

Text Mining to Save Lives

Timothy C. D'Auria
September 24th, 2008

Agenda

- Text, Text, and More Text!
- Example I: Predicting ER Usage
- Example II: Beyond the ER
- Classical Inference or Text Mining
- Saving Lives

- 80% of all valuable business data are in unstructured formats.¹

1. Christopher C. Shilakes and Julie Tylman, "Enterprise Information Portals." Merrill Lynch, 16 November, 1998.

Example I: ER Notes

Obs #	Nurse Soap Notes	ER Usage	Diabetes Pr...	Asthma Pro...
1	Patient presents with fever, chills, profuse sweating. BP 168/60. R 20 shallow and labored. P 120 bpm and thready. H...	1	0	0
2	CC: "I am dizzy." Male approx 70 yrs old is diaphoretic and tripodding. Took 2 nitro prior to onset of dizziness. I saw a pre...	0	1	0
3	Male 52 yo. BP: 110/74 P: 121, thready R: 24 shallow - rales. PO: 96%. Patient oriented x2, not oriented to date. Hi ever...	1	0	0
4	Patient fell down flight of stairs - 32 steps. No DCAP-BTLS apparent. Patient is disoriented, does not know time or plac...	1	0	0
5	7/12-year-old male c/o persistent cough x 4 wks. Child's mother reports that he has had a decrease in energy and ap...	1	0	0
6	BP: 102/58; P: 120; T: 98; Wt: 48.5 lb; Ht: 50.5 in; Peak flow: 150; O2 sat: 97%. Physical examination reveals a pale and t...	0	1	0
7	65yo Caucasian retired male schoolteacher presents for 6 mo S: 65yo Caucasian retired male schoolteacher presents f...	1	0	0
8	Been taking HCTZ 1 qd for 1 y. and added been taking HCTZ 1 qd for 1 y. and added atenolol atenolol 1 qd qd 6 mo ago t...	1	0	0
9	VS BP 160/96; T 98.0 f; RR=16/min.; HR=70/min and reg. Peripheral vascular peripheral vascular dis ln article < 1 qkgbulN...	1	0	0
10	Smoke x30y, smoke x30y, occas occas ETOH. No personal ETOH. No personal hx hx or or sx sx of CVA or TIA, diabet...	0	1	0
11	HEENT HEENT fundi fundi show moderate show moderate asvd asvd changes, no hems or exudates, changes, no hem...	0	0	1
12	discs sharp, carotid pulses full and equal, no bruits, no JVD. Chest: clear to Chest: clear to ausc ausc. No . No rales ral...	0	0	1
13	No Chest Pain or Shortness of Breath. "Feeling better today." Patient reports flatus. Is erythromycin effective in treating ...	0	0	1
14	Patient is a 37 year old man on post-operative day 2 for laparoscopic appendectomy, recently passed flatus. In article <...	0	0	1
15	This 40 year old woman presents to the office per chief complaint. 3 days ago patient began having fever, chills, sore t...	0	0	1
16	Rhinorrhea, myalgia, and fatigue. She denies cough, dyspnea, wheezing, otalgia, dysphagia, ln article < 1993Mar26.00...	0	0	1
17	Abdominal pain, nausea, vomiting, diarrhea or constipation. She has taken Advil, which improves symptoms. rgc3679...	1	0	0
18	Well-groomed, cheerful woman who appears stated age. T 97.5°F; P 78; R 12; BP 112/68; Wt 159 lbs.; Ht. 61 in. Pens-6...	0	1	0
19	Symmetrical auricles, canals nontender. Auditory canals clear and non-erythematous bilaterally. Tympanic membrane i...	1	0	0
20	No thrills or heaves felt. RRR. S1 and S2 of equal intensity. Supple, no lymphadenopathy, no masses, ln article < 1993A...	1	0	0
21	Erythematous bilaterally. Tympanic membrane intact and pearly with clear cone of light bilaterally. How long does it take...	0	0	1
22	While the differential diagnosis of acute pharyngitis in adults includes a number of bacterial and viral pathogens, the obj...	1	0	0
23	Pharyngitis is to identify those with group A streptococcus (GAS) pharyngitis. The patient's symptoms and physical ln ...	0	0	1
24	Although viruses are the most common cause of acute pharyngitis, Group A strep pharyngitis accounts Well, since so...	0	1	0
25	The absence of fever or presence of cough, conjunctivitis, hoarseness, coryza, viral exanthem, carl@SOL1.GPS.CAL...	0	0	1
26	The Centor criteria for predicting GAS have been the most widely used and accepted and include Just Testing!! Distribut...	0	1	0
27	Throat cultures have always been the "gold standard" for diagnosing GAS pharyngitis Probably within 50 years, a new...	0	0	1
28	5 mg/6.25 mg, two tablets in the morning, and denies any symptomatic side effects from the medication I sometimes see...	0	0	1
29	No chest pain, palpitations, dizziness, headaches, shortness-of-breath, wheezing, orthopnea, I'm starting an informal p...	0	1	0
30	On March 21, 1993 Roger Maynard wrote (in reply to an article by Graham Hudson): This was < 1993Mar21.223936.61...	0	1	0
31	Are complex bio-medical images available anywhere on the net for experimentation? By complex I mean that every sam...	1	0	0
32	Hello, and thank you for reading this request. I have a Mpeg viewer for x-windows and it did not run because I was run...	1	0	0
33	You can also swab the inside of your nose with Bacitracin using a Q tip. Bacitracin is an antibiotic that can be bought O...	0	0	1

SAS Analytics Platform



SAS® Enterprise Miner™ 5.2

The Power to Know.

Copyright © 2002-2005 by SAS Institute Inc., Cary, NC, USA. All Rights Reserved.

User name:

Password:

☒ Remember my password

☒ Personal Workstation

Welcome to Enterprise Miner



Help Topics



New Project...



Open Project...



Recent Projects...



Input ✕

 Diagram Name:

Patient Notes

Enterprise Miner - CCI Demo

File Edit View Actions Options Window Help

Sample Explore Modify Model Assess Utility Credit Scoring

Medical News

Medical News Sample

NEWS

Nurse Sample SOAP Notes

Nurse SOAP Notes

PRODUCTS

Diagrams

Property	Value
ID	EMWS16
Name	Patient Notes
Status	Open
Notes	...
History	...

Patient Notes

Enterprise Miner - CCI Demo

File Edit View Actions Options Window Help

Sample Explore Modify Model Assess Utility Credit Scoring

NEWS
Nurse Sample SOAP Notes
Nurse SOAP Notes
PRODUCTS
Diagrams
CATALOGUE CAMPAIGN

Patient Notes

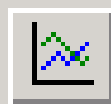
Nurse SOAP Notes

Property	Value
Node ID	lds3
Imported Data	...
Exported Data	...
Settings	
Variables	...
Decisions	...
Output Type	View
Role	Raw
Rerun	No
Metadata	
Data Source	Nurse SOAP Notes ...
Table	SOAPNOTES
Library	DEMOMINE
Description	
Number of Observations	600
Number of Columns	4
Scope	Local



Enterprise Miner - CCI Demo

File Edit View Actions Options Window Help



Sample

Explore

Modify

Model

Assess

Utility

Credit Scoring

Enterprise Miner - CCI Demo

File Edit View Actions Options Window Help



Sample

Data Partition

Partition data into separate tables.

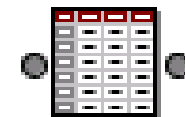
- News Mining
- News Mining Complete
- Nurse SOAP Notes Complete
- Patient Notes
- + Model Packages
- + Users

Utility

Credit Scoring

Patient Notes

	Property	Value
	ID	EMWS16
	Name	Patient Notes
	Status	Open
	Notes	...
	History	...



Nurse SOAP
Notes

Enterprise Miner - CCI Demo

File Edit View Actions Options Window Help

Sample Explore Modify Model Assess Utility Credit Scoring

medical news mining complete

News Mining

News Mining Complete

Nurse SOAP Notes Complete

Patient Notes

Model Packages

Users

Property	Value
Node ID	Part
Imported Data	...
Exported Data	...
Variables	...
Partitioning Method	Default
Random Seed	12345

Patient Notes

Nurse SOAP Notes

Data Partition

Enterprise Miner - CCI Demo

File Edit View Actions Options Window Help

Sample Explore Modify Model Assess Utility Credit Scoring

Medical News Mining Complete

News Mining

News Mining Complete

Nurse SOAP Notes Complete

Patient Notes

Model Packages

Users

Property

Value

From

Ids3

To

Part

Data export

Table

EMWS16.Ids3_DAT...

Variables

...

Role

Raw

Patient Notes

Nurse SOAP Notes

Data Partition

```
graph LR; A[Nurse SOAP Notes] --> B[Data Partition];
```

Enterprise Miner - CCI Demo

FileEditViewActionsOptionsWindowHelp

SampleExploreModifyModelAssessUtilityCredit Scoring

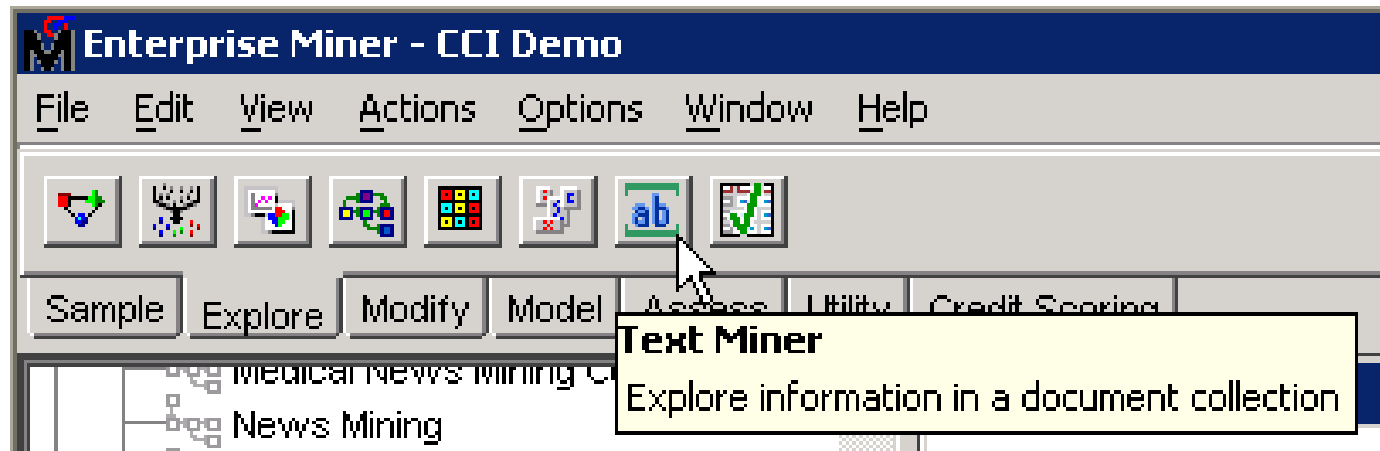
- Medical News Mining Complete
 - News Mining
 - News Mining Complete
 - Nurse SOAP Notes Complete
 - Patient Notes
- Model Packages
- Users

Property	Value
Node ID	Part
Imported Data	...
Exported Data	...
Variables	...
Partitioning Method	Default
Random Seed	12345
☒ Data Set Percentages	
Training	40.0
Validation	30.0
Test	30.0
☒ Status	
Time of Creation	9/21/08 7:36 PM
Run Id	
Last Error	
Last Status	
Needs Updating	Yes
Needs to Run	Yes
Time of Last Run	
Run Duration	
Grid Host	

Patient Notes

```

graph LR
    A[Nurse SOAP Notes] --> B[Data Partition]
    
```



Enterprise Miner - CCI Demo

File Edit View Actions Options Window Help

Sample Explore **Modify** Model Assess Utility Credit Scoring

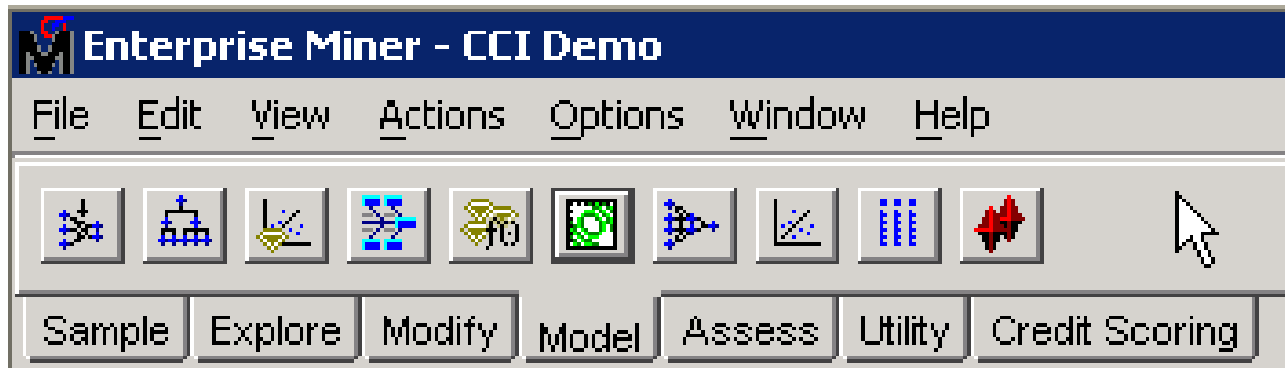
- Medical News Mining Complete
 - News Mining
 - News Mining Complete
 - Nurse SOAP Notes Complete
 - Patient Notes**
- Model Packages
- Users

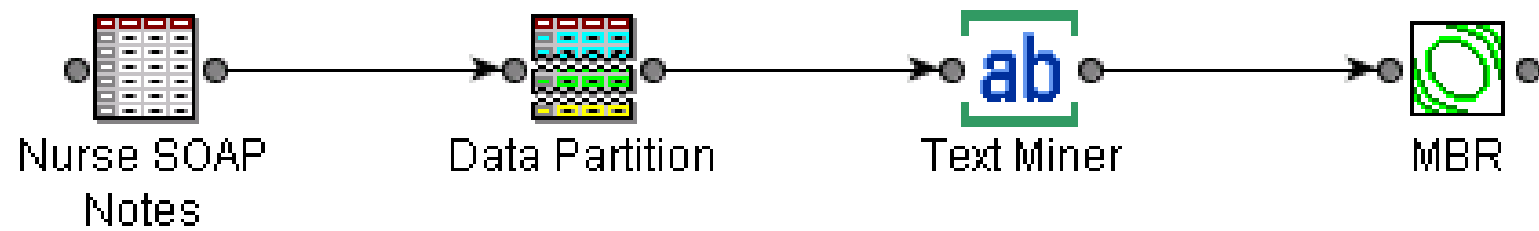
Property	Value
Node ID	TEXT
Imported Data	...
Exported Data	...
Variables	...
Interactive	...
Rerun	No
☒ Parse	
Parse Variable	
Language	ENGLISH
Stop List	SASHELP.STOPL
Start List	...
Stem Terms	Yes
Terms in Single Doc	No
Punctuation	No
Numbers	No
Different Parts of Sp	Yes
Ignore Parts of Spee	...
Noun Groups	Yes
Synonyms	SASHELP.ENG5\
Find Entities	No
Types of Entities	...
☒ Transform	
Compute SVD	Yes
SVD Resolution	Low
Max SVD Dimension	100
Scale SVD Dimensi	No
Frequency weighting	Log
Term Weight	Entropy
Roll up Terms	No

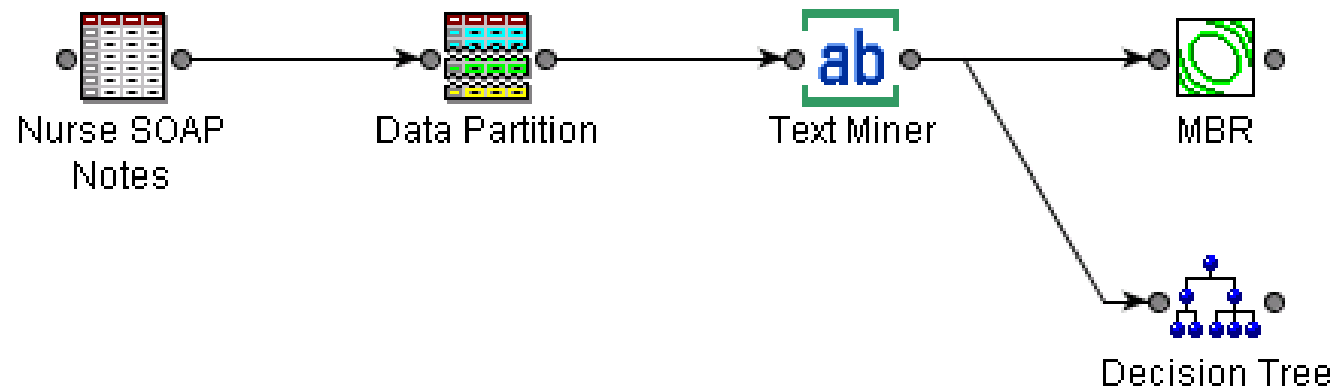
Patient Notes

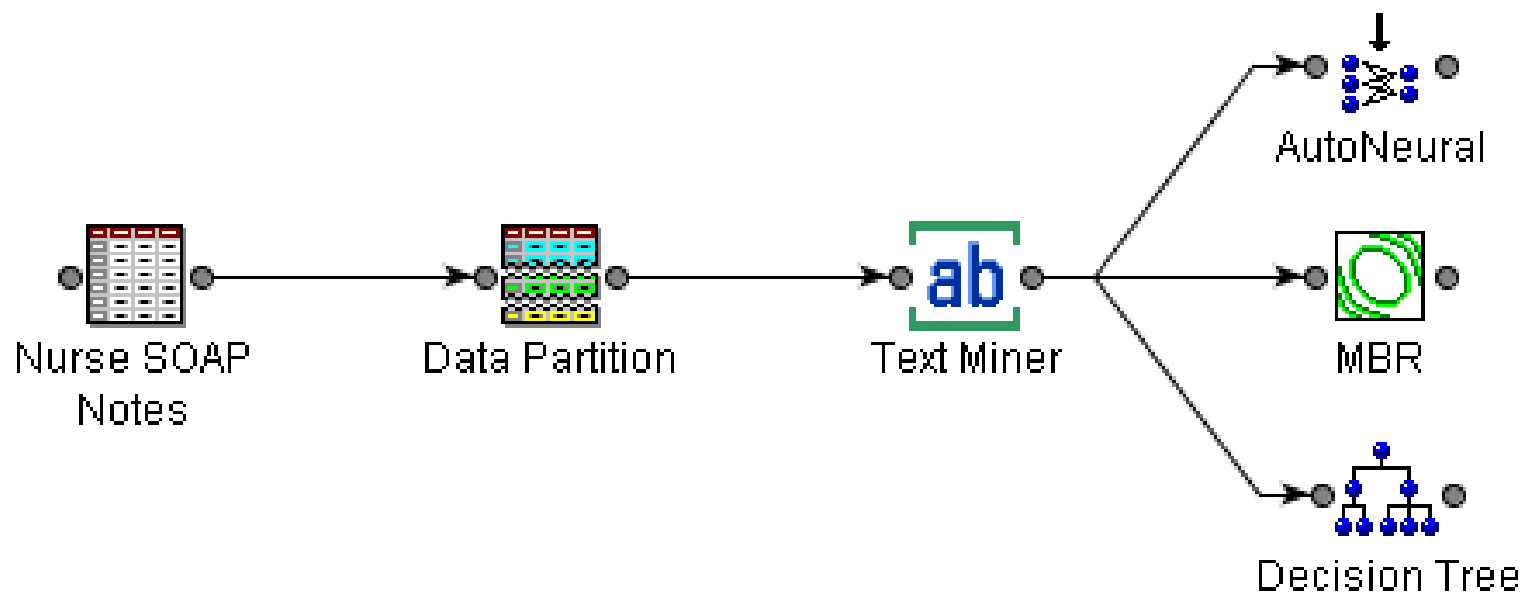
```

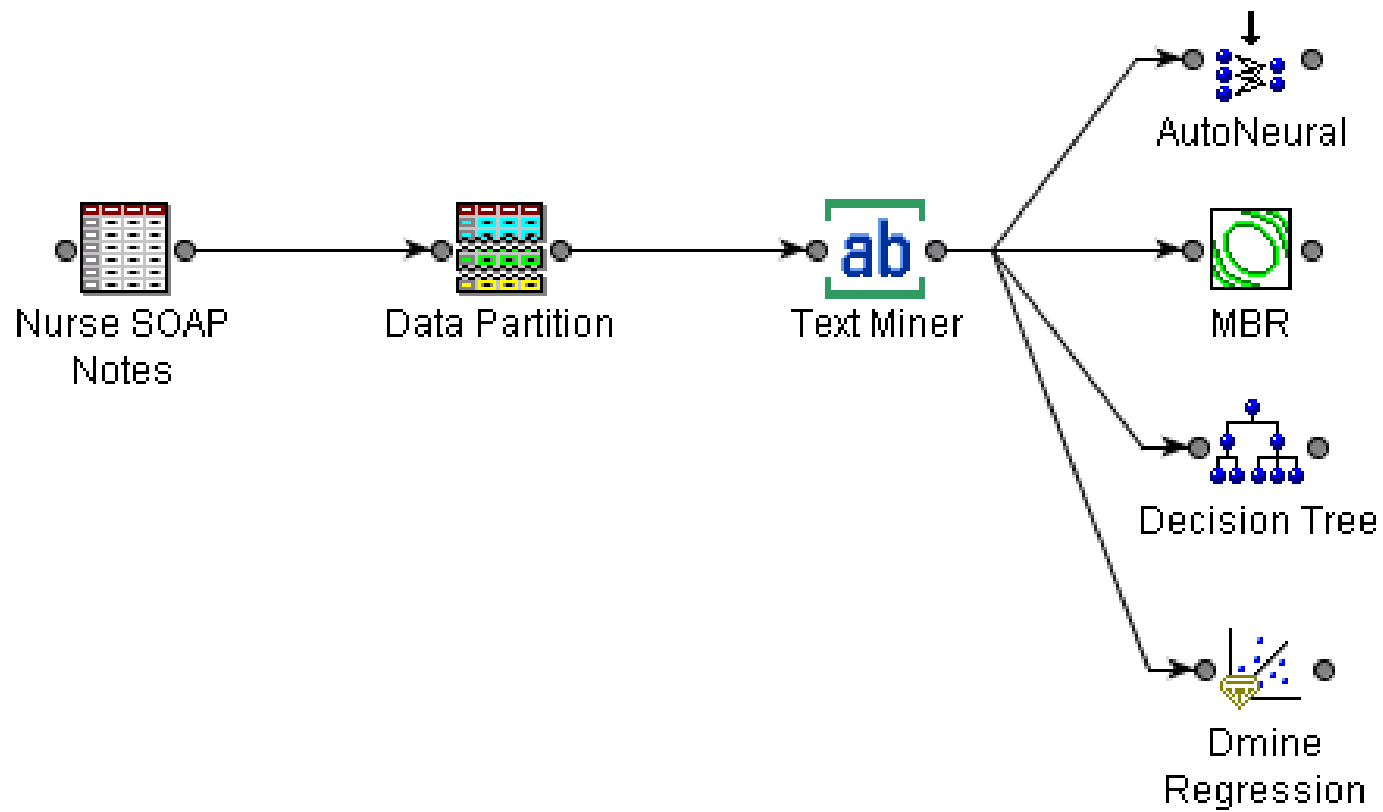
    graph LR
      A[Nurse SOAP Notes] --> B[Data Partition]
      B --> C[Text Miner]
  
```

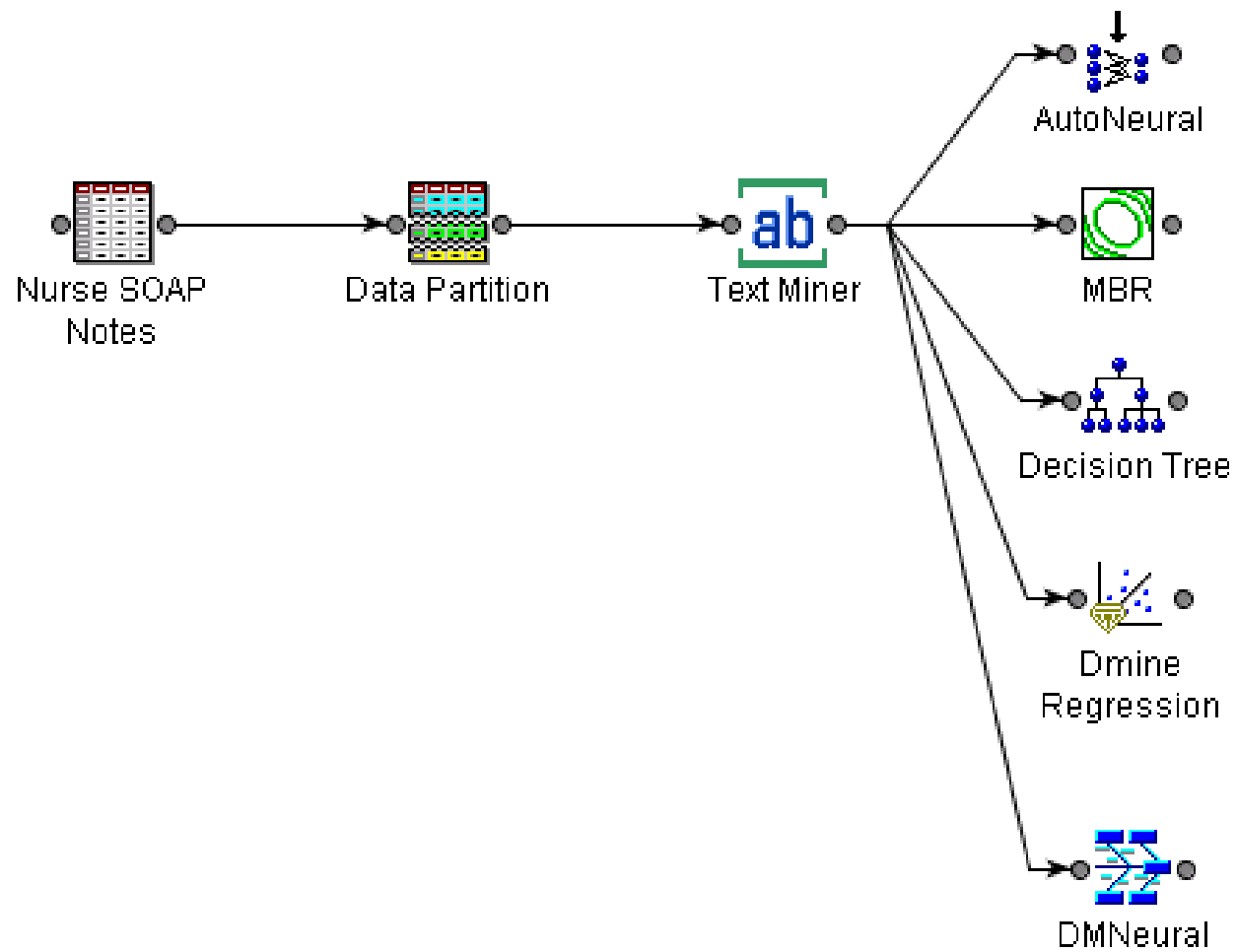


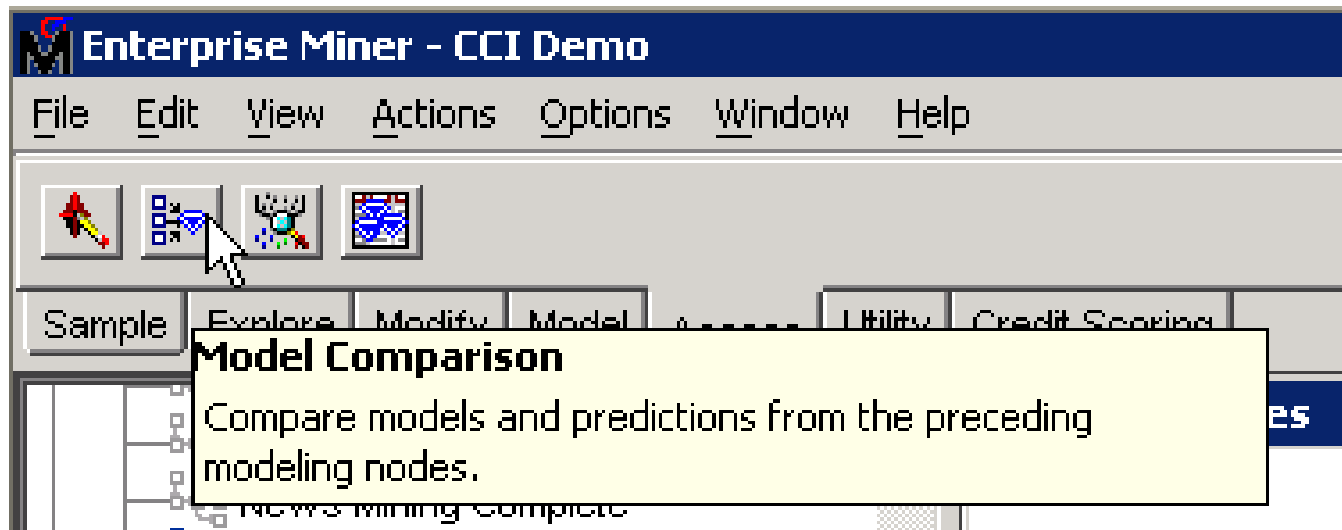


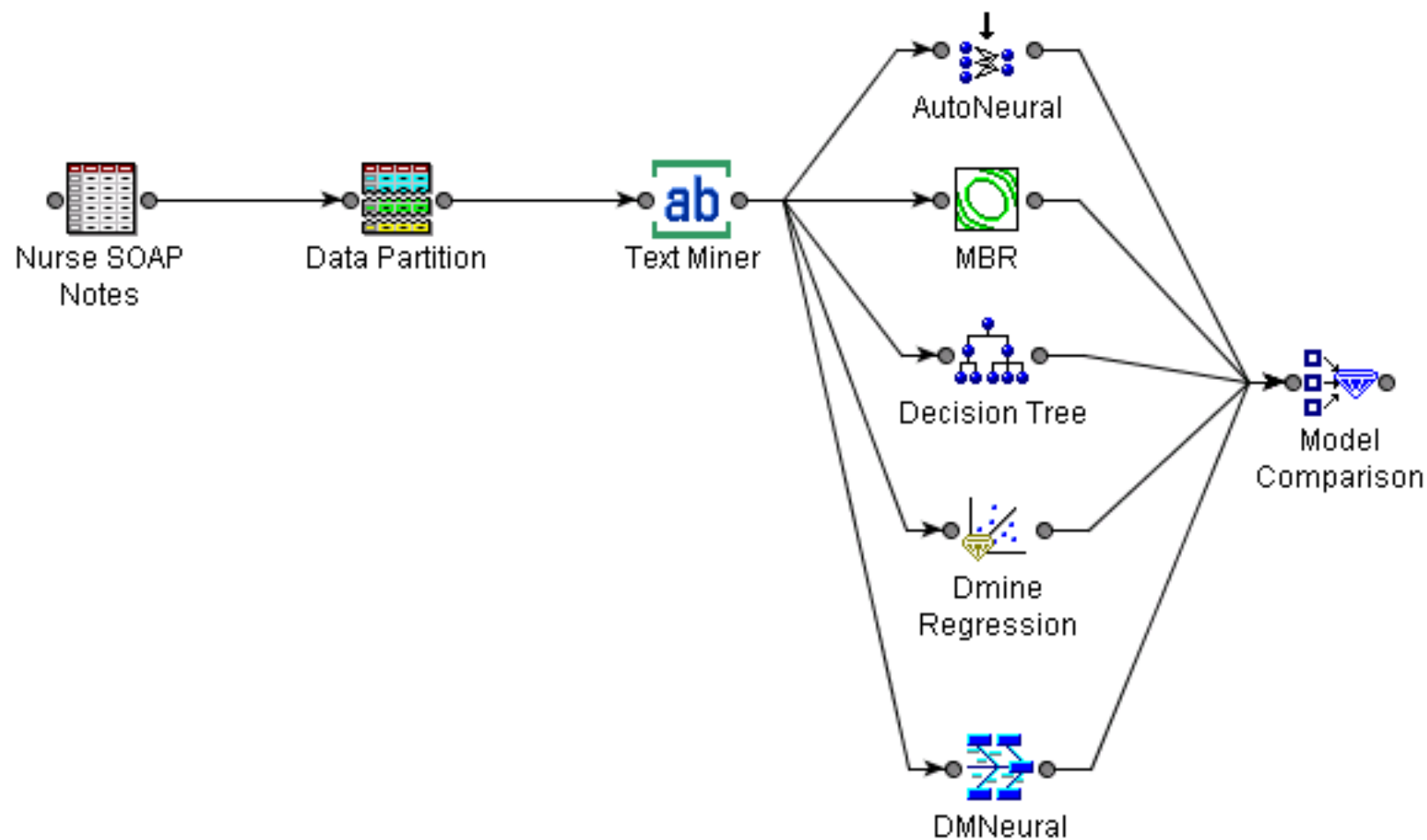


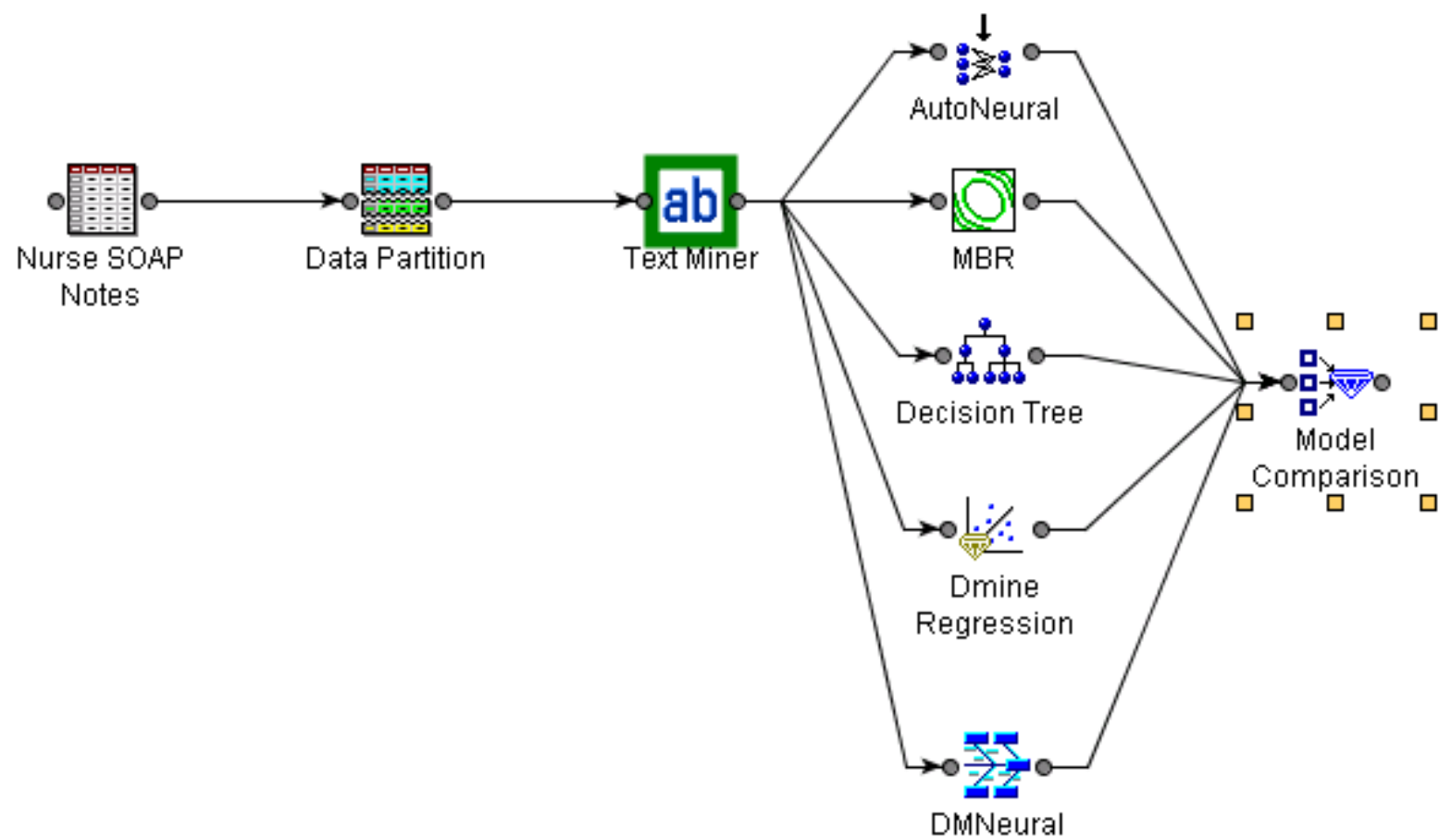


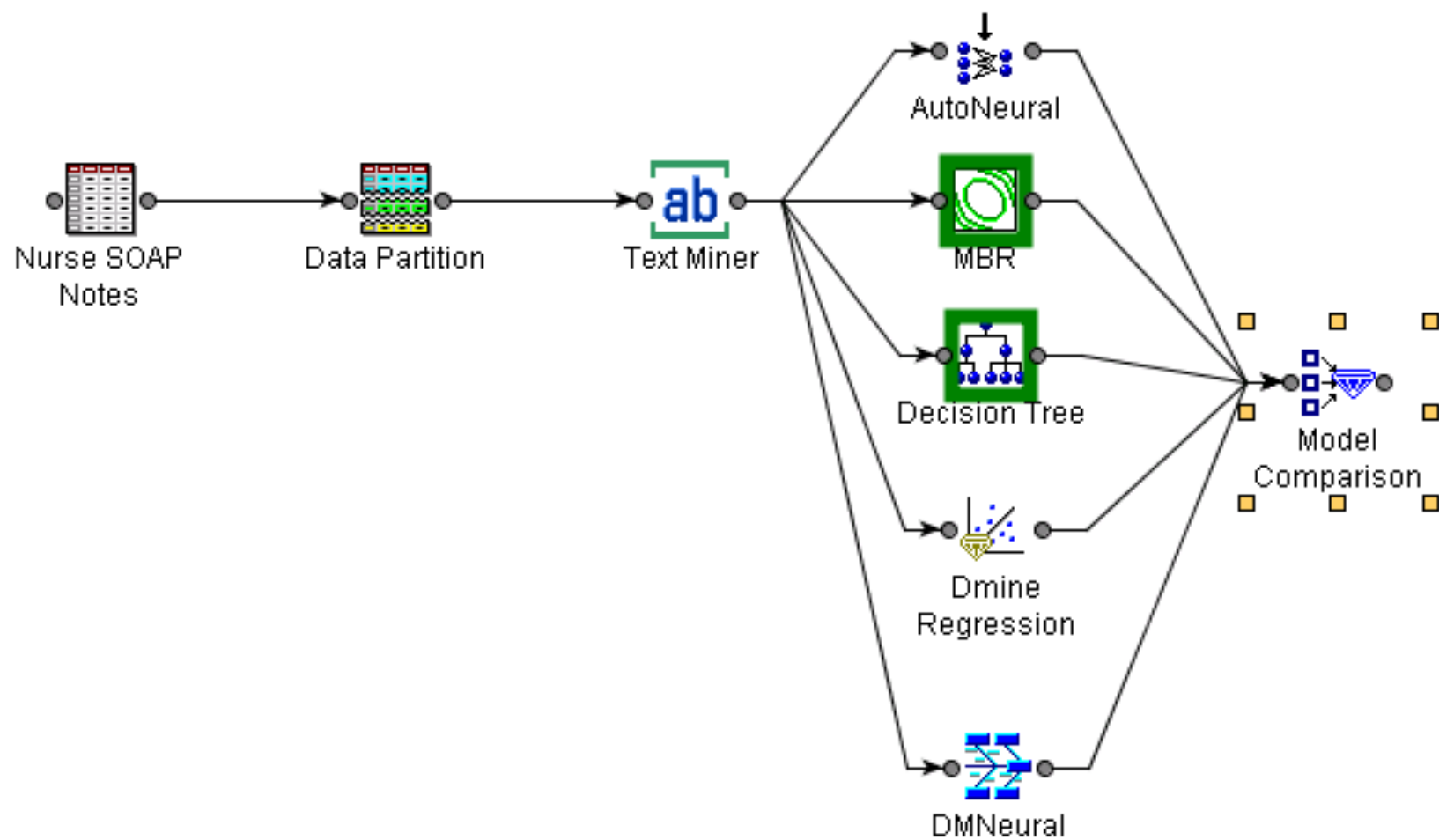






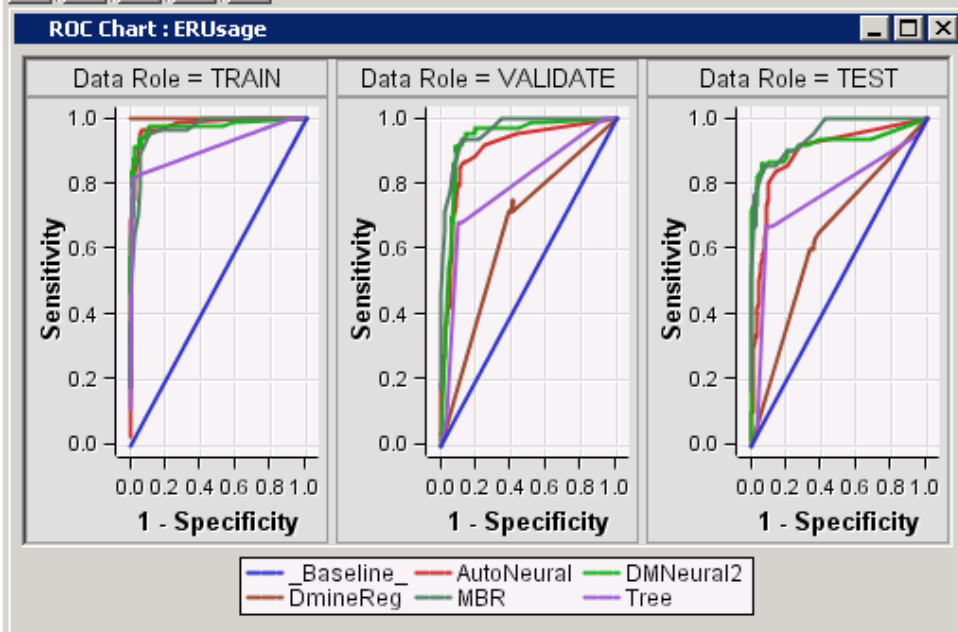






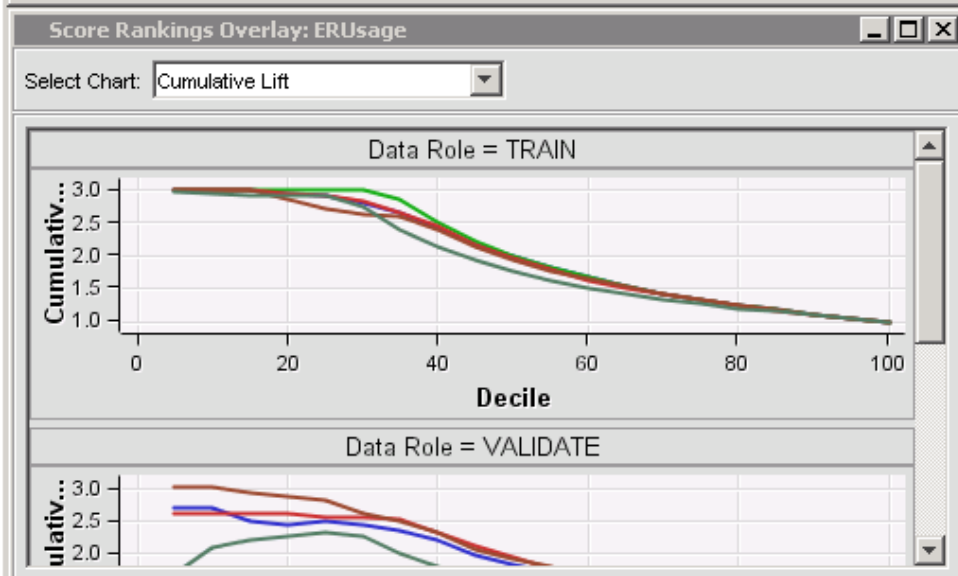
Results - Model Comparison

File Edit View Window



Fit Statistics

Selected Mo...	Predecesso...	Model Node	Target Vari...	Row Type	Number of E...	Train:
Y	AutoNeural	AutoNeural	ERUsage		229	
	DMNeural2	DMNeural2	ERUsage		21	
	DmineReg	DmineReg	ERUsage		.	
	MBR	MBR	ERUsage	PARMS	0	
	Tree	Tree	ERUsage		.	



Output

```

1
2
3
4
5 Variable Summary
6
7  ROLE      LEVEL      COUNT
8
9  TARGET    BINARY      1
10
11
12
13
14
15
16
    
```

Selected Model ▼	Predecessor Node
Y	DMNeural2
	AutoNeural
	DmineReg
	MBR
	Tree

Model Node	Test: Misclassification Rate ▲
DMNeural2	0.116022
MBR	0.143646
Tree	0.171271
AutoNeural	0.243094
DmineReg	0.370166

Variables - Nurse Sample SOAP Notes

Name	Role	Level	Report	Order	Drop	Lower Limit	Upper Limit	T
AsthmaProgram	Rejected	Binary	No		No	.	.	N
DiabetesProgram	Rejected	Binary	No		No	.	.	N
ERUsage	Rejected	Binary	No		No	.	.	N
NurseSOAPNotes	Text	Nominal	No		No	.	.	C

Explore...

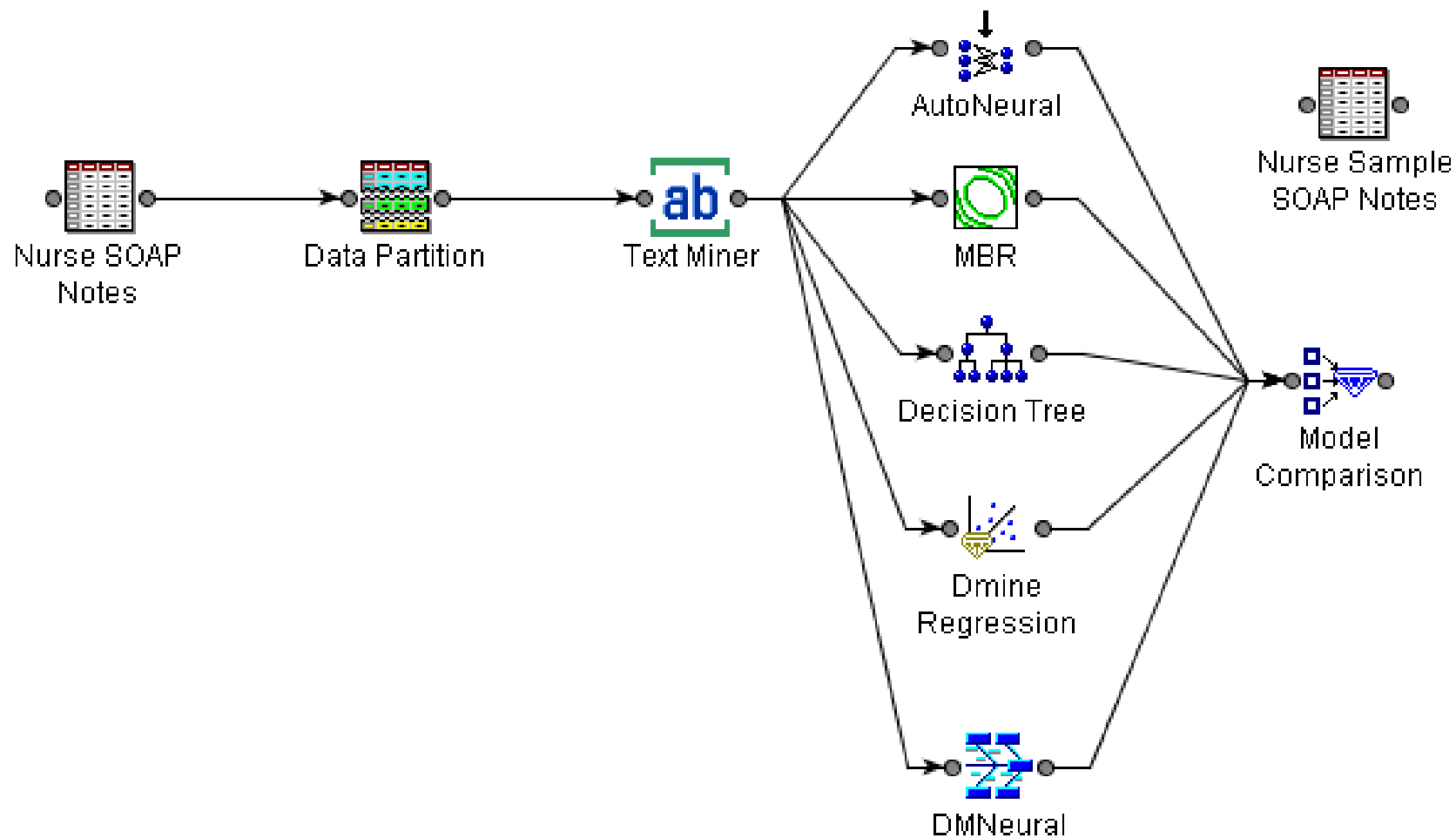
Edit Using SAS Code

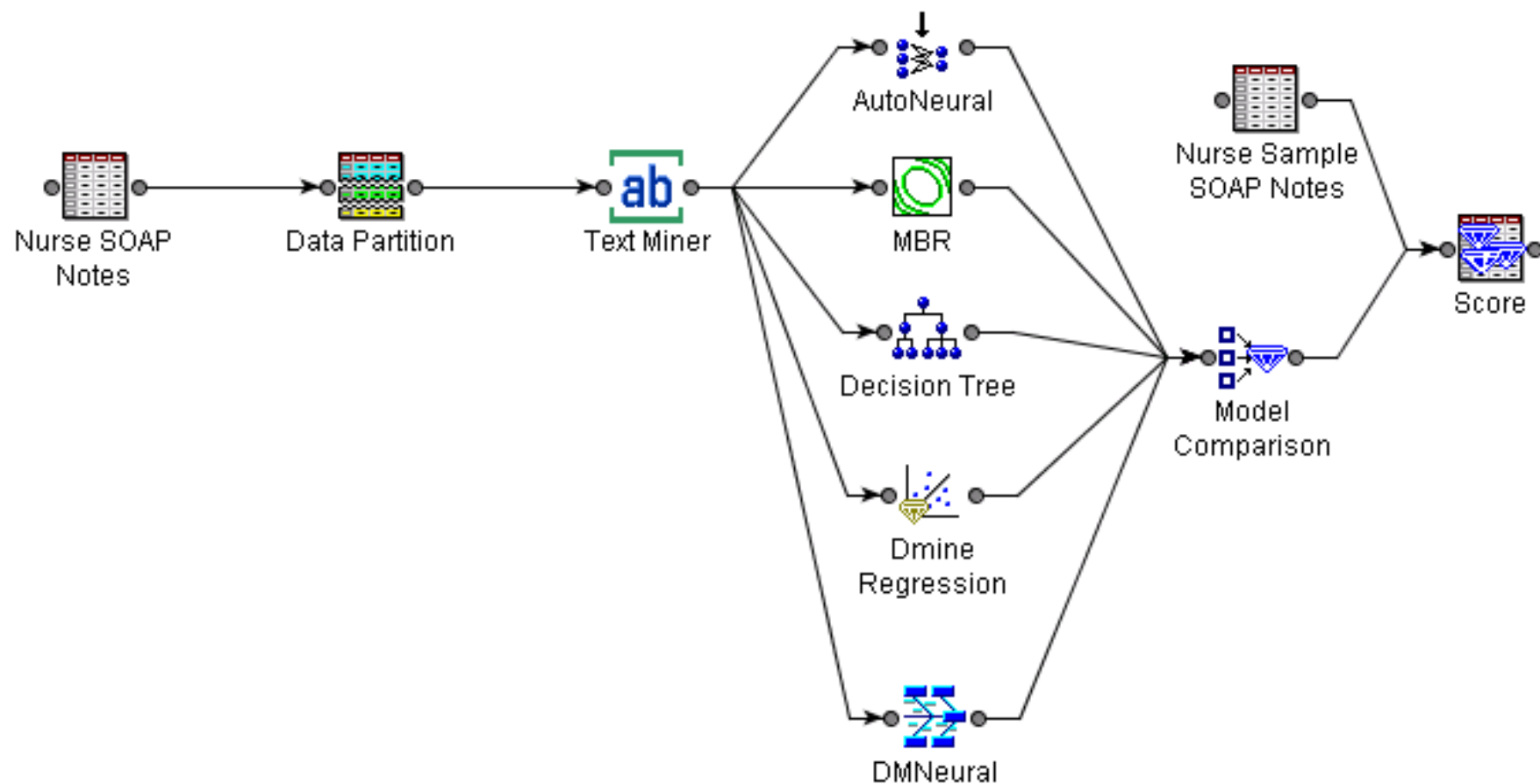
OK

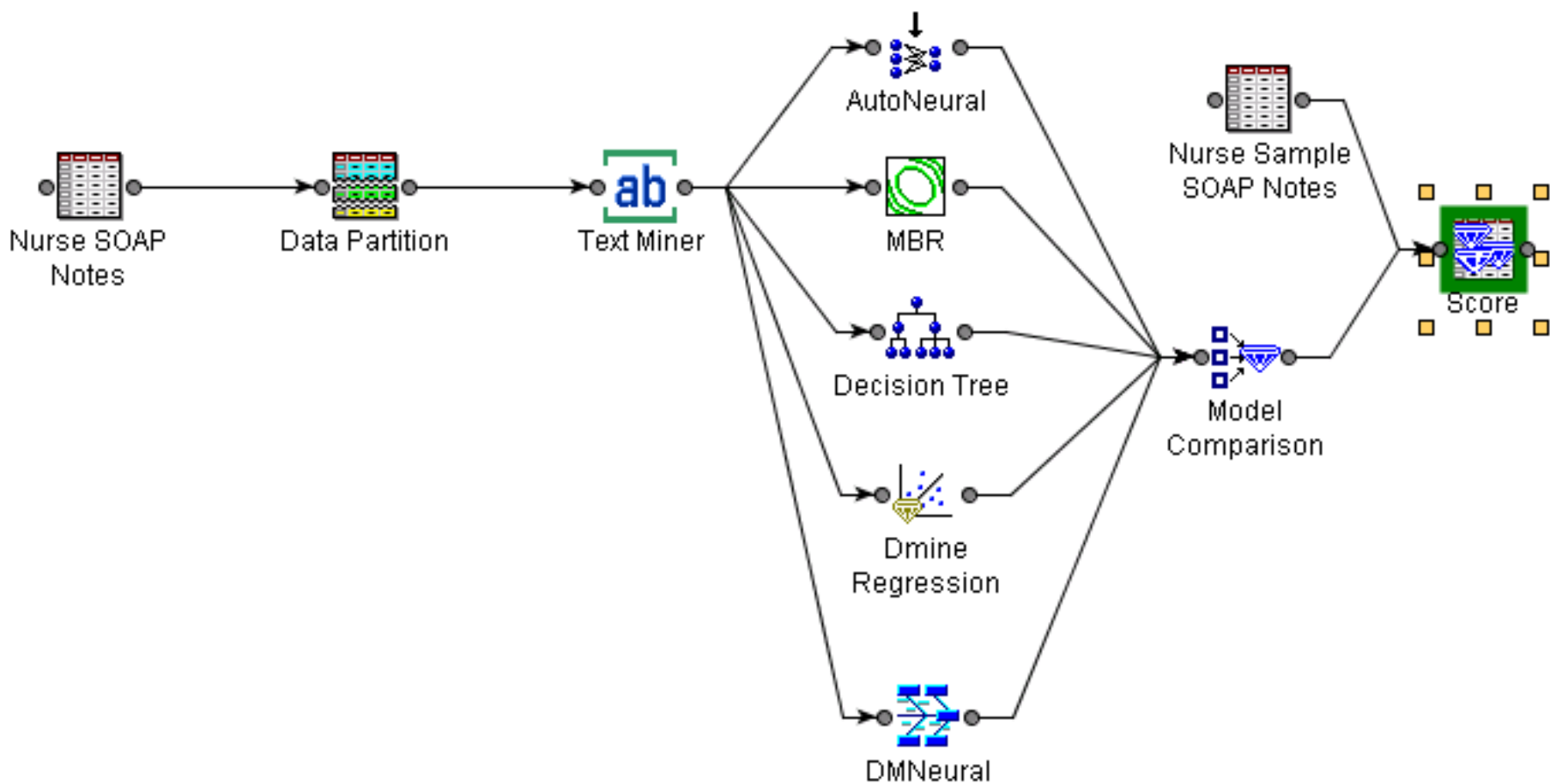
Cancel

Help

Obs #	Nurse Soap Notes
1	65yo Caucasian retired male schoolteacher presents for 6 mo S: 65yo Caucasian retired male schoolteacher ..
2	Been taking HCTZ 1qd for 1 y. and added been taking HCTZ 1qd for 1 y. and added atenolol atenolol 1 qd qd ...
3	VS BP 160/96;T 98.0 f ;RR=16/min ; HR=70/min and reg. Peripheral vascular peripheral vascular dis In article ...
4	Smoke x30y, smoke x30y, occas occas ETOH. No personal ETOH. No personal hx hx or or sx sx of CVA or ...
5	HEENT HEENT fundi fundi show moderate show moderate asvd asvd changes, no hems or exudates, chang...
6	discs sharp, carotid pulses full and equal, no bruits, no JVD. Chest: clear to Chest: clear to ausc ausc. No . N...
7	No Chest Pain or Shortness of Breath. "Feeling better today." Patient reports flatus. Is erythromycin effective ...
8	Patient is a 37 year old man on post-operative day 2 for laparoscopic appendectomy, recently passed flatus. ...
9	This 40 year old woman presents to the office per chief complaint. 3 days ago patient began having fever, c...
10	Rhinorrhea, myalgia, and fatigue. She denies cough, dyspnea, wheezing, otalgia, dysphagia, In article < 199...







Exported Data - Score			
Port	Table	Role	Data Exists
TRAIN	EM\WS16.Score_TRAIN	Train	Yes
VALIDATE	EM\WS16.Score_VALIDATE	Validate	Yes
TEST	EM\WS16.Score_TEST	Test	Yes
SCORE	EM\WS16.Score_SCORE	Score	Yes
Browse... Explore... Properties... OK			

EM_EVENTPROBABILITY	EM_PROBABILITY	EM_CLASSIFICATION
0.338537	0.661463	0
1	1	1
1	1	1
0	1	0
0.064571	0.935429	0
0	1	0
0.016075	0.983925	0
0.027931	0.972069	0
0.439546	0.560454	0
0.02152	0.97848	0

NURSES SOAP NOTES

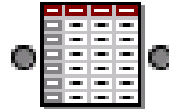
1	65yo Caucasian retired male schoolteacher presents for 6 mo S: 65yo Caucasian retired male ...
2	Been taking HCTZ 1qd for 1 y. and added been taking HCTZ 1qd for 1 y. and added atenolol at...
3	VS BP 160/96; T 98.0 f ; RR=16/min.; HR=70/min and reg. Peripheral vascular peripheral vascula...
4	Smoke x30y, smoke x30y, occas occas ETOH. No personal ETOH. No personal hx hx or or sx ...
5	HEENT HEENT fundi fundi show moderate show moderate asvd asvd changes, no hems or ex...
6	discs sharp, carotid pulses full and equal, no bruits, no JVD. Chest: clear to Chest: clear to aus...
7	No Chest Pain or Shortness of Breath. "Feeling better today." Patient reports flatus. Is erythrom...
8	Patient is a 37 year old man on post-operative day 2 for laparoscopic appendectomy, recently ...
9	This 40 year old woman presents to the office per chief complaint. 3 days ago patient began h...
10	Rhinorrhea, myalgia, and fatigue. She denies cough, dyspnea, wheezing, otalgia, dysphagia, l...

Example II

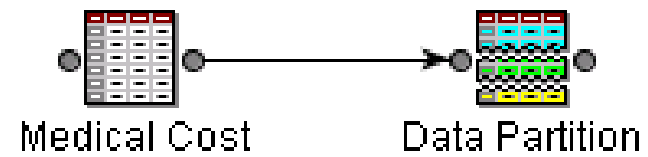
- Predicting patient health expenditures.
 - Lifetime expenditures for high-risk patient pool

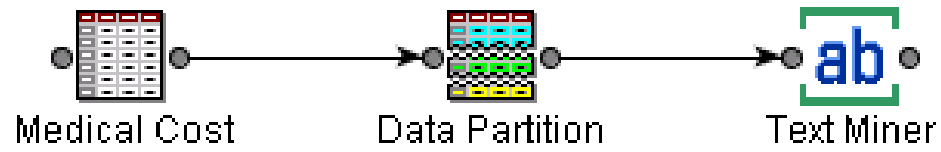
Name	Role	Level
ADM_SOC	Input	Binary
AFFECTIV	Input	Binary
AGE	Input	Interval
ALCOHOL	Input	Binary
ANXIETY	Input	Binary
COST	Target	Interval
HERED_DE	Input	Binary
ID	ID	Nominal
NURSE_PROGRESS_NOTES	Text	Nominal
OTHER_IN	Input	Binary
PERSONAL	Input	Binary
PREADULT	Input	Binary
SCHIZOPH	Input	Binary
SENILITY	Input	Binary
SEX	Input	Binary
STATE	Input	Nominal
SUBSTANC	Input	Binary

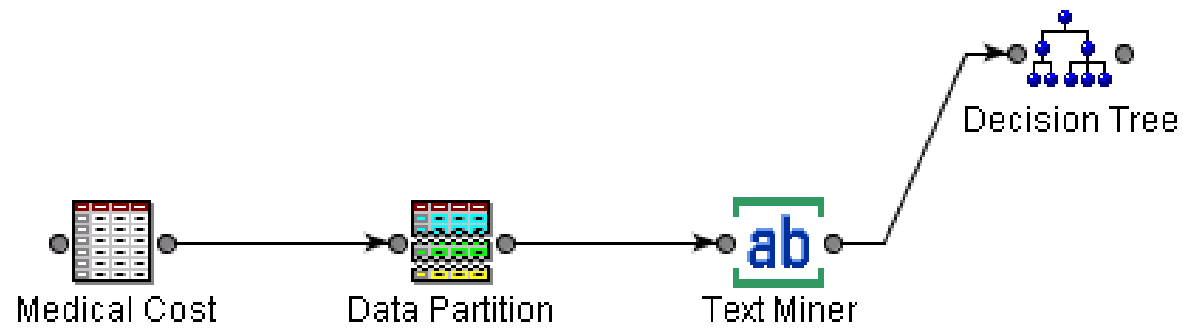
Obs #	ID	AGE	SEX	STATE	NURSE_PROGRESS_NOTES	SCHIZOPH	SUBSTANC	ADM_SOCI	HERED_DE	ANXIETY	ALCOHOL
1	1	1.828222M		RI	Patient presents with fever, chi...	0	0	0	1	0	0
2	2	60.1811M		RI	CC: "I am dizzy." Male approx 7...	0	0	0	0	1	0
3	3	80.85527F		RI	Male 52 yo. BP: 110/74 P: 121, t...	0	1	1	0	1	0
4	4	19.59926F		RI	Patient fell down flight of stairs ...	0	0	1	0	0	0
5	5	62.05862M		RI	7 7/12-year-old male c/o persis...	0	0	0	0	0	0
6	6	59.05946F		RI	BP: 102/58; P: 120; T: 98; Wt: 4...	0	0	0	0	1	0
7	7	69.99861M		RI	65yo Caucasian retired male sc...	0	0	0	0	0	0
8	8	45.66007M		RI	Been taking HCTZ 1qd for 1 y. ...	0	0	0	0	0	0
9	9	40.2132M		RI	VS BP 160/96;T 98.0 f ;RR=16/...	0	1	0	0	0	0
10	10	77.20401F		RI	Smoke x30y, smoke x30y, occ...	0	0	0	0	0	0
11	11	48.18755M		RI	HEENT HEENT fundi fundi show...	1	0	0	0	0	0
12	12	34.64023M		RI	discs sharp, carotid pulses full ...	0	0	0	1	0	0
13	13	16.25897M		RI	No Chest Pain or Shortness of ...	0	0	0	0	0	0
14	14	14.56568M		RI	Patient is a 37 year old man on ...	0	0	0	0	0	1
15	15	13.88355M		RI	This 40 year old woman prese...	0	0	0	1	0	0
16	16	79.78713M		RI	Rhinorrhea, myalgia, and fatigu...	0	1	0	0	0	0
17	17	56.99142F		RI	Abdominal pain, nausea, vomiti...	0	0	0	0	0	0
18	18	89.04197M		RI	Well-groomed, cheerful woman...	0	0	0	0	0	0
19	19	19.98871F		RI	Symmetrical auricles, canals no...	0	0	0	0	0	0
20	20	52.49282F		RI	No thrills or heaves felt. RRR. ...	0	0	0	0	0	0
21	21	11.4665F		RI	Erythematous bilaterally. Tymp...	0	0	0	1	0	0
22	22	98.68053M		RI	While the differential diagnosis ...	0	0	0	0	0	0
23	23	68.37665M		RI	Pharyngitis is to identify those ...	0	1	0	1	0	0
24	24	67.54003F		RI	Although viruses are the most ...	0	1	0	0	1	0
25	25	39.97747F		RI	The absence of fever or prese...	0	1	0	0	1	0
26	26	96.06926M		RI	The Centor criteria for predictin...	0	0	0	0	1	0
27	27	70.18458F		RI	Throat cultures have always b...	0	0	0	1	1	1
28	28	32.94335M		RI	5 mg/6.25 mg, two tablets in th...	0	0	0	0	1	0
29	29	17.91188M		RI	No chest pain, palpitations, dizz...	0	0	0	0	1	0
30	30	1.35045F		RI	On March 21, 1993 Roger May...	0	0	0	0	0	0
31	31	35.60011M		RI	Are complex bio-medical image...	0	0	0	0	0	0
32	32	63.03549F		RI	Hello, and thank you for readin...	0	0	0	0	0	0
33	33	6.3987M		RI	You can also swab the inside ...	0	0	0	0	0	1
34	34	48.94271M		RI	In article < C4IHM2.Gs9@wats...	0	0	0	0	0	0
35	35	1.406033F		RI	Archie told me the following sit...	0	1	0	0	0	0
36	36	88.23069F		RI	In article < 1993Apr15.180459....	0	0	0	0	0	0
37	37	95.50833M		RI	In article < 1993Mar29.181406....	0	0	0	0	0	0

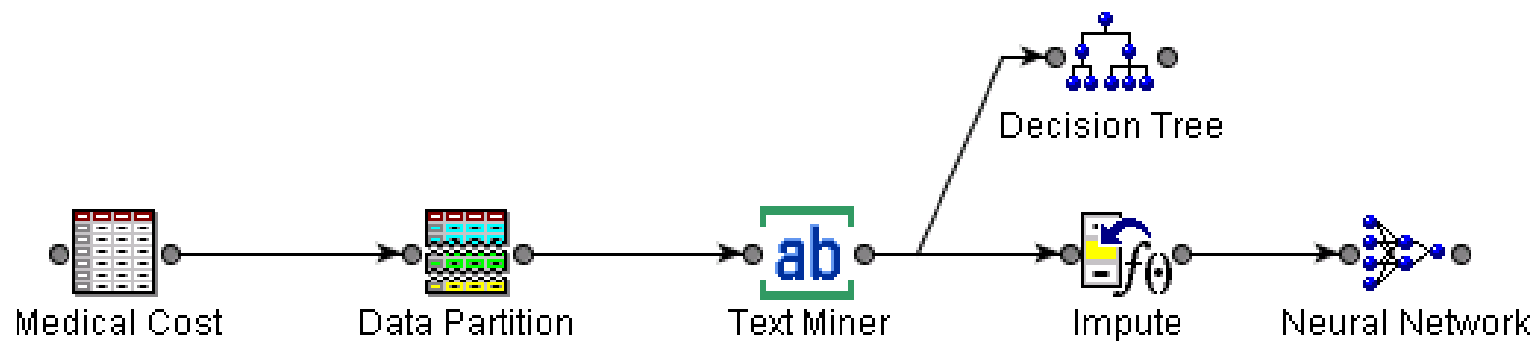


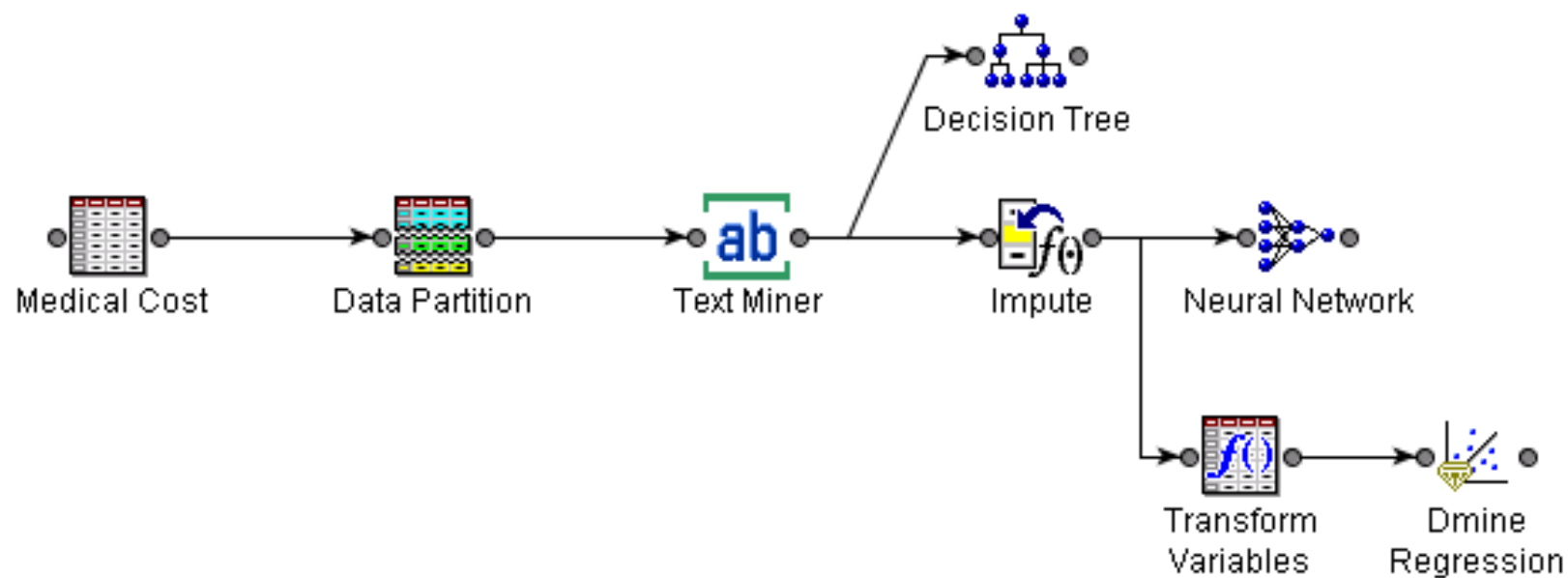
Medical Cost

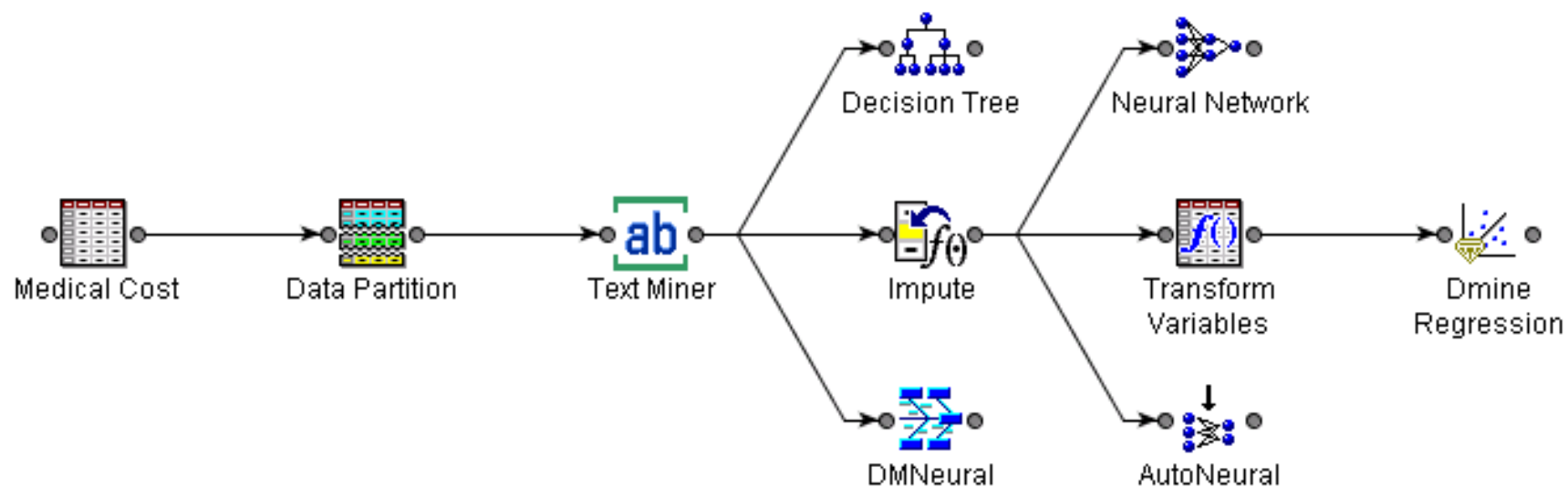












Why Mine?

- Logistic Regression is a common inferential technique when the dependent variable is a dichotomy.
- Why not use regression?

Assumption Check

- Logistic regression has many assumptions that should hold true for consistently valid results.
- Analytical assumptions are not aggressively assessed in literature.
- Assumptions may not hold.

ICD-9 Codes

- Logistic regression is commonly used to analyze ICD-9 and other diagnosis codes.
- Requires that all providers are entering codes the same way.
- Assumes there is no relationship between comorbidities.

Logistic Regression

- Data level
- Meaningful coding
- Inclusion of all relevant variables
- Exclusion of all irrelevant variables
- Independence of irrelevant alternatives
- Error terms independent
- Low error in the explanatory variables
- Linearity
- Additivity
- No multicollinearity
- Centered variables
- No outliers
- Large samples
- Sampling adequacy
- Expected dispersion

Pioneering Methods

- Dr. Patricia Cerrito from the University of Louisville pioneered the concept of viewing diagnostic codes as a language made exclusively of nouns.
- Diagnosis codes for a patient make up a single sentence of nouns.

Cerrito, P. (2007). Text Mining Coded Information. In H. A. D. Prado & E. Ferneda (Eds.), Emerging Technologies of Text Mining: Techniques and Applications. New York: Big Idea Group.

- Data mining relaxes many of the requirements of classical inferential technique.
- ...but at the cost of requiring more data.

- Text mining enables the assessment of much more information.
- Models may be ran in real-time on the web or from other delivery mechanisms.
- Provides up-to-the-minute predictive health insight.

Saving Lives

- Identify novel high-risk patients before they become symptomatic.
- Identify patients in need of managed care asthma, diabetes, or other program.
- Which ER patients will be admitted overnight?
- Which patients are at high-risk for a nosocomial infection?

Thank you

Timothy D'Auria
Creative Computing, A SAS Partner

tdauria@mycci.com

401-727-2400

** 3-Day Analytical Discovery Workshop **