Biostat 202: Opportunities and Challenges of “Big Data”.
Machine Learning 2: Classification

John Kornak
Division of Biostatistics
Department of Epidemiology and Biostatistics
08/17/2017

Overview

- Auto-numeric node Modeler Demo
- Classification
- Model evaluation
- Training-test
- Classification methods
  - Logistic Regression
  - K-nearest neighbors
- IBM SPSS Data Modeler examples left for todays lab

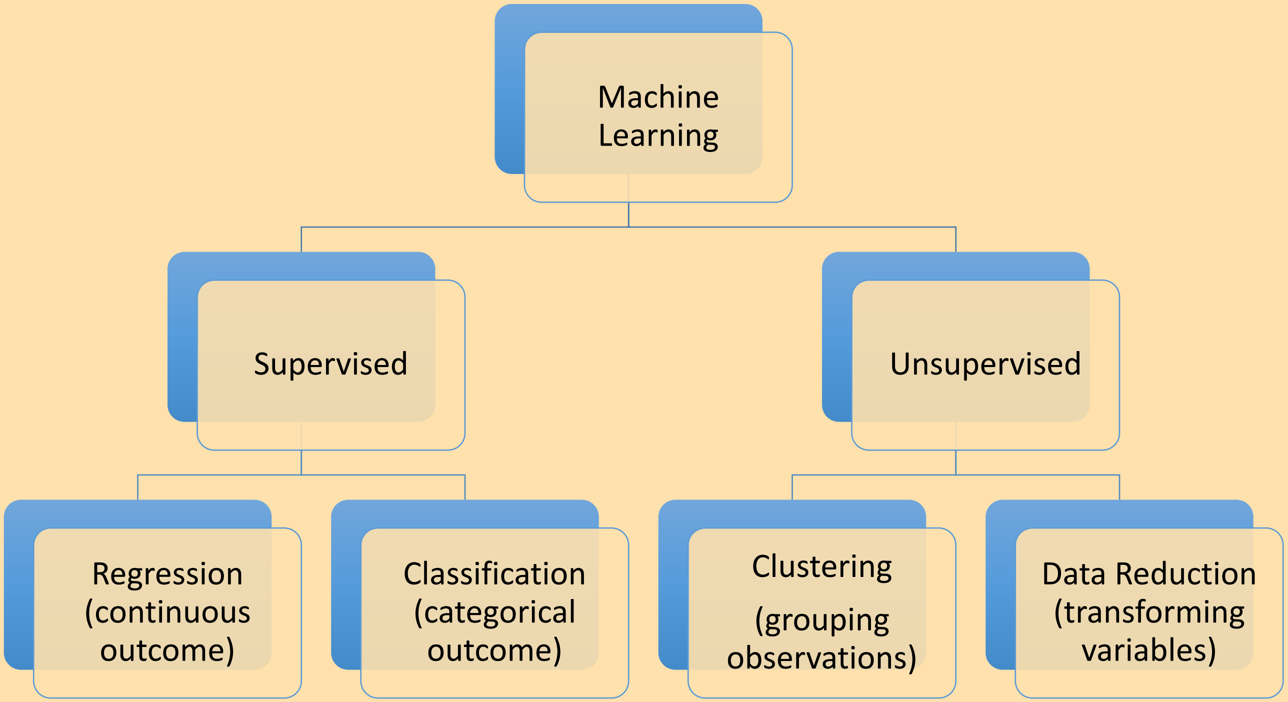
Brain volumes dataset Modeler demo

- Use regional brain volumes to predict age of subject in healthy controls with a range of ages – Target = Age, Predictors = regional volumes + demographic variables

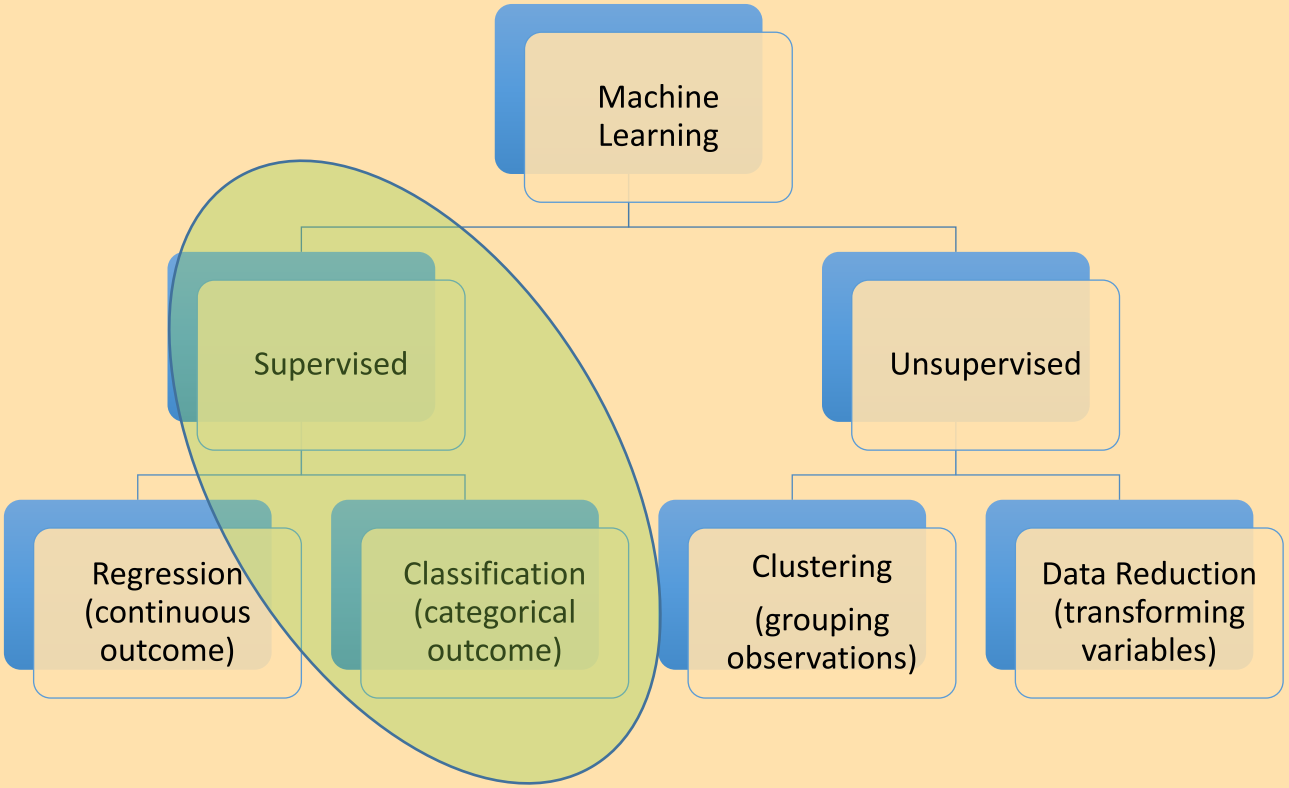
- ROI volumes are the main predictors: Extract 136 ROIs as predictors for each subject – big data?
- ROIs are defined in terms of anatomical regions and landmarks in the human brain that are common to all individuals – often have specific functional roles

Classification

Supervised vs. unsupervised



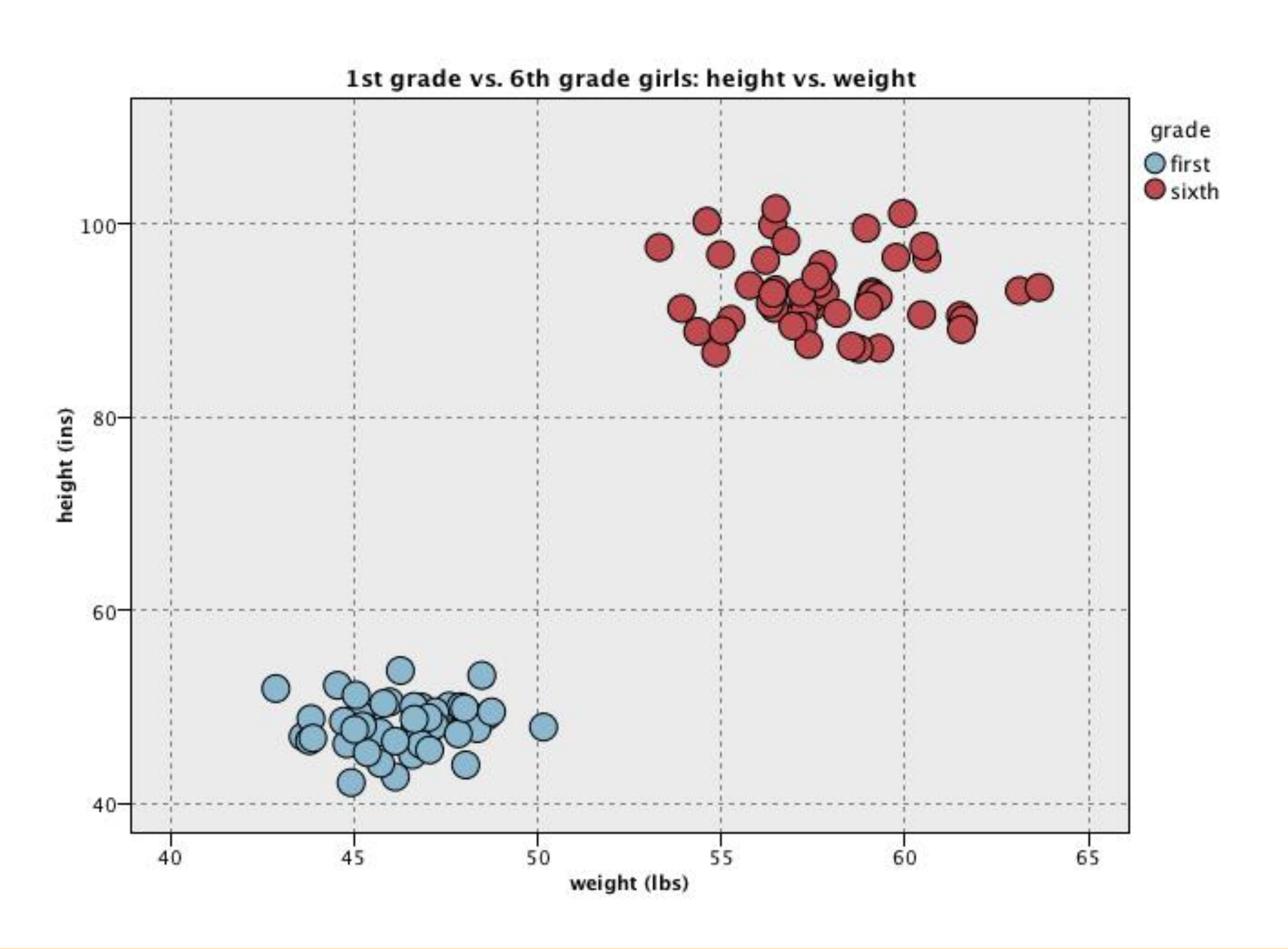
Supervised vs. unsupervised



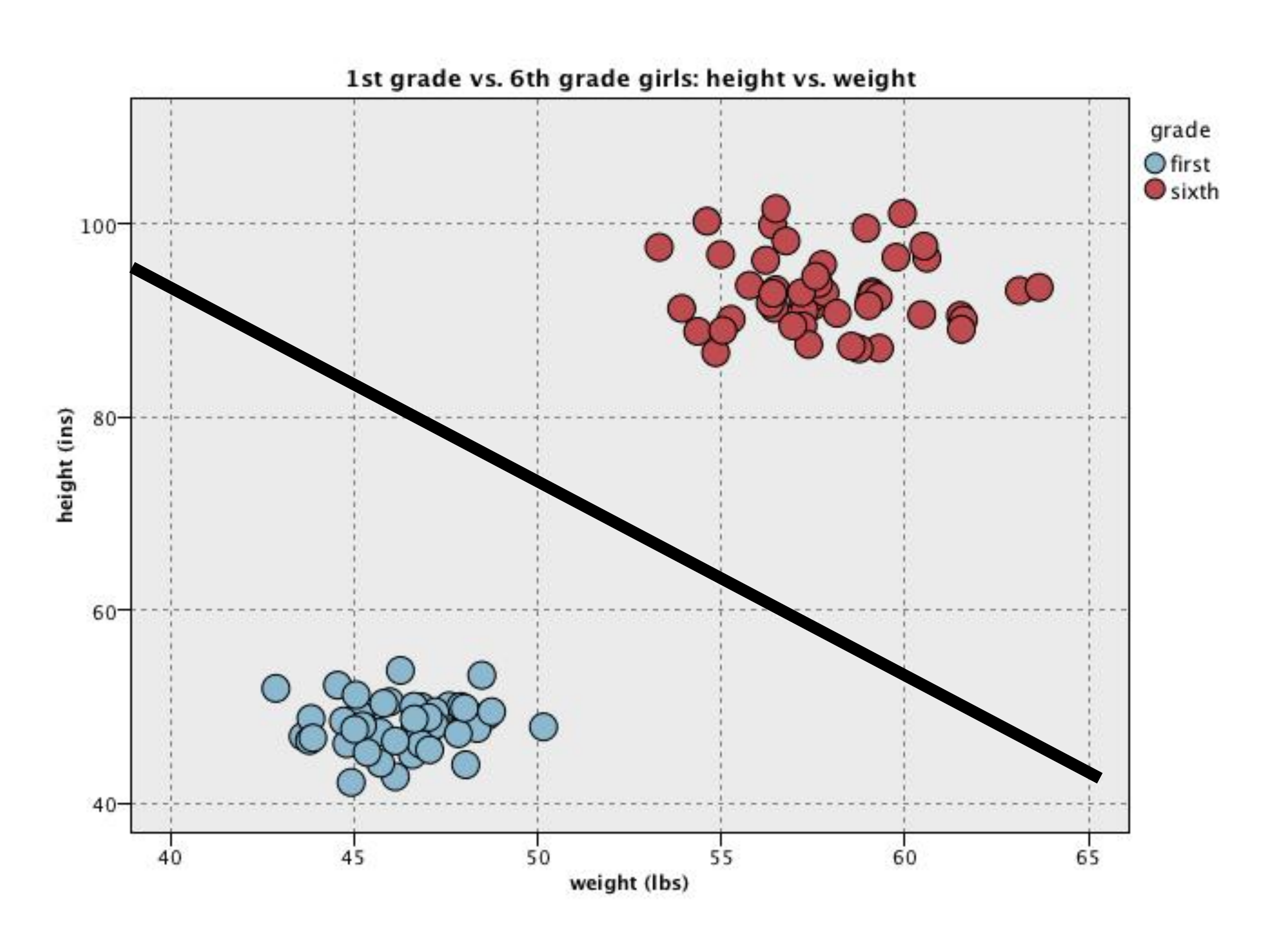
Classification

- Objective – given set of predictors and some training data, generate a rule (classifier) to predict which category a new instance/observation/subject belongs to
- Training data with observed predictors and outcomes (target) used to develop a classifier
- Classifier is applied to new data where the outcome (i.e. target group/class) is unknown
- Let's look at a toy example trying to classify 1^{st} vs. 6^{th} grade girls based on height and weight….

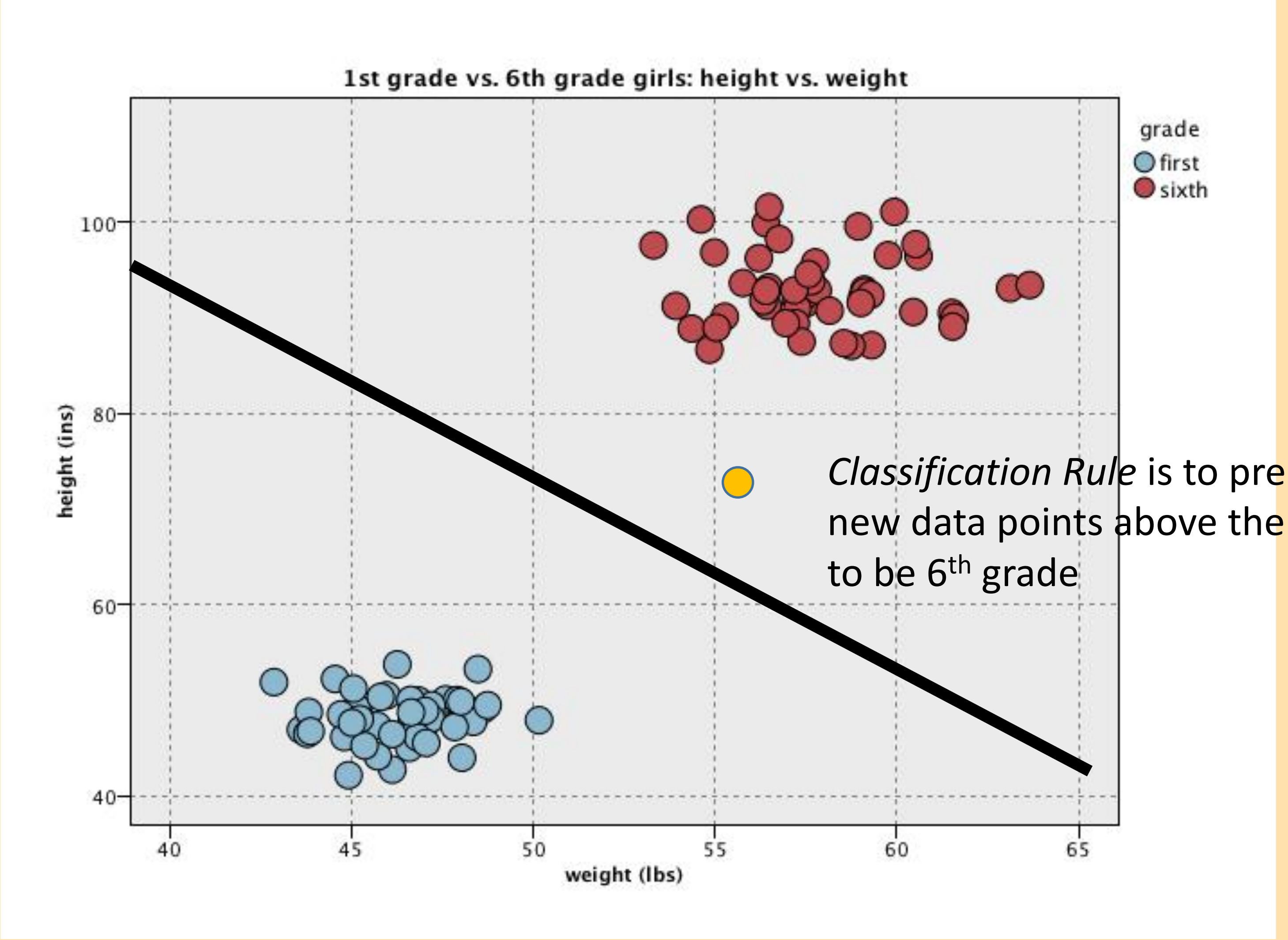
2D Classification example



2D Classification example

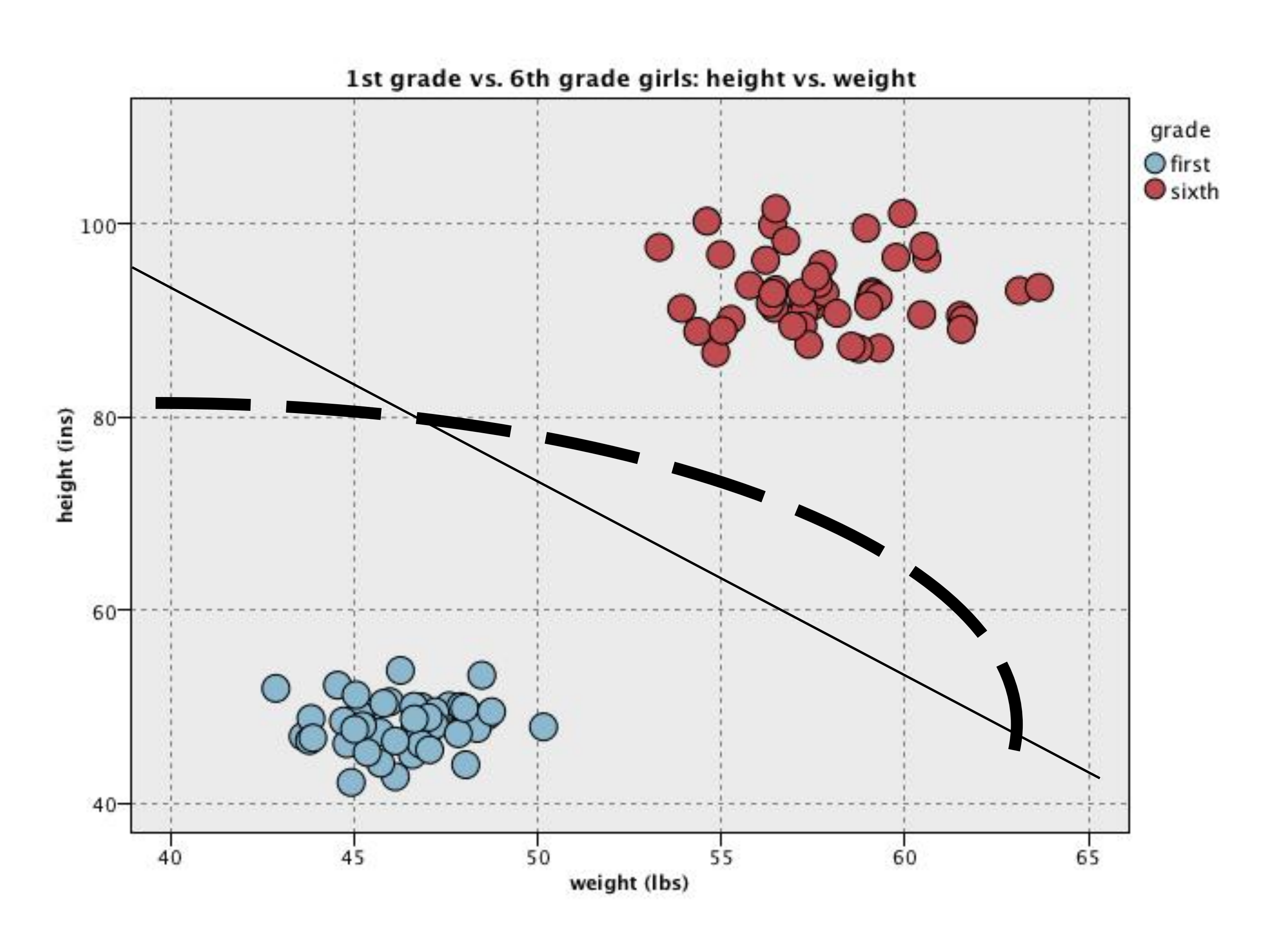


2D Classification example

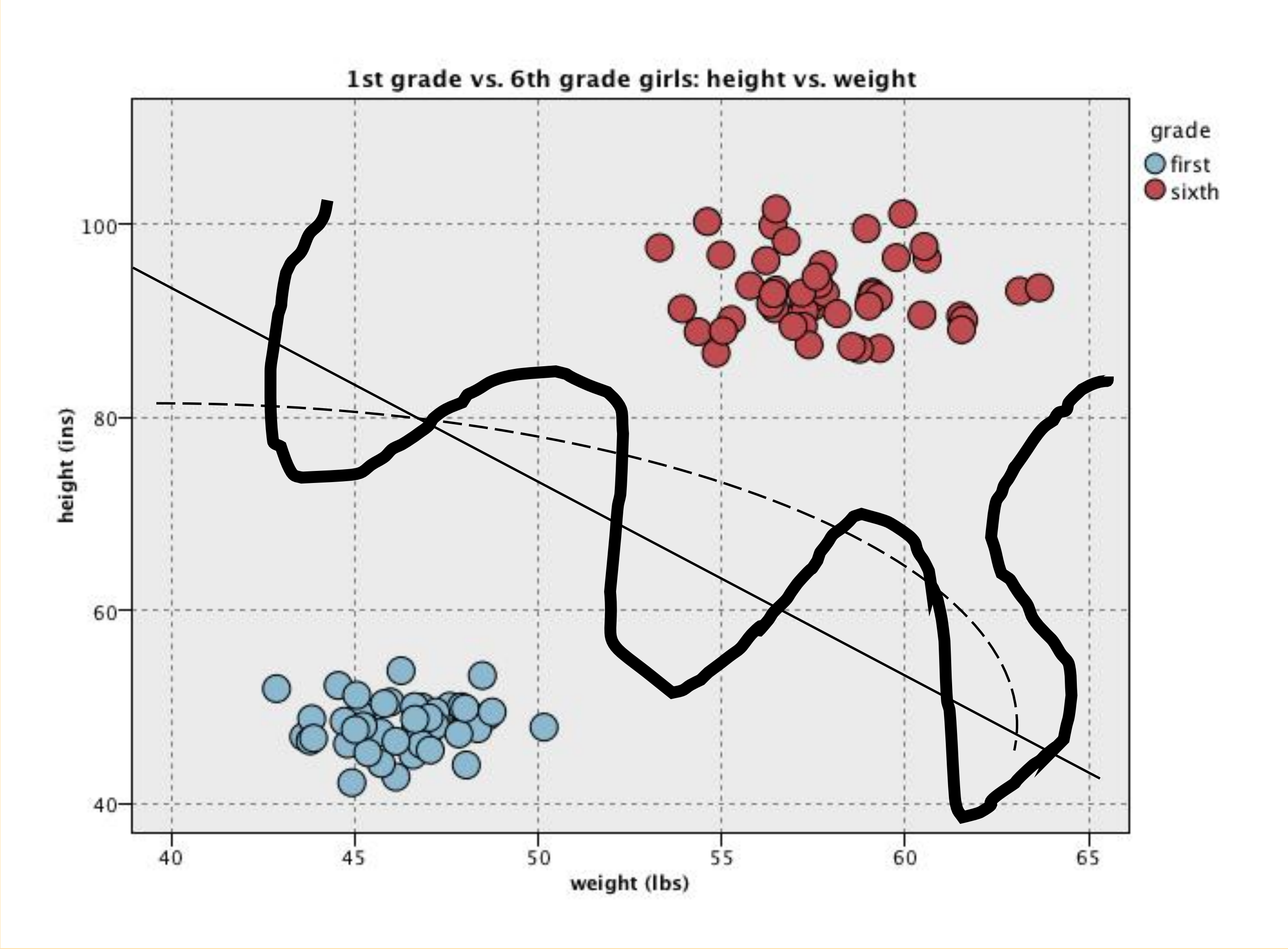


Classification Rule is to predict new data points above the line to be 6th grade

2D Classification example

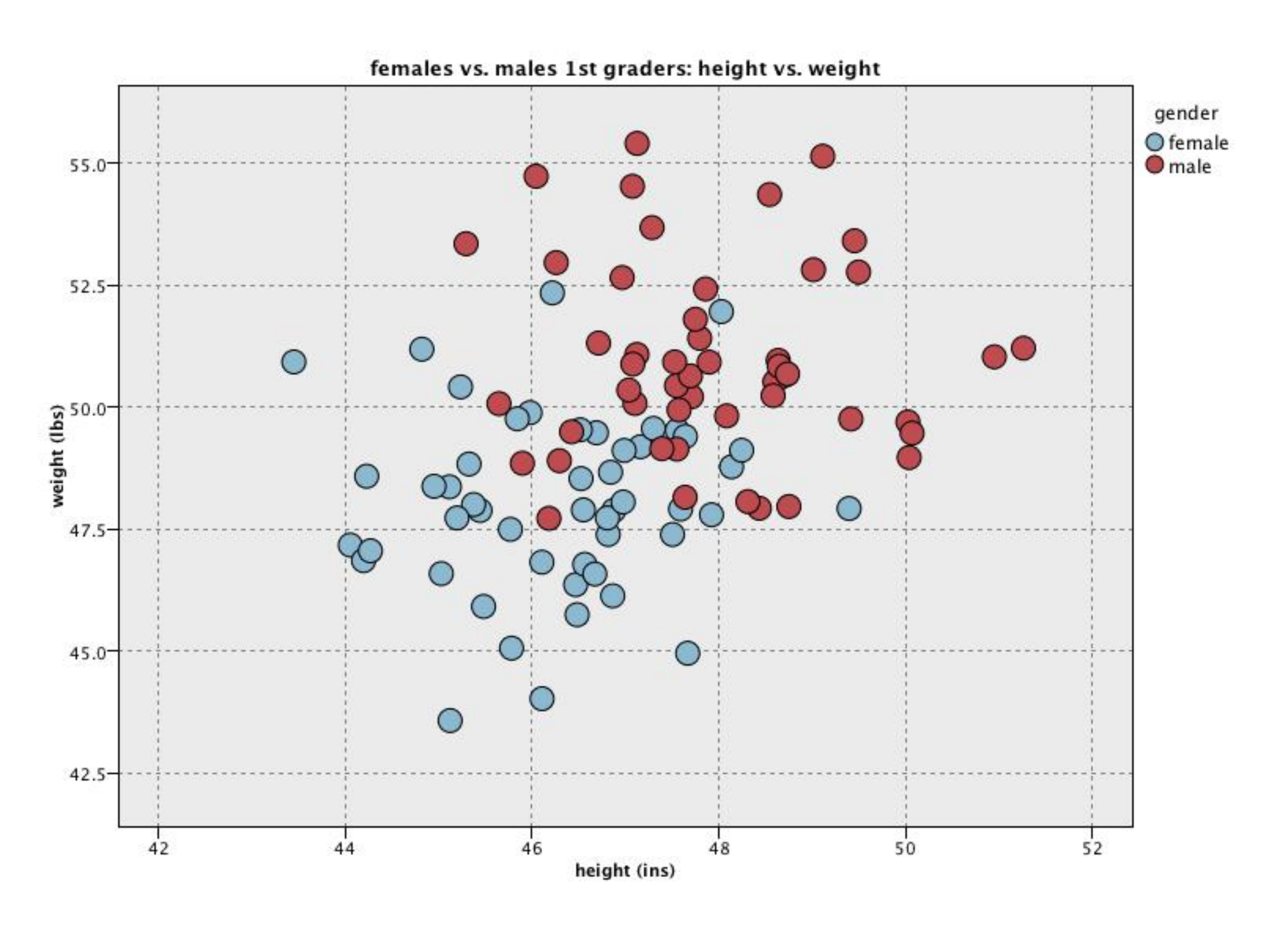


2D Classification example

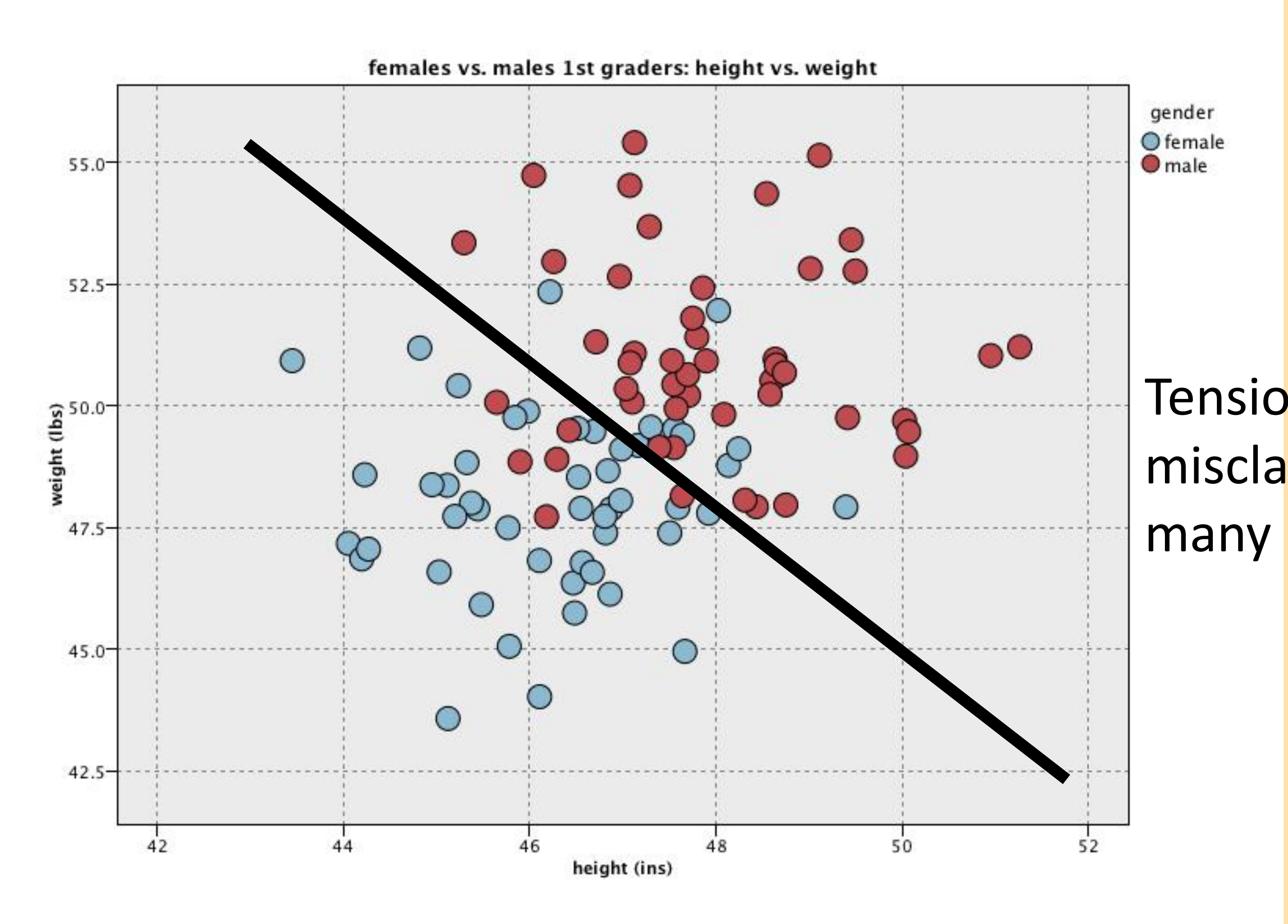


- But in real problems the separation is unlikely to be so neat
- Consider 1st grade boys vs. 1st grade girls

2D Classification example

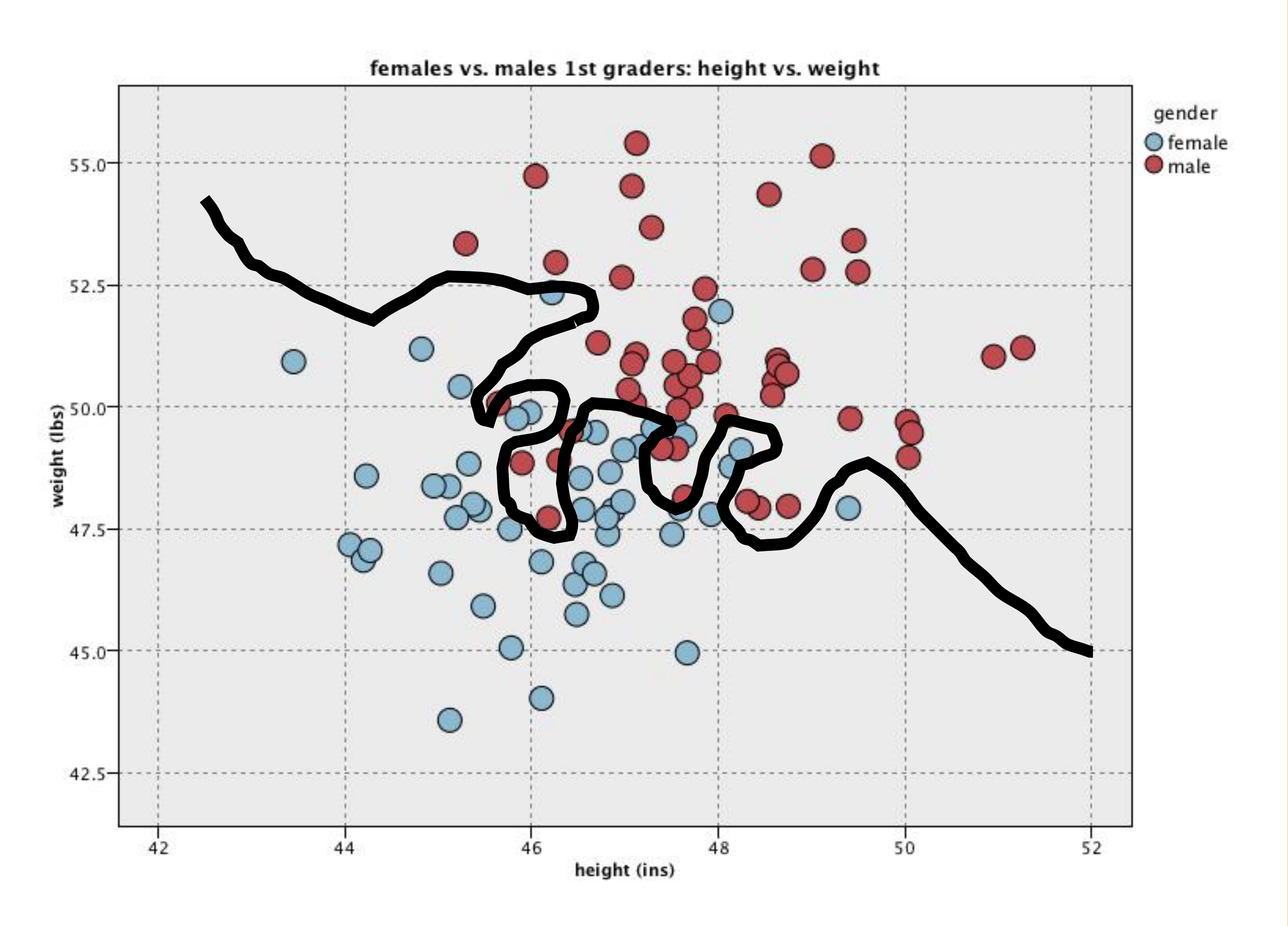


2D Classification example

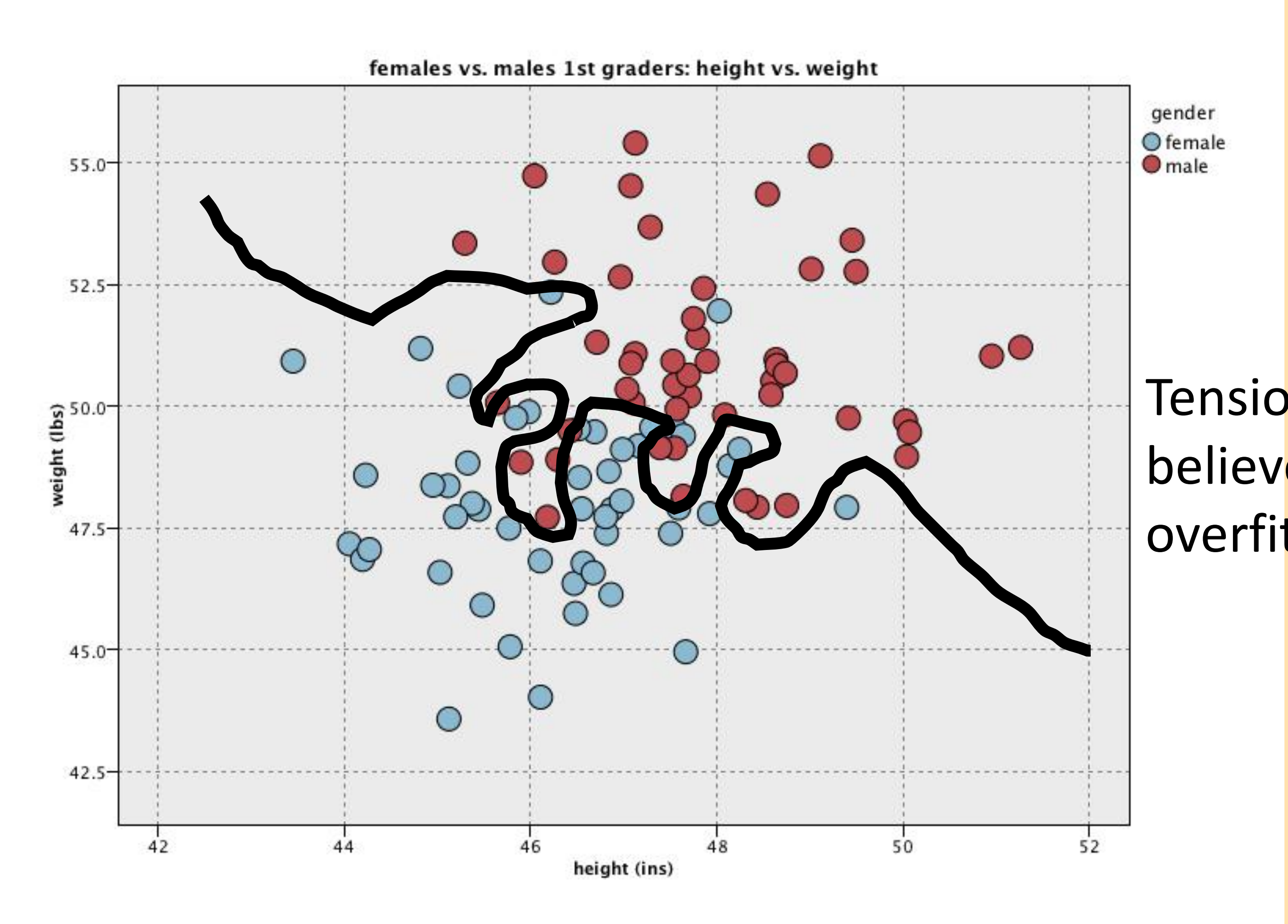


Tension: misclassifying many points

2D Classification example



2D Classification example



Tension: don't believe real - overfitting

The trade-off

- Want to find true predictive patterns in the data
- But, want to avoid overfitting spurious detail: i.e. don’t want to make the model capture apparent patterns due to noise - we want results that are likely to be *generalizable* to future observations
- Solution – independent testing – reserve some data that has not been used in model fitting for the purpose of model evaluation and comparison

Definitions

- True Positive (TP) – observation is positive and was predicted as positive
- True Negative (TN) – observation is negative and was predicted as negative
- False Positive (FP) – observation is negative and was predicted as positive
- False Negative (FN) – observation is positive and was predicted as negative
- Classification accuracy – proportion of observations classified correctly = (#TP + #TN)/N

Evaluation

- Possible evaluation measures:
  - *Classification Accuracy* – or conversely, the *classification error rate* = 1 – classification accuracy
  - *Differential penalties* – different errors may have different costs
  - Area Under Receiver-Operating Characteristic (ROC) Curve (AUC)

- How reliable are the predicted results ?

Classifier error rate

- Natural performance measure for classification problems: *error rate (or success rate)*
  - *Success*: instance's class is predicted correctly
  - *Error*: instance's class is predicted incorrectly
  - *Error rate*: proportion of errors made over the whole set of instances
  - *Success rate*: proportion of errors made over the whole set of instances = 1 – Error rate

Cost-Benefit

- Cost-benefit is similar to classification error, except that we might want to count some errors to be worse than others
- Consider a disease where treatment involves taking a simple cheap and benign medicine to cure it (e.g. taking vitamin C), but if the disease is unchecked, the consequences are severe.
- In that case, I would want to put much more weight on missing positives than missing negatives; i.e. consequences of false negatives are much worse than false positives.
- For example, we might make the penalty (cost) for a false negative be 10 or 100 times that of a false positive

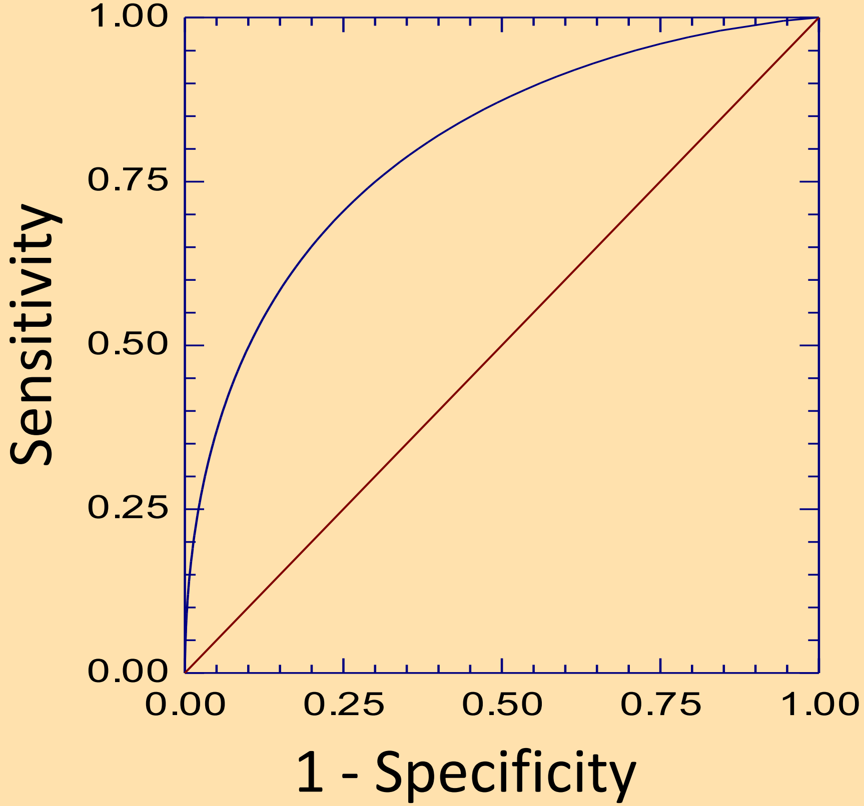
Sensitivity and Specificity

Test\Truth	Disease Free	Diseased	Total
Negative	TN	FN	FN+TN
Positive	FP	TP	TP+FP
Total	FP+TN	TP+FN	N

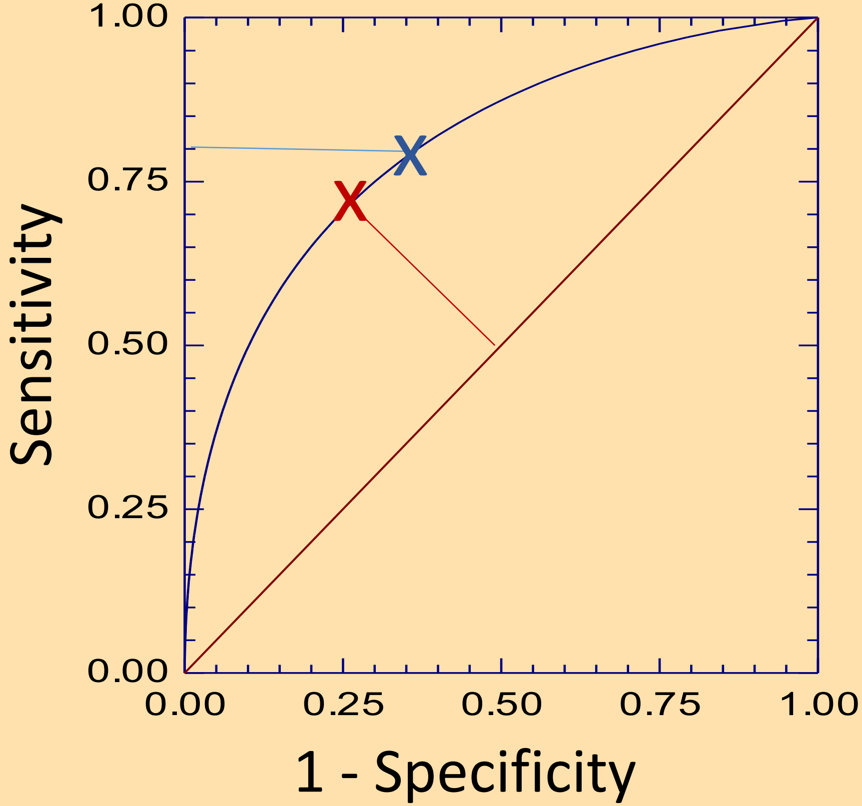
Sensitivity and Specificity

- $\text{Sensitivity} = \dfrac{\#\text{TP}}{\#\text{TP}+\#\text{FN}} =$ Proportion of positives correctly identified as positive
- $\text{Specificity} = \dfrac{\#\text{TN}}{\#\text{TN}+\#\text{FP}} =$ Proportion of negatives correctly identified as negative
- There is another trade-off here between sensitivity and specificity
  - Imagine a classification procedure that gives you a probability of being positive for a disease (such classifiers are common)
  - To classify individuals we need to choose a probability threshold to identify them as positive vs. negative
  - Increasing the threshold increases specificity but reduces sensitivity
  - Decreasing the threshold increases sensitivity but reduces specificity

Receiver operating characteristic (ROC) curve

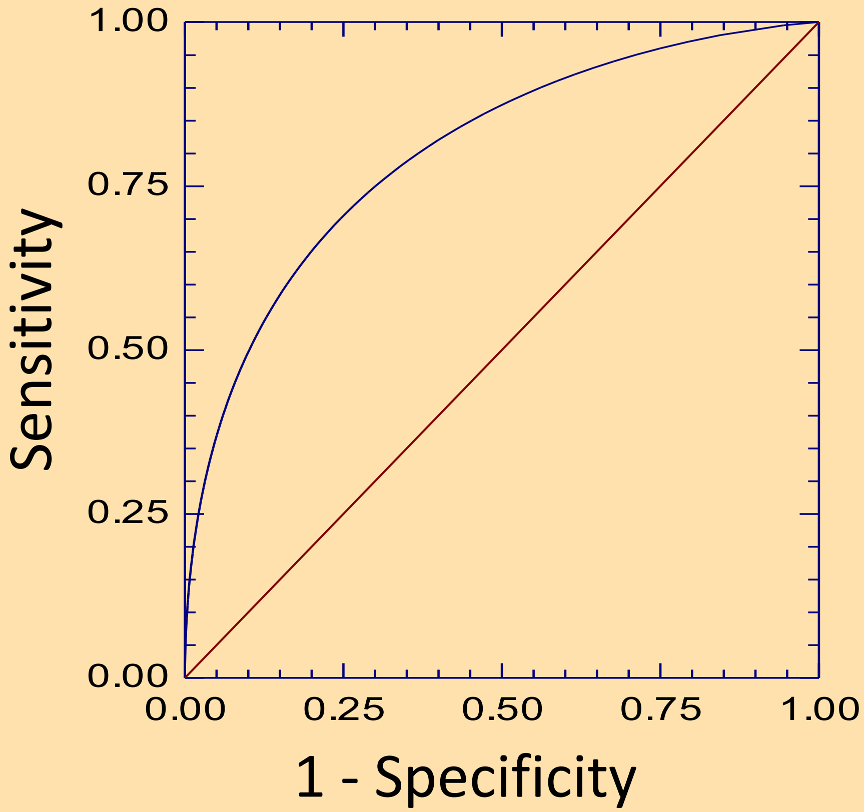


Choosing threshold based on ROC



- Many possible thresholds, each corresponding to one point on curve
- Choosing a single one depends on context
- Some studies say (for example) give me best specificity for a sensitivity of at least 0.8 (blue x)
- General alternative – maximal perpendicular from diagonal (red x)
- We will not be concerned with choosing thresholds, rather, we will use ROC to generate a metric

Area under the (ROC) curve (AUC)



- AUC gives a measure of how well the predictor does across all possible thresholds
- Ideal AUC is 1
- AUC for random predictor = 0.5
- Objective is to find classifier that maximizes AUC

Error rate vs. AUC

- When should we choose error rate and when AUC?
- Error rate (and cost-benefit) work well when your dataset is representative of the larger population (metric incorporates prevalence – e.g. proportion of people in population that have disease)
- AUC works better when your dataset is not representative of larger population, e.g. in “case-control” studies (metric is independent of prevalence)

Training set error rate: **is generally too optimistic!**

You can find patterns even in random data

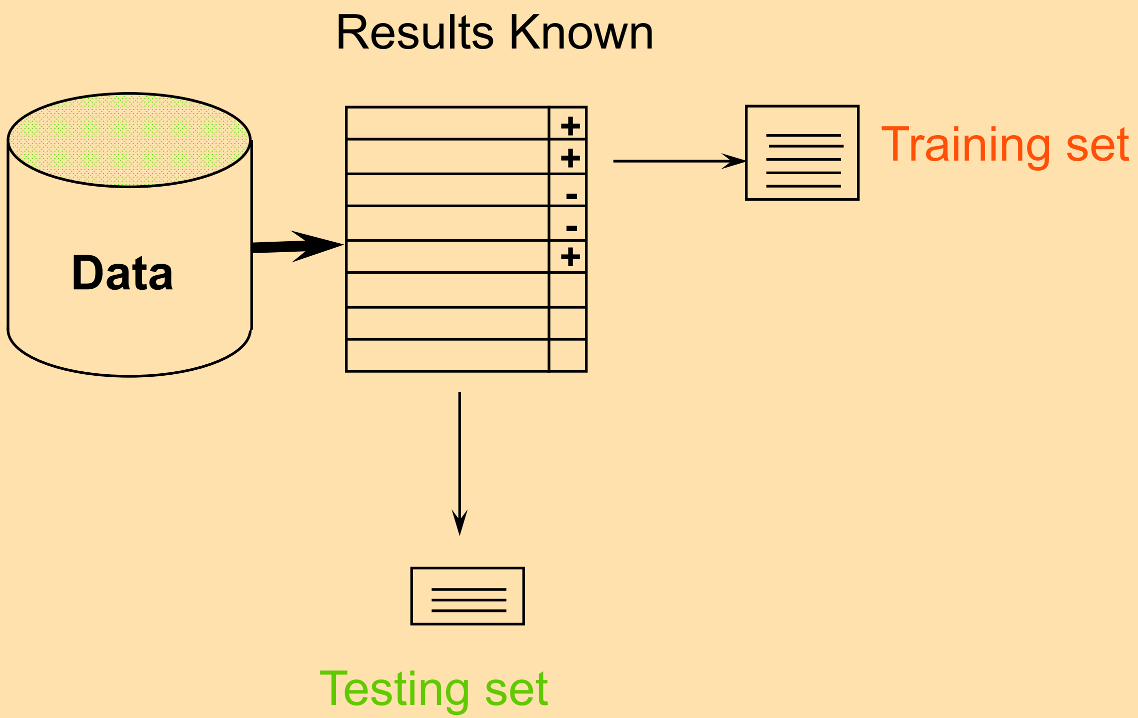
So we need to evaluate on separate independent (*Test*) set

When you have lots of “instances”

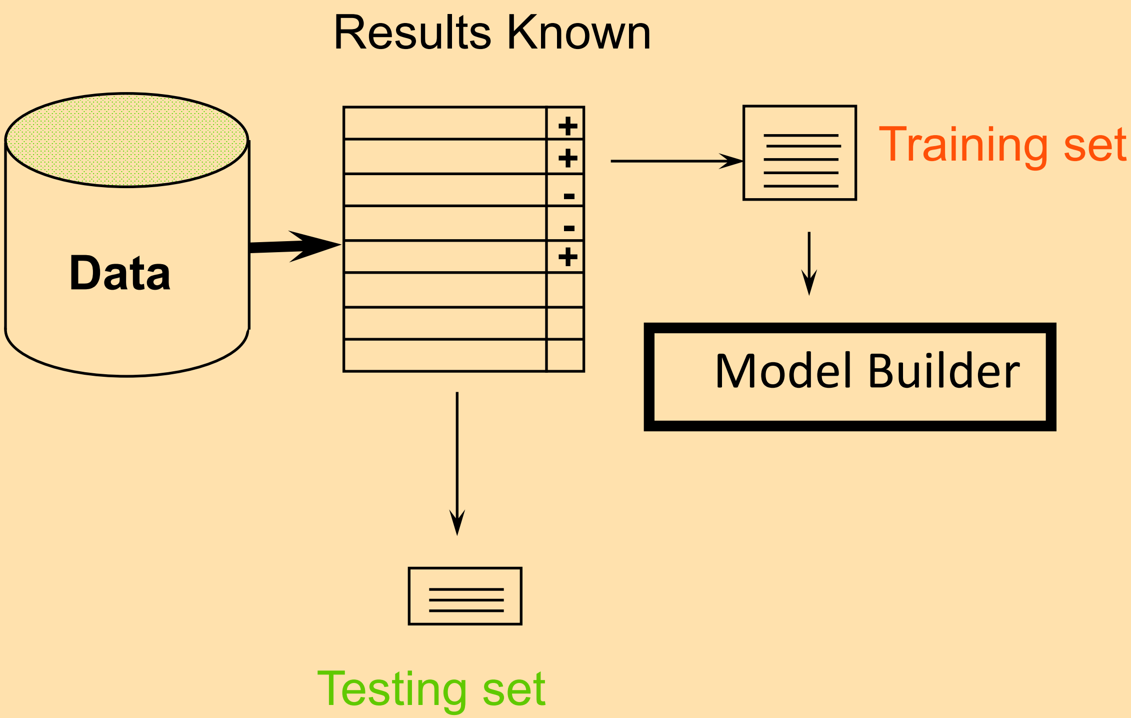
If many (>1000) instances are available, including >100 from each class

- A simple evaluation training/test evaluation will work
    - Randomly split data into training and test sets (often ~2/3 for train, 1/3 for test) – may want to *stratify* on outcome (have matching proportions of each class in each set)
- Build a classifier using the *training* set and evaluate it using the *test* set
- We will look at cross-validation later for when you do not have so many observations

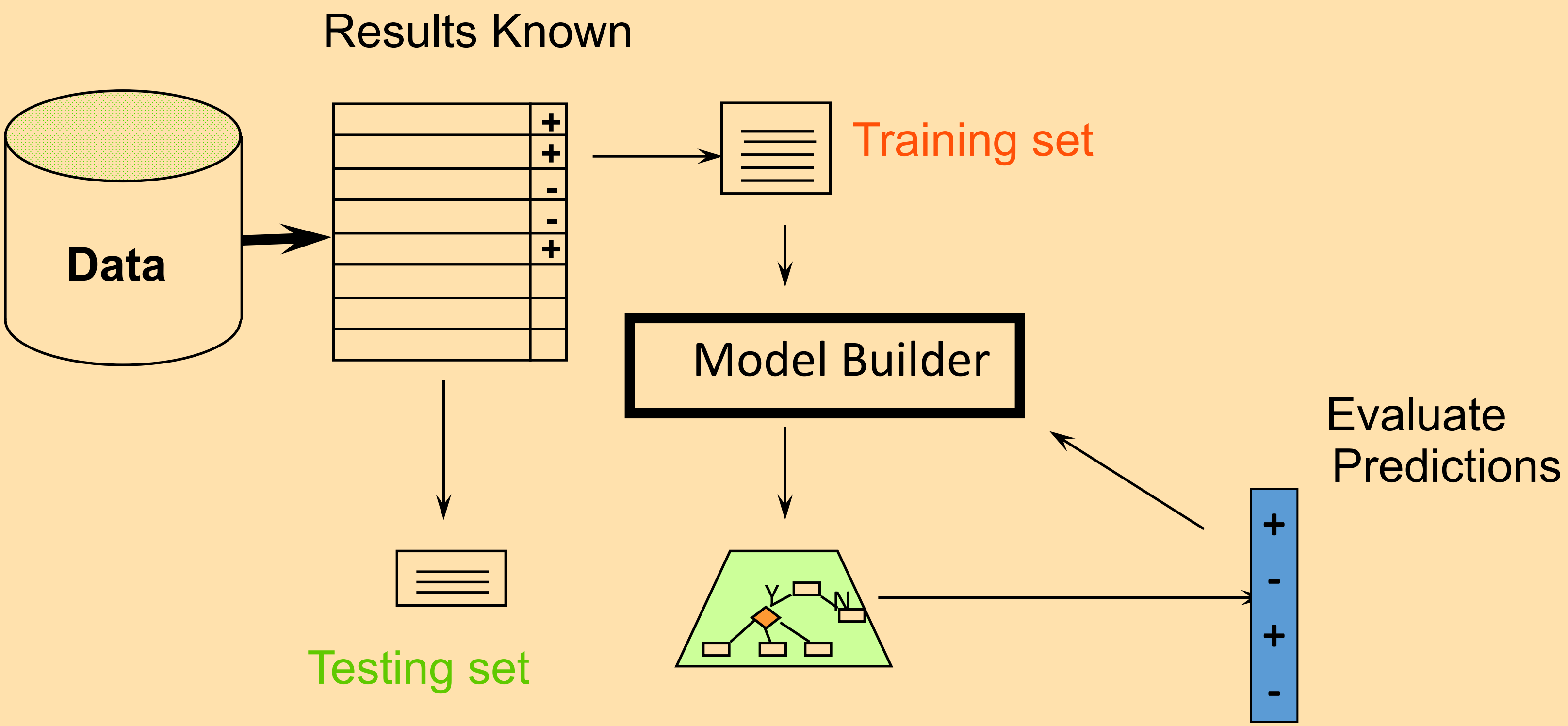
When you have lots of "instances"



When you have lots of "instances"

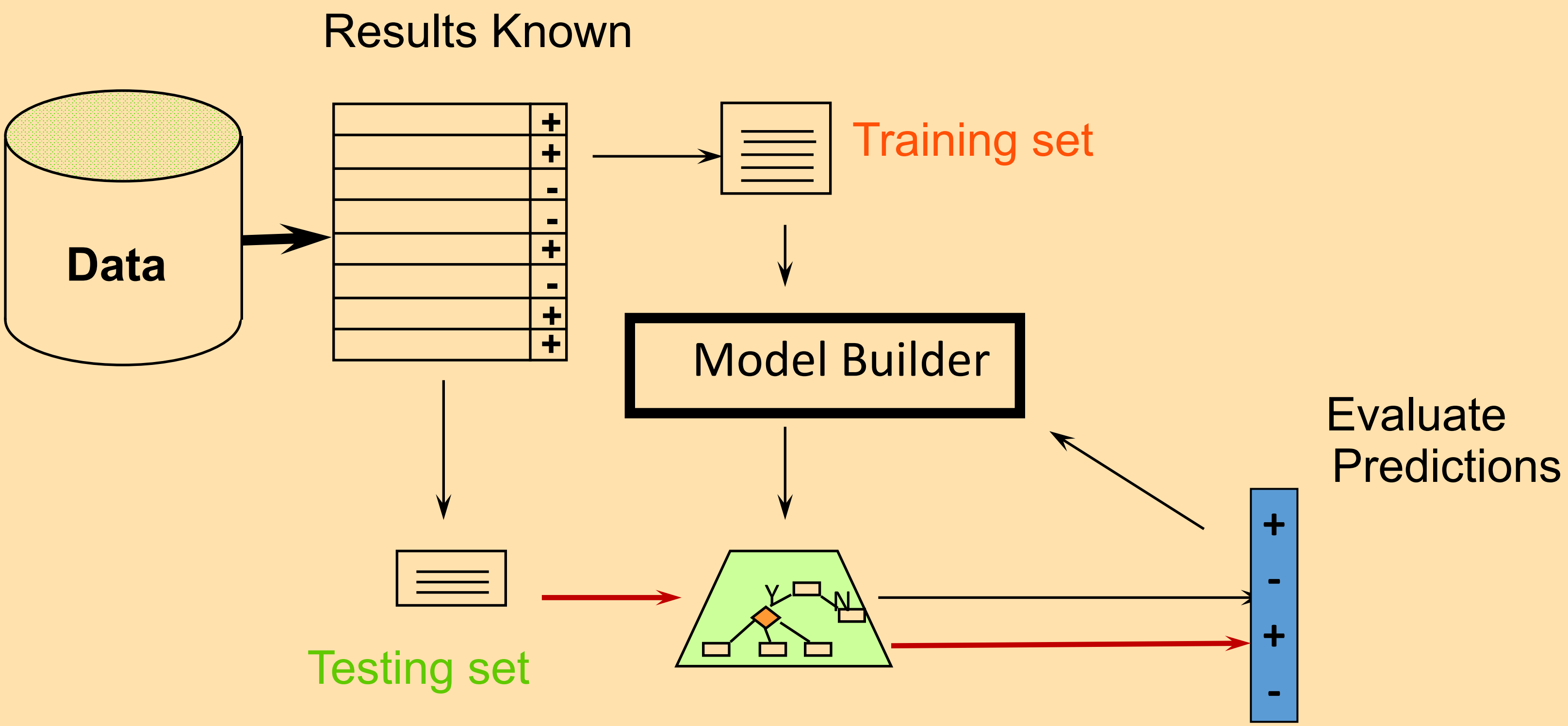


When you have lots of “instances”



At least greater than a few hundred instances and better >1000

When you have lots of "instances"



At least greater than a few hundred instances and better >1000

Classification methods

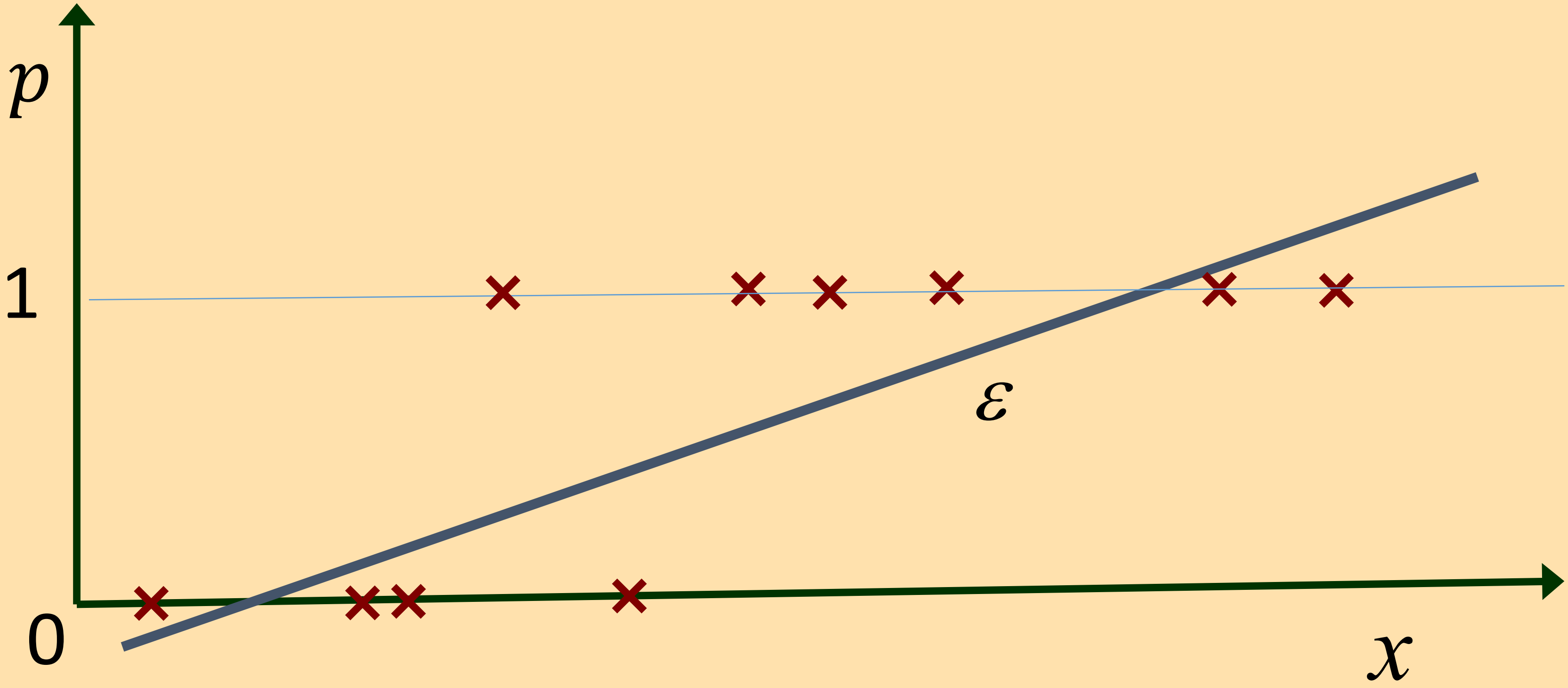
- Logistic regression
- K-nearest neighbors
- Classification trees
- Support vector machines (SVMs)
- Linear Discriminant Analysis (LDA)
- Neural Networks

Logistic Regression

Binary/flag outcome regression modeling

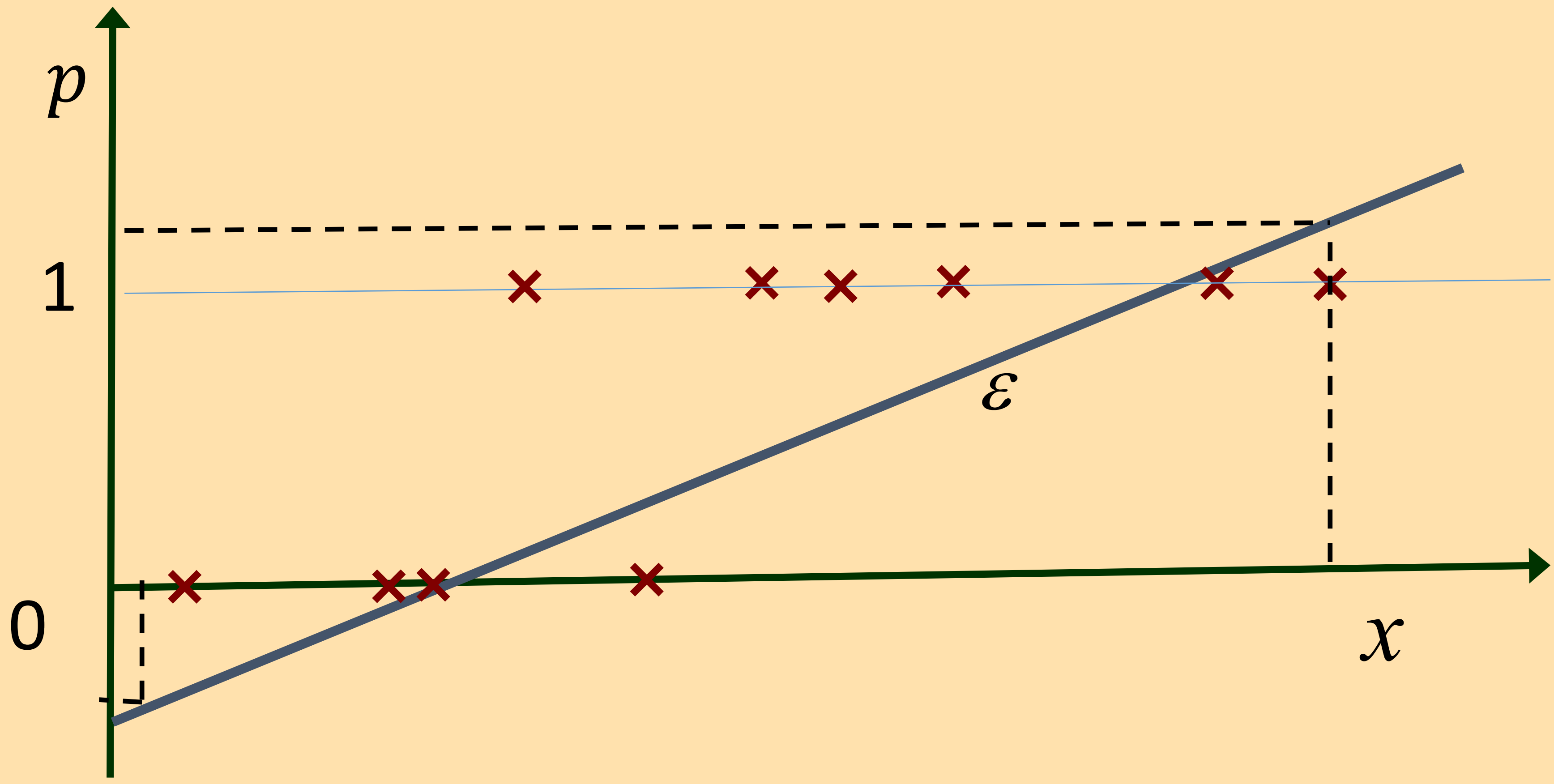
- We want a model where the probability of being in a class is based on the values of predictor(s)
- What about using linear regression?
- $p = ax + c$, where p is the probability of being in class 1 vs. 0, e.g. $p =$ probability boy (=1) vs. girl (=0)
- What might be the concern here?

Linear regression?



Linear regression

Problem predictions – outside range 0 to 1



Binary/flag outcome regression modeling

- We want a model where the probability of being in a class depends on predictor(s)
- What about using linear regression?
- $p = ax + c$, where p is the probability of being in class 1 vs. 0, e.g. $p =$ probability boy (=1) vs. girl (=0)
- What might be the concern here?
- Possible values for p are not restricted to be between 0 and 1 (as required for probabilities) – can be >1 or <0

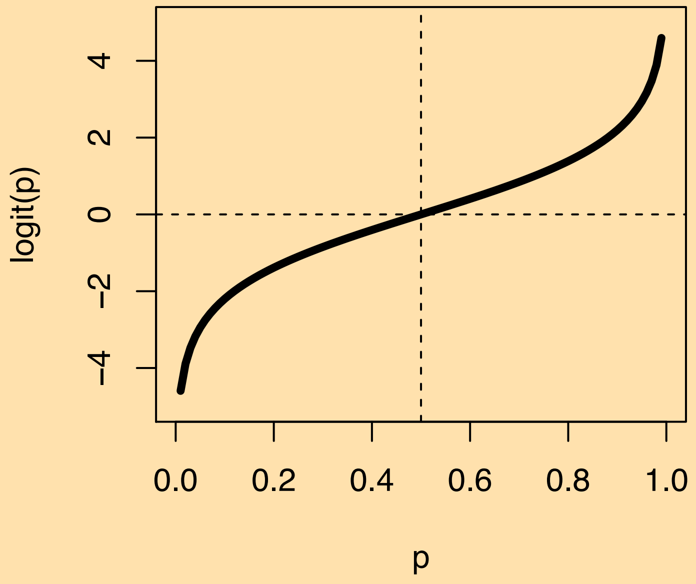
Logistic regression

p must be between 0 and 1 (probability)

We can instead take the *odds* which is between 0 and infinity: $\frac{p}{1-p} > 0$ but still not quite what we need…

to get values in the range 0 to 1, we take the log of the odds (*logit function*) and get the logistic regression equation:

$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right) = ax + c$$



We can extend logistic regression to multiple predictors just as we did for linear regression.

Logistic regression for classification

- When fitting logistic regression we get a model that relates predictors to a class probability
- To convert the probability to a class we need to threshold the probability
- E.g. if probability $p > 0.5$ then class as boy, otherwise if $p \leq 0.5$ class as girl

Logistic regression

- A long history (relatively) – well known
- Much statistical theory
- Works well for many problems
- Can be extended to multiple categories
- Can use step-wise/subset selection techniques
- Can provide interpretable models

Running example: UCSF Memory and Aging Clinic (MAC) Dataset

- Structural (volumetric) MRI – multiple dementia studies – but in same scanner and same protocol
- Single time point
- Gold standard Dx – autopsy (if dead) UCSF clinical expert joint decision (otherwise)

CN	SV	BV	AD	MCI	DLB	PSP	nfPPA
291	44	68	180	68	0	44	38

- Objective is to classify patients as having or not having dementia (binary) and if yes, what type of dementia (multi-class)?

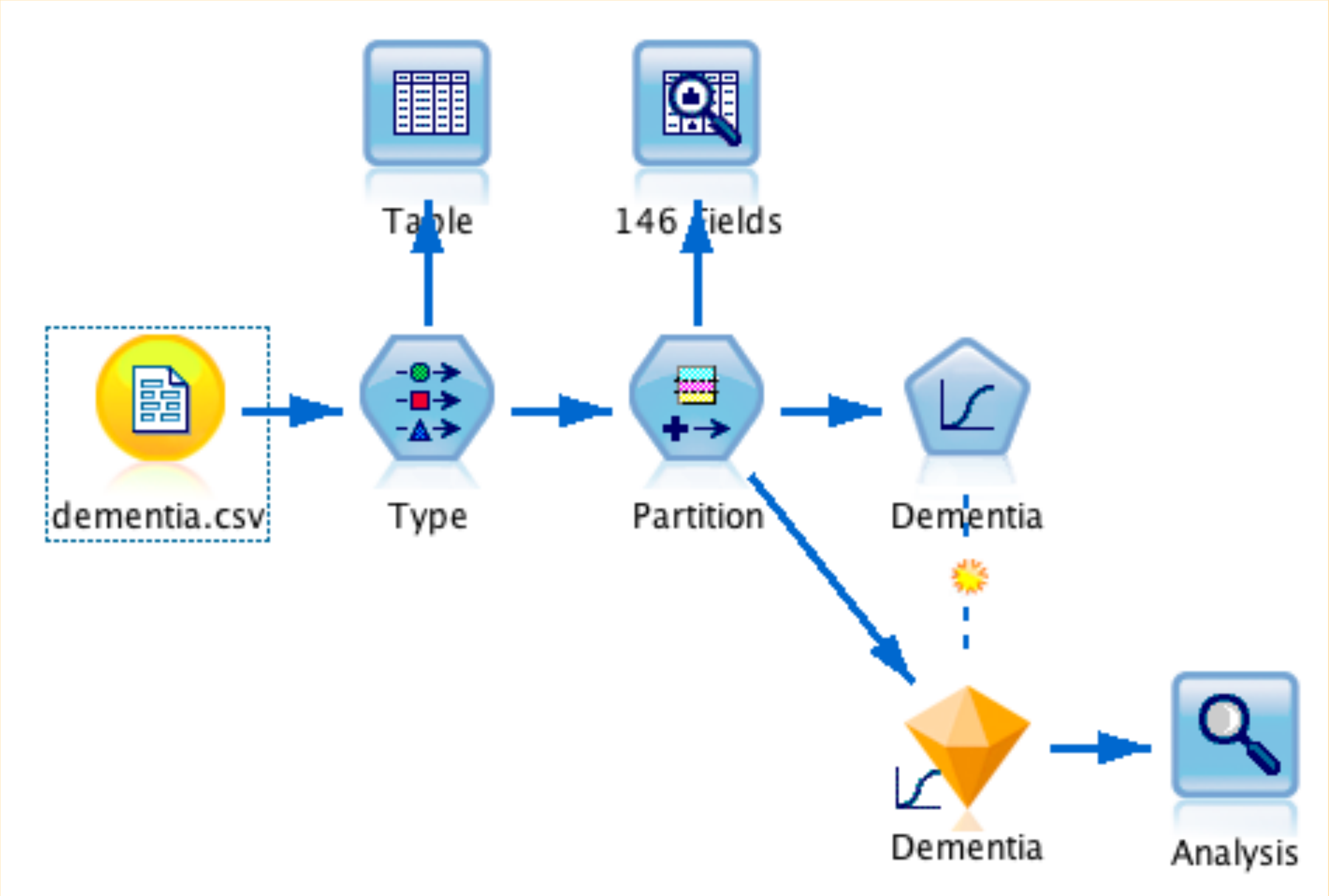
ROI predictors

- ROI volumes are the main predictors: Extract 136 ROIs as predictors for each subject – big data?
- ROIs are defined in terms of anatomical regions and landmarks in the human brain that are common to all individuals – often have specific functional roles

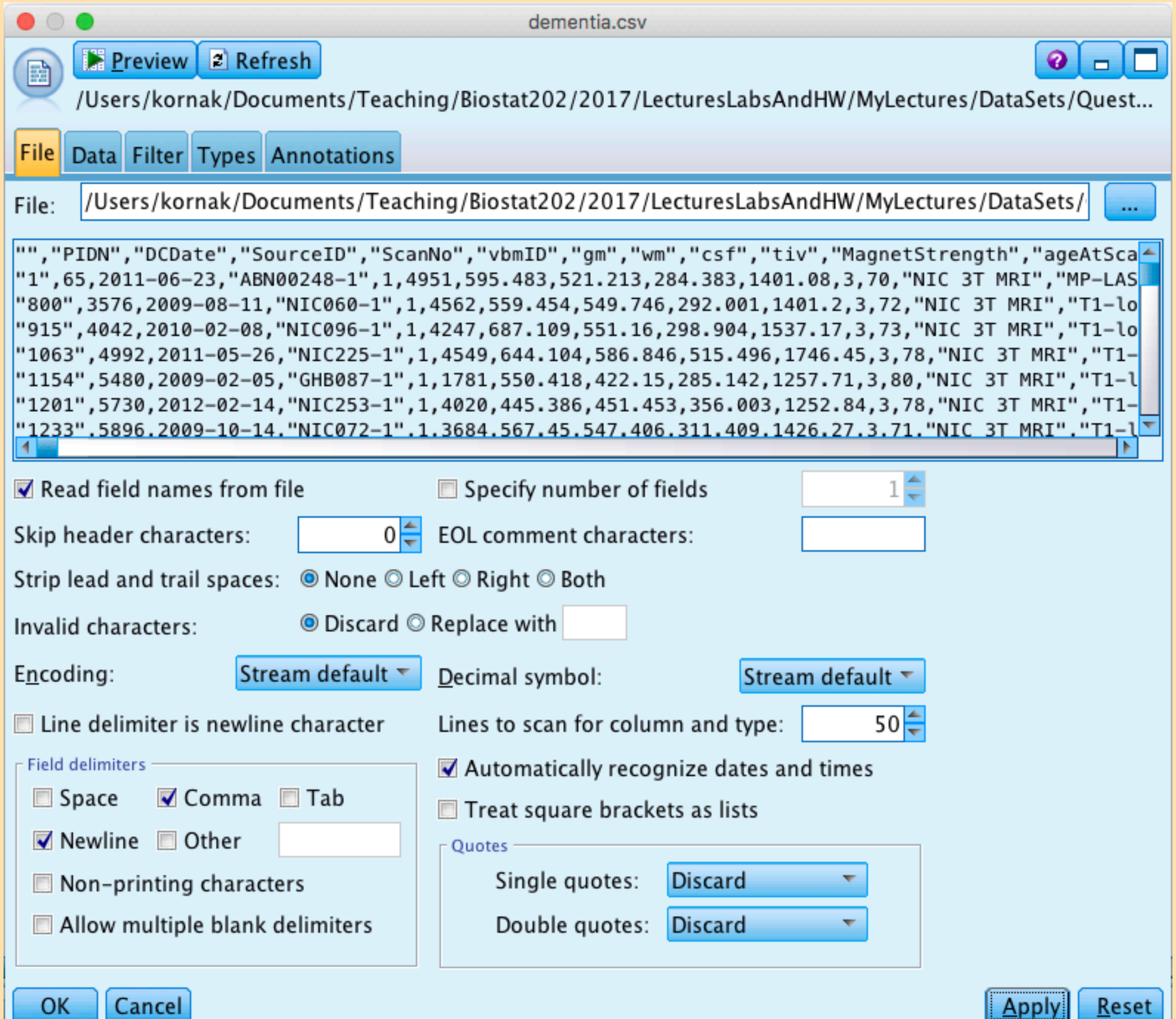
IBM Modeler analysis

- Classification analysis – predicting dementia existence and type based on voxel-based-morphometry gray matter maps
- Using region of interest (ROI) estimates as predictors
- So let’s look at logistic regression predicting yes/no for dementia

IBM Modeler stream



Var. File node



Var. File node – Preview

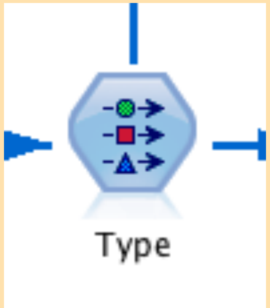
Preview from dementia.csv Node (163 fields, 10 records)

File Edit Generate

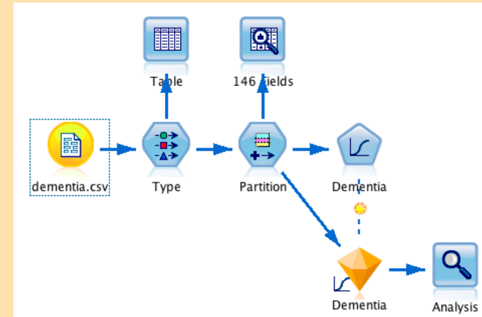
Table Annotations

	field1	PIDN	DCDate	SourceID	ScanNo	vbmID	gm	wm	csf	tiv	MagnetStrength	ageAtScan	ScannerID	ScanType	ROWNUM	InstrID	ScannerID_2
1	1	65	2011-06-23	ABN00248-1	1	4951	595.483	521.213	284.383	1401.0...	3	70	NIC 3T MRI	MP-LAS-ABN	1	2106...	NIC 3T MRI
2	800	3576	2009-08-11	NIC060-1	1	4562	559.454	549.746	292.001	1401.2...	3	72	NIC 3T MRI	T1-long	1493	1486...	NIC 3T MRI
3	915	4042	2010-02-08	NIC096-1	1	4247	687.109	551.160	298.904	1537.1...	3	73	NIC 3T MRI	T1-long	1696	1618...	NIC 3T MRI
4	1063	4992	2011-05-26	NIC225-1	1	4549	644.104	586.846	515.496	1746.4...	3	78	NIC 3T MRI	T1-long	1972	2078...	NIC 3T MRI
5	1154	5480	2009-02-05	GHB087-1	1	1781	550.418	422.150	285.142	1257.7...	3	80	NIC 3T MRI	T1-long	1153	1328...	NIC 3T MRI
6	1201	5730	2012-02-14	NIC253-1	1	4020	445.386	451.453	356.003	1252.8...	3	78	NIC 3T MRI	T1-long	2105	2306...	NIC 3T MRI
7	1233	5896	2009-10-14	NIC072-1	1	3684	567.450	547.406	311.409	1426.2...	3	71	NIC 3T MRI	T1-long	2165	1552...	NIC 3T MRI
8	1237	5927	2010-07-08	ADRC0054-1	1	2697	504.363	391.272	336.735	1232.3...	3	82	NIC 3T MRI	T1-long	2172	1777...	NIC 3T MRI
9	1321	6438	2010-01-05	GHB212-1	1	2583	589.887	468.907	264.074	1322.8...	3	65	NIC 3T MRI	T1-long	1320	1578...	NIC 3T MRI
10	1327	6565	2008-04-18	GHB005-1	1	835	558.285	435.448	215.101	1208.8...	3	66	NIC 3T MRI	T1-long	1326	1264...	NIC 3T MRI

OK



Type node – initial



Type

Preview

Types | Format | Annotations

Read Values | Clear Values | Clear All Values

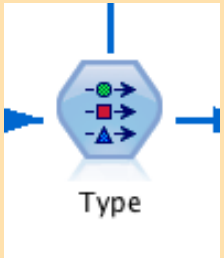
Field	Measurement	Values	Missing	Check	Role
field1	Continuous	<Read>		None	Input
PIDN	Continuous	<Read>		None	Input
DCDate	Continuous	<Read>		None	Input
SourceID	Categorical	<Read>		None	Input
ScanNo	Continuous	<Read>		None	Input
vbmID	Continuous	<Read>		None	Input
gm	Continuous	<Read>		None	Input
wm	Continuous	<Read>		None	Input
csf	Continuous	<Read>		None	Input
tiv	Continuous	<Read>		None	Input
MagnetStreng...	Continuous	<Read>		None	Input
ageAtScan	Continuous	<Read>		None	Input
ScannerID	Categorical	<Read>		None	Input
ScanType	Categorical	<Read>		None	Input
ROWNUM	Continuous	<Read>		None	Input
InstrID	Continuous	<Read>		None	Input
ScannerID_2	Categorical	<Read>		None	Input
MacCohort_kr	Categorical	<Read>		None	Input
PIBPos	Categorical	<Read>		None	Input
Sex	Categorical	<Read>		None	Input
Educ	Continuous	<Read>		None	Input
Race	Categorical	<Read>		None	Input
DateofDeath	Continuous	<Read>		None	Input
autopsy.	Categorical	<Read>		None	Input
PIBDone.	Categorical	<Read>		None	Input
X3rd_Ventricle	Continuous	<Read>		None	Input
X4th_Ventricle	Continuous	<Read>		None	Input
Right_Accum...	Continuous	<Read>		None	Input
Left_Accumbe...	Continuous	<Read>		None	Input
Right_Amygd...	Continuous	<Read>		None	Input
Left_Amygdala	Continuous	<Read>		None	Input
Brain_Stem	Continuous	<Read>		None	Input
Right_Caudate	Continuous	<Read>		None	Input
Left_Caudate	Continuous	<Read>		None	Input
Right_Cerebel...	Continuous	<Read>		None	Input
Left_Cerebell...	Continuous	<Read>		None	Input
Right_Cerebel...	Continuous	<Read>		None	Input
Left_Cerebell...	Continuous	<Read>		None	Input
Right_Cerebr...	Continuous	<Read>		None	Input

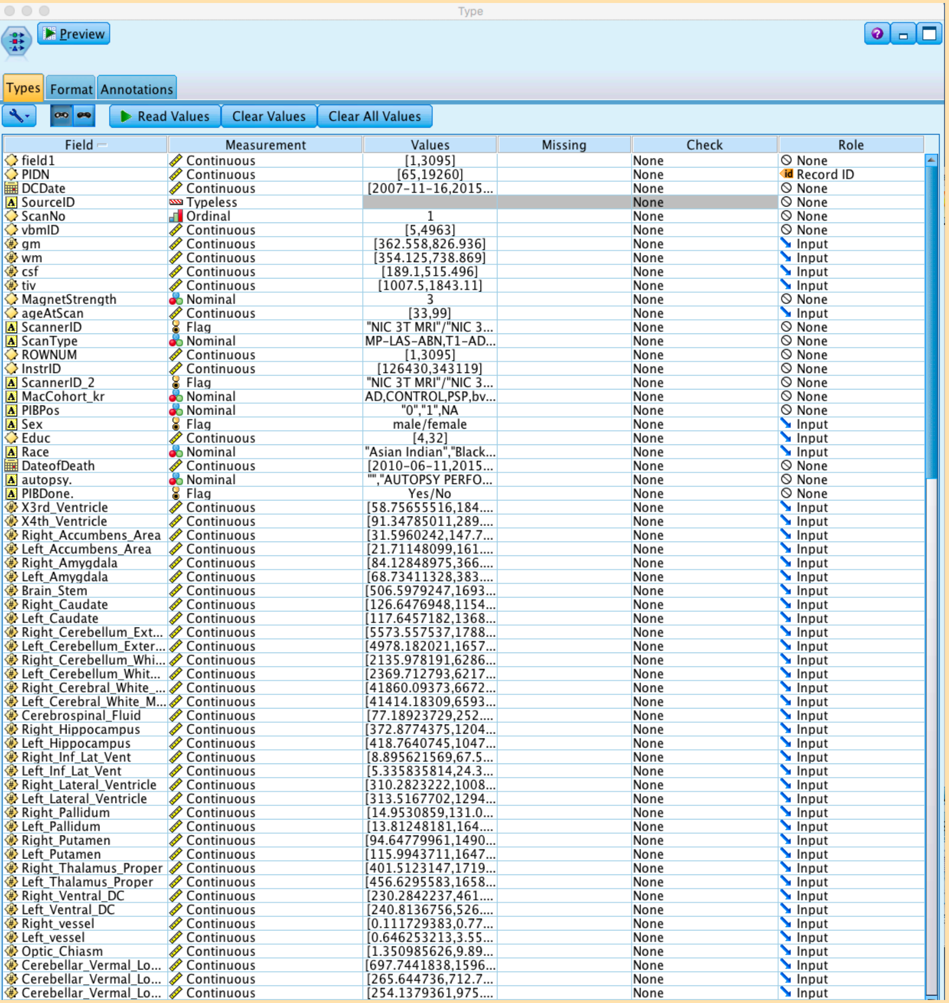
Left_PCu_prec...	Continuous	<Read>		None	Input
Right_PHG_pa...	Continuous	<Read>		None	Input
Left_PHG_par...	Continuous	<Read>		None	Input
Right_PIns_po...	Continuous	<Read>		None	Input
Left_PIns_post...	Continuous	<Read>		None	Input
Right_PO_pari...	Continuous	<Read>		None	Input
Left_PO_parie...	Continuous	<Read>		None	Input
Right_PoG_po...	Continuous	<Read>		None	Input
Left_PoG_post...	Continuous	<Read>		None	Input
Right_POrG_p...	Continuous	<Read>		None	Input
Left_POrG_po...	Continuous	<Read>		None	Input
Right_PP_plan...	Continuous	<Read>		None	Input
Left_PP_planu...	Continuous	<Read>		None	Input
Right_PrG_pre...	Continuous	<Read>		None	Input
Left_PrG_prec...	Continuous	<Read>		None	Input
Right_PT_plan...	Continuous	<Read>		None	Input
Left_PT_planu...	Continuous	<Read>		None	Input
Right_SCA_su...	Continuous	<Read>		None	Input
Left_SCA_sub...	Continuous	<Read>		None	Input
Right_SFG_su...	Continuous	<Read>		None	Input
Left_SFG_sup...	Continuous	<Read>		None	Input
Right_SMC_su...	Continuous	<Read>		None	Input
Left_SMC_sup...	Continuous	<Read>		None	Input
Right_SMG_su...	Continuous	<Read>		None	Input
Left_SMG_sup...	Continuous	<Read>		None	Input
Right_SOG_su...	Continuous	<Read>		None	Input
Left_SOG_sup...	Continuous	<Read>		None	Input
Right_SPL_su...	Continuous	<Read>		None	Input
Left_SPL_supe...	Continuous	<Read>		None	Input
Right_STG_su...	Continuous	<Read>		None	Input
Left_STG_sup...	Continuous	<Read>		None	Input
Right_TMP_te...	Continuous	<Read>		None	Input
Left_TMP_tem...	Continuous	<Read>		None	Input
Right_TrIFG_t...	Continuous	<Read>		None	Input
Left_TrIFG_tri...	Continuous	<Read>		None	Input
Right_TTG_tra...	Continuous	<Read>		None	Input
Left_TTG_tran...	Continuous	<Read>		None	Input
drop	Categorical	<Read>		None	Input
Dementia	Categorical	<Read>		None	Input

View current fields   View unused field settings

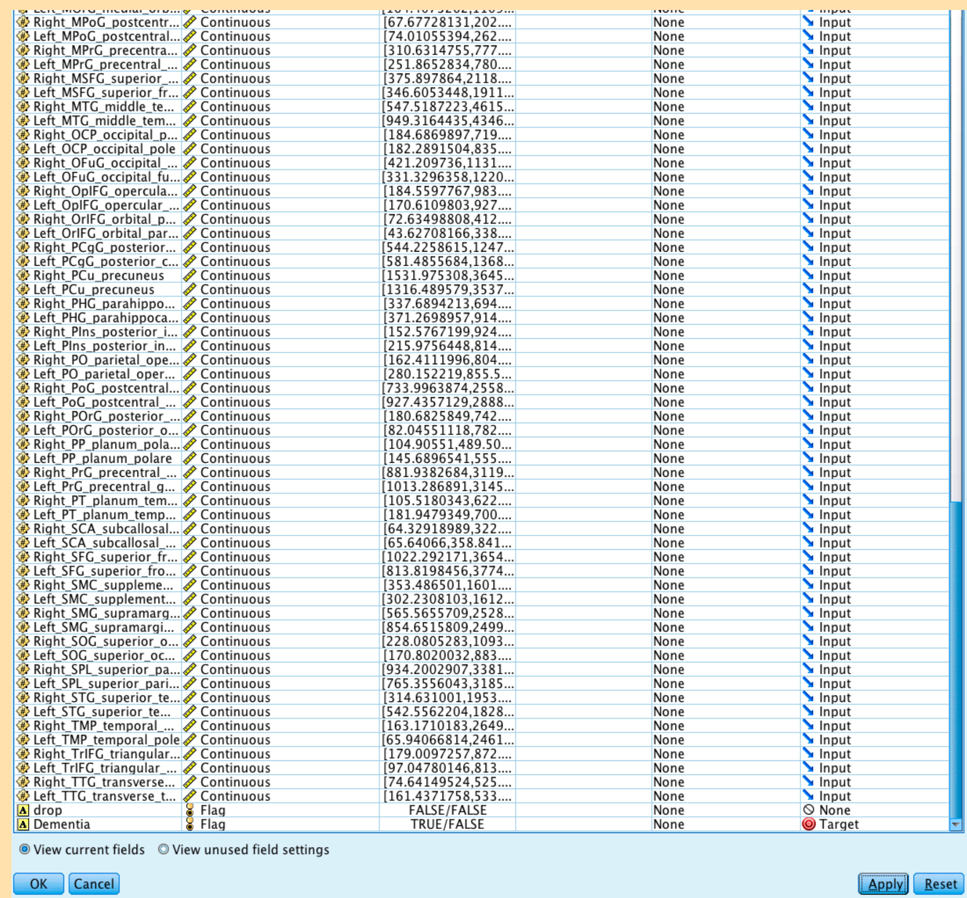
OK | Cancel | Apply | Reset

Type node – after correcting



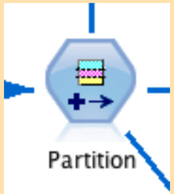


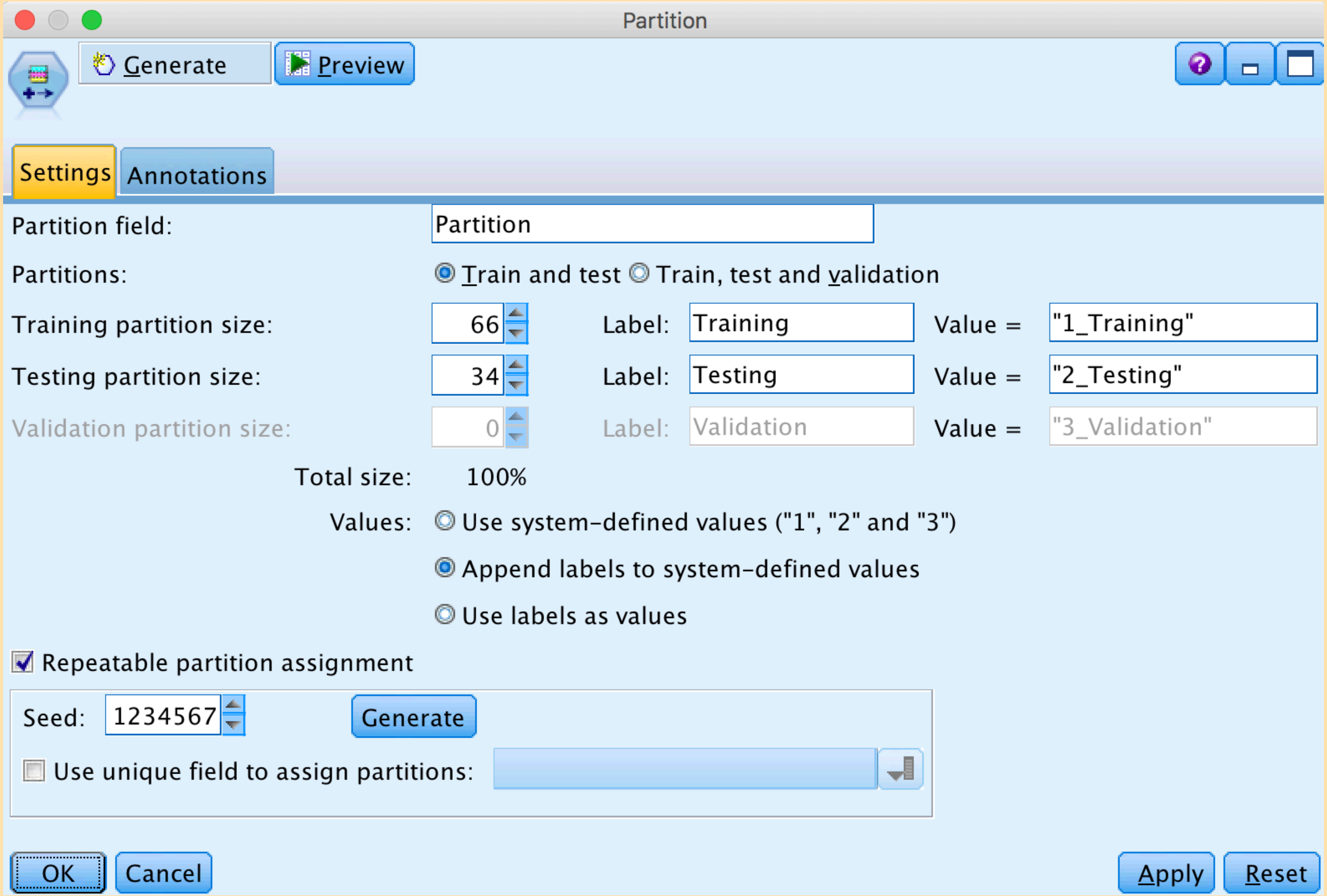
Field	Measurement	Values	Missing	Check	Role
field1	Continuous	[1,3095]		None	None
PIDN	Continuous	[65,19260]		None	Record ID
DCDate	Continuous	[2007-11-16,2015...		None	None
SourceID	Typeless			None	None
ScanNo	Ordinal	1		None	None
vbmID	Continuous	[5,4963]		None	None
gm	Continuous	[362.558,826.936]		None	Input
wm	Continuous	[354.125,738.869]		None	Input
csf	Continuous	[189.1,515.496]		None	Input
tiv	Continuous	[1007.5,1843.11]		None	Input
MagnetStrength	Nominal	3		None	None
ageAtScan	Continuous	[33,99]		None	Input
ScannerID	Flag	"NIC 3T MRI"/"NIC 3...		None	None
ScanType	Nominal	MP-LAS-ABN,T1-AD...		None	None
ROWNUM	Continuous	[1,3095]		None	None
InstrID	Continuous	[126430,343119]		None	None
ScannerID_2	Flag	"NIC 3T MRI"/"NIC 3...		None	None
MacCohort_kr	Nominal	AD,CONTROL,PSP,bv...		None	None
PIBPos	Nominal	"0","1",NA		None	None
Sex	Flag	male/female		None	Input
Educ	Continuous	[4,32]		None	Input
Race	Nominal	"Asian Indian","Black...		None	Input
DateofDeath	Continuous	[2010-06-11,2015...		None	None
autopsy.	Nominal	"","AUTOPSY PERFO...		None	None
PIBDone.	Flag	Yes/No		None	None
X3rd_Ventricle	Continuous	[58.75655516,184....		None	Input
X4th_Ventricle	Continuous	[91.34785011,289....		None	Input
Right_Accumbens_Area	Continuous	[31.5960242,147.7...		None	Input
Left_Accumbens_Area	Continuous	[21.71148099,161....		None	Input
Right_Amygdala	Continuous	[84.12848975,366....		None	Input
Left_Amygdala	Continuous	[68.73411328,383....		None	Input
Brain_Stem	Continuous	[506.5979247,1693...		None	Input
Right_Caudate	Continuous	[126.6476948,1154...		None	Input
Left_Caudate	Continuous	[117.6457182,1368...		None	Input
Right_Cerebellum_Ext...	Continuous	[5573.557537,1788...		None	Input
Left_Cerebellum_Exter...	Continuous	[4978.182021,1657...		None	Input
Right_Cerebellum_Whi...	Continuous	[2135.978191,6286...		None	Input
Left_Cerebellum_Whit...	Continuous	[2369.712793,6217...		None	Input
Right_Cerebral_White_...	Continuous	[41860.09373,6672...		None	Input
Left_Cerebral_White_M...	Continuous	[41414.18309,6593...		None	Input
Cerebrospinal_Fluid	Continuous	[77.18923729,252....		None	Input
Right_Hippocampus	Continuous	[372.8774375,1204...		None	Input
Left_Hippocampus	Continuous	[418.7640745,1047...		None	Input
Right_Inf_Lat_Vent	Continuous	[8.895621569,67.5...		None	Input
Left_Inf_Lat_Vent	Continuous	[5.335835814,24.3...		None	Input
Right_Lateral_Ventricle	Continuous	[310.2823222,1008...		None	Input
Left_Lateral_Ventricle	Continuous	[313.5167702,1294...		None	Input
Right_Pallidum	Continuous	[14.9530859,131.0...		None	Input
Left_Pallidum	Continuous	[13.81248181,164....		None	Input
Right_Putamen	Continuous	[94.64779961,1490...		None	Input
Left_Putamen	Continuous	[115.9943711,1647...		None	Input
Right_Thalamus_Proper	Continuous	[401.5123147,1719...		None	Input
Left_Thalamus_Proper	Continuous	[456.6295583,1658...		None	Input
Right_Ventral_DC	Continuous	[230.2842237,461....		None	Input
Left_Ventral_DC	Continuous	[240.8136756,526....		None	Input
Right_vessel	Continuous	[0.111729383,0.77...		None	Input
Left_vessel	Continuous	[0.646253213,3.55...		None	Input
Optic_Chiasm	Continuous	[1.350985626,9.89...		None	Input
Cerebellar_Vermal_Lo...	Continuous	[697.7441838,1596...		None	Input
Cerebellar_Vermal_Lo...	Continuous	[265.644736,712.7...		None	Input
Cerebellar_Vermal_Lo...	Continuous	[254.1379361,975....		None	Input



Right_MPoG_postcentr...	Continuous	[67.67728131,202....		None	Input
Left_MPoG_postcentral...	Continuous	[74.01055394,262....		None	Input
Right_MPrG_precentra...	Continuous	[310.6314755,777....		None	Input
Left_MPrG_precentral_...	Continuous	[251.8652834,780....		None	Input
Right_MSFG_superior_...	Continuous	[375.897864,2118....		None	Input
Left_MSFG_superior_fr...	Continuous	[346.6053448,1911....		None	Input
Right_MTG_middle_te...	Continuous	[547.5187223,4615....		None	Input
Left_MTG_middle_tem...	Continuous	[949.3164435,4346....		None	Input
Right_OCP_occipital_p...	Continuous	[184.6869897,719....		None	Input
Left_OCP_occipital_pole	Continuous	[182.2891504,835....		None	Input
Right_OFuG_occipital_...	Continuous	[421.209736,1131....		None	Input
Left_OFuG_occipital_fu...	Continuous	[331.3296358,1220....		None	Input
Right_OpIFG_opercula...	Continuous	[184.5597767,983....		None	Input
Left_OpIFG_opercular_...	Continuous	[170.6109803,927....		None	Input
Right_OrIFG_orbital_p...	Continuous	[72.63498808,412....		None	Input
Left_OrIFG_orbital_par...	Continuous	[43.62708166,338....		None	Input
Right_PCgG_posterior...	Continuous	[544.2258615,1247....		None	Input
Left_PCgG_posterior_c...	Continuous	[581.4855684,1368....		None	Input
Right_PCu_precuneus	Continuous	[1531.975308,3645....		None	Input
Left_PCu_precuneus	Continuous	[1316.489579,3537....		None	Input
Right_PHG_parahippo...	Continuous	[337.6894213,694....		None	Input
Left_PHG_parahippoca...	Continuous	[371.2698957,914....		None	Input
Right_PIns_posterior_i...	Continuous	[152.5767199,924....		None	Input
Left_PIns_posterior_in...	Continuous	[215.9756448,814....		None	Input
Right_PO_parietal_ope...	Continuous	[162.4111996,804....		None	Input
Left_PO_parietal_oper...	Continuous	[280.152219,855.5...		None	Input
Right_PoG_postcentral...	Continuous	[733.9963874,2558....		None	Input
Left_PoG_postcentral_...	Continuous	[927.4357129,2888....		None	Input
Right_POrG_posterior_...	Continuous	[180.6825849,742....		None	Input
Left_POrG_posterior_o...	Continuous	[82.04551118,782....		None	Input
Right_PP_planum_pola...	Continuous	[104.90551,489.50...		None	Input
Left_PP_planum_polare	Continuous	[145.6896541,555....		None	Input
Right_PrG_precentral_...	Continuous	[881.9382684,3119....		None	Input
Left_PrG_precentral_g...	Continuous	[1013.286891,3145....		None	Input
Right_PT_planum_tem...	Continuous	[105.5180343,622....		None	Input
Left_PT_planum_temp...	Continuous	[181.9479349,700....		None	Input
Right_SCA_subcallosal...	Continuous	[64.32918989,322....		None	Input
Left_SCA_subcallosal_...	Continuous	[65.64066,358.841...		None	Input
Right_SFG_superior_fr...	Continuous	[1022.292171,3654....		None	Input
Left_SFG_superior_fro...	Continuous	[813.8198456,3774....		None	Input
Right_SMC_suppleme...	Continuous	[353.486501,1601....		None	Input
Left_SMC_supplement...	Continuous	[302.2308103,1612....		None	Input
Right_SMG_supramarg...	Continuous	[565.5655709,2528....		None	Input
Left_SMG_supramargi...	Continuous	[854.6515809,2499....		None	Input
Right_SOG_superior_o...	Continuous	[228.0805283,1093....		None	Input
Left_SOG_superior_oc...	Continuous	[170.8020032,883....		None	Input
Right_SPL_superior_pa...	Continuous	[934.2002907,3381....		None	Input
Left_SPL_superior_pari...	Continuous	[765.3556043,3185....		None	Input
Right_STG_superior_te...	Continuous	[314.631001,1953....		None	Input
Left_STG_superior_te...	Continuous	[542.5562204,1828....		None	Input
Right_TMP_temporal_...	Continuous	[163.1710183,2649....		None	Input
Left_TMP_temporal_pole	Continuous	[65.94066814,2461....		None	Input
Right_TrIFG_triangular...	Continuous	[179.0097257,872....		None	Input
Left_TrIFG_triangular_...	Continuous	[97.04780146,813....		None	Input
Right_TTG_transverse...	Continuous	[74.64149524,525....		None	Input
Left_TTG_transverse_t...	Continuous	[161.4371758,533....		None	Input
drop	Flag	FALSE/FALSE		None	None
Dementia	Flag	TRUE/FALSE		None	Target

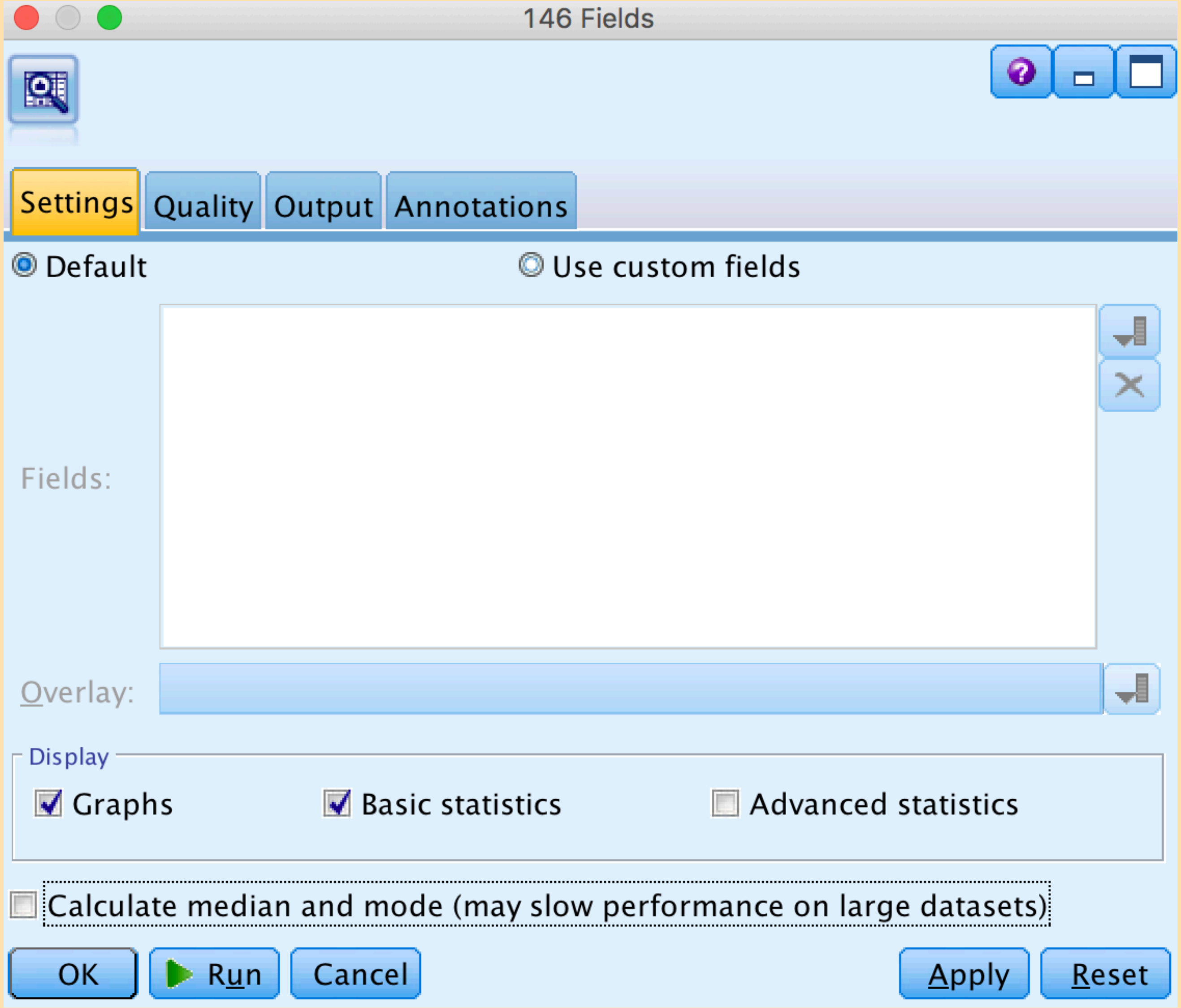
Partition node





Audit node


147 Fields


146 Fields
Settings
Quality
Output
Annotations
Default
Use custom fields
Fields:
Overlay:
Display
Graphs
Basic statistics
Advanced statistics
Calculate median and mode (may slow performance on large datasets)
OK
Run
Cancel
Apply
Reset

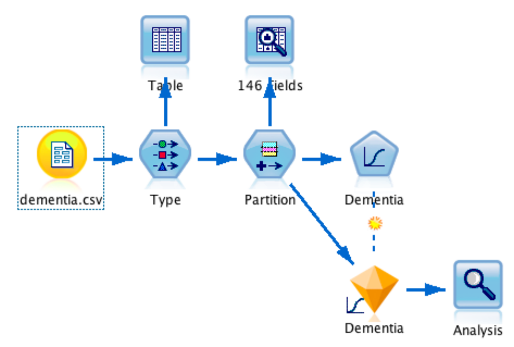


Table
146 Fields
dementia.csv
Type
Partition
Dementia
Dementia
Analysis

Audit node

Data Audit of [146 fields] #1

File Edit Generate

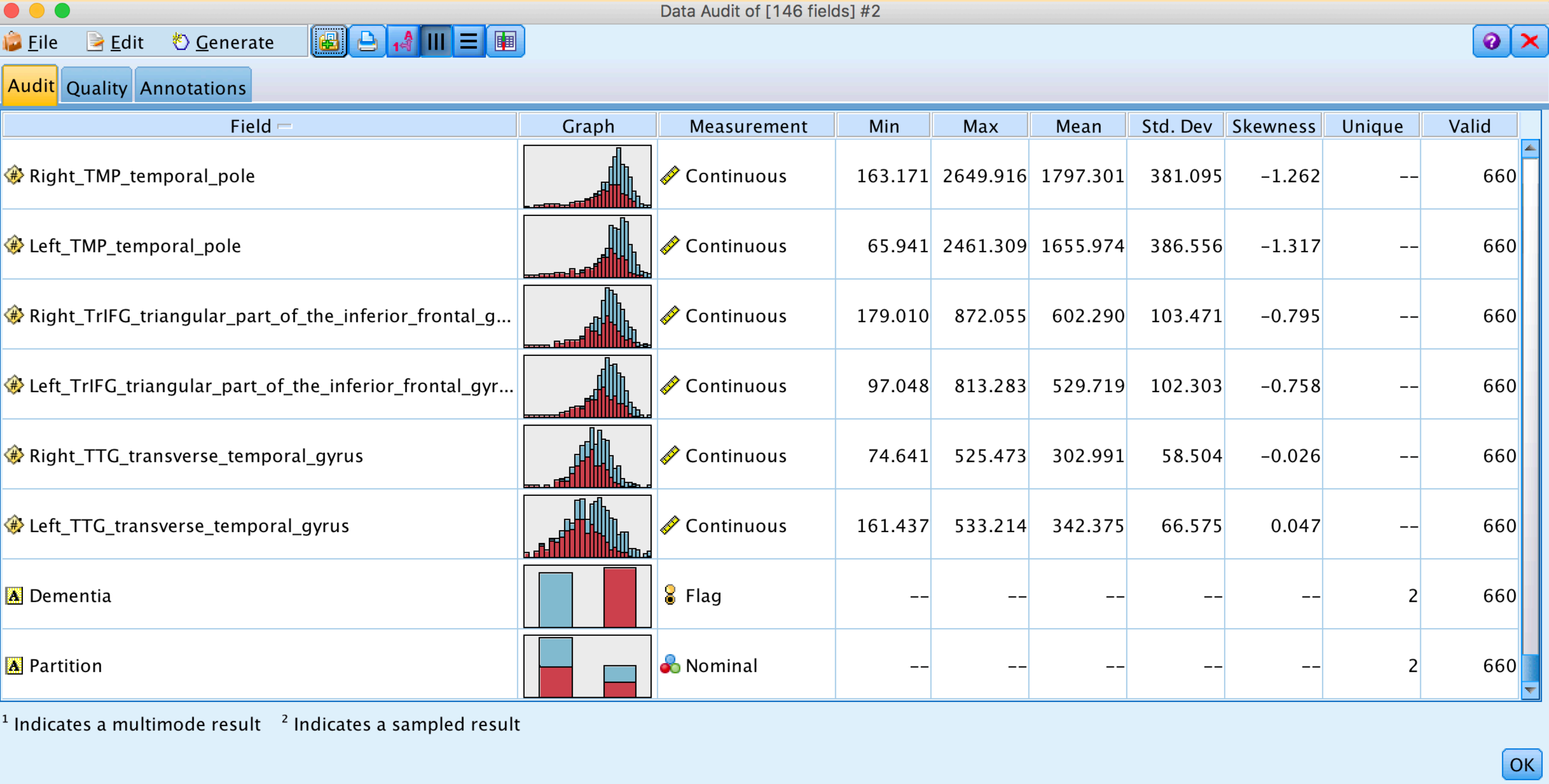
Audit Quality Annotations

Field	Graph	Measurement	Min	Max	Mean	Std. Dev	Skewness	Unique	Valid
PIDN		Continuous	65	19260	12372.7...	3514.176	-0.021	--	660
gm		Continuous	362.558	826.936	568.708	65.967	0.031	--	660
wm		Continuous	354.125	738.869	492.911	61.072	0.478	--	660
csf		Continuous	189.100	515.496	300.274	55.397	0.496	--	660
tiv		Continuous	1007.500	1843.110	1361.893	132.946	0.437	--	660
ageAtScan		Continuous	33	99	66.698	8.945	0.082	--	660
Sex		Flag	--	--	--	--	--	2	660
Educ		Continuous	4	32	16.693	3.023	0.108	--	639
X3rd_Ventricle		Continuous	58.757	184.767	119.459	20.945	0.051	--	660
X4th_Ventricle		Continuous	91.348	289.200	163.998	22.729	0.554	--	660

[1] Indicates a multimode result [2] Indicates a sampled result

OK

Audit node



Field	Graph	Measurement	Min	Max	Mean	Std. Dev	Skewness	Unique	Valid
Right_TMP_temporal_pole		Continuous	163.171	2649.916	1797.301	381.095	-1.262	--	660
Left_TMP_temporal_pole		Continuous	65.941	2461.309	1655.974	386.556	-1.317	--	660
Right_TrIFG_triangular_part_of_the_inferior_frontal_g...		Continuous	179.010	872.055	602.290	103.471	-0.795	--	660
Left_TrIFG_triangular_part_of_the_inferior_frontal_gyr...		Continuous	97.048	813.283	529.719	102.303	-0.758	--	660
Right_TTG_transverse_temporal_gyrus		Continuous	74.641	525.473	302.991	58.504	-0.026	--	660
Left_TTG_transverse_temporal_gyrus		Continuous	161.437	533.214	342.375	66.575	0.047	--	660
Dementia		Flag	--	--	--	--	--	2	660
Partition		Nominal	--	--	--	--	--	2	660

[1] Indicates a multimode result [2] Indicates a sampled result

Audit node

Data Audit of [146 fields] #1

File Edit Generate

Audit Quality Annotations

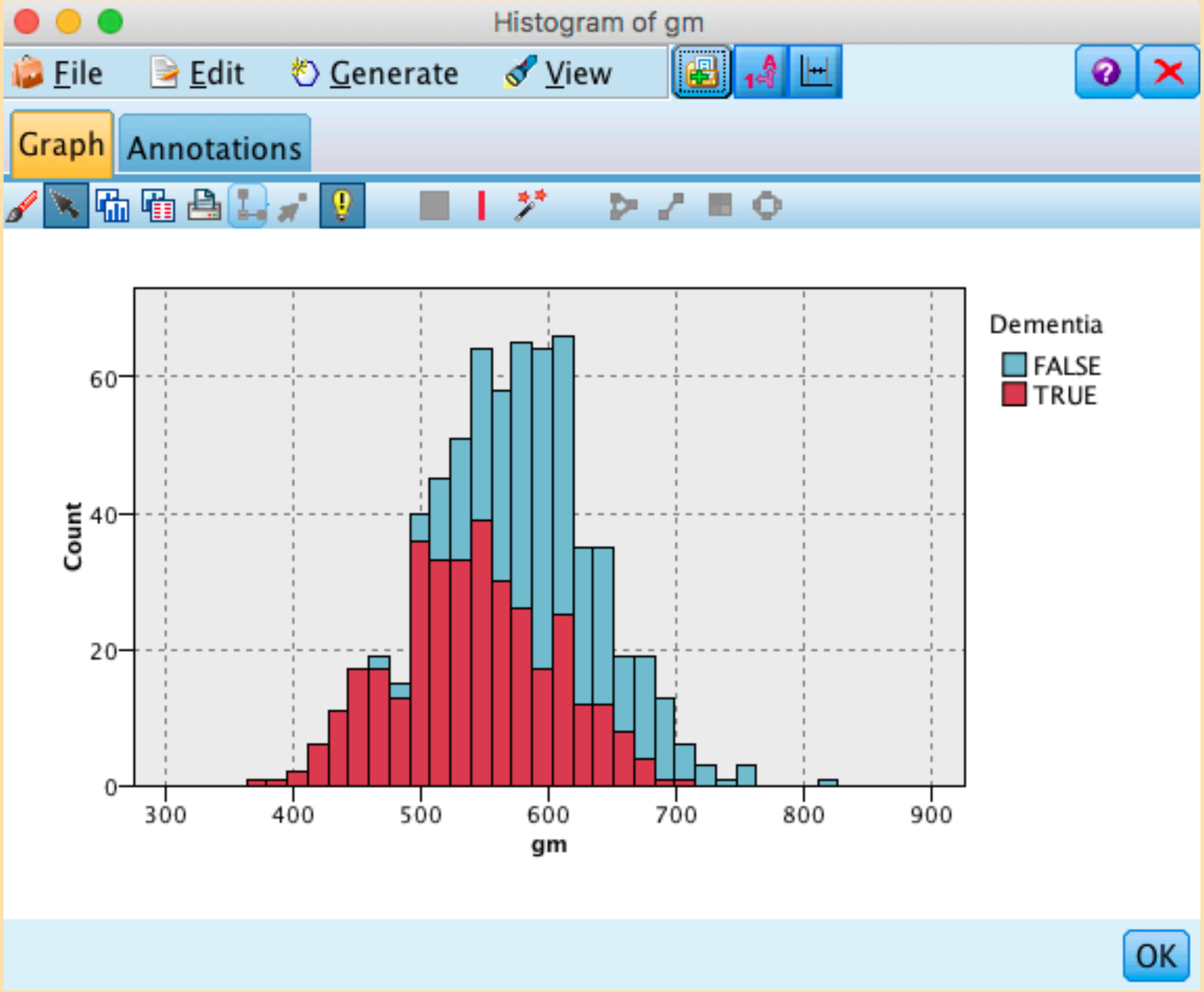
Field	Graph	Measurement	Min	Max	Mean	Std. Dev	Skewness	Unique	Valid
PIDN		Continuous	65	19260	12372.7...	3514.176	-0.021	--	660
gm		Continuous	362.558	826.936	568.708	65.967	0.031	--	660
wm		Continuous	354.125	738.869	492.911	61.072	0.478	--	660
csf		Continuous	189.100	515.496	300.274	55.397	0.496	--	660
tiv		Continuous	1007.500	1843.110	1361.893	132.946	0.437	--	660
ageAtScan		Continuous	33	99	66.698	8.945	0.082	--	660
Sex		Flag	--	--	--	--	--	2	660
Educ		Continuous	4	32	16.693	3.023	0.108	--	639
X3rd_Ventricle		Continuous	58.757	184.767	119.459	20.945	0.051	--	660
X4th_Ventricle		Continuous	91.348	289.200	163.998	22.729	0.554	--	660

[1] Indicates a multimode result [2] Indicates a sampled result

OK

Audit node


147 fields


Histogram of gm
File
Edit
Generate
View
Graph
Annotations
Dementia
FALSE
TRUE
Count
0
20
40
60
300
400
500
600
700
800
900
gm
OK

Audit node

Data Audit of [146 fields] #2

File Edit Generate

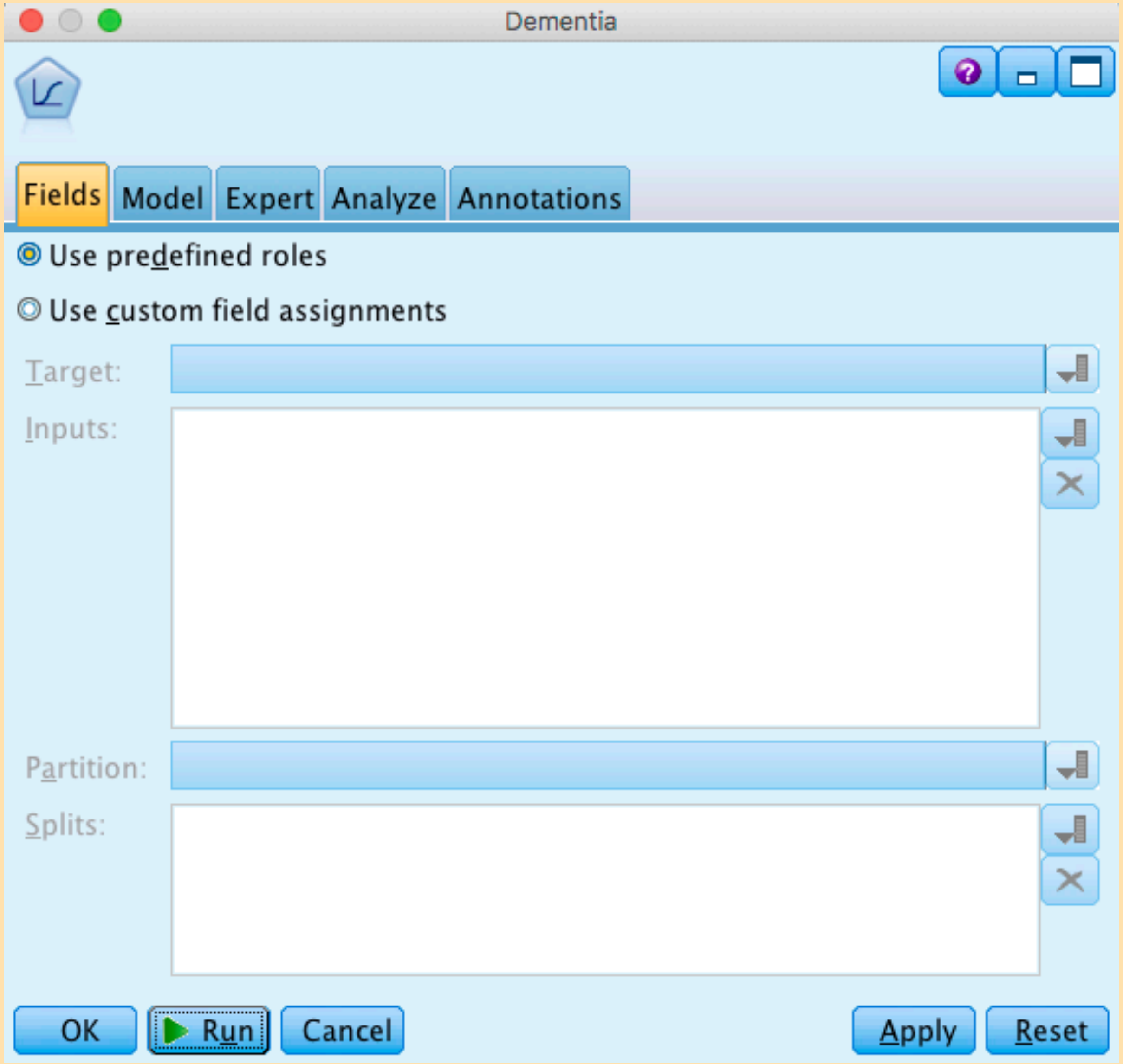
Audit | Quality | Annotations

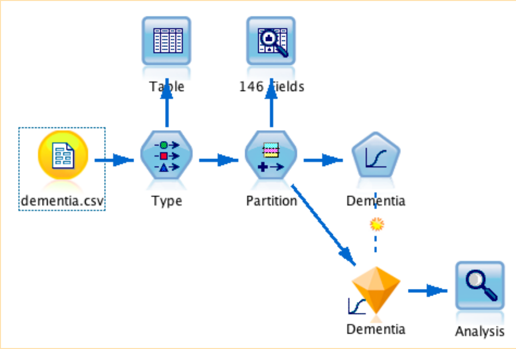
Complete fields (%): 99.32% Complete records (%): 96.82%

Field	Measurement	Outliers	Extremes	Action	Impute Missi...	Method	% Complete	Valid Records	Null Value	Emp
PIDN	Continuous	1	0	None	Never	Fixed	100	660	0	
gm	Continuous	2	0	None	Never	Fixed	100	660	0	
wm	Continuous	5	0	None	Never	Fixed	100	660	0	
csf	Continuous	4	0	None	Never	Fixed	100	660	0	
tiv	Continuous	4	0	None	Never	Fixed	100	660	0	
ageAtScan	Continuous	2	0	None	Never	Fixed	100	660	0	
Sex	Flag	--	--	--	Never	Fixed	100	660	0	
Educ	Continuous	10	1	None	Never	Fixed	96.818	639	21	
X3rd_Ven...	Continuous	1	0	None	Never	Fixed	100	660	0	
X4th_Vent...	Continuous	5	1	None	Never	Fixed	100	660	0	
Right_Acc...	Continuous	7	0	None	Never	Fixed	100	660	0	
Left_Accu...	Continuous	1	0	None	Never	Fixed	100	660	0	
Right_Am...	Continuous	10	0	None	Never	Fixed	100	660	0	
Left_Amy...	Continuous	4	0	None	Never	Fixed	100	660	0	
Brain_Stem	Continuous	7	4	None	Never	Fixed	100	660	0	
Right_Cau...	Continuous	5	0	None	Never	Fixed	100	660	0	
Left_Caud...	Continuous	5	0	None	Never	Fixed	100	660	0	
Right_Cer...	Continuous	7	1	None	Never	Fixed	100	660	0	
Left_Cere...	Continuous	6	1	None	Never	Fixed	100	660	0	
Right_Cer...	Continuous	4	1	None	Never	Fixed	100	660	0	
Left_Cere...	Continuous	6	1	None	Never	Fixed	100	660	0	
Right_Cer...	Continuous	1	0	None	Never	Fixed	100	660	0	
Right_SO...	Continuous	10	0	None	Never	Fixed	100	660	0	
Left_SOG ...	Continuous	8	0	None	Never	Fixed	100	660	0	
Right_SPL...	Continuous	11	0	None	Never	Fixed	100	660	0	
Left_SPL_s...	Continuous	9	0	None	Never	Fixed	100	660	0	
Right_STG...	Continuous	6	0	None	Never	Fixed	100	660	0	
Left_STG ...	Continuous	5	0	None	Never	Fixed	100	660	0	
Right_TM...	Continuous	14	0	None	Never	Fixed	100	660	0	
Left_TMP ...	Continuous	15	0	None	Never	Fixed	100	660	0	
Right_TrIF...	Continuous	7	0	None	Never	Fixed	100	660	0	
Left_TrIFG...	Continuous	8	0	None	Never	Fixed	100	660	0	
Right_TT...	Continuous	4	0	None	Never	Fixed	100	660	0	
Left_TTG ...	Continuous	0	0	None	Never	Fixed	100	660	0	
Dementia	Flag	--	--	--	Never	Fixed	100	660	0	
Partition	Nominal	--	--	--	Never	Fixed	100	660	0	

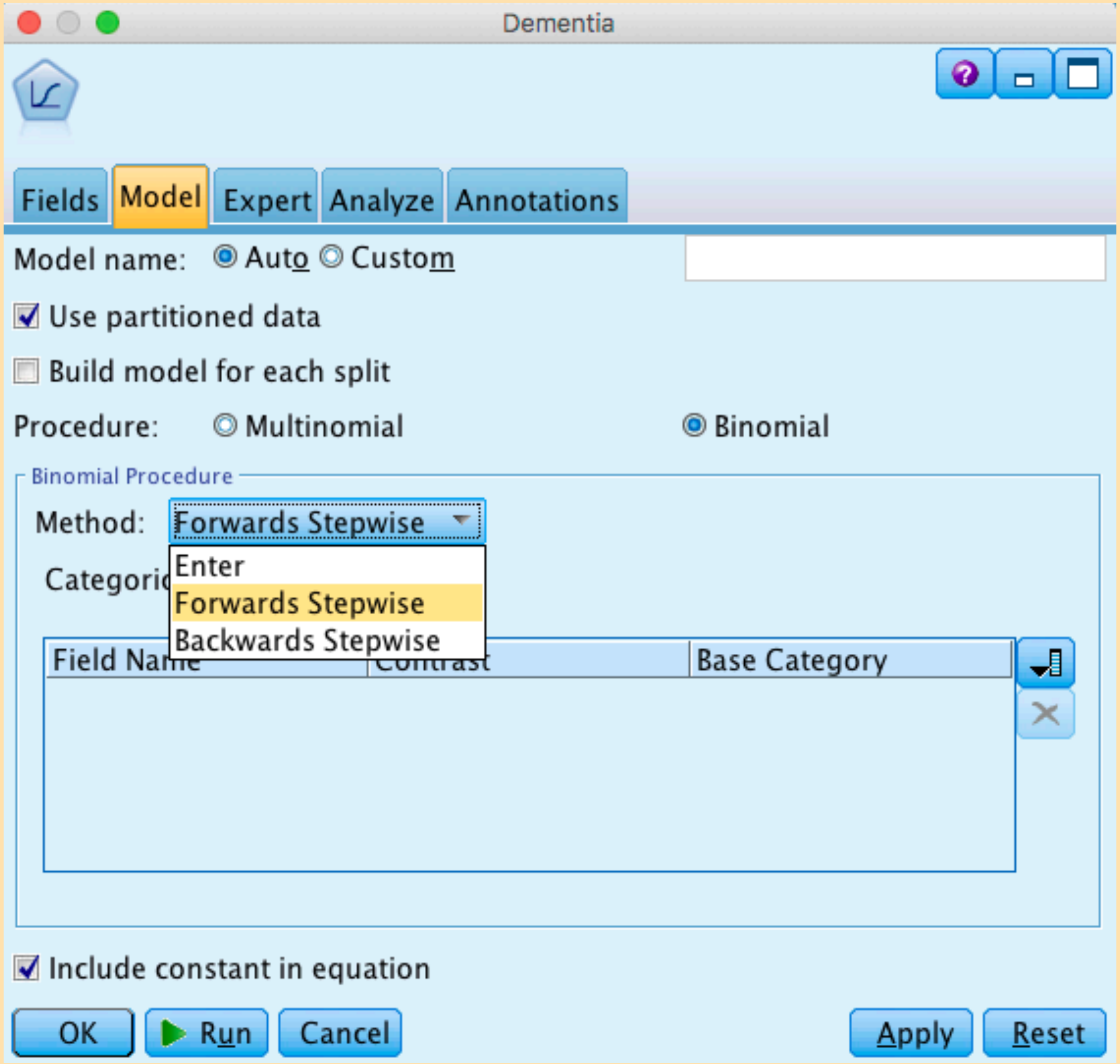
OK

Logistic node

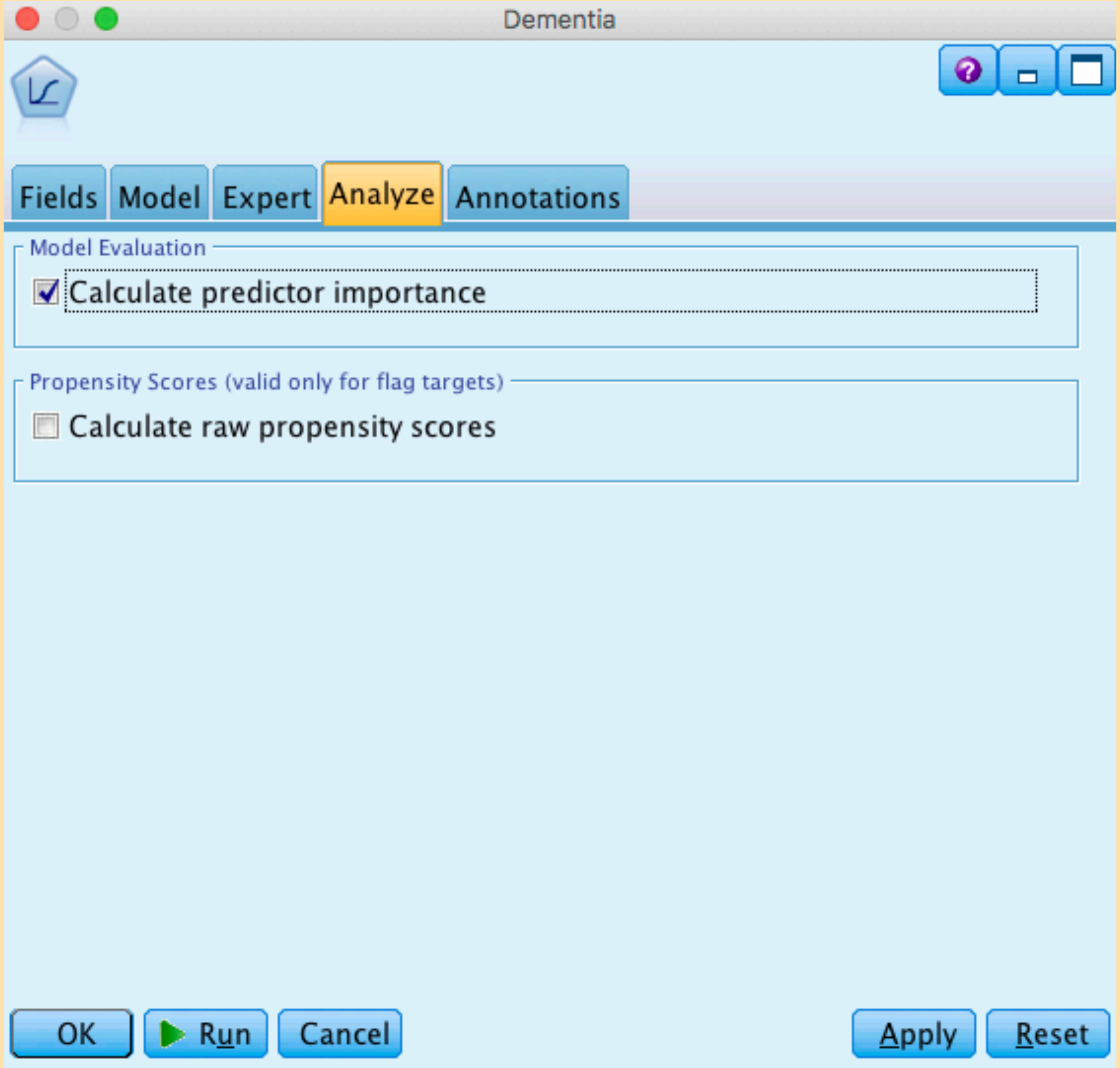




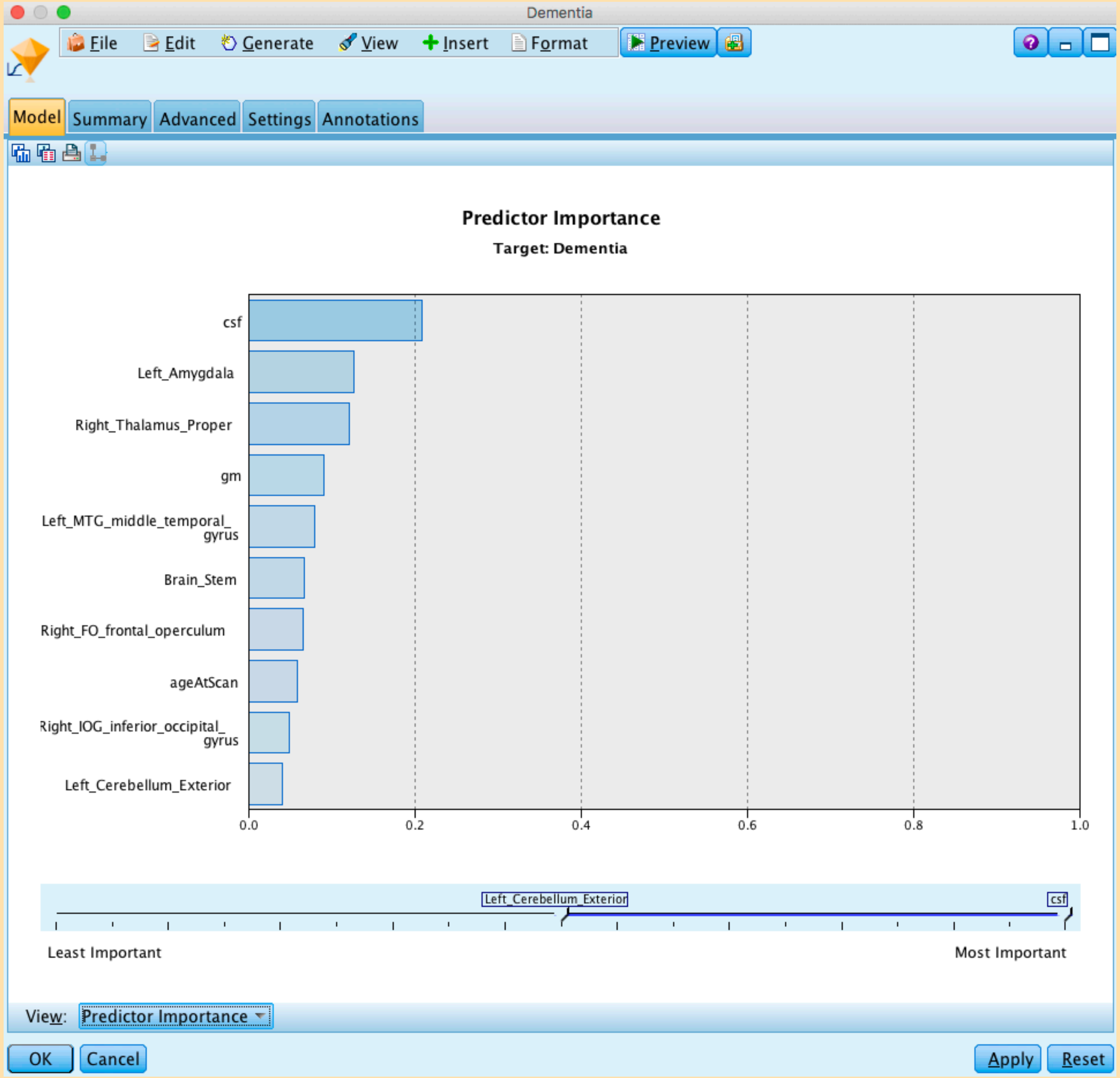
Logistic node

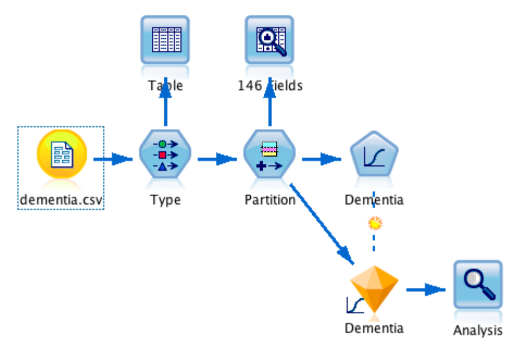


Logistic node

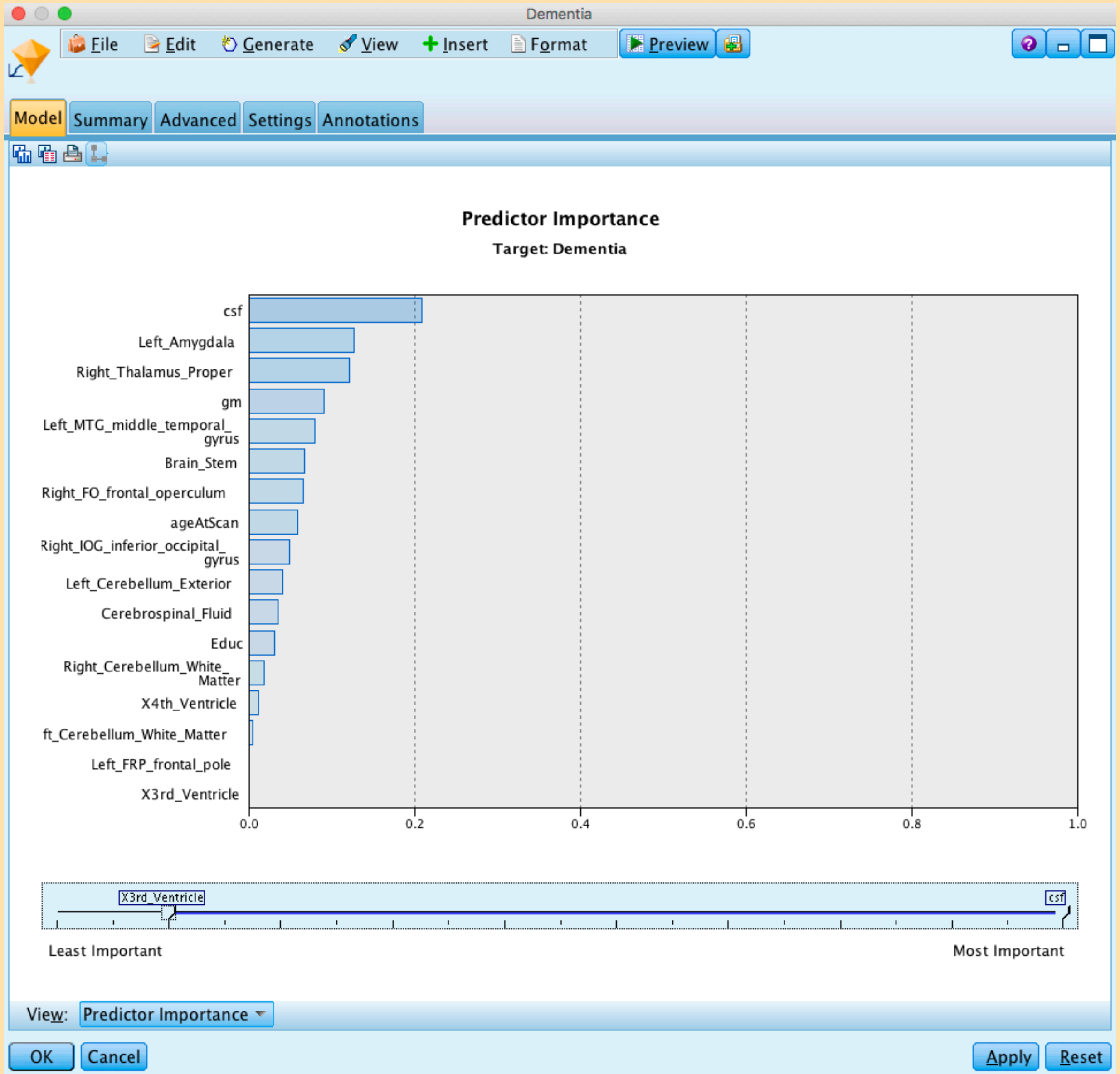


Golden Nugget



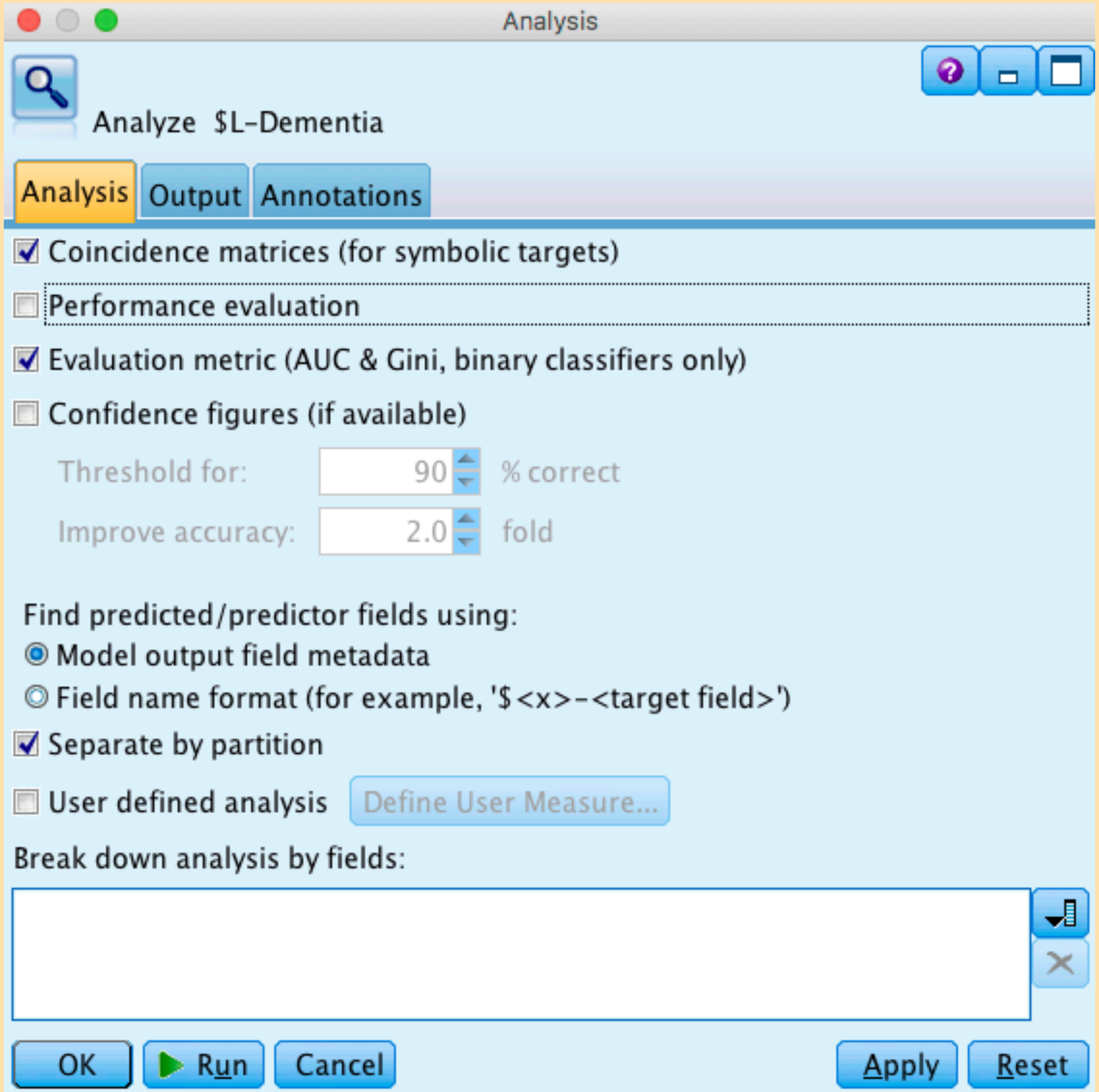


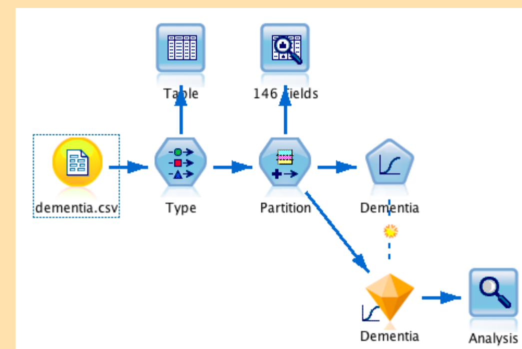
Golden Nugget



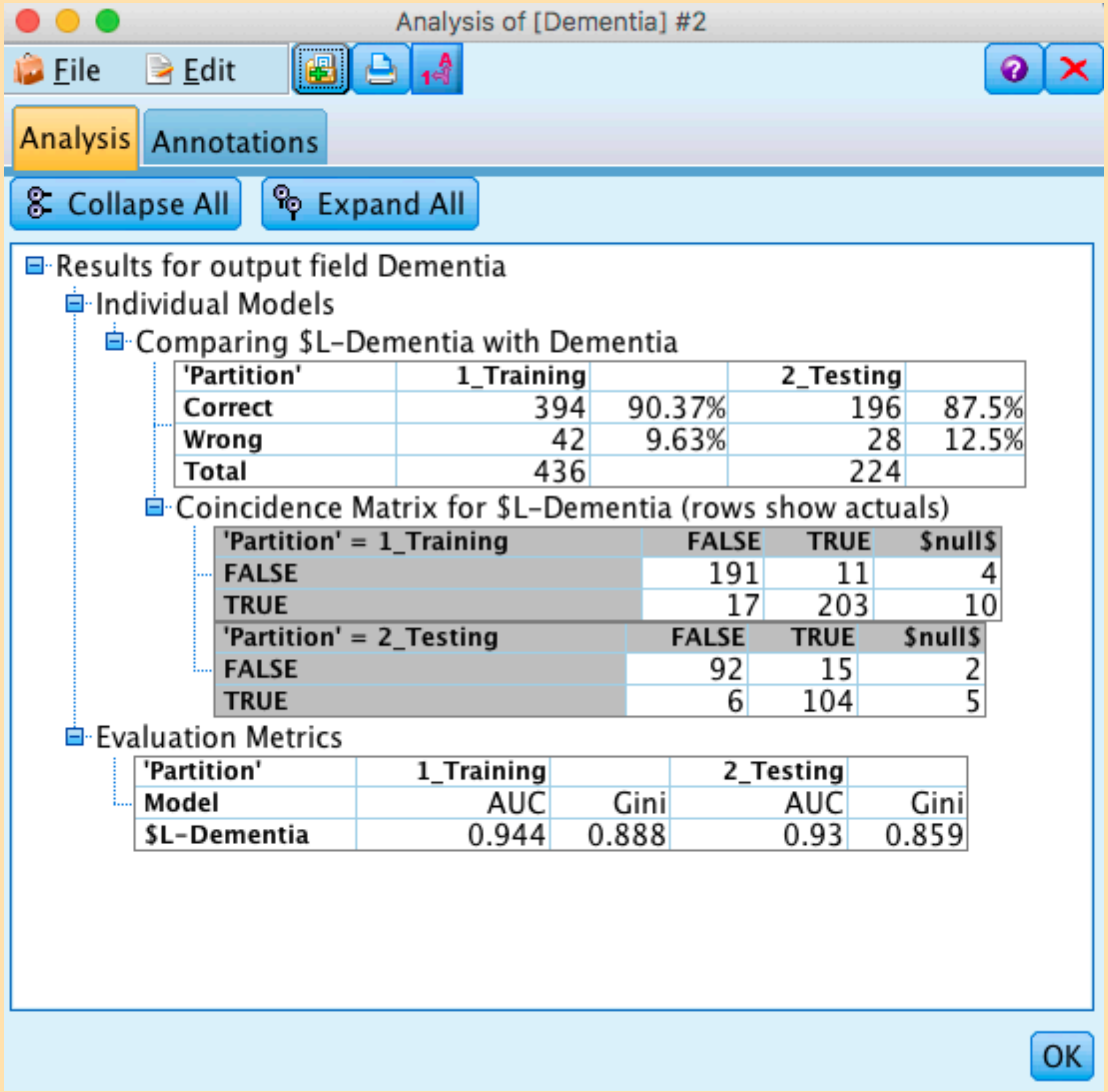
Also, the Advanced tab has a lot of information about what happened in the forward step-wise procedure, which predictors were retained etc.

Analysis node





Analysis node



Monday's lecture – Machine Learning 3

- K-nearest neighbors
- Cross-validation
- Classification Trees
- Neural Networks (Neural Nets)

Note: HW #4 is on CLE
– due 11.55pm on Thursday August 24