

sampsio

# Suggested Use of the SAMPSIO Data Sets

## Introduction

The SAMPSIO library contains both real and fictitious data sets. You use the DMA[xxxx] data sets and a Data Partition node to create your training, validation, and test data. The DML[xxxx] data sets contains input and target values and can be used for training. You can the DMT\_ data sets as test data for comparing models. The are few data sets available as validation (DMV[xxxx]) and score (DMS[xxxx]) data sets.

Note: The DMD[xxxx] data sets are not suitable to be used in the Enterprise Miner graphic user interface environment.



## List of Data Sets

Here is a list of the data set that are available to use in the Enterprise Miner.

Name of Data Set	Number of Observations	Number of Variables
<a href="#">ASSOCS</a>	7007	3
<a href="#">DMABASE</a>	322	23
<a href="#">DMABOOL</a>	4096	24
<a href="#">DMACUSH</a>	27	6
<a href="#">DMAGECR</a>	1000	21
<a href="#">DMAGESCR</a>	75	21
<a href="#">DMAHART</a>	303	14
<a href="#">DMAHMEQ</a>	5960	13
<a href="#">DMAIRIS</a>	150	5
<a href="#">DMAMUSH</a>	8124	23
<a href="#">DMEXA1</a>	1966	50
<a href="#">DMLBALL</a>	500	6
<a href="#">DMLBASE</a>	163	23
<a href="#">DMLBLDG</a>	2926	13
<a href="#">DMLBLOC</a>	2000	7
<a href="#">DMLBOOL</a>	2048	24

<a href="#">DMLCENS</a>	32561	17
<a href="#">DMLCUBE</a>	500	6
<a href="#">DMLCUSH</a>	21	6
<a href="#">DMLGECR</a>	600	21
<a href="#">DMLHART</a>	203	14
<a href="#">DMLHMEQ</a>	3576	13
<a href="#">DMLIRIS</a>	120	5
<a href="#">DMLMAGL</a>	3000	6
<a href="#">DMLMNK3</a>	122	7
<a href="#">DMLMUSH</a>	4062	23
<a href="#">DMLRING</a>	180	3
<a href="#">DMLSINE</a>	127	5
<a href="#">DMLSONA</a>	104	61
<a href="#">DMLSPIR</a>	194	3
<a href="#">DMLXOR</a>	4	3
DMSCUSH	2665	2
<a href="#">DMSRING</a>	510	2
<a href="#">DMSSPIR</a>	4225	2
<a href="#">DMTBALL</a>	1000	6
<a href="#">DMTBASE</a>	100	23
<a href="#">DMTBLDG</a>	1282	13
<a href="#">DMTBLOC</a>	1000	7
<a href="#">DMTBOOL</a>	2048	24
<a href="#">DMTCENS</a>	16281	17
<a href="#">DMTCUBE</a>	1000	6
<a href="#">DMTCUSH</a>	6	6
<a href="#">DMTGECR</a>	400	21
<a href="#">DMTHART</a>	100	14
<a href="#">DMTHMEQ</a>	1192	13
<a href="#">DMTIRIS</a>	30	5

<a href="#">DMTMAGL</a>	500	6
<a href="#">DMTMNK3</a>	432	7
<a href="#">DMTMUSH</a>	4062	23
<a href="#">DMTSINE</a>	1881	2
<a href="#">DMTSONA</a>	104	61
<a href="#">DMVHMEQ</a>	1192	13
<a href="#">HMEQ</a>	5960	13
<a href="#">MSEQ2</a>	564	3

ASSOCS

Use this data to run the Association node. For an association analysis, set the model role of TIME to rejected. For a sequence analysis, set the model role of TIME to sequence.

Variable	Model Role	Measurement	Description
CUSTOMER	id	interval	Customer ID
PRODUCT	target	nominal	Product of purchase
TIME	sequence or rejected	ordinal	Time of purchase

DMABASE, DMLMBASE, DMTBASE

Variable	Model Role	Measurement	Description
cr_atbat	input	interval	Career times at bat
cr_bb	input	interval	Career Walks
cr_hits	input	interval	Career Hits
cr_home	input	interval	Career Home Runs
cr_rbi	input	interval	Career RBIs
cr_runs	input	interval	Career runs
division	input	binary	Division at the end of 1986
league	input	binary	League at the end of 1986
logsalar	target	interval	Log Salary
name	id	nominal	Player's name
no_assts	input	interval	Assists in 1986

no_atbat	input	interval	Times at Bat in 1986
no_bb	input	interval	Walks in 1986
no_error	input	interval	Errors in 1986
no_hits	input	interval	Hits in 1986
no_home	input	interval	Home Runs in 1986
no_outs	input	interval	Put Outs in 1986
no_rbi	input	interval	RBIs in 1986
no_runs	input	interval	Runs in 1986
position	input	nominal	Position(s) in 1986
salary	rejected	interval	1987 Salary in \$Thousands
team	input	nominal	Team at the end of 1986
yr_major	input	interval	Years in the Major Leagues

DMABOOL, DMLBOOL, DMTBOOL

Variable	Model Role	Measurement	Description
and	target	binary	x1 AND x2
dnf3	target	binary	3-term dnf function
dnf4	target	binary	4-term dnf function
dnf5	target	binary	5-term dnf function
linear	target	binary	a linear function
linearb	target	binary	another linear function
parhalf	target	binary	mod(floor(sum/2),2)
parity	target	binary	parity
random	target	binary	random p=.5
ransum	target	binary	random function of sum
sum	target	interval	summation of x's
x1	input	binary	bit
x2	input	binary	bit
x3	input	binary	bit
x4	input	binary	bit

x5	input	binary	bit
x6	input	binary	bit
x7	input	binary	bit
x8	input	binary	bit
x9	input	binary	bit
x10	input	binary	bit
x11	input	binary	bit
x12	input	binary	bit
xor	target	binary	x1 XOR x2

DMACUSH, DMLCUSH, DMTCUSH

Variable	Model Role	Measurement	Description
label	id	nominal	Identification number
lpregnan	input	interval	log(pregnan)
ltetra	input	interval	log(tetra)
pregnan	input	interval	pregnanetriol mg/24h
tetra	input	interval	tetrahydrocortisone mg/24h
type	target	nominal	type of syndrome: a, b, c, or other

Reference:

Aitchison, J. and Dunsmore, I.R. (1975), Statistical Prediction Analysis, Cambridge: Cambridge University Press.  
Ripley (1996)

DMAGECR, DMLGECR, DMTGECR, DMAGESCR

DMAGESCR data does not contain the good\_bad variable and can be used as a score data.

Variable	Model Role	Measurement	Description
age	input	interval	age in years
amount	input	interval	credit amount
checking	input	nominal or ordinal	status of existing checking account  1: ... < 0 DM  2: 0 <= ... < 200 DM

			3: ... >= 200 DM 4: no checking account
coapp	input	nominal	other debtors/guarantors  1: none  2: co-applicant  3: guarantor
depends	input	interval	number of dependents
durations	input	interval	duration in months
employed	input	ordinal	present employment since  1: unemployed  2: ... < 1 year  3: 1 <= ... < 4 years  4: 4 <= ... < 7 years  5: ... >= 7 years
exister	input	interval	number of existing credits at this bank
foreign	input	binary	foreign worker  1: yes  2: no
good_bad	target	binary	credit rating
history	input	ordinal	credit history  0: no credits taken / all credits paid back duly  1: all credits at this bank paid back duly  2: existing credits paid back duly till now  3: delay in paying off in the past  4: critical account / other credits existing (not at this bank)
housing	input	nominal	housing  1: rent

			2: own 3: for free
installp	input	interval	installment rate in percentage of disposable income
job	input	ordinal	job 1: unemployed / unskilled non-resident 2: unskilled resident 3: skilled employee / official 4: management / self-employed / highly qualified employee / officer
marital	input	nominal	personal status and sex 1: male -- divorced / separated 2: female -- divorced / separated / married 3: male -- single 4: male -- married / widowed 5: female -- single
other	input	nominal	other installment plans 1: bank 2: stores 3: none
property	input	nominal or ordinal	property 1: real estate 2: if not 1, building society savings agreement / life insurance 3: if not 1 or 2, car or others 4: unknown / no property
purpose	input	nominal	purpose 0: new car 1: used car

			2: furniture / equipment 3: radio / television 4: domestic appliances 5: repairs 6: education 7: vacation 8: retraining 9: business x: others
resident	input	interval	present residence since
savings	input	nominal or ordinal	status of existing saving account or bonds  1: ... < 100 DM 2: 100 <= ... < 500 DM 3: 500 <= ... < 1,000 DM 3: ... >= 1,000 DM 4: unknown / no saving account
telephon	input	binary	telephone  1: none  2: yes, registered under the customer's name

DMAHART, DMLHART, DMTHART

Variable	Model Role	Measurement	Description
age	input	interval	age
bpress	input	interval	resting blood pressure
bsugar	input	binary	fasting blood sugar > 120 mg/dl
ca	input	ordinal	number of major vessels (0-3) colored by flourosopy
chol	input	interval	serum cholestoral in mg/dl



ekg	input	nominal	resting electrocardiographic results
exang	input	binary	exercise induced angina
oldpeak	input	interval	ST depression induced by exercise relative to rest
pain	input	nominal	chest pain type
sex	input	binary	sex
slope	input	ordinal	slope of the peak exercise ST segment
target	target	ordinal	number of major vessels (0-4) reduced in diameter by more than 50%
thal	input	nominal	thal
thalach	input	interval	maximum heart rate achieved

DMAHMEQ, DMLHMEQ, DMVHMEQ, DMTHMEQ, HMEQ

Variable	Model Role	Measurement	Description
bad	target	binary	default or seriously delinquent
clage	input	interval	age of oldest trade line in months
clno	input	interval	number of trade (credit) lines
debtinc	input	interval	debt to income ratio
delinq	input	interval	number of delinquent trade lines
derog	input	interval	number of major derogatory reports
job	input	nominal	job category
loan	input	interval	amount of current loan request
mortdue	input	interval	amount due on existing mortgage
ninq	input	interval	number of recent credit inquiries
reason	input	binary	home improvement or debt consolidation
value	input	interval	value of current property
yoj	input	interval	years on current job

DMAIRIS, DMLIRIS, DMTIRIS

--	--	--	--

Variable	Model Role	Measurement	Description
petallen	input	interval	petal length in mm
petalwid	input	interval	petal width in mm
sepalen	input	interval	sepal length in mm
sepalwid	input	interval	sepal width in mm
species	target	nominal	species of iris

DMAMUSH, DMLMUSH, DMTMUSH

Variable	Model Role	Measurement	Description
bruises	input	nominal	bruises
capcolor	input	nominal	cap color
capshape	input	nominal	cap shape
capsurf	input	nominal	cap surface
gillatta	input	nominal	gill attachment
gillcolo	input	nominal	gill color
gillsize	input	nominal	gill size
gillspac	input	nominal	gill spacing
habitat	input	nominal	habitat
odor	input	nominal	odor
populat	input	nominal	population
ringnumb	input	nominal	ring number
ringtype	input	nominal	ring type
sporepc	input	nominal	spore print color
stalkcar	input	nominal	stalk color above ring
stalkcbr	input	nominal	stalk color below ring
stalkroo	input	nominal	stalk root
stalksar	input	nominal	stalk surface above ring
stalksbr	input	nominal	stalk surface below ring
stalksha	input	nominal	stalk shape
target	target	binary	poisonous or edible

veilcolo	input	nominal	veil color
ceiltype	input	nominal	veil type

DMEXA1

Variable	Model Role	Measurement	Description
ACCTNUM	id	nominal	account Number
AGE	input	interval	age
AMOUNT	input	interval	amount of money spent
APPAREL	input	interval	apparel purchase
APRTMNT	input	binary	rent apartment: Yes or No
BLANKETS	input	interval	blankets purchase
COATS	input	interval	coats purchase.
COUNTY	input	interval	county code
CUSTDATE	input	interval	date of first Order
DISHES	input	interval	dishes purchase
DOMESTIC	input	interval	domestic prod.
DPM12	input	interval	monetary value per mailing
EDLEVEL	input	ordinal	education level
FLATWARE	input	interval	flatware purchase.
FREQUENT	input	interval	order frequency
GENDER	input	binary	gender
HEAT	input	nominal	heat utility
HHAPPAR	input	interval	his/her apparel.
HOMEACC	input	interval	home furniture
HOMEVAL	input	interval	home value
INCOME	input	interval	annual income
JEWELRY	input	interval	jewelry purchase
JOB	input	ordinal	job category
KITCHEN	input	interval	kitchen product
LAMPS	input	interval	lamps purchase

LEISURE	input	interval	leisure product
LINES	input	interval	linens purchase
LUXURY	input	binary	luxury items
MARITAL	input	binary	matial staus)
MENSWARE	input	interval	mens apparel
MOBILE	input	binary	Occupied <1 yr : Yes or No
NTITLE	input	nominal	name prefix
NUMCARS	input	ordinal	number of cars
NUMKIDS	input	interval	number of kids
ORIGIN	input	nominal	orgin
OUTDOOR	input	interval	outdoor product
PROMO7	input	interval	promotion: 1-7 months
PROMO13	input	interval	promotion: 8-13 months
PURCHASE	target	binary	purchase: Yes or No
RACE	input	nominal	race
RECENCY	input	interval	recency of purchase
RETURN	input	interval	total returns
SNGLMOM	input	binary	single mother: Yes or No
STATECOD	input	nominal	state code
TELIND	input	binary	telemarket industry
TMKTORD	input	ordinal	telemarket ord.
TOWELS	input	interval	towels purchase
TRAVTIME	input	interval	travel time
WAPPAR	input	interval	ladies apparel
WCOAT	input	interval	ladies coats

DMLBALL, DMTBALL

The input points (x, y, z) are uniformaly distributed on a unit cube. The target vairalbes t0 indicates whether an input point is within a sphere that has the same center as the cube and occupies half the volumn of the cube. Targets t10 and t20 are corrupted by noise for 10% and 20% of the data, respectively.

Variable	Model Role	Measurement	Description
----------	------------	-------------	-------------

t0	target	binary	whether the input point is within a sphere
t10	target	binary	whether the input point (with 10% noise) is within a sphere
t20	target	binary	whether the input point (with 20% noise) is within a sphere
x	input	interval	first dimension
y	input	interval	second dimension
z	input	interval	third dimension

DMLBLDG, DMTBLDG

Variable	Model Role	Measurement	Description
day	input	interval	day of month
dow	input	ordinal	day of week
hour	input	interval	hour of day
humid	input	interval	outside humidity
month	input	interval	month
solar	input	interval	solar radiation
tdate	input	interval	SAS date value for the date of today
temp	input	interval	outside temperature
wbcw	target	interval	cold water consumption
wbe	target	interval	electricity consumption
wbhw	target	interval	hot water consumption
wind	input	interval	wind speed
year	rejected	unary	year (1989)

Reference:

"The Great Energy Predictor Shootout" -- The First Building Data Analysis and Prediction Competition; ASHRAE Meeting; Denver, Colorado; June, 1993; co-chaired by Jan F. Kreider and Jeff S. Haberl; active period: December 1, 1992 -- April 30, 1993

DMLBLOC, DMTBLOC

The input points (x, y, z) are uniformly distributed on a unit cube. The target variable t0 indicates whether an input point is within any of three blocks of varying shape, that occupies half the volume of the cube. Targets t10 and t20 are

corrupted by noise for 10% and 20% of the data, respectively.

Variable	Model Role	Measurement	Description
t0	target	binary	whether the input point is within any of three blocks
t10	target	binary	whether the input point (with 10% noise) is within any of three blocks
t20	target	binary	whether the input point (with 20% noise) is within any of three blocks
x	input	interval	first dimension
y	input	interval	second dimension
z	input	interval	third dimension

DMLCENS, DMTCENS

Variable	Model Role		
Measurement	Description		
age	input	interval	age
cap_gain	input	interval	capital gains
cap_loss	input	interval	capital losses
class	target	binary	income level: <=50k or >50k
country	input	nominal	native country
country2	input	nominal	native country with rare categories collapsed
educ	input	nominal	education
educ_num	input	interval or ordinal	education (numeric)
fnlwgt	input	interval	
hourweek	input	interval	hours worked per week
marital	input	nominal	marital status
occupatn	input	nominal	occupation
race	input	nominal	race
relation	input	nominal	relationship
sex	input	binary	sex
workcla2	input	nominal	employer category with rare

			categories collapsed
workclas	input	nominal	employer category

References:

<http://ftp.ics.uci.edu/pub/machine-learning-databases/adult>

Kohavi, R. (1996). Scaling Up the Accuracy of Native-Bayes Classifiers: a Decision-Tree Hybrid, Proceedings of the Second International Conference on Knowledge Discovery and Data Mining

Merz, C.J., & Murphy, P.M. (1996). UCI Repository of machine learning databases [http://www.ics.uci.edu/~mllearn/MLRepository.html]. Irvine, CA: University of California, Department of Information and Computer Science.

DMLCUBE, DMTCUBE

The input points (x, y, z) are uniformly distributed on a unit cube. The target vairalbes t0 indicates whether an input point is within a smaller cube, that has the same center as the cube and occupiies half the volumn of the cube. Targets t10 and t20 are corrupted by noise for 10% and 20% of the data, respectively.

Variable	Model Role	Measurement	Description
t0	target	binary	whether the input point is within a smaller cube
t10	target	binary	whether the input point (with 10% noise) is within a smaller cube
t20	target	binary	whether the input point (with 20% noise) is within a smaller cube
x	input	interval	first dimension
y	input	interval	second dimension
z	input	interval	third dimension

DMLMAGL, DMTMAGL

Variable	Model Role	Measurement	Description
xm0	input	interval	$x_t$
xm6	input	interval	$x_{t-6}$
xm12	input	interval	$x_{t-12}$
xm18	input	interval	$x_{t-18}$
xp6	target	interval	$x_{t+6}$

xp85	target	interval	$x_{t+85}$
------	--------	----------	------------

DMLMNK3, DMTMNK3

Variable	Model Role	Measurement	Description
Body	input	nominal	shape of body
Head	input	nominal	shape of head
Holding	input	nominal	object holding
Jacket	input	nominal	color of jacket
Smiling	input	binary	smiling: Yes or No
Tie	input	binary	tie: Yes or No
target	target	binary	(jacket is green and holding a sword) or (jacket is not blue and body is not octagon)

DMLRING, DMTRING, DMSRING

Variable	Model Role	Measurement	Description
c	target	nominal	
x	input	interval	
y	input	interval	

DMLSINE, DMTSINE

Variable	Model Role	Measurement	Description
angle	input	interval	angle
sine	target	interval	interval sine of angle, no noise
sine1	target	interval	interval sine of angle, low noise
sine2	target	interval	interval sine of angle, medium noise
sine3	target	interval	interval sine of angle, high noise

DMLSONA, DMTSONA

	Model Role	Measurement	Description
--	------------	-------------	-------------



Variable			
target	binary	target	target: Mine or Rock
x1	input	interval	sonar measurement
x2	input	interval	sonar measurement
x3	input	interval	sonar measurement
x4	input	interval	sonar measurement
x5	input	interval	sonar measurement
x6	input	interval	sonar measurement
x7	input	interval	sonar measurement
x8	input	interval	sonar measurement
x9	input	interval	sonar measurement
x10	input	interval	sonar measurement
x11	input	interval	sonar measurement
x12	input	interval	sonar measurement
x13	input	interval	sonar measurement
x14	input	interval	sonar measurement
x15	input	interval	sonar measurement
x16	input	interval	sonar measurement
x17	input	interval	sonar measurement
x18	input	interval	sonar measurement
x19	input	interval	sonar measurement
x20	input	interval	sonar measurement
x21	input	interval	sonar measurement
x22	input	interval	sonar measurement
x23	input	interval	sonar measurement
x24	input	interval	sonar measurement
x25	input	interval	sonar measurement
x26	input	interval	sonar measurement
x27	input	interval	sonar measurement
x28	input	interval	sonar measurement

x29	input	interval	sonar measurement
x30	input	interval	sonar measurement
x31	input	interval	sonar measurement
x32	input	interval	sonar measurement
x33	input	interval	sonar measurement
x34	input	interval	sonar measurement
x35	input	interval	sonar measurement
x36	input	interval	sonar measurement
x37	input	interval	sonar measurement
x38	input	interval	sonar measurement
x39	input	interval	sonar measurement
x40	input	interval	sonar measurement
x41	input	interval	sonar measurement
x42	input	interval	sonar measurement
x43	input	interval	sonar measurement
x44	input	interval	sonar measurement
x45	input	interval	sonar measurement
x46	input	interval	sonar measurement
x47	input	interval	sonar measurement
x48	input	interval	sonar measurement
x49	input	interval	sonar measurement
x50	input	interval	sonar measurement
x51	input	interval	sonar measurement
x52	input	interval	sonar measurement
x53	input	interval	sonar measurement
x54	input	interval	sonar measurement
x55	input	interval	sonar measurement
x56	input	interval	sonar measurement
x57	input	interval	sonar measurement
x58	input	interval	sonar measurement

x59	input	interval	sonar measurement
x60	input	interval	sonar measurement

References:

Gorman, R. P. and Sejnowski, T.J. (1988). Analysis of Hidden Units in a Layered Network Trained to Classify Sonar Targets, Neural Networks, 1, 75-89.

DMLSPIR, DMSSPIR

Variable	Model Role	Measurement	Description
c	target	binary	0 or 1
x	input	interval	
y	input	interval	

DMLXOR

Variable	Model Role	Measurement	Description
x1	input	binary	x1: 0 or 1
x2	input	binary	x2: 0 or 1
y	target	binary	x1 XOR x2

MSEQ2 (no info about this data)

Variable	Type
ACTION	Char
CUSTOMER	Num
TIME	Num