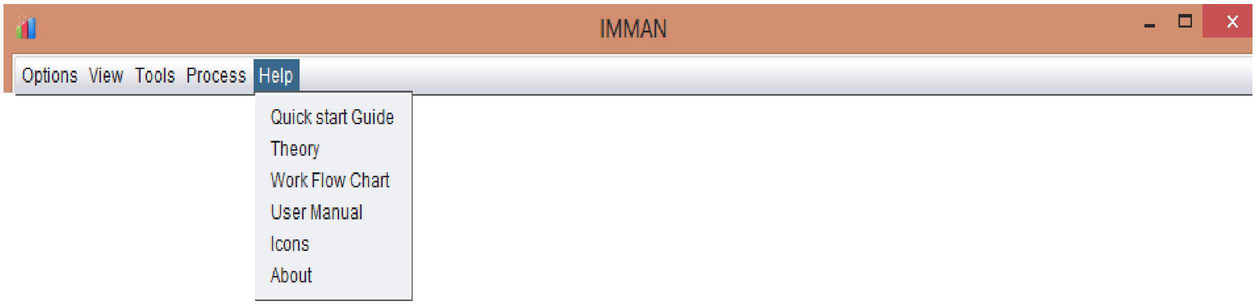


Figure 27 The Help menu

User Manual

Provides extensive documentation on how to work with the program, as well as tutorials and practices outlined in this user-oriented manual.

Theory

Present a wide explanation for features selections techniques implemented in IMMAN.

Work Flow Chart

Shows an illustrative procedure of how use the program. The flowcharts are presented below:

Flowcharts

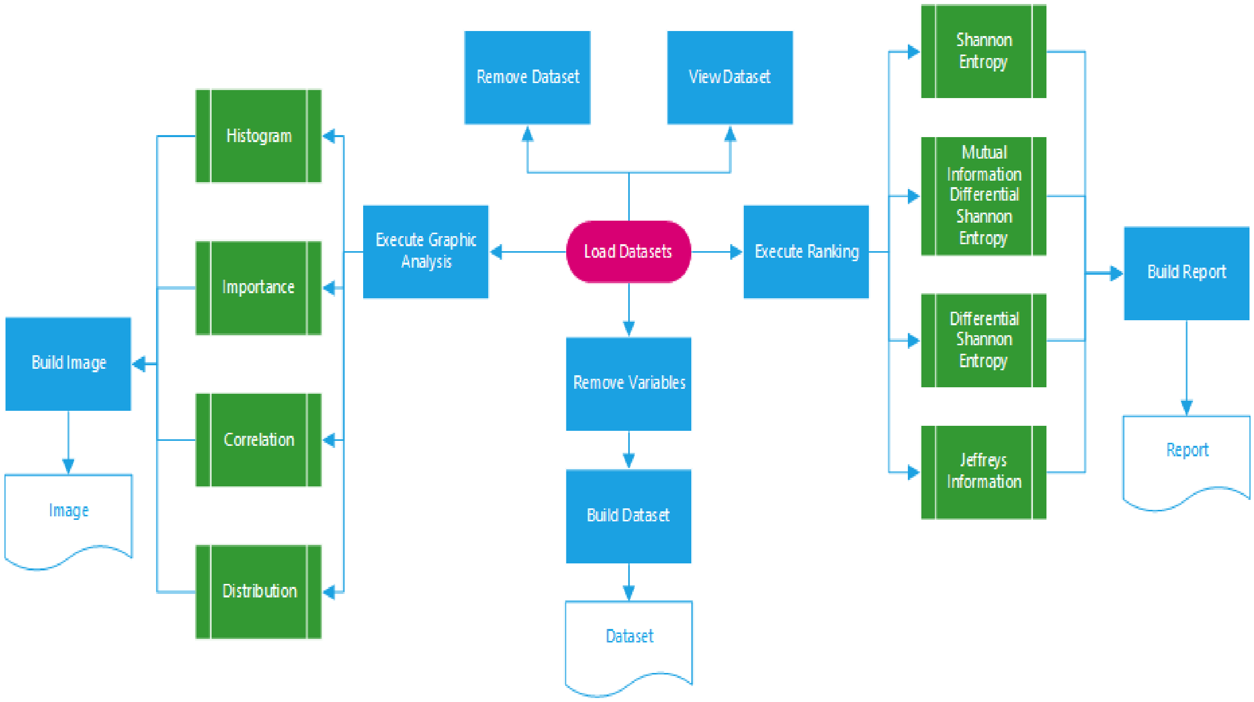


Figure 28 Flow Charts

Quick start Guide

Launches the window for a practical guide for IMMAN, in which the program's general characteristics are explained

Icons

The functionality of all Icon used in the program is described, so that the user learns the meaning of **IMMAN's** icons naturally.

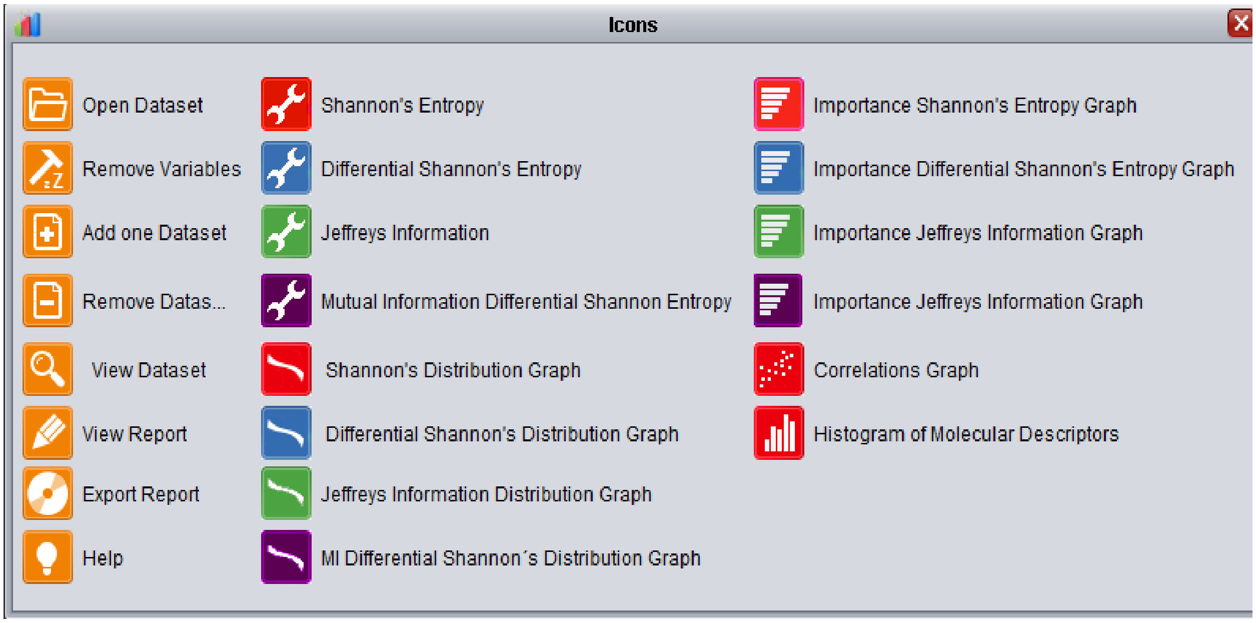

Icons
Open Dataset
Remove Variables
Add one Dataset
Remove Datas...
View Dataset
View Report
Export Report
Help
Shannon's Entropy
Differential Shannon's Entropy
Jeffreys Information
Mutual Information Differential Shannon Entropy
Shannon's Distribution Graph
Differential Shannon's Distribution Graph
Jeffreys Information Distribution Graph
MI Differential Shannon´s Distribution Graph
Importance Shannon's Entropy Graph
Importance Differential Shannon's Entropy Graph
Importance Jeffreys Information Graph
Importance Jeffreys Information Graph
Correlations Graph
Histogram of Molecular Descriptors

Figure 29 Icons description

Home

Avails useful information about **CAMD-BIR Unit**, how to contact us and cite.

Thanks

Recognition to different contributions to the success of this project is offered.

About

Information about **IMMAN** software and publications.

Tool Bar

Quick access shortcuts for most relevant option and tools. That is, the toolbar icons replace the most important and frequently used **IMMAN** menu commands. Clicking on toolbar icons enables the user to perform the following commands:

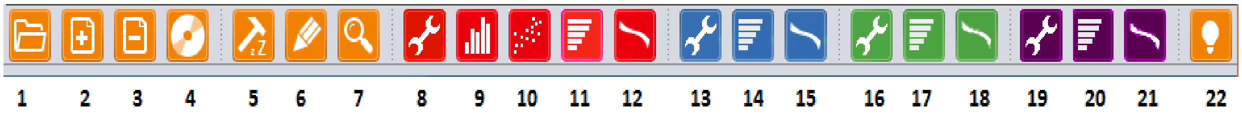


Figure 30 Tool Bar elements

1. Load Datasets
2. Add one Dataset
3. Remove Datasets
4. Export Report
5. Remove Variables
6. View Report

7. View Datasets
8. Shannon’s Entropy Ranking
9. Shannon’s Histogram Graph
10. Shannon’s Correlations Graph
11. Shannon’s Importance Graph
12. Shannon’s Distribution Graph
13. Differential Shannon’s Entropy Ranking
14. Differential Importance Graph
15. Differential Distribution Graph
16. Jeffrey’s Information Ranking
17. Jeffrey’s Importance Graph
18. Jeffrey’s Distribution Graph
19. Mutual Information Differential Shannon’s Entropy Ranking
20. Mutual Information Differential Importance Graph
21. Mutual Information Differential Distribution Graph
22. User Manual

Status Bar

The status bar located at the bottom of main windows shows the percentage of completion, also displays the name for running process.

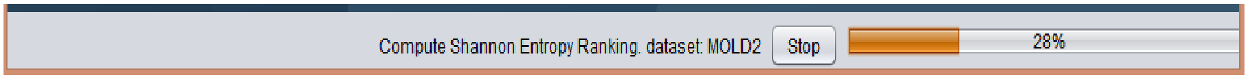
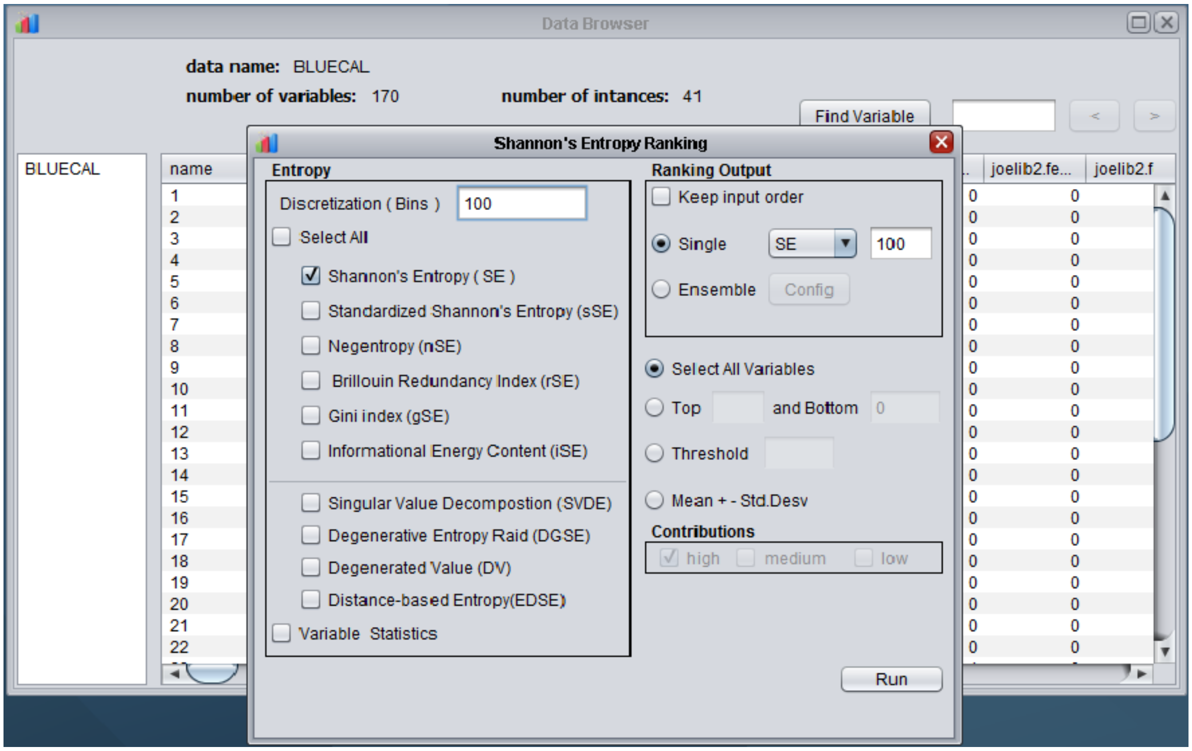

Compute Shannon Entropy Ranking. dataset: MOLD2

Figure 31 Status Bar

Work Area

The work area is where all the configuration windows and outputs for the program are displayed.



Data Browser

data name: BLUECAL

number of variables: 170 **number of intances:** 41

Shannon's Entropy Ranking

Entropy

Discretization (Bins) 100

- ☐ Select All
- ☑ Shannon's Entropy (SE)
- ☐ Standardized Shannon's Entropy (sSE)
- ☐ Negentropy (nSE)
- ☐ Brillouin Redundancy Index (rSE)
- ☐ Gini index (gSE)
- ☐ Informational Energy Content (iSE)
- ☐ Singular Value Decompostion (SVDE)
- ☐ Degenerative Entropy Raid (DGSE)
- ☐ Degenerated Value (DV)
- ☐ Distance-based Entropy(EDSE)
- ☐ Variable Statistics

Ranking Output

- ☐ Keep input order
- ● Single SE 100
- ○ Ensemble Config
- ● Select All Variables
- ○ Top and Bottom 0
- ○ Threshold
- ○ Mean + - Std.Desv

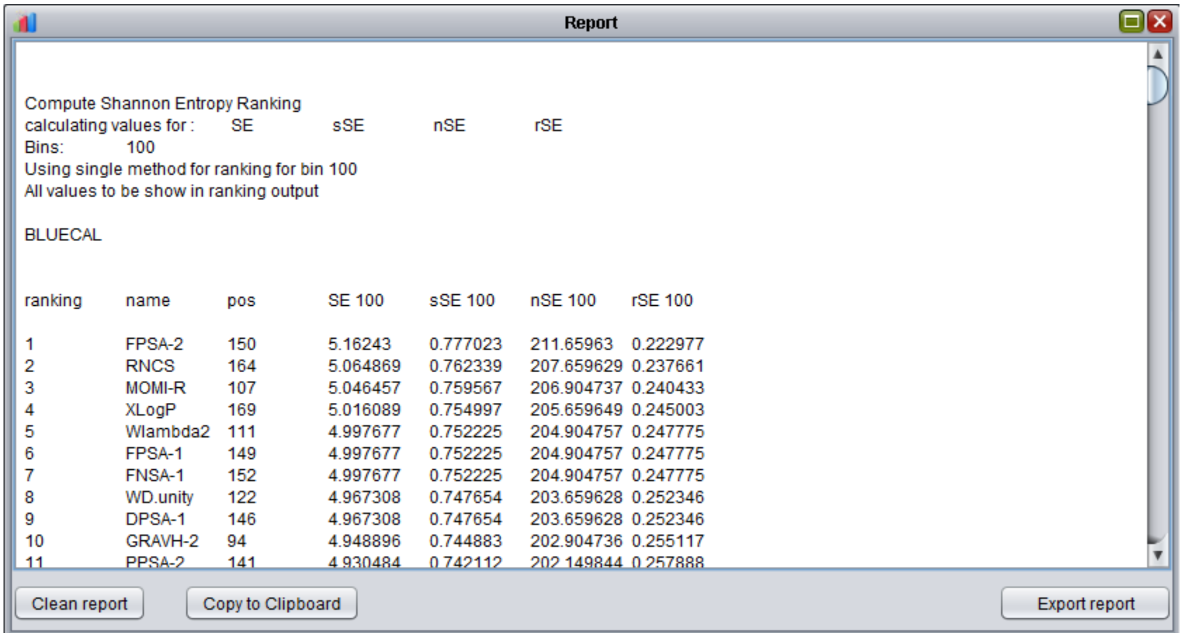
Contributions

☑ high ☐ medium ☐ low

Figure 32 Work Area. Shannon’s Entropy Ranking Window

Report Window

Logging windows for all operations and task. Besides, after the calculation is finished, the **Report History (log file)** shows some details and statistics of the run.



Report

```
Compute Shannon Entropy Ranking
calculating values for :     SE          sSE          nSE          rSE
Bins:          100
Using single method for ranking for bin 100
All values to be show in ranking output

BLUECAL


ranking      name        pos         SE 100      sSE 100     nSE 100     rSE 100

1            FPSA-2      150         5.16243     0.777023    211.65963   0.222977
2            RNCS        164         5.064869    0.762339    207.659629  0.237661
3            MOMI-R      107         5.046457    0.759567    206.904737  0.240433
4            XLogP       169         5.016089    0.754997    205.659649  0.245003
5            Wlambda2    111         4.997677    0.752225    204.904757  0.247775
6            FPSA-1      149         4.997677    0.752225    204.904757  0.247775
7            FNSA-1      152         4.997677    0.752225    204.904757  0.247775
8            WD.unity    122         4.967308    0.747654    203.659628  0.252346
9            DPSA-1      146         4.967308    0.747654    203.659628  0.252346
10           GRAVH-2     94          4.948896    0.744883    202.904736  0.255117
11           PPSA-2      141         4.930484    0.742112    202.149844  0.257888
```

Figure 33 Report window

Exit System

Describe the actions necessary to properly exit the system.

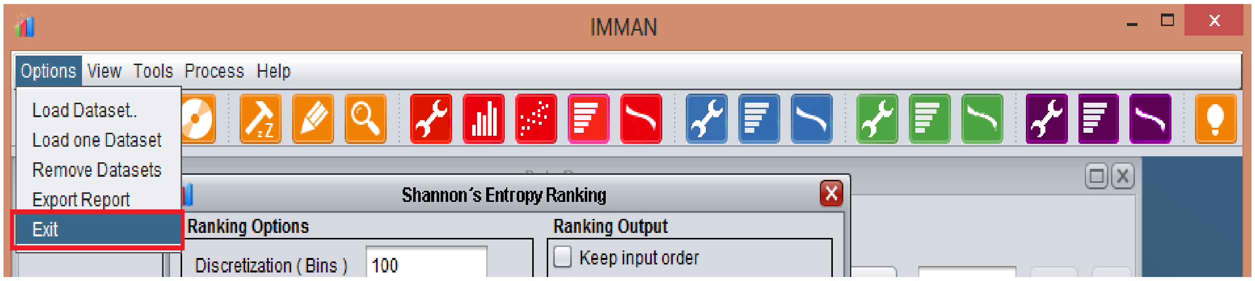
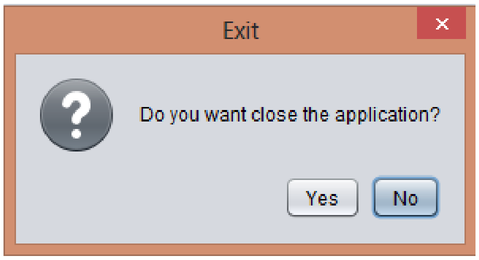


Figure 344 Program Exit Options



Do you want close the application?

Figure 35 Safe Exit Action prompt

4.0 USING THE SYSTEM

4.0 USING THE SYSTEM

This section provides a detailed description of the **IMMAN** Software from the initial to the final steps, explaining in detail the characteristics of the required input and system-produced output. It covers both calculations of single and multiple molecular datasets and batch mode calculations.

Each **IMMAN** function is under a separate section header, and corresponds sequentially to the system functions (menu items) listed in subsections of chapters above.

Starting IMMAN

IMMAN is launched by clicking on the icon on the desktop (e.g., on Microsoft Windows platforms). For advanced users other launching methods are available in order to improve the performance and speed, these are tweaks for the Java™ VM (JVM), that increase the maximum default limit of JVM heap memory.

Preconfigured command line scripts are provided in the root directory of IMMAN program folder. The configurations of JVM heap memory limit are:

- 1024MB
- 2048MB
- 4096MB
- 8192MB

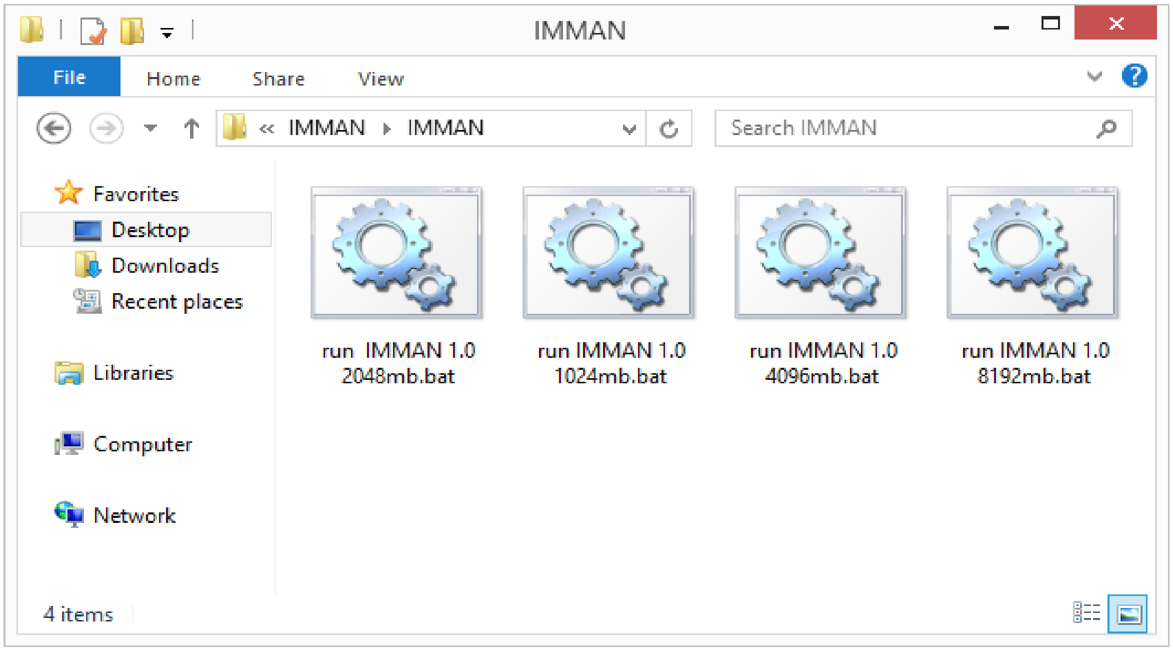


Figure 36 IMMAN Windows Batch Files (.bat)

Indeed, for each heap memory limit, a command line scripts were targeted for two different kind of platform:

- Windows Command Script (*.cmd*) and Windows Batch File (*.bat*)
- Linux Shell Script (*.sh*).

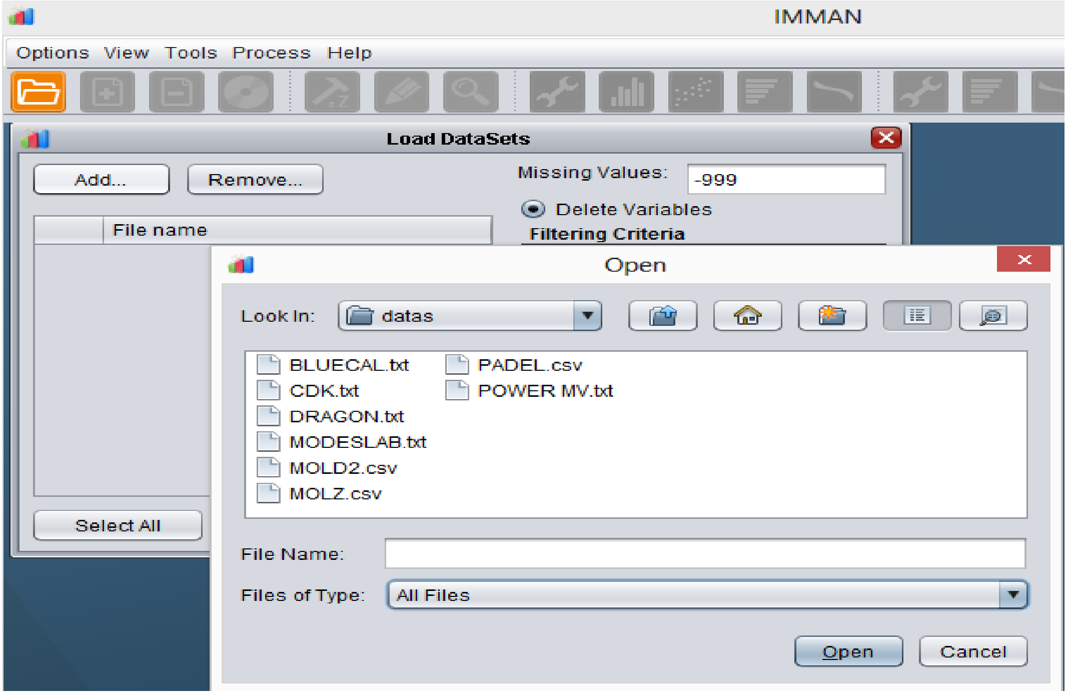
Otherwise, if the preconfigured command line scripts do not suit your hardware preferences, users can modify the scripts for both platforms to adjust the program JVM heap memory limit according to their system hardware properties, editing these scripts with a text editor program, (i.e. NotePad or WordPad in Windows, and GEdit or Vi in Linux). The following example limits de JVM heap memory up to 1024 megabytes:

```
java -Xms256m -Xmx1024m -jar IMMAN.jar
```

After splash (Figure 13), the main window (GUI) will be displayed on the screen (see Figure 14).

Loading Datasets

To import a file, click on the open data icon to display the datasets window (also accessible by selecting the open dataset option from the process menu).



Missing Values: -999

Look In: datas

BLUECAL.txt
CDK.txt
DRAGON.txt
MODESLAB.txt
MOLD2.csv
MOLZ.csv
PADEL.csv
POWER MV.txt

File Name:

Files of Type: All Files

Figure 37 loading files

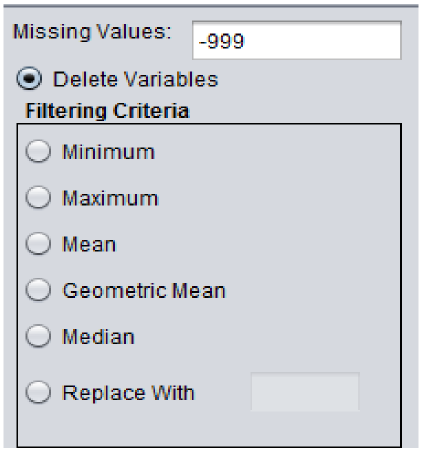
Edit Datasets

Missing values

To process missing dataset values, click the options button in the Dataset window or select the missing values option in the tools menu. By default, IMMAN uses the numeric code –999 as the identification for missing values. This code could be altered for any other value(s) (“;” separator) code of choice.

The editing operations for missing values in a dataset are:

- Delete variable: the variables with missing values are deleted from the dataset.
- Minimum: the missing values in a variable are replaced by the minimum value of the variable.
- Maximum: the missing values in a variable are replaced by the maximum value of the variable.
- Mean: the missing values in a variable are replaced by the mean value of the variable.
- Geometric mean: the missing values in a variable are replaced by the geometric mean value of the variable
- Median: the missing values in a variable are replaced by the median value of the variable.
- Replace with: a user defined numeric value is used as replacement for the missing data in a variable.



Missing Values: -999

● Delete Variables

Filtering Criteria

○ Minimum

○ Maximum

○ Mean

○ Geometric Mean

○ Median

○ Replace With

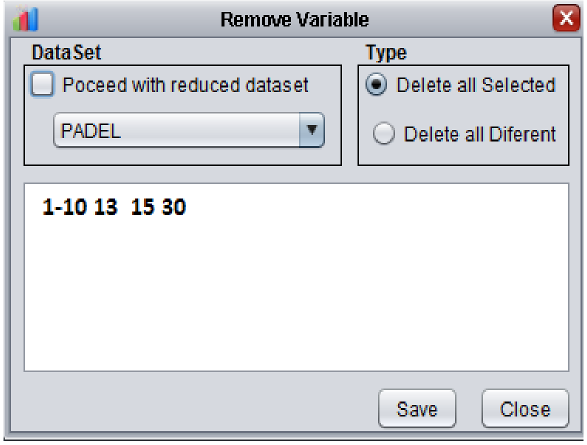
Figure 38 loading files

The identities of the variables with missing values are automatically sent to the report. Note that IMMAN automatically recognizes as errors in file structure the presence of non-numeric codes, tabs, carriage returns, character spaces and line spaces.

Delete Variables

IMMAN also permits deleting variables in a dataset file, which may not necessarily contain missing values. Select the *Delete Variables* option, accessible from the tools menu bar drop-down list, to display the delete variables window. Choose the dataset file to edit in the DataSet drop-down list and enter in text field the variables (in terms of their positions) to be removed or retained. IMMAN provides two options for removing variables, configurable in the *Type* pane. The *Delete Selection* option permits the user to remove the variables entered in the text field, while *Delete all*

Different permits the user to eliminate the all variables in the Dataset other than the ones entered in the text field. Below is an example of this operation.

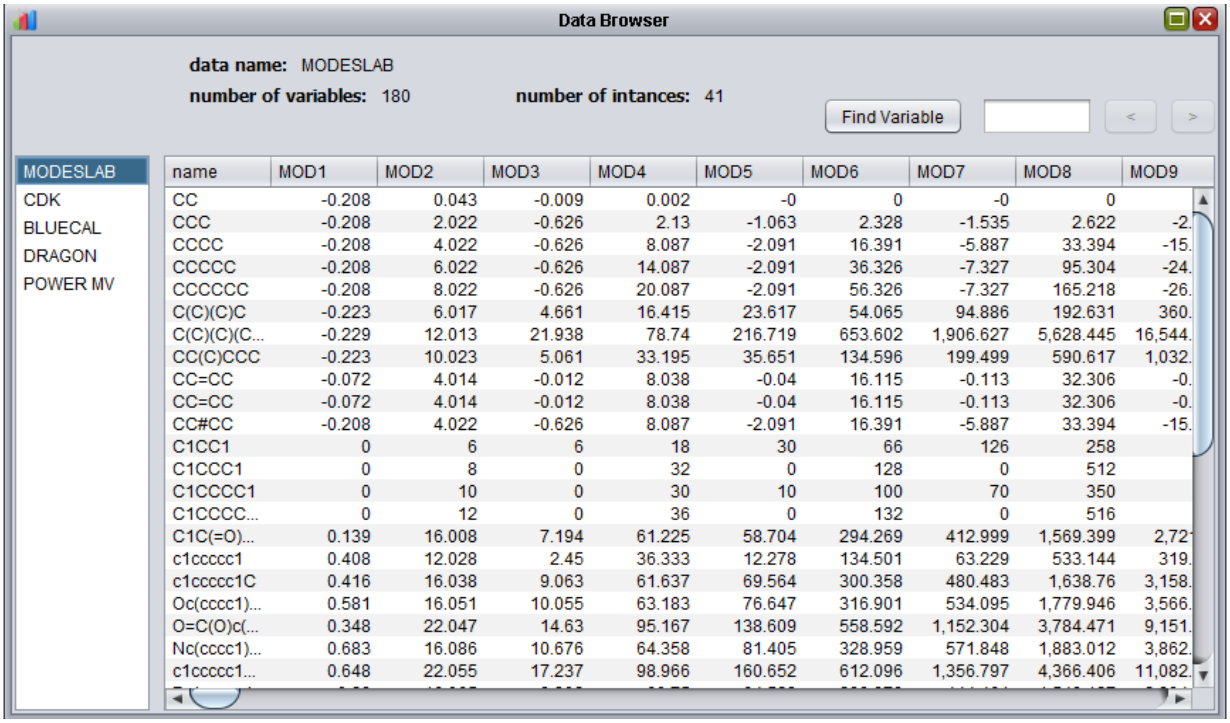


1-10 13 15 30

Figure 39 View Data Window

View Data

The user may wish to analyze the loaded dataset file(s), by clicking the View Data icon (🔍) on the menu toolbar, which displays a Data Browser window. Click on the dataset file name to display the corresponding data matrix. Note that this data matrix is not editable.



data name: MODESLAB

number of variables: 180

number of intances: 41

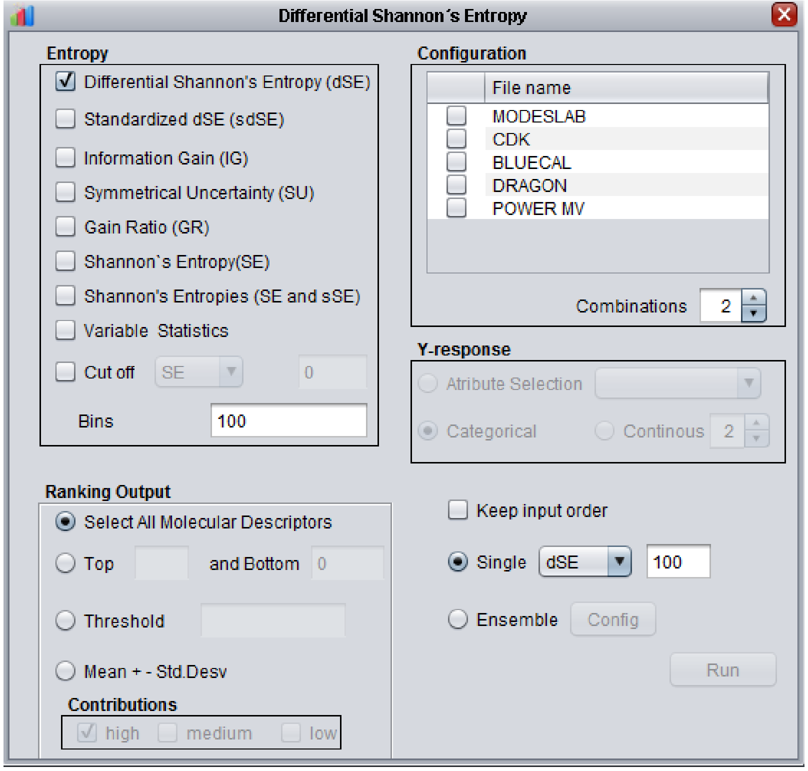
name	MOD1	MOD2	MOD3	MOD4	MOD5	MOD6	MOD7	MOD8	MOD9
CC	-0.208	0.043	-0.009	0.002	-0	0	-0	0	
CCC	-0.208	2.022	-0.626	2.13	-1.063	2.328	-1.535	2.622	-2.
CCCC	-0.208	4.022	-0.626	8.087	-2.091	16.391	-5.887	33.394	-15.
CCCCC	-0.208	6.022	-0.626	14.087	-2.091	36.326	-7.327	95.304	-24.
CCCCCC	-0.208	8.022	-0.626	20.087	-2.091	56.326	-7.327	165.218	-26.
C(C)(C)C	-0.223	6.017	4.661	16.415	23.617	54.065	94.886	192.631	360.
C(C)(C)(C...	-0.229	12.013	21.938	78.74	216.719	653.602	1,906.627	5,628.445	16,544.
CC(C)CCC	-0.223	10.023	5.061	33.195	35.651	134.596	199.499	590.617	1,032.
CC=CC	-0.072	4.014	-0.012	8.038	-0.04	16.115	-0.113	32.306	-0.
CC=CC	-0.072	4.014	-0.012	8.038	-0.04	16.115	-0.113	32.306	-0.
CC#CC	-0.208	4.022	-0.626	8.087	-2.091	16.391	-5.887	33.394	-15.
C1CC1	0	6	6	18	30	66	126	258	
C1CCC1	0	8	0	32	0	128	0	512	
C1CCCC1	0	10	0	30	10	100	70	350	
C1CCCC...	0	12	0	36	0	132	0	516	
C1C(=O)...	0.139	16.008	7.194	61.225	58.704	294.269	412.999	1,569.399	2,72
c1ccccc1	0.408	12.028	2.45	36.333	12.278	134.501	63.229	533.144	319.
c1ccccc1C	0.416	16.038	9.063	61.637	69.564	300.358	480.483	1,638.76	3,158.
Oc(cccc1)...	0.581	16.051	10.055	63.183	76.647	316.901	534.095	1,779.946	3,566.
O=C(O)c(...	0.348	22.047	14.63	95.167	138.609	558.592	1,152.304	3,784.471	9,151.
Nc(cccc1)...	0.683	16.086	10.676	64.358	81.405	328.959	571.848	1,883.012	3,862.
c1ccccc1...	0.648	22.055	17.237	98.966	160.652	612.096	1,356.797	4,366.406	11,082.

Figure 40 Remove Variables

Feature Selection Ranking

Shannon's Entropy Ranking

Select Simple Ranking from the Process menu to display the drop-down list, from which Shannon's entropy is selected, also accessible by clicking the configure SE ranking icon (🔧) in toolbar menu. This opens a ranking window where the necessary parameters are defined.



Differential Shannon´s Entropy

Entropy
☑ Differential Shannon's Entropy (dSE)
☐ Standardized dSE (sdSE)
☐ Information Gain (IG)
☐ Symmetrical Uncertainty (SU)
☐ Gain Ratio (GR)
☐ Shannon`s Entropy(SE)
☐ Shannon's Entropies (SE and sSE)
☐ Variable Statistics
☐ Cut off SE ▼ 0
Bins 100

Configuration
	File name
☐	MODESLAB
☐	CDK
☐	BLUECAL
☐	DRAGON
☐	POWER MV

Combinations 2

Y-response
○ Atribute Selection ▼
● Categorical ○ Continous 2

Ranking Output
● Select All Molecular Descriptors
○ Top and Bottom 0
○ Threshold
○ Mean + - Std.Desv
Contributions
☑ high ☐ medium ☐ low

☐ Keep input order
● Single dSE ▼ 100
○ Ensemble Config
Run

Figure 41 Shannon's Entropy Ranking window

Entropies Panel

Discretization (Bins): In this field, the user enters the number of discrete intervals (bins) to be used (default- 100). Various intervals could be used simultaneously, separated by commas.

Shannon's Entropy

Quantifies the degree of uncertainty of features in a dataset. For a discrete information source, Shannon's entropy (o entropy of a finite set) is expressed using the following formula (Shannon 1948):

$$SE(X) = -\sum_{i=1}^{N} P_i \log P_i$$

where p_i is the probability that a randomly selected instance belongs to discrete interval i and N is the number of discrete intervals.

Standardized Shannon's Entropy

Normalized variant of Shannon's entropy with respect to the maximum entropy and is expressed by the following equation (Godden & Bajorath 2001):

$$sSE(X) = \frac{SE(X)}{\log N}$$

where N is the number of discrete intervals. This entropy represents a measure of the relative efficiency of the attained information

Gini Index (gSE)

Evaluates the diversity of the information contained in a feature, characterized by an increase in its value with an increase in the diversity of the instances. The Gini index is mathematically defined as follows (Breiman et al. 1984):

$$gSE(X) = \sum_{i=1}^{N} p_i * p_{1+1} \qquad 0 \le gSE \le \frac{N-1}{2N}$$

Singular Value Decomposition Entropy (VDSEi)

Evaluates the contribution of the i^{th} feature to the dataset entropy (Alter et al. 2000). Let S_j represent the singular values of the matrix A $_{[nxm]}$ of n instances and m features. Then S_j^2 denotes the eigenvalues of the n x n matrix $A_* A^t$. The dataset entropy is defined by:

$$\mathrm{E(A)} = -\frac{1}{\log \boldsymbol{N}} * \sum_{\boldsymbol{j}=1}^{\boldsymbol{N}} \frac{\boldsymbol{S_j}^2}{\boldsymbol{S_T}} \log \frac{\boldsymbol{S_j}^2}{\boldsymbol{S_T}}$$

where $\boldsymbol{S_T}$ denotes the total sum of the S_j^2 values. Therefore the contribution of the i^{th} feature to the dataset entropy is defined as follows:

$$DSE_i = E(A) - E(A^{'})$$

where A′ denotes the matrix A without the analyzed feature.

Redundancy Index

Measures information redundancy of the features and is mathematically defined as follows(Brillouin 1962):

$$rSE(X) = 1 - sSE(X)$$

The interpretation of the redundancy index is in terms of the degree of loss of information and its values lie between 0 and 1.

Negentropy

Quantifies the residual information contained in a feature after the statistical distribution of the instances is analyzed (Kier 1980). The negentropy or total information content is mathematically defined as follows:

$$nSE(X) = SE_{\max}(X) - SE(X)$$

Information Energy Content

This parameter is a complementary quantity to the Gini index is the informational defined as (Onicescu 1966):

$$iSE(X) = \sum_{i=1}^{N} p_i^2 \qquad \frac{1}{N} \le iSE \le 1$$

Degenerated Value

Diversity measure that indicates the number of instances characterized differently by features in a data matrix and is defined as:

$$DV(X) = \text{Instances}_{\text{total}} - \text{Instances}_{\text{different}}$$

Degenerated Value Entropy

Evaluates the feature variability according to the degenerated value:

$$DVSE(X) = DV(X) * \frac{SE(X)_{DV(X)}}{SE(X)_{inst}}$$

where DV (X) is the degenerated value for variable X.

Distance based Entropy

Diversity measure computed on the distance among instances and is expressed as follows:

$$DBE(X) = \sum_i \sum_j \left[D_{ij} \log_2 D_{ij} + (1 - D_{ij}) \log_2 (1 - D_{ij}) \right]$$

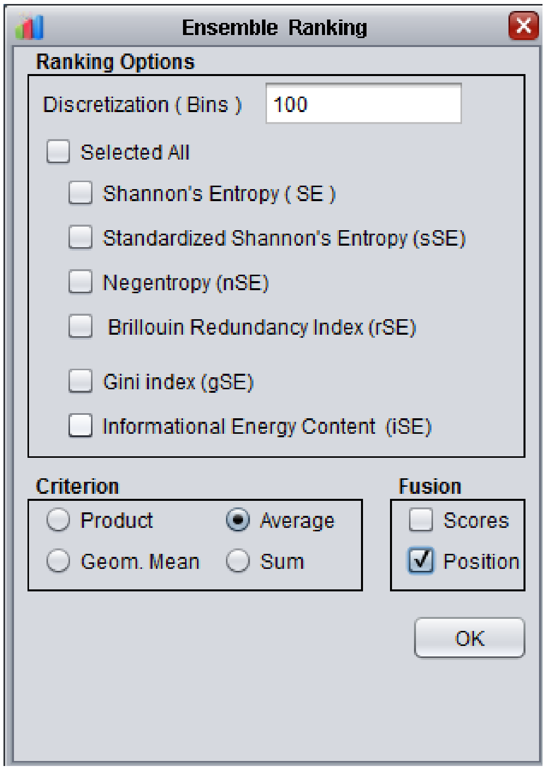
where D_{ij} is the normalized distance in the range [0.0 – 1.0] between the instances Xi and Xj.

Ranking Output Panel

Single: unicriteria(use only one parameter) is used to rank the variables (click on the drop-down list to select the parameter to use).

Ensemble: multicriteria(more than one parameter) is used to rank the variables. Selecting this option activates the Config(Configuration) button. Click the Config button to display the Ensemble Ranking window, in which the parameters, criterion, and fusion (multiple ranking type) to use are defined.

- Criterion: defines the mathematical operation employed to unify various entropy parameters into a single value.
- Fusion: Select Scores to use the numeric values for a variable or Rank to use the positions of a variable in a set of variables, obtained according to the selected parameters.



Ensemble Ranking

Ranking Options

Discretization (Bins) 100

□ Selected All

□ Shannon's Entropy (SE)

□ Standardized Shannon's Entropy (sSE)

□ Negentropy (nSE)

□ Brillouin Redundancy Index (rSE)

□ Gini index (gSE)

□ Informational Energy Content (iSE)

Criterion

○ Product ● Average

○ Geom. Mean ○ Sum

Fusion

□ Scores

☑ Position

Figure 42 Ensemble window

Features Selection Options

By default, the selected parameters are calculated for all the variables in an imported dataset file and sent to a report. The user may however wish to have a reduced report, i.e. one that contains a particular number of the best and/or worst variables, in terms of their entropy values. This option is configurable in the Top and Bottom text fields of the Ranking window. Selecting the threshold option permits the user to determine the minimum entropy value that the variables sent to the report should have.

Contribution options.

Permits the user to select variables according to their distribution with respect to the average. Three ranges are considered:

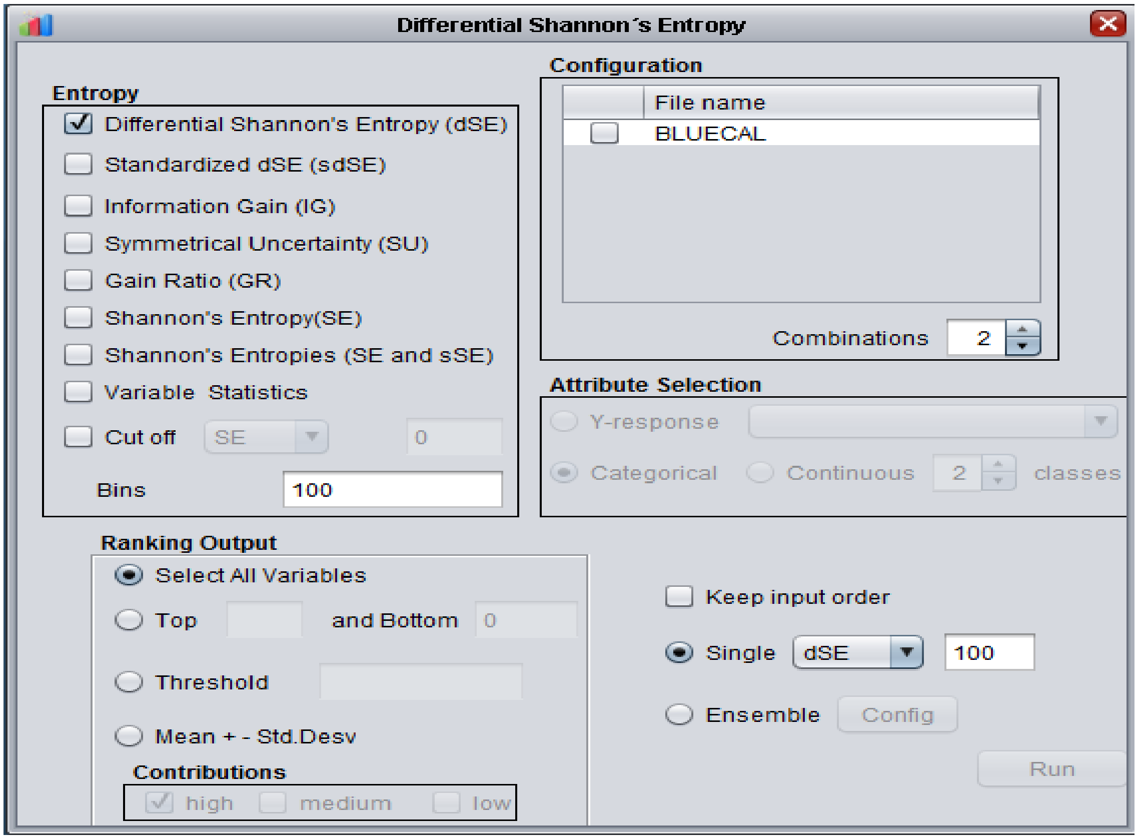
- *High*- variables with values greater than the arithmetic mean (AVG) + Standard Deviation (Std Dev).
- *Medium*- variables with values within the range [AVG + Std Dev; AVG – Std Dev].
- *Low*- variables with values less than the AVG - Std Dev.

Once all the necessary configurations are performed, click the run button to execute the calculations.

Mean - Std Dev < Mean < Mean + Std Dev

Differential Shannon’s Entropy Ranking

Differential Shannon entropy (dSE) ranking permits comparing Shannon’s Entropy variability for two or more compound populations. Click the process menu to select the Simple Ranking option, which in turn leads to the entropy type drop-down list from which Differential Shannon’s Entropy is selected to display the corresponding window. The Differential Shannon’s Entropy window is also accessible by clicking the configure DSE ranking icon (🔧) in toolbar menu.



Differential Shannon´s Entropy

Entropy
- ☑ Differential Shannon's Entropy (dSE)
- ☐ Standardized dSE (sdSE)
- ☐ Information Gain (IG)
- ☐ Symmetrical Uncertainty (SU)
- ☐ Gain Ratio (GR)
- ☐ Shannon's Entropy(SE)
- ☐ Shannon's Entropies (SE and sSE)
- ☐ Variable Statistics
- ☐ Cut off SE 0

Bins 100

Configuration

	File name
☐	BLUECAL

Combinations 2

Attribute Selection
- ○ Y-response
- ● Categorical ○ Continuous 2 classes

Ranking Output
- ● Select All Variables
- ○ Top and Bottom 0
- ○ Threshold
- ○ Mean + - Std.Desv

Contributions ☑ high ☐ medium ☐ low

- ☐ Keep input order
- ● Single dSE 100
- ○ Ensemble Config

Run

Figure 43 Differential Shannon’s Entropy Window

In the configuration Panel, two options are provided:

Database comparison: for this at least 2 datasets should be selected. In the Combinations field the user introduces the number of datasets to be compared simultaneously. For example, if the user selects 12 datasets and enters the number 2 in the combinations field, it means that all possible pair-wise comparisons of the 12 datasets are to be performed. Note that it is IMPERATIVE that the selected files have the same features calculated.

Attribute Selection: attribute selection in this context is related to the evaluation of the performance of a particular feature in different sets of the same datasets, formed on basis of a specific property or activity, particularly for classification purposes. Click the drop-down list to select the feature that you wish to evaluate. Note that for this option, the user must work with only one dataset file and have the classification criteria defined in the last column.

Entropies Panel

Differential Shannon's entropy

Compares the variability of features in terms of their degree of discrimination of datasets or classes and is defined as(Godden & Bajorath 2001):

$$DSE = SE_{1,2,3\ldots n} - (SE_1 + SE_2 + SE_3 + \ldots SE_n)/n$$

where, "SE1, 2, 3…n" is the SE calculated for the combination of n datasets or classes under consideration.

Standardized Shannon's Differential Entropy

Normalization of Differential Shannon entropy with respect to the maximum obtained SE value and it is calculated as (Godden & Bajorath 2001):

$$sdSE = \frac{DSE(U)}{\log_2 N}$$

where, N is the number of bins.

Information Gain (IG)

The IG of attribute X refers to the reduction in uncertainty about class attribute Y given that X is known and is mathematically defined as follows(Ian et al. 2011):

$$IG(X|Y) = H(X) - H(X|Y)$$

Gain Ratio (GR)

Normalization of the IG to compensate for the preference for the attribute with large number of values and is defined as(Quinlan 1983):

$$GR(X,Y) = \frac{IG(X|Y)}{H(X,Y)}$$

where H(X, Y) is the intrinsic information (entropy of distribution of instances into branches).

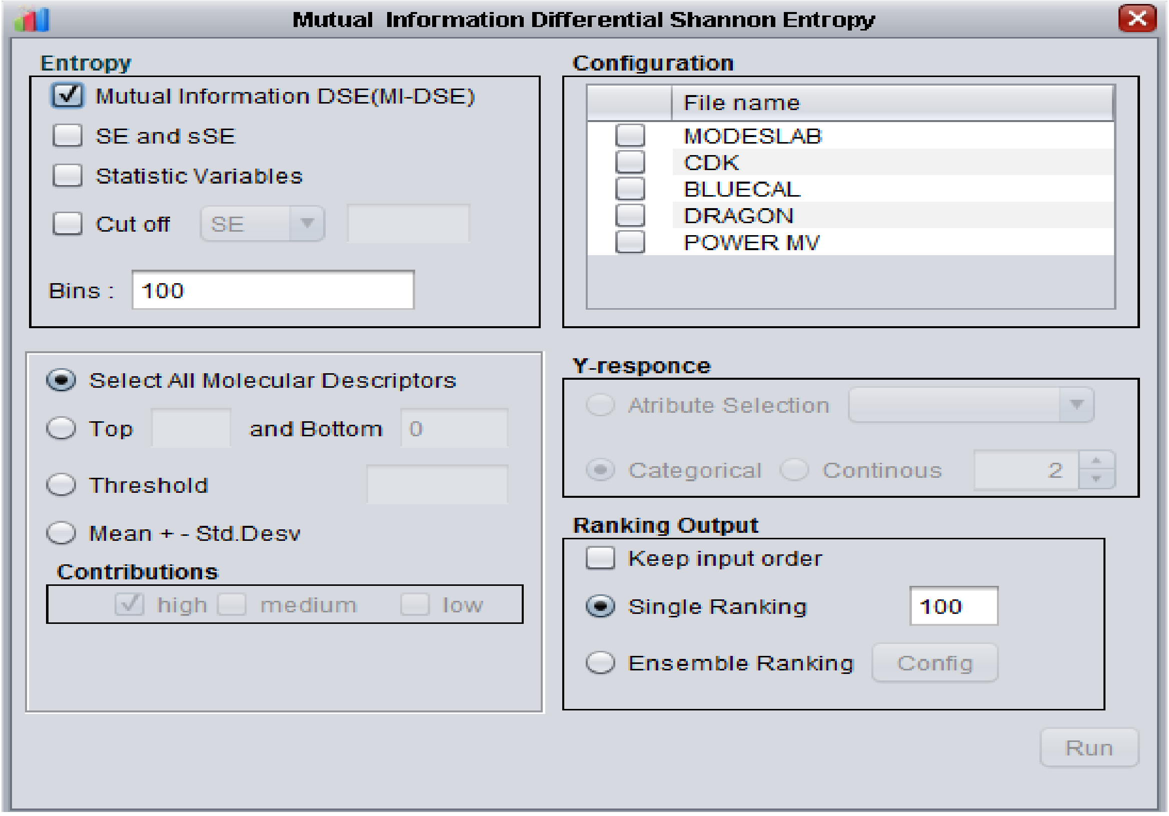
Symmetrical Uncertainty (SU)

The SU compensates for IG's bias towards attributes with more values and is mathematically expressed as (W H Press et al. 1988):

$$SU(X,Y) = \frac{IG(X|Y)}{SE(X) + SE(Y)}$$

Symmetrical Uncertainty normalizes its value to the range [0, 1], where 0 indicates that the attributes are completely independent while 1 indicates that each attribute predicts the values of the other.

Mutual Information Differential Shannon's Entropy Ranking



Mutual Information Differential Shannon Entropy

Entropy

☑ Mutual Information DSE(MI-DSE)

□ SE and sSE

□ Statistic Variables

□ Cut off SE

Bins : 100

Configuration

	File name
□	MODESLAB
□	CDK
□	BLUECAL
□	DRAGON
□	POWER MV

● Select All Molecular Descriptors

○ Top and Bottom 0

○ Threshold

○ Mean + - Std.Desv

Contributions

☑ high □ medium □ low

Y-responce

○ Atribute Selection

● Categorical ○ Continous 2

Ranking Output

□ Keep input order

● Single Ranking 100

○ Ensemble Ranking

Figure 44 Mutual Information Differential Shannon's Entropy Window

Entropies Panel

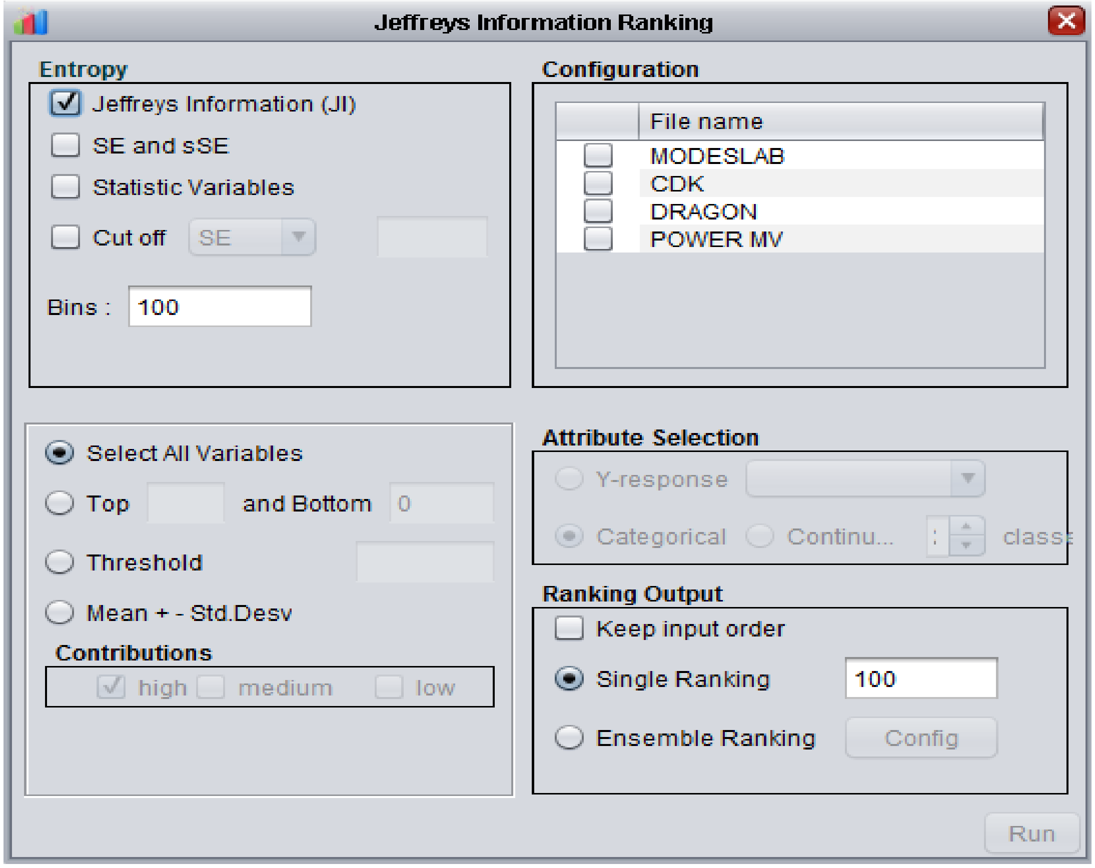
Mutual Information Differential Shannon's Entropy (MI-DSE)

An extension of the DSE to compare the variability of attributes with different number of instances (Anne Mai et al. 2010). The MI-DSE is mathematically expressed as follows:

$$MIDSE(X) = SE_{norm}(X,Y) - \frac{SE(X) - SE(Y)}{2}$$

Jeffrey's Information Ranking

Jeffreys Information, also known as Symmetrical Kullback-Leibler Entropy is a measure of the difference between two probability distributions. Click the process menu to select the Simple Ranking option, which in turn leads to the entropy type pull-down list from which Jeffreys Information is selected to display the corresponding window. The Jeffreys Information window is also accessible by clicking the configure JI ranking icon (🔧) in toolbar menu.



Jeffreys Information Ranking

Entropy

☑ Jeffreys Information (JI)

□ SE and sSE

□ Statistic Variables

□ Cut off SE

Bins : 100

Configuration

	File name
□	MODESLAB
□	CDK
□	DRAGON
□	POWER MV

● Select All Variables

○ Top and Bottom 0

○ Threshold

○ Mean + - Std.Desv

Contributions

☑ high □ medium □ low

Attribute Selection

○ Y-response

● Categorical ○ Continu... class

Ranking Output

□ Keep input order

● Single Ranking 100

○ Ensemble Ranking Config

Run

Figure 45 Jeffrey's Information Window

Entropies Panel

Jeffrey's Information (JI)

Symmetric definition for Kullback-Leibler entropy (or divergence) aimed at dealing with the subjectivity that accompanies the categorization of features as experimental or theoretical, a requirement for the Kullback-Leibler entropy computation (William H Press et al. 2007). Therefore, the JI is an unbiased measure for similarity (or dissimilarity) of probability distributions and mathematically expressed as:

$$JI(X\|Y) = \sum_{i=0}^{N} p_i(X)\log p_i(\frac{X}{Y}) + p_i(Y) * \log p_i(\frac{Y}{X})$$

Graphic Analysis

Once the datasets are loaded, the user may wish to perform graphic distribution analyses of the variables in terms of their entropy (or information) values. To carry out the graphic analysis, click the process menu to display the drop-down list , from which the *graphic* options, also accesible in the toolbar menu, are selected. IMMAN provides a comprehensive selection of graphic methods. These include: Shannon's Distribution Graphs, Histograms, Importance and Correlations Graphs, selected depending on the selected entropy type.

Shannon's Distribution Graph

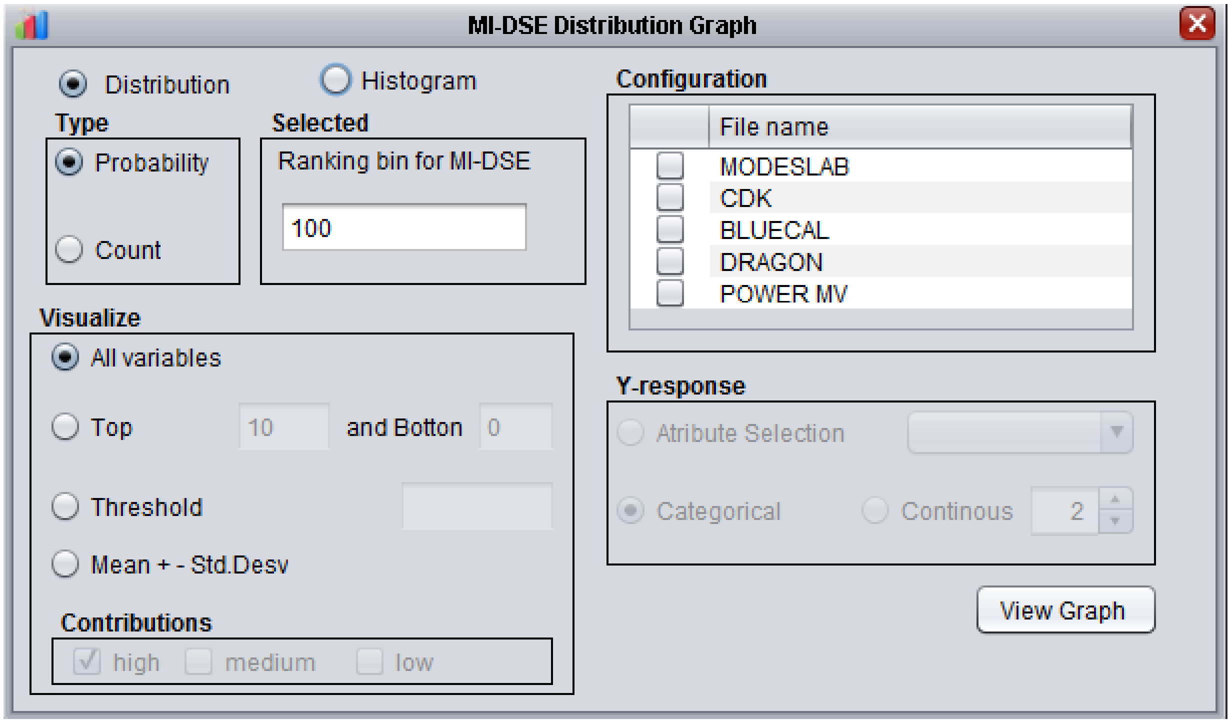
Two options are provided for Shannon's Distribution Graphs, i.e. a cummulative line plots and bar graphs. The *Type* panel allows the user to determine the Y-axis scaling scheme for the number of variables in the datasets. Two alternatives are provided:

- Probability: this is a normalization procedure for the number of variables. Given an entropy value of x bits, zero probability means none of the variables has entropy values equal or greater than x, while a probability value of one means all the variables have entropy values equal or more than x.
- Count: refers to a simple tally of the number of variables with entropy values equal or superior to a particular value.

In the *Entropy* panel, the user defines the discretization scheme and desired entropy parameter, selected from the drop-down list.

The *Dataset Selection* panel displays the loaded dataset files. The user must check the dataset(s) to be analyzed in the graphic representation.

The *Visualize* panel is used define which variables are to be observed in the graphic representation.



MI-DSE Distribution Graph

● Distribution ○ Histogram

Type: ● Probability ○ Count

Selected: Ranking bin for MI-DSE 100

Configuration

	File name
□	MODESLAB
□	CDK
□	BLUECAL
□	DRAGON
□	POWER MV

Visualize: ● All variables ○ Top 10 and Botton 0 ○ Threshold ○ Mean + - Std.Desv

Contributions: ☑ high □ medium □ low

Y-response: ○ Atribute Selection ● Categorical ○ Continous 2

Figure 46 Mutual Information Differential Shannon'a Entropy Distribution Graph Window

Below is an example of cummulative line plot and a bar graph of variable distribution.

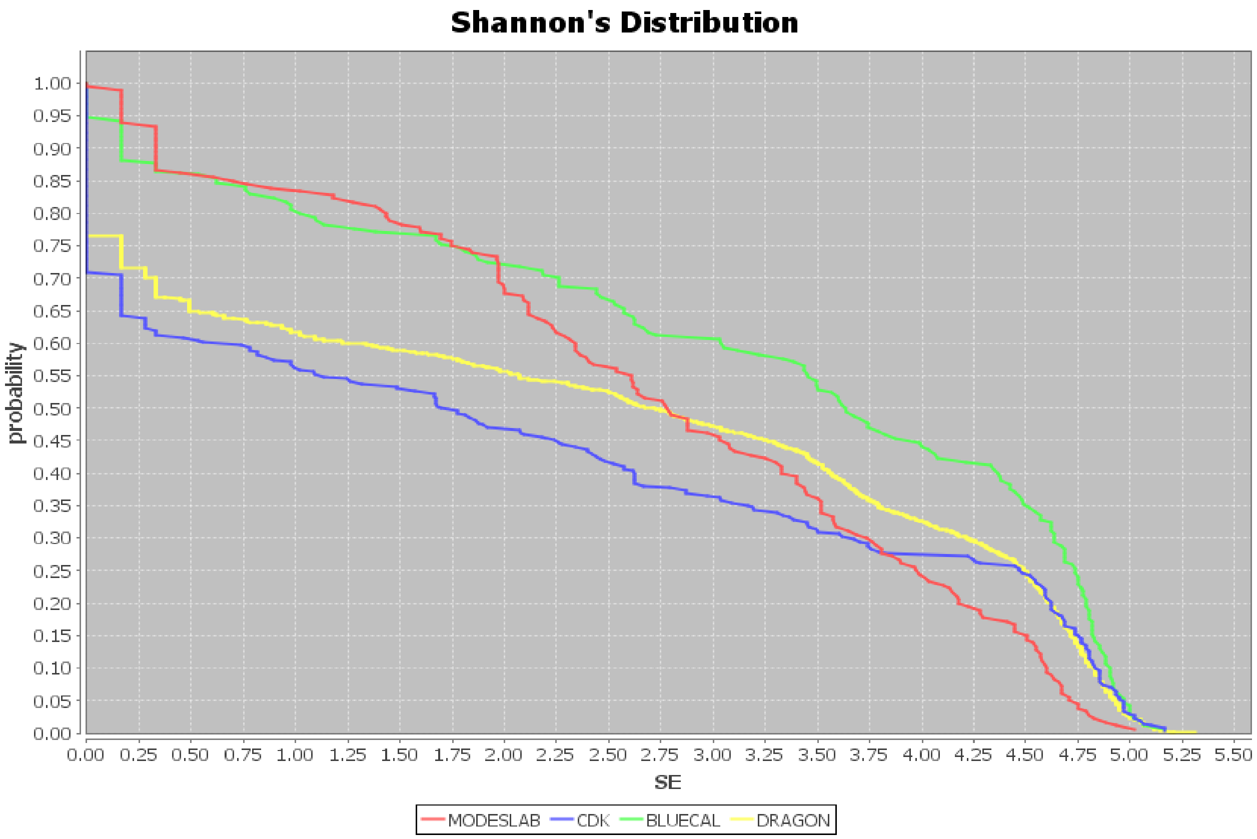


Figure 47 Shannon Distribution Graph

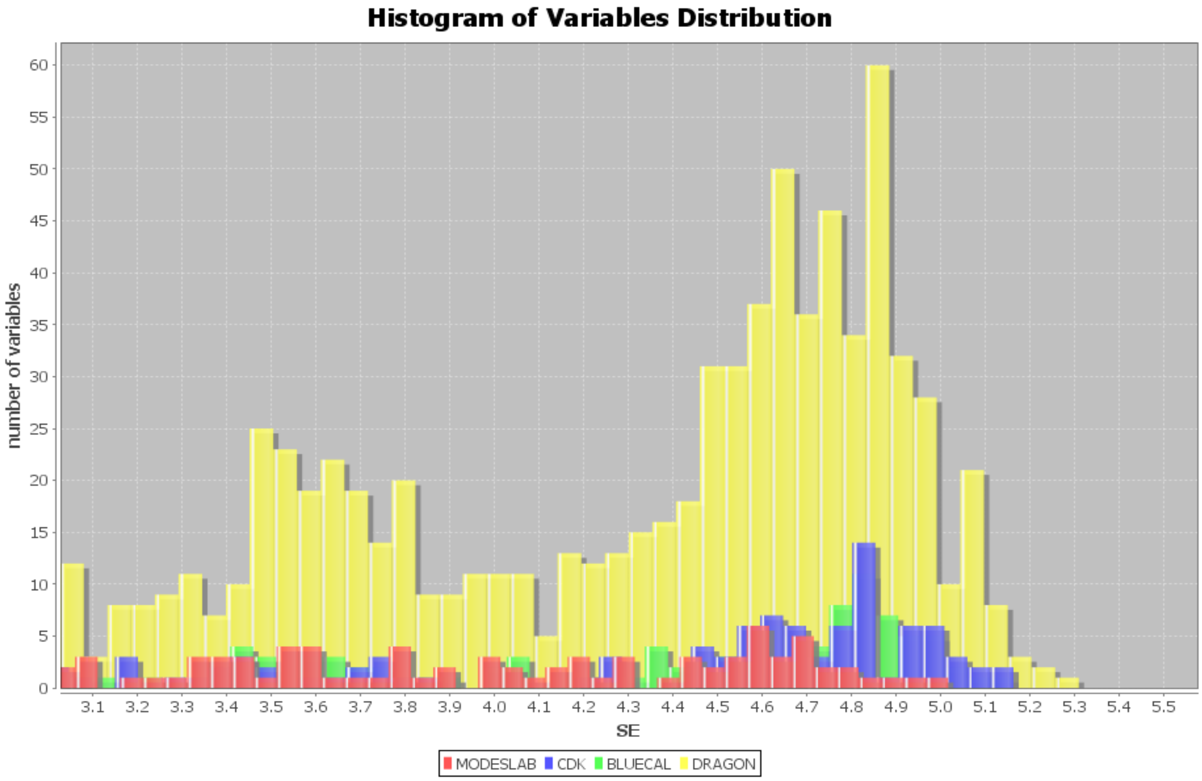
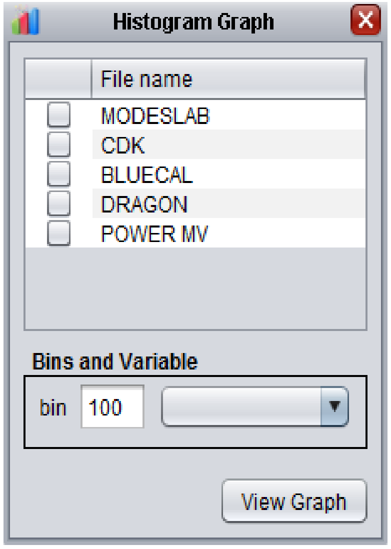


Figure 48 Histogram of Shannon Distribution Graph

Histogram Graph

This Graphic method is used to represent the distribution of features values over discrete intervals.



Histogram Graph

	File name
□	MODESLAB
□	CDK
□	BLUECAL
□	DRAGON
□	POWER MV

Bins and Variable

bin 100

Figure 49 Histogram Graph Window

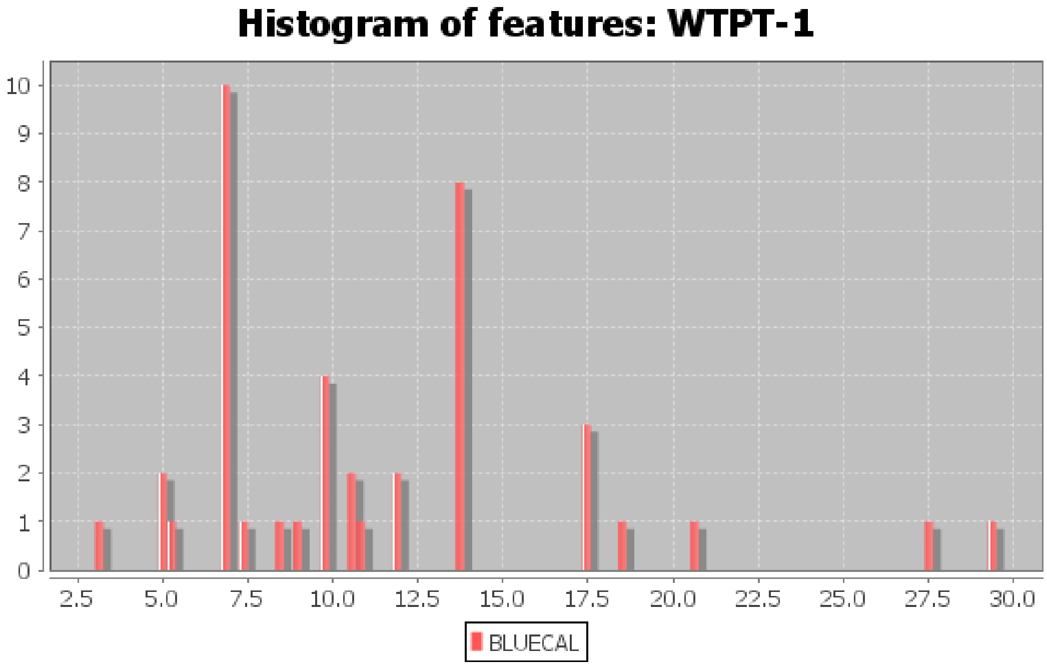


Figure 50 Histogram of feature

Importance Graph

This graphic representation shows the performance of the variables, arranged in a descending order in terms of their entropy or information values . Below is an an example of the Importance Graph.

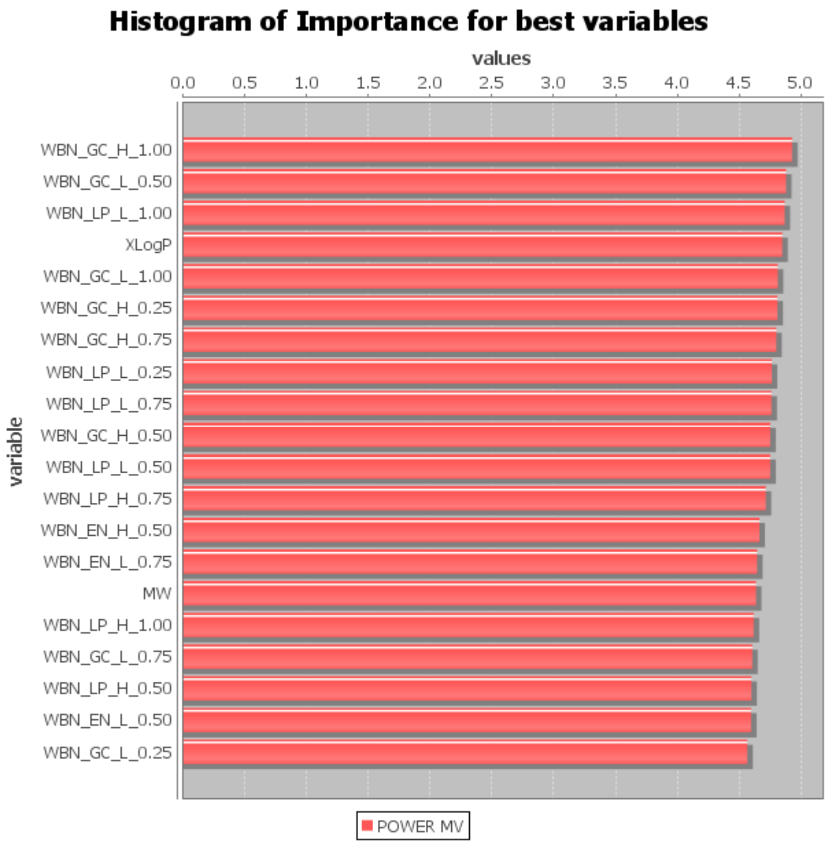


Figure 51 Importance of Best Features

Correlations Graph

Shows the homogeneity between two dataset files. For this graphic representation, the datasets files must contain the same number of variables. Below is an example of a correlations graph.

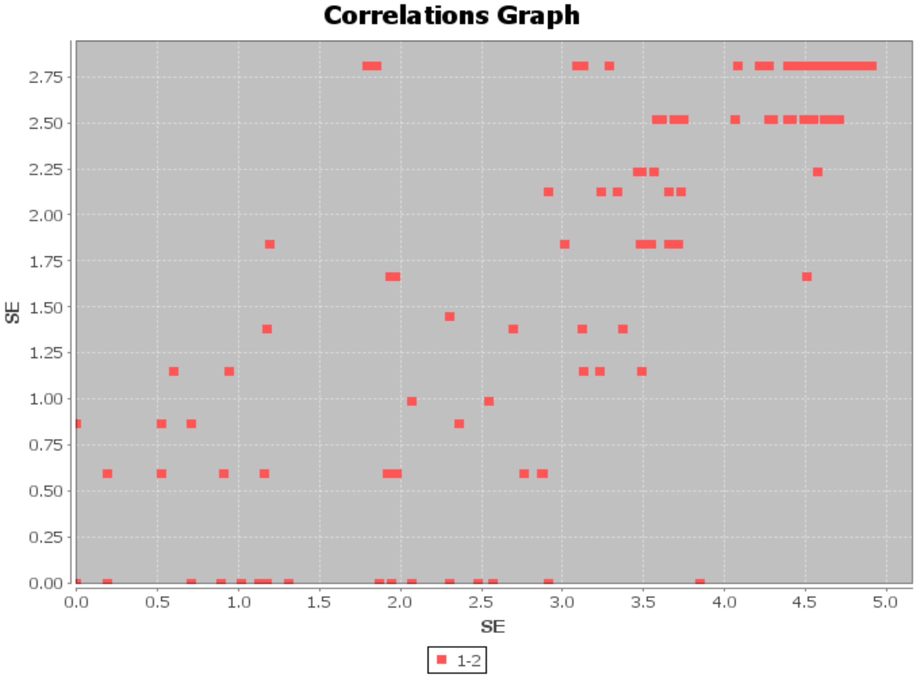


Figure 52 Correlation of values Features

Input and Output Files

The following is a description of the **Input** and **Output** section of the **IMMAN** GUI. In these sections, input structure files can be loaded (see Figure 53).

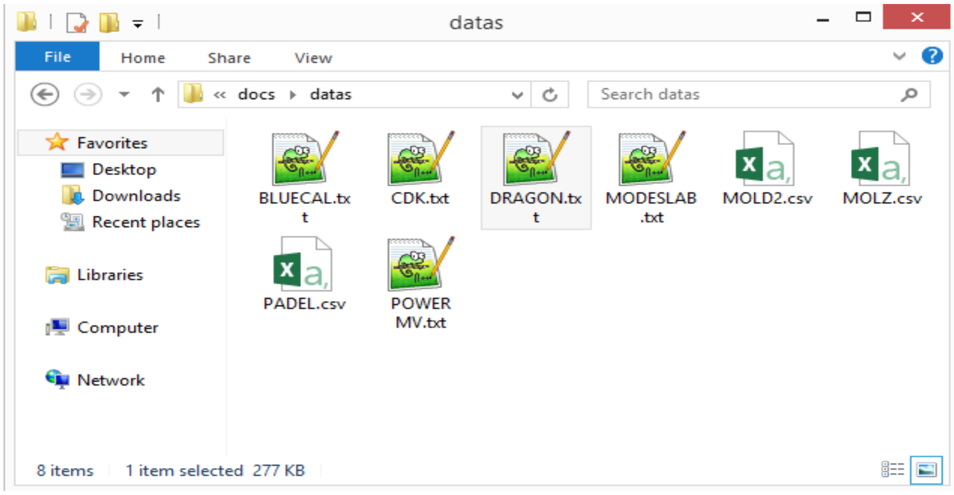


Figure 53 Browsing Input and Output files

Supported File Formats

Tab and Comma Separated Value Files (TXT, CSV)

A **tab-separated values** file is a simple text format for a database table. Each record in the table is one line of the text file. Each field value of a record is separated from the next by a space (or blank) character, it is a form of the more general delimiter-separated values format.

As file extension for this output file we choose TXT, because it is a simple file format that is widely supported, so it is often used to move spaced data between different computer programs that support the format. For example, a space-separated file might be used to transfer information from a database program to a spreadsheet.

TXT is an alternative to the common comma-separated values (CSV) format, which often causes difficulties because of the need to escape commas. Literal commas are very common in text data.

A **comma-separated value (CSV)** file stores tabular data (numbers and text) in plain-text form. A plain text form means that the file is a sequence of characters, with no data that has to be interpreted instead, as binary numbers. A CSV file consists of any number of records, separated by line breaks of some kind; each record consists of fields, separated by some other character or string, most commonly a literal comma or tab. Usually, all records have an identical sequence of fields.

CSV is a common, relatively simple file format that is widely supported by consumer, business, and scientific applications. Among its most common uses is moving tabular data between programs that natively operate on incompatible (often proprietary and/or undocumented) formats. This works because so many programs support some variation of CSV at least as an alternative import/export format.

Files Created for IMMAN

IMMAN produce an output file containing the values of the selected features, together with the additional information imported by the user, this file is the saved report.

The **standard IMMAN format** for the output dataset(.txt) is organized as follows (this format, namely, array of features blocks, cannot be changed by the user, see Table 6 for a simple example):

- The *first record* (column) contains the name of instance.
- The following records (columns) contain the **variable labels** *(features headers), i.e.* **Attribute1**.

Table 1 An example Output file

Cases name	Attribute1	Attribute2
STEsteroids2COR3D.sdf_aldosterone	55.57641698	191.2661335
STEsteroids2COR3D.sdf_androstanediol	69.5912	235.2749
STEsteroids2COR3D.sdf_5-androstenediol	69.5912	235.2749
STEsteroids2COR3D.sdf_4-androstenedione	58.7208	204.4986
STEsteroids2COR3D.sdf_androsterone	63.92535441	219.3476411
STEsteroids2COR3D.sdf_corticosterone	63.92535441	219.3476411
STEsteroids2COR3D.sdf_cortisol	65.02036164	222.4444088
STEsteroids2COR3D.sdf_cortisone	62.84878819	216.2939852
STEsteroids2COR3D.sdf_dehydroepiandrosterone	63.92535441	219.3476411
STEsteroids2COR3D.sdf_11-deoxycorticosterone	62.1411688	214.2818688
STEsteroids2COR3D.sdf_11-deoxycortisol	63.92535441	219.3476411
STEsteroids2COR3D.sdf_dihydrotestosterone	63.92535441	219.3476411
STEsteroids2COR3D.sdf_estradiol	69.5912	235.2749

Example Data

Example data sets are provided to enable the user to carry out trial operations with IMMAN for familiarization purposes. These example files are stored in the Data subdirectory of the IMMAN installation directory, and contain variables calculated for a dataset of 41 heterogeneous molecules using MD calculating programs well-known in the literature.

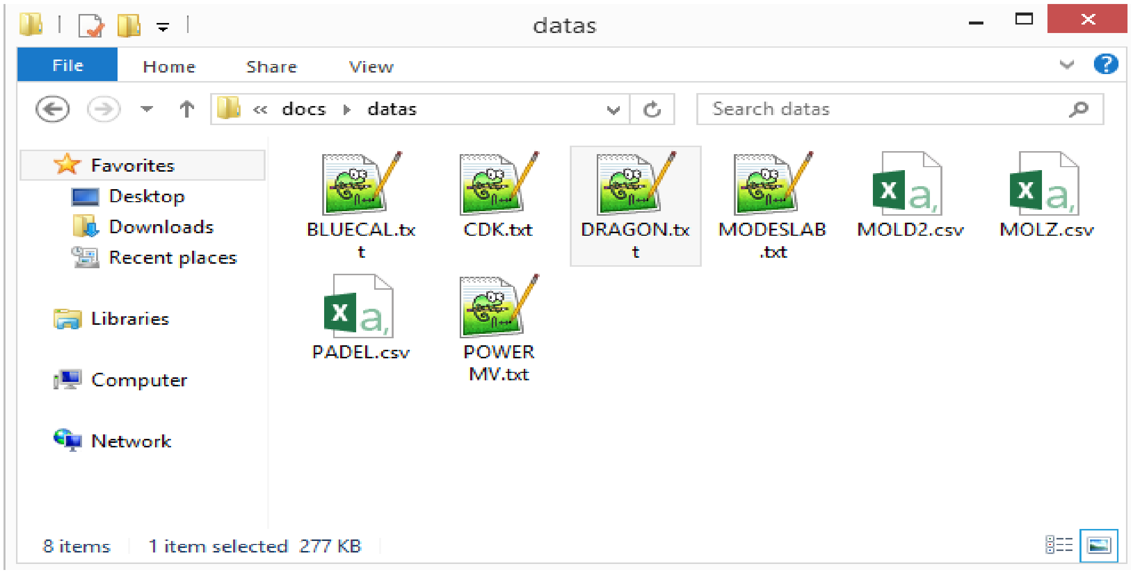


Figure 54 Browsing example data: DRAGON file

References

Alter, O., Brown, P.O. & Botstein, D., 2000. Singular value decomposition for genome-wide expression data processing and modeling.

Anne Mai, W. et al., 2010. Identification of Descriptors Capturing Compound Class-Specific Features by Mutual Information Analysis.

Breiman, L. et al., 1984. Classification and Regression Trees. *Wadsworth International Group*.

Brillouin, L., 1962. *Science and information theory*, Academic Press.

Godden, J.W. & Bajorath, J., 2001. Chemical Descriptors with Distinct Levels of Information Content and Varying Sensitivity to Differences between Selected Compound Databases Identified by SE-DSE Analysis. *Computational Chemistry and Informatics, New Chemical Entities*, p.6.

Ian, H.W., Eibe, F. & Mark, A.H., 2011. *Data Mining. Practical Machine Learning Tools and Techniques* ,

Kier, L.B., 1980. Use of molecular negentropy to encode structure governing biological activity. . *J. Pharm. Sci.*, pp.807–810pp.

Onicescu, O., 1966. Energie informationelle. *Comp.Rend*, pp.841–842.

Press, W H et al., 1988. Numerical Recipes in C. *Cambridge University Press*.

Press, William H et al., 2007. *Numerical Recipes.The Art of Scientific Computing* 3rd ed., Cambridge University Press.

Quinlan, J., 1983. Learning efficient classification procedures and their application to chess end games. *Machine Learning: An Artificial Intelligence Approach*.

Shannon, C.E., 1948. A Mathematical Theory of Communication. *The Bell System Technical Journal,System*, 27, pp.379–423.