DatasetDivisionGUI 1.0:

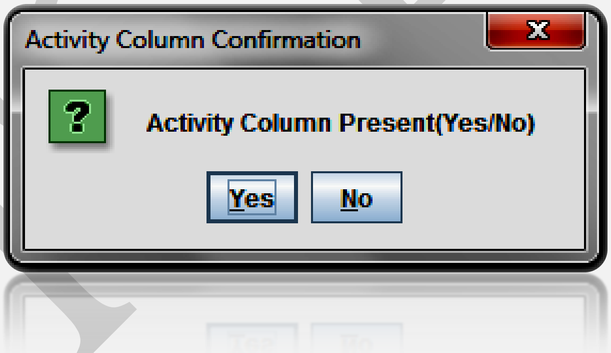
The purpose of this application tool is to perform rational selection of training and test set using Kennard Stone algorithm, diversity based and activity based division method.

Input file: Single *.csv file (saved as 'comma delimited' in Microsoft excel), consist of all compounds in dataset with their descriptor information.
Input file format: Columns containing descriptor and their values for each compound present in dataset set. *Header**(e.g. name of respective descriptor, activity title) for each column is must. *Serial numbers** must be present in the first column. If user wants to keep activity column, keep it in last column. So that activity column is avoided during calculation, and only printed in output file along with respective training set and test set compounds. This avoids entry of activity values manually after training and test set selection. To further clarify file format, sample input file is provided.

To run the program: Download *zip file*, extract and just do single/double click on jar file to run the program.
Note: For user convenience, user may keep dataset file (.csv) in "data" folder and also may select "output" folder for storing output files.

Dialog boxes reference:

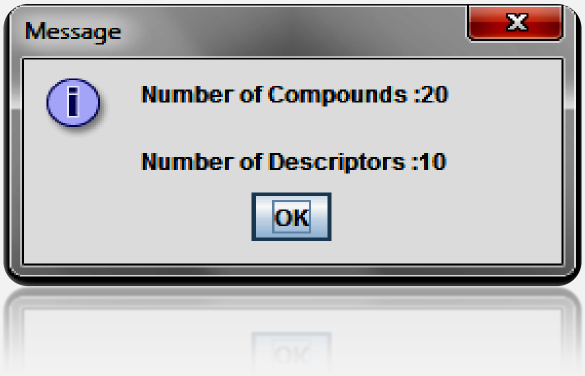
1. User can keep activity column (*last column in input file*) to avoid entry of activity values manually after training and test set selection. This step is to verify whether activity column is present or not.



Activity Column Confirmation

Activity Column Present(Yes/No)

2. This is an important step. Please check both values carefully and see if both are correct. If not, check your input file. If you find your input file correct and error still exist, then create input file freshly once again (by copying all observations and paste it in new excel sheet and save it as csv file). This should eliminate the error.

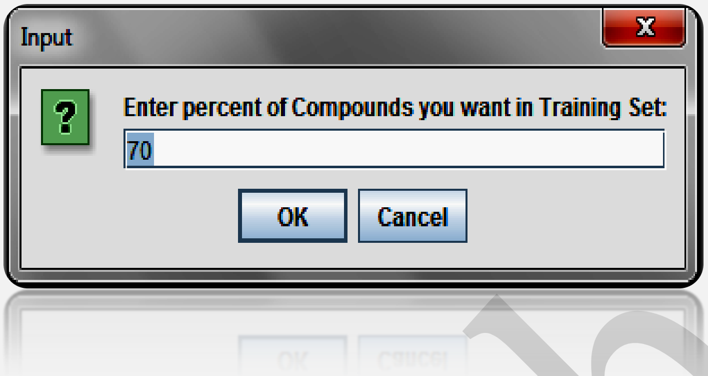


Message

Number of Compounds :20

Number of Descriptors :10

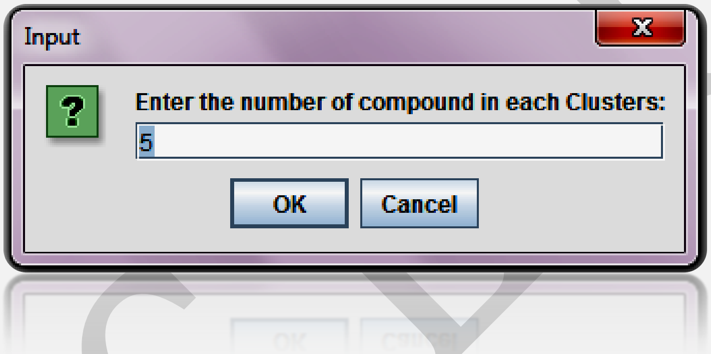
3. Enter a number which will corresponds to the percentage of compounds that will get selected in training set (default value is 70 that means *70 %*).



Enter percent of Compounds you want in Training Set:

70

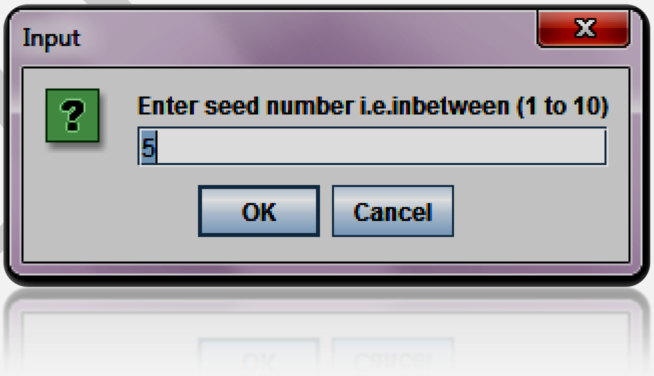
4. Enter the number of compounds in each cluster depending on which clusters are formed.



Enter the number of compound in each Clusters:

5

5. Enter the seed number in-between the range displayed (which is based on number of compound present in each cluster). Different seed number leads to different division and for reproducing the division, you have to use same seed number.
   Note: You can try different seed number and/or number of compounds in each cluster, till you find a better division.



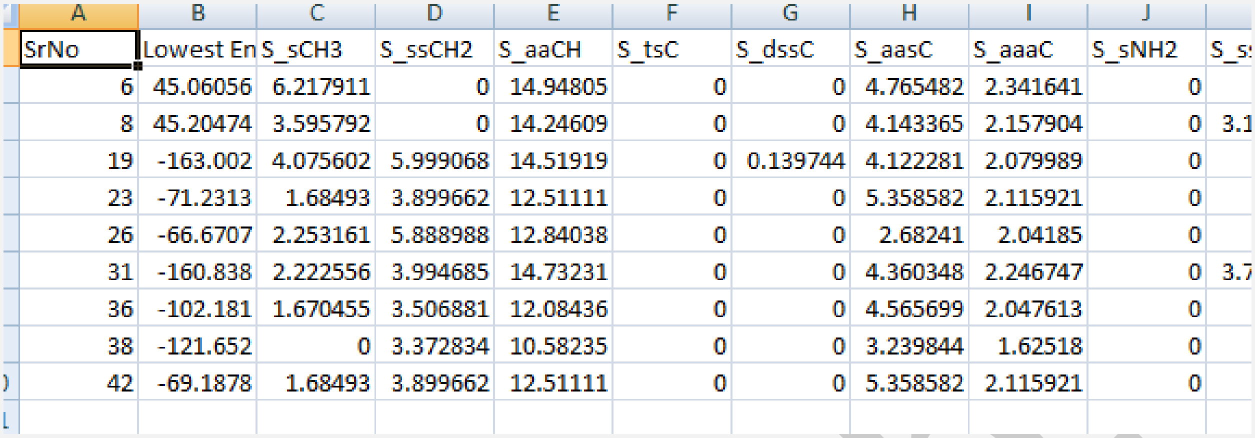
Enter seed number i.e.inbetween (1 to 10)

5

Output files:

This output files consists of resulting training and test set with all information which were present in input file.

Snapshot: [InputFileName]_Train.csv/ [InputFileName]_Test.csv in Microsoft Excel



A	B	C	D	E	F	G	H	I	J	
SrNo	Lowest En	S_sCH3	S_ssCH2	S_aaCH	S_tsC	S_dssC	S_aasC	S_aaaC	S_sNH2	S_s
6	45.06056	6.217911	0	14.94805	0	0	4.765482	2.341641	0	
8	45.20474	3.595792	0	14.24609	0	0	4.143365	2.157904	0	3.1
19	-163.002	4.075602	5.999068	14.51919	0	0.139744	4.122281	2.079989	0	
23	-71.2313	1.68493	3.899662	12.51111	0	0	5.358582	2.115921	0	
26	-66.6707	2.253161	5.888988	12.84038	0	0	2.68241	2.04185	0	
31	-160.838	2.222556	3.994685	14.73231	0	0	4.360348	2.246747	0	3.7
36	-102.181	1.670455	3.506881	12.08436	0	0	4.565699	2.047613	0	
38	-121.652	0	3.372834	10.58235	0	0	3.239844	1.62518	0	
42	-69.1878	1.68493	3.899662	12.51111	0	0	5.358582	2.115921	0	

Disclaimer

For academic purpose only.

The program DatasetDivisionGUI 1.0 has been developed (in Java) and validated on known data sets by Pravin Ambure (ambure.pharmait@gmail.com) of Drug Theoretics & Cheminformatics (DTC) Laboratory, Jadavpur University (2013).
Note: For any query/suggestions, please feel free to contact us via email.

References:

1. Kennard, R. W.; Stone, L. A., Computer aided design of experiments. *Technometrics* 1969, 11, (1), 137-148.

2. Martin, T. M.; Harten, P.; Young, D. M.; Muratov, E. N.; Golbraikh, A.; Zhu, H.; Tropsha, A., Does Rational Selection of Training and Test Sets Improve the Outcome of QSAR Modeling? *J. Chem. Inf. Model.* 52, (10), 2570-2578.