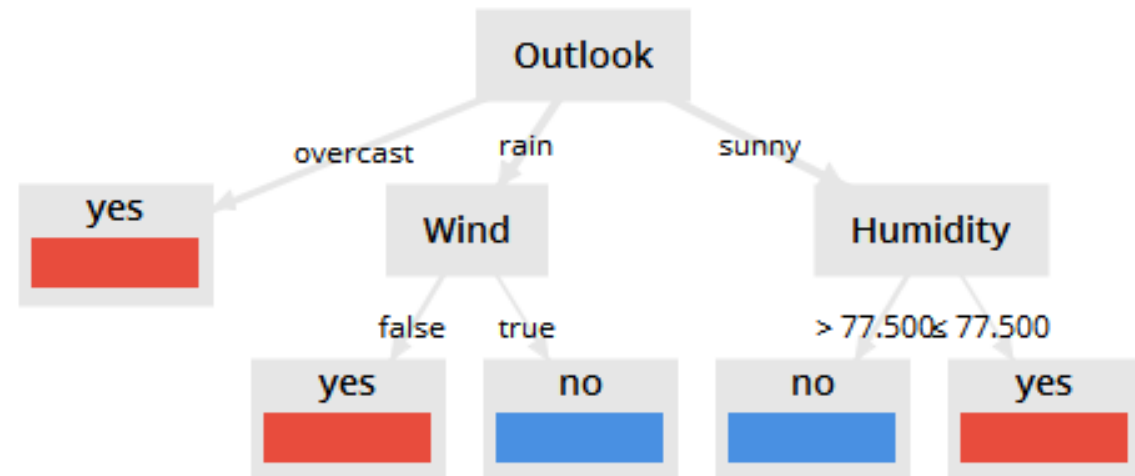


# USING DECISION TREES TO ANALYZE ONLINE LEARNING DATA

(updated)

# PLAYING GOLF TODAY?

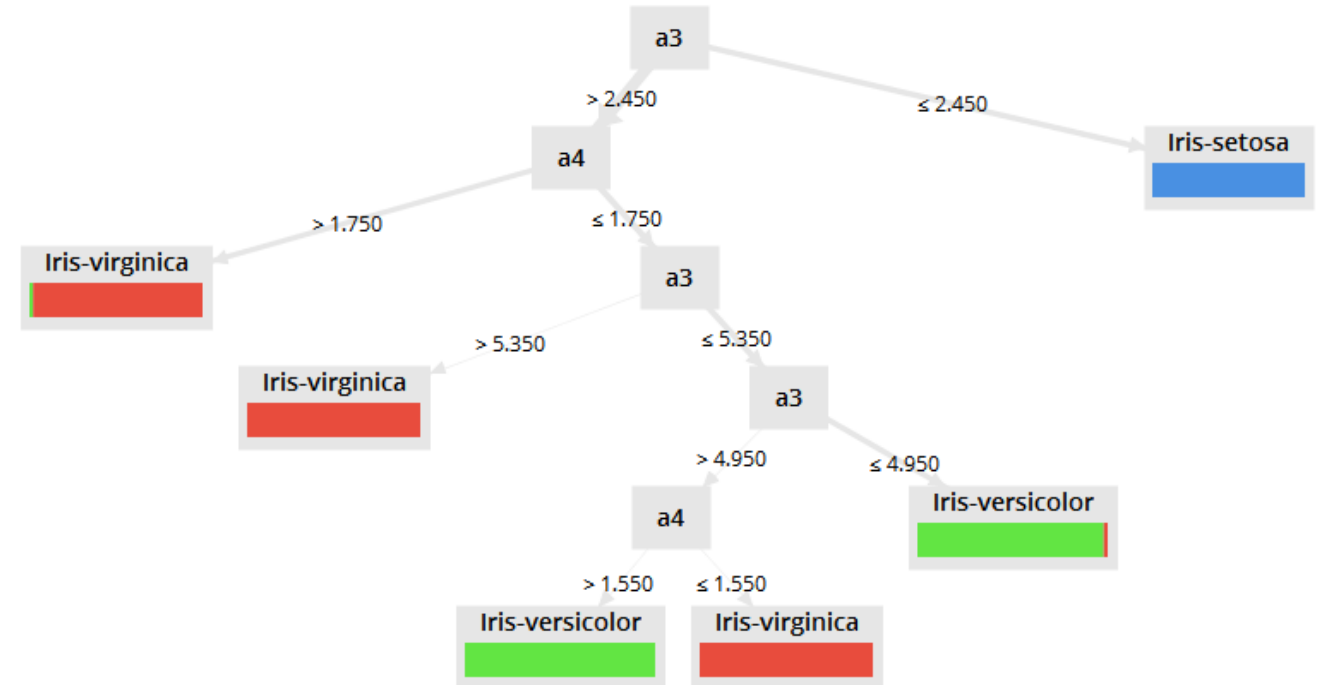
(from sample faux data in RapidMiner Studio)



relative humidity as a % with  
100% as fully saturated

# IRIS DATASET

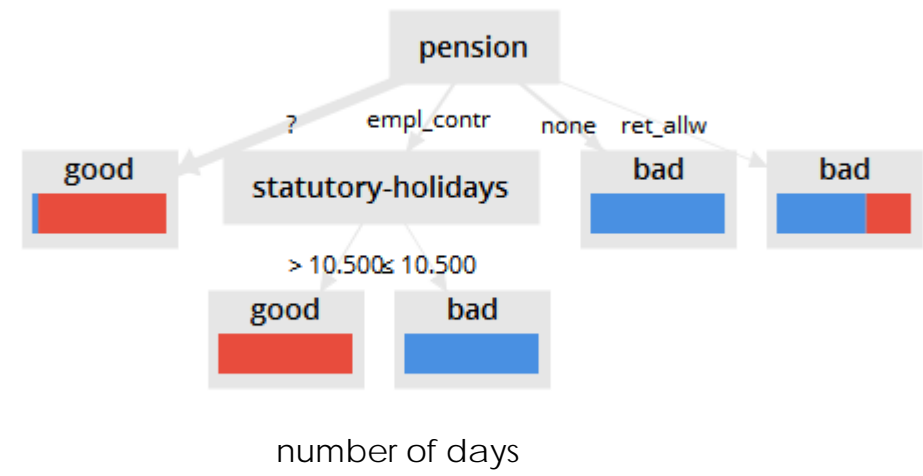
(from sample open-source data in RapidMiner Studio)



length and width of petals in centimeters

# GOOD AND BAD WORKPLACE PRACTICES IN "LABOR-NEGOTIATIONS" DATASET

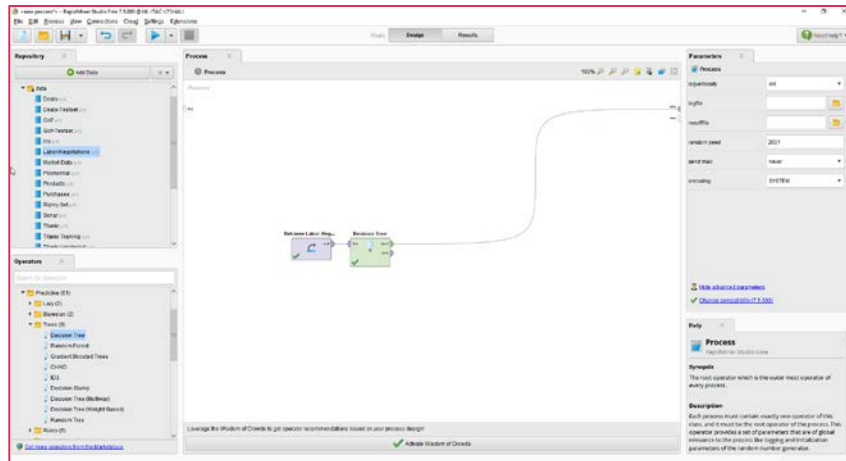
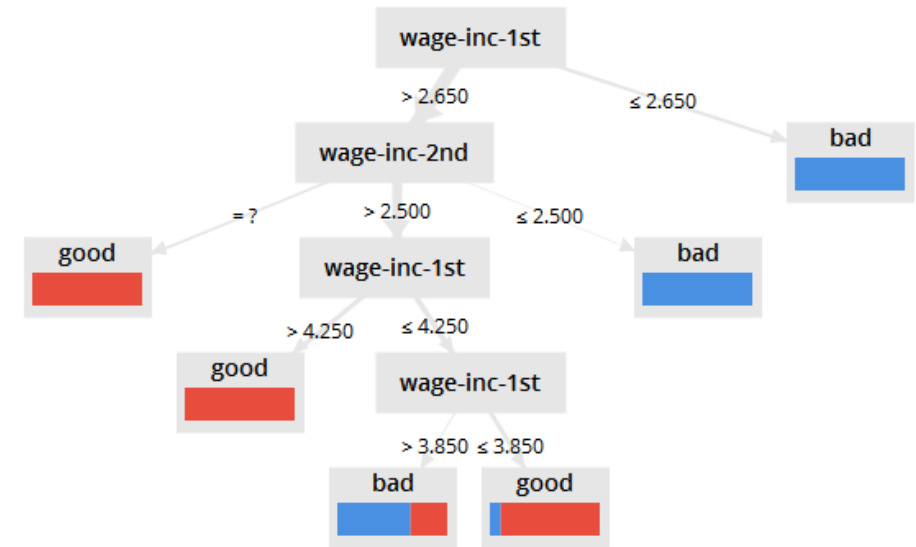
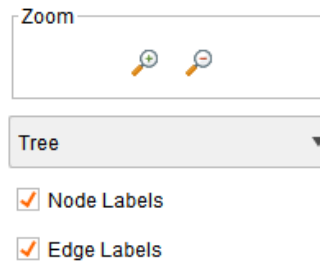
(from sample data in RapidMiner Studio)



Row No.	class	duration	wage-inc-1st	wage-inc-2nd	wage-inc-3rd	col-adj	working-hou...	pension	standby-pay	shift-differe...	education-eL...	statutory-ho...	vacation
1	good	1	5	?	?	?	40	?	?	2	?	11	average
2	good	2	4.500	5.800	?	?	35	ret_allw	?	?	yes	11	below-avera
3	good	?	?	?	?	?	38	empl_contr	?	5	?	11	generous
4	good	3	3.700	4	5	tc	?	?	?	?	yes	?	?
5	good	3	4.500	4.500	5	?	40	?	?	?	?	12	average
6	good	2	2	2.500	?	?	35	?	?	6	yes	12	average
7	good	3	4	5	5	tc	?	empl_contr	?	?	?	12	generous
8	good	3	6.900	4.800	2.300	?	40	?	?	3	?	12	below-avera
9	good	2	3	7	?	?	38	?	12	25	yes	11	below-avera
10	good	1	5.700	?	?	none	40	empl_contr	?	4	?	11	generous
11	good	3	3.500	4	4.600	none	36	?	?	3	?	13	generous
12	good	2	6.400	6.400	?	?	38	?	?	4	?	15	?
13	bad	2	3.500	4	?	none	40	?	?	2	no	10	below-avera
14	good	3	3.500	4	5.100	tcf	37	?	?	4	?	13	generous
15	good	1	3	?	?	none	36	?	?	10	no	11	generous
16	good	2	4.500	4	?	none	37	empl_contr	?	?	?	11	average
17	good	1	2.800	?	?	?	35	?	?	2	?	12	below-avera
18	bad	1	2.100	?	?	tc	40	ret_allw	2	3	no	9	below-avera
19	bad	1	2	?	?	none	38	none	?	?	yes	11	average
20	good	2	4	5	?	tcf	35	?	13	5	?	15	generous
21	good	2	4.300	4.400	?	?	38	?	?	4	?	12	generous
22	bad	2	2.900	3	?	?	40	none	?	?	?	11	below-avera
23	good	3	3.500	4	4.600	tcf	27	?	?	?	?	?	?
24	good	2	4.500	4	?	?	40	?	?	4	?	10	generous
25	good	1	6	?	?	?	38	?	8	3	?	9	generous
26	bad	3	2	2	2	none	40	none	?	?	?	10	below-avera
27	good	?	4.500	4.500	?	tcf	?	?	?	?	yes	10	below-avera

# GOOD AND BAD WAGE PRACTICES IN "LABOR-NEGOTIATIONS" DATASET

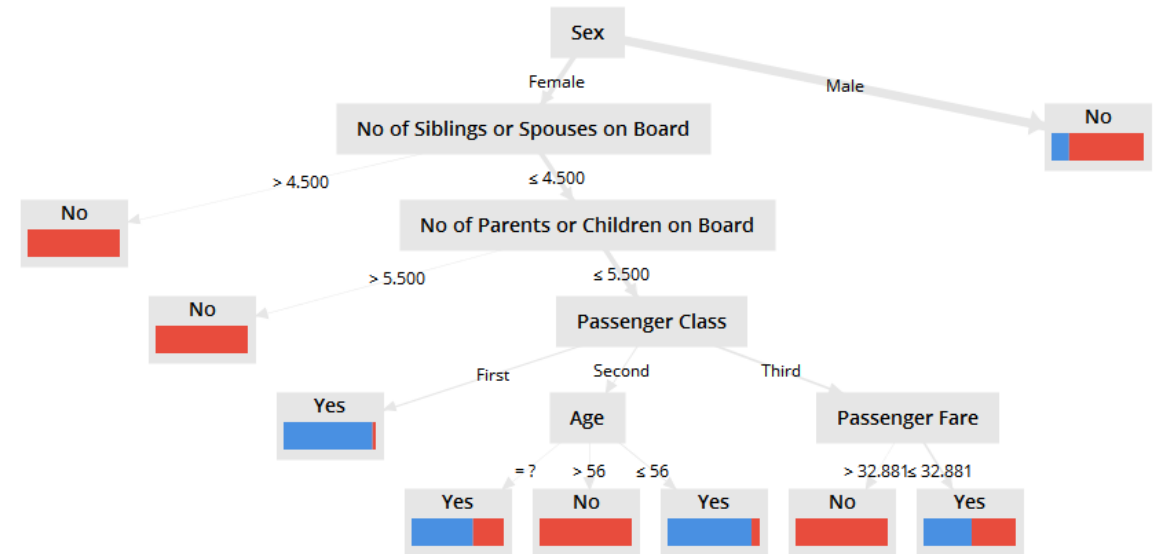
(from sample data in RapidMiner Studio)



# SURVIVORS OF THE TITANIC SINKING

(from sample open-source data in RapidMiner Studio)

A decision tree from a random forest process



**Parameters** ×

Random Forest

number of trees: 10

criterion: gain\_ratio

maximal depth: 20

apply pruning

confidence: 0.25

apply prepruning

minimal gain: 0.1

minimal leaf size: 2

minimal size for split: 4

number of prepruning altern...: 3

guess subset ratio

[Hide advanced parameters](#)

[Change compatibility \(7.5.000\)](#)

Row No.	Passenger ...	Name	Sex	Age	No of Siblings...	No of Parents...	Ticket Num...	Passenger F...	Cabin	Port of Emb...	Life Boat	Survived
1	First	Allen, Mrs. E.	Female	29	0	0	24160	211336	B5	Southampton	2	Yes
2	First	Alvarez, Mrs.	Male	8.917	1	0	112091	951600	C22-C26	Southampton	11	Yes
3	First	Alvarez, Mrs.	Female	2	1	2	112091	951600	C22-C26	Southampton	7	No
4	First	Alvarez, Mr. H.	Male	39	1	2	112091	951600	C22-C26	Southampton	7	No
5	First	Alvarez, Mrs.	Female	29	1	2	112091	951600	C22-C26	Southampton	7	No
6	First	Jackson, Mr.	Male	48	0	0	18652	26460	F10	Southampton	9	Yes
7	First	Jackson, Mr.	Female	63	1	0	13562	77898	D3	Southampton	10	Yes
8	First	Jackson, Mr.	Male	39	0	0	112091	0	A8	Southampton	7	No
9	First	Agoston, Mr.	Female	53	2	0	112091	0	C104	Southampton	0	Yes
10	First	Agoston, Mr.	Male	71	0	0	PC 17626	40884	?	Cherbourg	7	No
11	First	Asie, Col. J.	Male	47	1	0	PC 17717	327835	CE3-CE4	Cherbourg	7	No
12	First	Asie, Mrs. J.	Female	18	1	0	PC 17757	327835	CE3-CE4	Cherbourg	4	Yes
13	First	Asie, Mrs.	Female	24	0	0	PC 17477	58380	B05	Cherbourg	8	Yes
14	First	Barnes, Mrs.	Female	25	0	0	18077	76650	?	Southampton	5	Yes
15	First	Barnes, Mr.	Male	60	0	0	27542	20	A02	Southampton	6	Yes
16	First	Barnes, Mr.	Male	?	0	0	PC 17338	25825	?	Southampton	7	No
17	First	Baker, Mr. O.	Male	24	0	1	PC 17338	247521	B09-B10	Cherbourg	7	No
18	First	Baker, Mr. J.	Female	50	0	1	PC 17338	247521	B09-B10	Cherbourg	8	Yes
19	First	Baccari, Mrs.	Female	32	0	0	11013	76292	D15	Cherbourg	8	Yes
20	First	Baile, M. T.	Male	38	0	0	13099	76242	06	Cherbourg	4	No
21	First	Bechtholtz, Mr.	Male	37	1	1	11701	52554	D05	Southampton	5	Yes
22	First	Bond, Mrs.	Female	47	1	1	11701	52554	D08	Southampton	5	Yes
23	First	Bon, Mr. Karl	Male	25	0	0	11588	50	C148	Cherbourg	5	Yes
24	First	Bon, Mrs.	Female	42	0	0	PC 17757	327835	?	Cherbourg	4	Yes
25	First	Bon, Mrs. Ek.	Female	29	0	0	PC 17483	221779	C87	Southampton	8	Yes
26	First	Bonham, Mr.	Male	25	0	0	13805	20	?	Cherbourg	7	No
27	First	Bonham, Mr. G.	Male	20	1	0	13807	91079	B09	Cherbourg	7	Yes

# SURVIVORS OF THE TITANIC SINKING

(CONT.)

(from sample open-source data in RapidMiner Studio)

A "random forest" sequence of decision trees



# SURVIVORS OF THE TITANIC SINKING (CONT.)

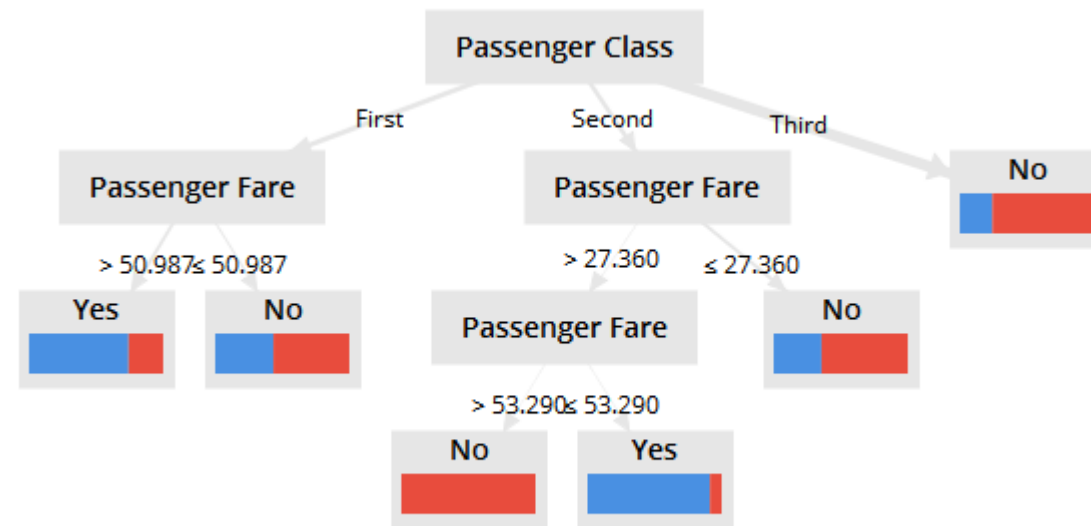
(from sample open-source data in RapidMiner Studio)

A "random forest" sequence without auto-pre-pruning

- More nuance

- More subdividing and branching

- Higher levels of accuracy





# SESSION DESCRIPTION

- In machine learning, decision trees enable researchers to identify possible indicators (variables) that are important in predicting classifications, and these offer a sequence of nuanced groupings. For example, are there “tells” which would suggest that a particular student will achieve a particular grade in a course? Are there indicators that would identify learners who would select a particular field of study vs. another?
- This session will introduce how decision trees are used to model data based on supervised machine learning (with labeled training set data) and how such models may be evaluated for accuracy with test data, with the open-source tool, RapidMiner Studio. Several related analytical data visualizations will be shared: 2D spatial maps, decision trees, and others. Attendees will also experience how 2x2 contingency tables work with Type 1 and Type 2 errors (and how the accuracy of the machine learning model may be assessed) to represent model accuracy, and the strengths and weaknesses of decision trees applied to some use cases from higher education. In this session, various examples of possible outcomes will be discussed and related pre-modeling theorizing (vs. post-hoc) about what may be seen in terms of particular variables. The basic data structure for running the decision tree algorithms will be described. If time allows, relevant parameters for a decision tree model will be discussed: criterion (gain\_ratio, information\_gain, gini\_index, and accuracy), minimal size for split, minimal leaf size, minimal gain, maximal depth (based on the need for human readability of decision trees), confidence, and pre-pruning (and the desired level).

# PRESENTATION ORDER

1. Decision trees, classification, and predictive analytics
2. Inducing decision trees from multivariate data
3. Data structures
4. Model accuracy and decision tree validation: Type 1 and type 2 errors
5. RapidMiner Studio and drawing decision trees; parameters; sequential walk-throughs
6. Educational data with defined outcomes and candidate co-variates
7. Summary

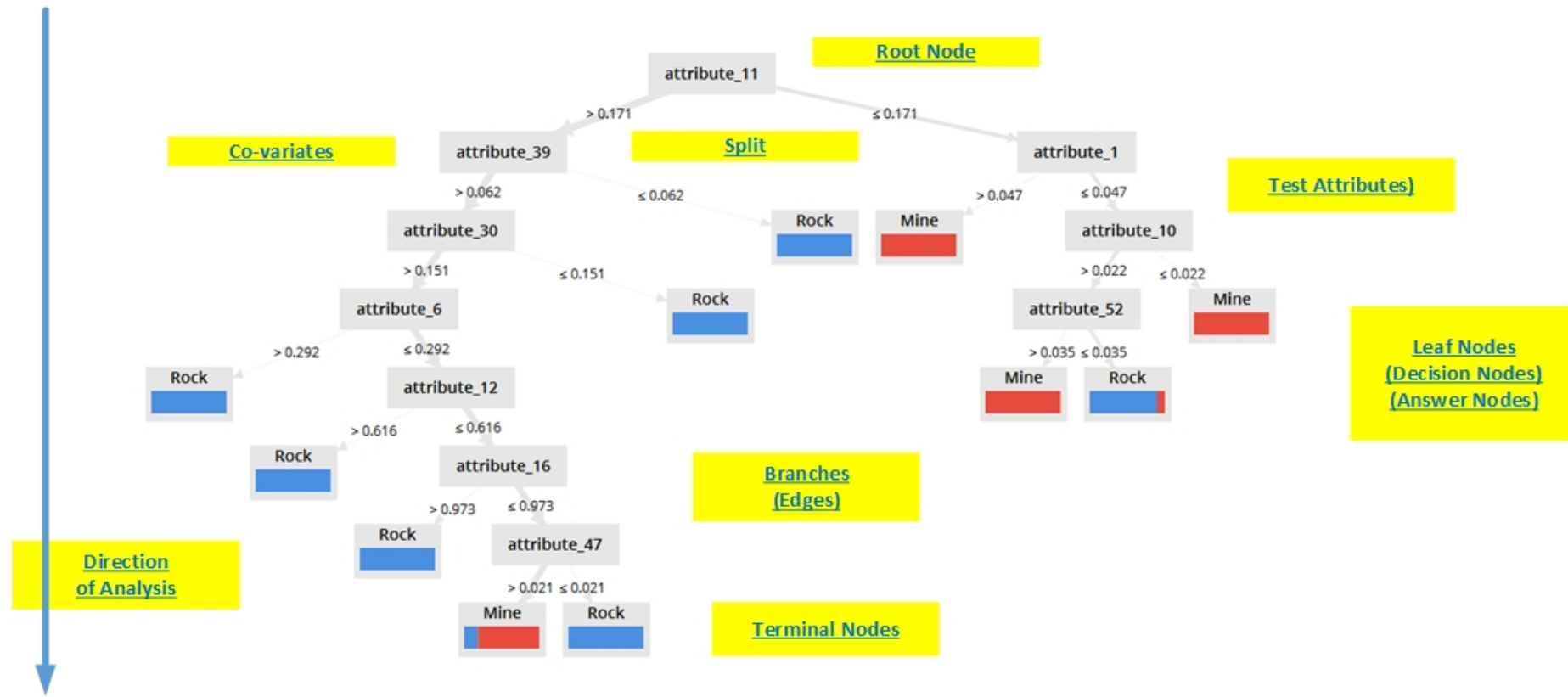
# 1. DECISION TREES, CLASSIFICATION, AND PREDICTIVE ANALYTICS



# DECISION TREES FOR MACHINE LEARNING

- Decision trees—in machine learning (formerly “data mining”)—can be applied for
  - (1) classification (labeled categories, optimally mutually exclusive classes, un-pre-labeled categories), and / or
  - (2) regression (continuous number outcomes, with potentially small numerical differences between categories).
- In our case, we are using decision trees to extract main factors which influence the ultimate classification of a population into respective classes.
- Decision trees may be “induced” from particular datasets. Rules are extracted based on the inherent structure in data, and they form the basis for classification models. (These do **not** require prior theorizing and framing in order to code for meaning.)
- These induced decision trees may be used to predict what a particular datapoint (or single record) or new set of (test) data may look like in terms of class, assuming the population is similar to the analyzed dataset used to “train” the particular decision tree (and the predictive power and accuracy of that decision tree).

# BASIC PARTS OF A DECISION TREE



Basic Parts of a Decision Tree (from a Sonar Sample Dataset from RapidMiner Studio)

# TOP-DOWN INDUCTION OF A DECISION TREE (TDIDT)

- The Top-Down Induction of a Decision Tree (TDIDT) refers to the drawing of a decision tree from the root node on down.
  - “Top-Down” refers to the identification of the best attribute / indicator first.
- A target attribute (classification) is indicated as select column data.
- The most informative input attribute(s) that indicate a particular class is the one that enables the first and top-most split based on the numerical measure that best identifies classification (along that dimension).
- The attributes are all part of a “feature vector” for the particular examples in the dataset.
  - They all describe facets of the individuals (animate) or objects / phenomena (inanimate) in the dataset.

# INTERPRETING A DECISION TREE: READING FROM ROOTS TO LEAVES

- Start with the root node of the tree at the top. This contains an attribute which is highly influential in the categorizing of the target population.
- In the next level of the tree, analyze the split (and the split attributes) to understand how the data may be categorized at the coarsest level with the most highly influential variable. Look at the leaf node labels. Read the edges to understand the test attribute values that determine the split.
- Work down the tree, through each successive splitting of the training set. (Data is processed iteratively and sequentially, so prior splits inform future ones.)
- Explore branch by branch, towards the terminal nodes (leaves).
- It helps to understand the underlying data well.
  - Of all the possible co-variates, which affect each other (no naive assumptions), which ones were the most influential in “predicting” or “identifying” class categorization?

# HOW DECISION TREES MAY BE USED

- **Growing decision trees:** Decision trees are “induced” from data **to enable**
  - data exploration
    - What’s in the data, broadly speaking?
  - discovery
    - What latent data patterns are in the data?
    - What attributes or variables are the most informative of classification?
  - the capturing of “rules” from the data that may be applied to similar (~)data in the future in a predictive way
    - What extracted rules are apparent? What do these rules say about similar data?
    - What are accurate ways to label or classify ~ data in a future context of uncertainty?
  - decision-making and action-taking
    - What does the decision tree suggest about ways to intervene in particular situations for desirable outcomes?
    - What action-taking should be done?



# HOW DECISION TREES MAY BE USED (CONT.)

- **Labeled data for (human-)supervised machine learning:** What is already known in many cases are the classes / categorizations (labeled data in supervised machine learning)...
- These categorization labels may be treated as “outcomes” data.

# HOW DECISION TREES MAY BE USED (CONT.)

- **Unlabeled data for unsupervised machine learning:** In other cases, the decision trees themselves are used to extract the respective classes / bins / categories...and then to further identify the co-variates that are most predictive of membership in the particular classes.
- The categories are identified based on similarity analyses (like attracting like, ~ attracting ~), which can be built on a number of dimensions (depending on data dimensionality, built on attributes). Extracted categories are unlabeled by the computer but should be coherently human-interpretable in many cases.
- There are various clustering algorithms which identify similar objects across different dimensions.

# HOW DECISION TREES MAY BE USED (CONT.)

- For numerical data, clustering may be done around close or similar values...or may be done based on the number of identified "k" desired by the researcher.
- Decision trees can also run on unlabeled data to see if the outcome categories are the same as the labeled data.
  - Decision trees may be applied to "uncertain contexts" in which an instance's classes are not known prior, with classes represented by "belief functions," and with resulting "belief decision trees."



# DECISION TREES: STRENGTHS AND WEAKNESSES

## Strengths

- Human readable, human interpretable
- Enables finding of patterns in data
- Reproducible, same results every time (with same parameters)
- Involves various types and may be compounded in “random forests” (with iterated refinements)

## Weaknesses

- Are limits to predictivity based on the data and based on the method
  - Predictive accuracy range is limited
- May output no findings (no root nodes, no splits), depending on the data
- May miss data nuances (because the focus is on the main variables that affect the targeted outcomes)

# DECISION TREES: STRENGTHS AND WEAKNESSES (CONT.)

## Strengths

- Some with numeric values and others not for outcomes [CHAID, ID3, and decision tree (weight based)]
- Does not require complex data cleaning or data preparation
- Is aesthetically pleasing as a visualization

## Weaknesses

- May output very different trees based on input parameters ("tuning")
- Requires some intimacy with the data
- Requires human interpretation of the outcomes

## 2. INDUCING DECISION TREES FROM MULTIVARIATE DATA



# GENERAL SEQUENCE: EIGHT STEPS

1. Collect data. Or select existent data. Combine or “union” datasets as needed.
2. Clean the data.
  - If data columns are not particularly relevant, these should be removed. The idea is not to pre-empt the process by removing data columns unthinkingly or doing other things that may change up the findings.
  - If numeric indicators are required, depending on the decision tree type, change nominal variables to numeric dummy variables. These can sometimes work with multimodal variables, but it helps to have the data aligned for easier decision tree runs.
  - Re-name column headers to information-rich labels (usually in camelCase).
  - Sometimes, running the decision tree initially helps surface what changes need to be made.
  - If there is missing data, decide how that will be handled (either fill empty cells with averages of the other cells, or use some other method to handle this).



# BUT NOT SO MANY VARIABLES THAT THERE IS INCOHERENCE

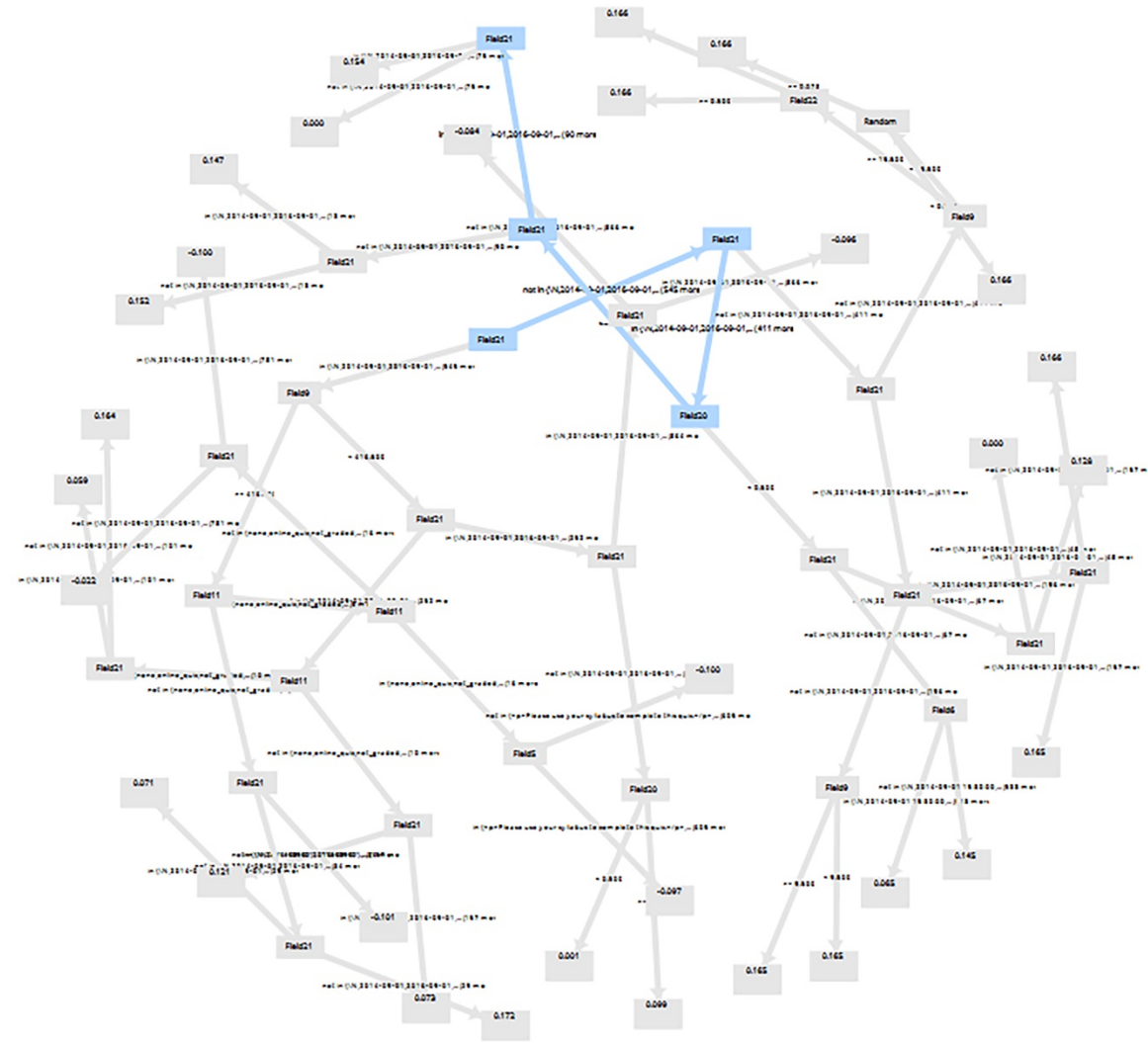
Ensemble of multiple trees induced from training data

Tend to be shallow trees (aka "weak learners")

Makes fewer mistakes in each iteration;  
non-random tree building

If high learning rate set, corrects more errors iteratively resulting in more complex trees (but may overfit to the training data and be less generalizable to other similar data)

If low learning rate, simpler trees result but may be more generalizable and applicable to new test data

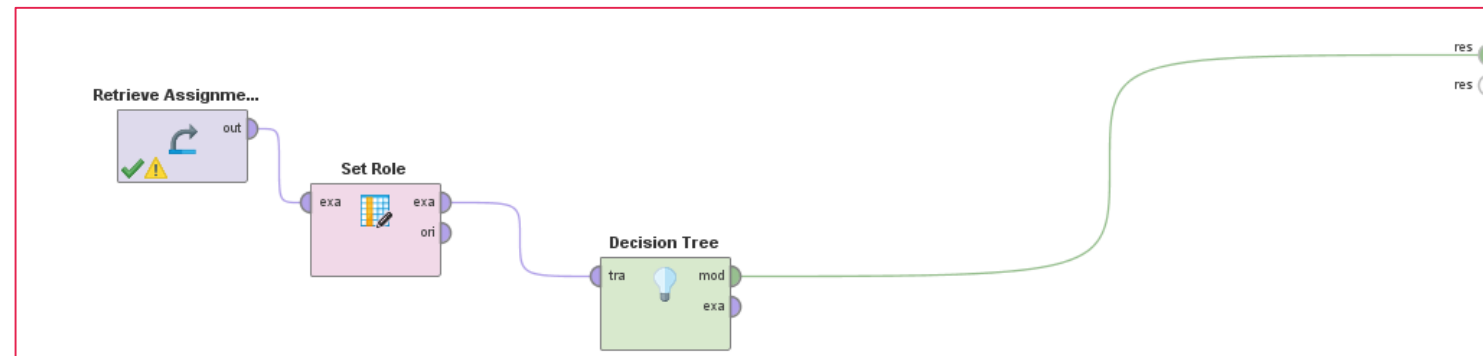
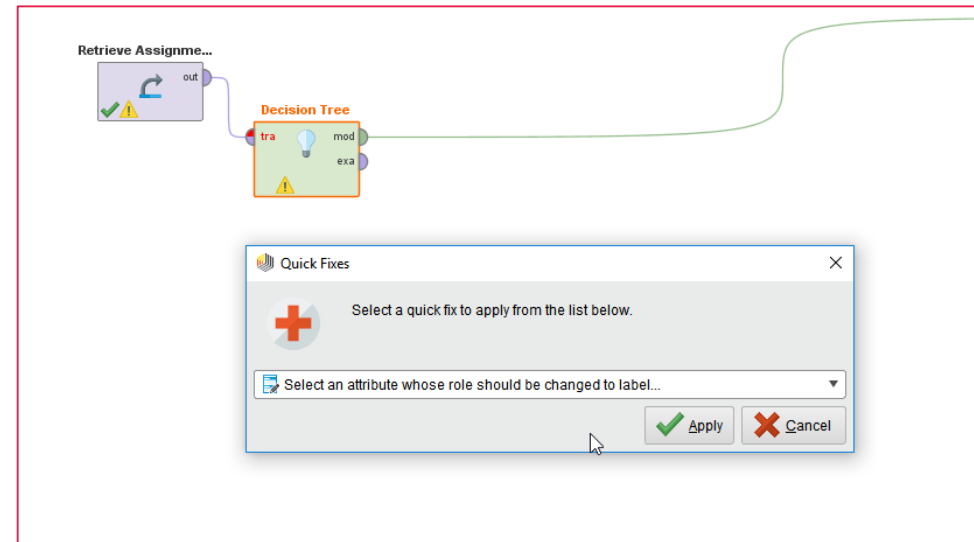


a gradient boosted tree

# GENERAL SEQUENCE: EIGHT STEPS

(CONT.)

3. Identify a column with the class identifier (for supervised machine learning, with labeled data).
  - Use the Set Role Operator feature in RapidMiner Studio.
4. Set up the decision tree process with the desired parameters.
  - Record the parameters for later reference, if needed.



# GENERAL SEQUENCE: EIGHT STEPS

(CONT.)

5. Run the decision tree induction process.
  - The first phase of the decision tree process is the “growth” stage.
  - A follow-on phase is the “pruning” phase. (Pre-pruning runs simultaneous or parallel to the tree growth in the prior stage.)
6. Analyze the decision tree visualization(s), from the root node through the branches to the terminal leaf nodes (aka “answer nodes”). Examine the induced rules to see what those suggest about the particular data in the particular domain.
  - Hypothesize about why particular associations and relationships are seen in the data. (In inductive logic, the person still has to take the collected data and make an assertion from it. The idea is to use clean logic.)

# GENERAL SEQUENCE: EIGHT STEPS

(CONT.)

7. Re-run the analysis with different parameters, if desired.
8. Test the efficacy of the induced decision tree on new data to see how accurate the decision tree is for predictive analytics.

# WHAT IS HAPPENING ON THE BACK END?

- Decision trees analyze the variables in the dataset to see which ones are most informative and predictive of classes in the dataset (the ones which are the best cues to category or class).
- The data is split recursively from the most coarse categories to the most nuanced.
  - Once labeled, the data can only belong to that one leaf or node. There is a general assumption of mutual exclusivity.
  - The more data that can be processed, the more informative the model can be.
  - The closer a decision tree is to its terminal leaves, the smaller and smaller the segments. The decision tree ends when subsets cannot be partitioned further based on predefined parameter criteria.

# WHAT IS HAPPENING ON THE BACK END? (CONT.)

- Decision tree depth (layers) are generally stopped at about 20 layers in default settings in this tool but can actually be iterated many more times.
  - The data will likely often limit decision tree depth.
- The co-variates that affect classifications (outcomes) are inherently within the data, and there are other machine learning methods that can surface the same factors and data patterns. Machine learning methods enable the surfacing of otherwise-latent or hidden data patterns.

Import Data - Select the cells to import.

Select the cells to import.

Sheet: Assignment\_dim\_00000\_7f909300 Cell range: A:AB Select All Define header row: 1

Field1	Field2	Field3	Field4	Field5	Field6	Field7	Field8	Field9	Field10	Field11	Field12	Field13
393504	colorime...	W	W	W	10,000	points	none	published	20			
393506	preparati...	W	W	W	10,000	points	none	published	20			
393507	antacid B...	W	W	W	10,000	points	none	published	20			
396729	fatty acid	W	W	W	10,000	points	none	published	20			
396260	Saponin...	W	W	W	10,000	points	none	published	20			
402288	vic	W	W	W	10,000	points	none	published	20			
402289	CHO pa...	W	W	W	10,000	points	none	published	20			
402292	iodine n...	W	W	W	10,000	points	none	published	20			
405409	Two dim...	W	W	W	10,000	points	none	published	20			
405410	Lactose L...	W	W	W	10,000	points	none	published	20			
409914	isolation	W	W	W	10,000	points	none	published	20			
409915	TLC	W	W	W	10,000	points	none	published	20			
409916	gel filtra...	W	W	W	10,000	points	none	published	20			
416254	lowry me...	W	W	W	10,000	points	none	published	20			
416480	carbonic...	W	W	W	10,000	points	none	published	20			
412984	enzyme...	W	W	W	10,000	points	none	published	20			
412995	uricase	W	W	W	10,000	points	none	published	20			
415258	SDS gel	W	W	W	10,000	points	none	published	20			
395359	Exam #1	W	W	W	100,000	points	on_paper	published	20			

Sheet content too large  
Only a preview of the first 5,000 rows is shown. Data range selection within the table is disabled. Please use the text fields to specify the data range.  
Got it!

<new process\*> - RapidMiner Studio Free 7.5.000 @ HL-ITAC-LTSHALI

File Edit Process View Connections Cloud Settings Extensions

Views: Design Results

Repository

- Samples
- DB
- Local Repository (shalin)
  - data (shalin)
  - processes (shalin)
    - Assignment\_dim-00000-7f909300 (shalin - v1, 8/14/17 12:51)
    - FauxDatasheet1 (shalin - v1, 7/31/17 1:15 PM - 4 kB)
    - FauxishData (shalin - v1, 7/31/17 5:39 P)
    - RunonNutritionData (shalin - v1, 7/31/17 5:39 P)
- Cloud Repository (disconnected)

Process

Process

inp

Retrieve Assignme... out

Decision Tree tra mod exa

Data limit reached

The process is trying to use 421,943 rows of data, but your product license only supports up to 10,000 rows.

We can automatically **downsample** the data to comply with the limit, or you can **upgrade** your license to increase the limit.

Downsample data Upgrade license

[Learn more about license limits](#)

Operators

Search for Operators

Predictive (61)

# ASSIGNMENT DATA FROM THE K-STATE ONLINE CANVAS LMS DATA PORTAL

May 2017 data dump

10,000/421,943 rows of data (2% of data analyzed)

RapidMiner Studio (free educational version) only allows 10,000 rows of data at a time to be analyzed





# RANDOMIZING A SMALLER DATASET

Capture the .gz download from the LMS data portal.

Expand zipped files with 7Zip.

Open files in MS Access.

Export table to Access format .accdb.

Export to Excel from there.

Randomize data in Excel by creating a column of random numbers and then sorting by random.

Use =RAND()

Then sort so the top 10,000 records are actually pseudo-randomized.

Use "down-sampled" data for the induction of decision trees.

Refer to the Canvas Data Portal's "Schema Docs" (data dictionary) for understandings.

The screenshot shows an Excel spreadsheet with the following data:

	A	B	C	D	E	F
1	Random	Field4	Field5	Field6	Field7	Field8
2	0.924292	colorimetic of aspirin	\N	\N	\N	\N
3	0.757549	preparation of aspirin	\N	\N	\N	\N
4	0.828464	antacid titration	\N	\N	\N	\N
5	0.165602	fatty acid	\N	\N	\N	\N
6	0.518662	Saponification num	\N	\N	\N	\N
7	0.343342	vitC	\N	\N	\N	\N
8	0.565081	CHO paper chromat	\N	\N	\N	\N
9	0.531485	iodine number	\N	\N	\N	\N
10	0.390369	Two dimensional an	\N	\N	\N	\N
11	0.286059	Lactose in milk	\N	\N	\N	\N
12	0.58918	isolation cholesterol	\N	\N	\N	\N
13	0.399266	TLC	\N	\N	\N	\N
14	0.041749	gel filtration	\N	\N	\N	\N
15		ry method and b	\N	\N	\N	\N

# TURNING "TRUE" TO 1; TURNING "FALSE" TO 0

- 153,901 true-s
- 3,223,550 false-s

The image shows two overlapping Excel 'Find and Replace' dialog boxes. The top dialog is set to find 'false' and replace it with '0'. The bottom dialog is set to find 'true' and replace it with '1'. Both dialogs show a list of found cells with columns for Book, Sheet, Name, Cell, Value, and Formula. The top dialog shows 3,223,550 cells found, and the bottom dialog shows 153,901 cells found.

Book	Sheet	Name	Cell	Value	Formula
Edite...	Assig...		SNS2	false	
Edite...	Assig...		SOS2	false	
Edite...	Assig...		SPS2	false	

Book	Sheet	Name	Cell	Value	Formula
Edite...	Assig...		SQ...	true	
Edite...	Assig...		SQ...	true	
Edite...	Assig...		SQ...	true	

EditedAssignmentDataforDecisionTreeRendering - Excel

Shalin Hai-Jew | Share

File Home Insert Page Layout Formulas Data Review View ACROBAT Tell me what you want to do...

Clipboard Font Alignment Number Styles Cells Editing

Clipboard: Cut, Copy, Paste, Format Painter  
 Font: Calibri, 11, Bold, Italic, Underline, Text Color, Background Color  
 Alignment: Wrap Text, Merge & Center  
 Number: General, Currency, Percentage, Decimals  
 Styles: Normal, Bad, Good, Neutral, Calculation, Check Cell, Explanatory, Input, Linked Cell, Note  
 Cells: Insert, Delete, Format  
 Editing: AutoSum, Fill, Clear, Sort & Filter, Find & Select

A1: Random

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC		
1	Random	Field4	Field5	Field6	Field7	Field8	Field9	Field10	Field11	Field12	Field13	Field14	Field15	Field16	Field17	Field18	Field19	Field20	Field21	Field22	Field23	Field24	Field25	Field26	Field27	Field28					
2	0.432292	Semana 11	\N	\N	\N		20	points	none	published	2014-08-1'	2014-08-1'		0	2013-11-0'	0	0	0	0	\N	0	0	0	0		12	everyone				
3	0.955874	EXAM1 VERSION2	\N	\N	\N		100	points	online_qu	published	2016-09-2'	2016-12-2'		0	\N	0	0	0	0	\N	0	0	0	0		2	only_visible_to_overrides				
4	0.934174	Syllabus C	<p>Please	2014-09-0'	2014-08-2'	2014-09-0'		10	points	online_qu	deleted	2014-12-2'	2014-12-2'		0	2014-09-0'	0	0	0	0	2014-09-0'	0	0	0	0		1	everyone			
5	0.432223	Lab 1 Instr	<p><a id='	2016-09-0'	\N	\N			not_grade	not_grade	deleted	2016-12-1'	2016-12-1'		0	2016-09-0'	0	0	1	2016-09-0'	0	0	0	0		4	everyone				
6	0.838449	Quiz #3	\N	\N	\N	\N		30	points	none	published	2016-02-2'	2016-03-0'		0	\N	0	0	0	\N	0	0	0	0		3	everyone				
7	0.224446	What exp	<p style='	\N	\N	\N		1	points	discussio	published	2017-01-1'	2017-02-0'		0	\N	0	0	0	\N	0	0	0	0		1	everyone				
8	0.729038	Session 4	<div class='	2016-09-2'	\N	\N		100	points	online_up	published	2016-08-2'	2016-09-2'		0	2016-09-2'	0	0	0	0	2016-09-2'	1	0	0	0		3	everyone			
9	0.714982	Homewor	<ul>\r\n<	\N	\N	\N			not_grade	not_grade	unpublish	2016-09-2'	2016-11-1'		0	\N	0	0	0	\N	0	0	0	0		26	everyone				
10	0.814093	Participation Week 1	\N	\N	\N	\N		6	points	none	published	2016-02-0'	2016-04-2'		0	\N	0	0	0	\N	0	0	0	0		14	everyone				
11	0.10864	Week 4 Ré	<p><span	2016-02-1'	\N	\N		0	pass_fail	online_te	published	2016-01-1'	2016-02-1'		0	2016-02-1'	0	0	0	1	2016-02-1'	1	0	0	0		4	everyone			
12	0.198245	Attendance Bonus	\N	\N	\N	\N		0	points	none	published	2016-12-1'	2017-05-0'		0	\N	0	0	0	\N	0	0	0	0		4	everyone				
13	0.26461	Exam 1 test	\N	\N	\N	\N		72	points	online_qu	deleted	2017-02-1'	2017-02-1'		0	\N	0	0	0	\N	0	0	0	0		19	everyone				
14	0.725889	Tan & Cot Graph Onl	2017-02-2'	2017-02-1'	2017-03-0'		6	points	online_qu	published	2017-01-2'	2017-02-0'		0	2017-02-2'	0	0	0	0	2017-02-2'	0	0	0	0		17	everyone				
15	0.234438	#19 Styles	<p><a id='	2015-12-1'	2015-11-1'	2015-12-1'		5	points	online_up	published	2015-11-1'	2015-12-1'		0	2015-12-1'	0	0	1	2015-12-0'	0	0	0	0		20	everyone				
16	0.063835	MPR/Thes	<p>Please	2016-05-0'	\N	\N		100	points	online_up	published	2016-04-2'	2016-05-1'		0	2016-05-0'	0	0	0	2016-05-0'	0	0	0	0		5	everyone				
17	0.936924	FRH Chapt	<p>This is	2016-04-0'	\N	\N		50	points	online_qu	published	2016-01-2'	2016-01-2'		0	2016-04-0'	0	0	0	2016-04-0'	0	0	0	0		6	everyone				
18	0.081061	CH 21 LQ Giovanni B	\N	\N	\N	\N		5	points	online_qu	published	2014-07-2'	2014-07-2'		0	\N	0	0	0	\N	0	0	0	0		107	everyone				
19	0.630569	Group Project - Indu	2016-03-1'	\N	\N	\N		10	points	online_up	published	2016-03-0'	2016-03-0'		0	2016-03-1'	0	0	0	0	2016-03-1'	0	0	0	0		15	everyone			
20	0.826155	Extra Cred	<p>Write	2016-10-1'	2016-10-1'	2016-10-1'		0	points	online_te	deleted	2017-01-1'	2017-01-1'		0	2016-10-1'	0	0	0	2016-10-1'	1	0	0	0		18	everyone				
21	0.316608	Answers To the T/F	\N	\N	\N	\N		0	points	none	published	2015-12-1'	2015-12-1'		0	\N	0	0	0	\N	1	0	0	0		7	everyone				
22	0.325894	Peer Eval	<p>To be	2016-10-1'	2016-10-0'	2016-10-1'		25	points	none	published	2016-08-2'	2016-10-1'		0	2016-10-1'	0	0	0	2016-10-1'	0	0	0	0		4	everyone				
23	0.379636	Quiz 9	\N	\N	\N	\N		5	points	online_qu	published	2014-11-1'	2014-11-1'		0	\N	0	0	0	\N	0	0	0	0		11	everyone				
24	0.983654	Assignment #3	2016-02-1'	\N	\N	\N		20	points	on_paper	published	2016-02-1'	2016-02-1'		0	2016-02-1'	0	0	0	2016-02-1'	0	0	0	0		15	everyone				
25	0.525818	Chapter 28. Causes c	2015-11-1'	\N	\N	\N		5	points	none	published	2015-11-1'	2016-01-0'		0	2015-11-1'	0	0	1	2015-11-1'	0	0	0	0		0	everyone				
26	0.871872	Time for Action 4.2	\N	\N	\N	\N		1	percent	external_un	unpublish	2016-07-1'	2016-08-1'		0	\N	0	0	0	\N	1	0	0	0		45	everyone				
27	0.54396	Post Asse	<p><span	\N	\N	\N		15	points	online_qu	published	2016-10-2'	2016-10-2'		0	\N	0	0	0	\N	0	0	0	0		9	everyone				
28	0.340241	Unnamed	\N	\N	\N	\N			points	online_qu	deleted	2016-12-1'	2016-12-2'		0	\N	0	0	0	\N	0	0	0	0		6	everyone				
29	0.537834	Assignment 6	\N	\N	\N	\N		1	points	on_paper	published	2015-10-1'	2015-10-1'		0	\N	0	0	0	\N	0	0	0	0		8	everyone				
30	0.300671	Topic Sele	<p>In Can	2015-09-1'	2015-08-1'	2015-09-1'		10	points	online_te	published	2015-08-0'	2015-09-1'		0	2015-09-1'	0	0	0	2015-09-1'	0	0	0	0		5	everyone				
31	0.508755	Off-Site Fi	<p>Take t	\N	\N	\N		3	points	online_qu	published	2017-03-0'	2017-03-0'		0	\N	0	0	0	\N	0	0	0	0		3	everyone				
32	0.413189	Speaking Final	\N	\N	\N	\N		100	points	none	published	2016-01-1'	2016-01-1'		0	\N	0	0	0	\N	0	0	0	0		7	everyone				
33	0.009466	Case 2: Co	<p>Dear s	2016-09-2'	2016-09-1'	2016-09-2'		25	points	online_up	published	2016-07-2'	2016-07-2'		0	2016-09-2'	0	0	0	2016-09-2'	0	0	0	0		4	everyone				
34	0.168286	AccuScan Cards	\N	\N	\N	\N		5	points	none	published	2015-11-1'	2015-11-1'		0	\N	0	0	0	\N	0	0	0	0		5	everyone				
35	0.204193	Think Nati	\N	2015-10-0'	\N	\N		15	points	none	published	2015-10-0'	2015-10-0'		0	2015-10-0'	0	0	0	1	2015-09-3'	0	0	0		12	everyone				
36	0.576105	*on-line check in	2015-04-0'	\N	\N	\N		8	points	none	deleted	2015-01-2'	2015-03-0'		0	2015-04-0'	0	0	0	0	2015-04-0'	0	0	0		11	everyone				
37	0.245275	Participation Week 5	\N	\N	\N	\N		10	points	none	published	2016-01-1'	2016-03-2'		0	\N	0	0	0	\N	0	0	0	0		5	everyone				

Assignment\_dim\_00000\_7f909300

Ready | Average: 0.4994089 | Count: 421944 | Sum: 210722.0896 | 100%

# MORE NUANCED DATA PREPARATION FOR DECISION TREES

- Combine co-variates that are co-linear to reduce multi-collinearity.
- Rework the data types to align with the types of decision trees to run (some can deal with numeric target classes and others cannot, for example).
- Omit variables which may not have an effect on the target outcomes classifications, or which may not make (any) sense in that context.
  - Sometimes, oftentimes, less is more.

# SETTING PARAMETERS FOR DECISION TREES

**Pre-pruning (during decision tree growth; | | to decision tree growth)**

- **Criterion:** “which attributes will be selected for splitting”
  - **information\_gain:** selecting attributes with minimum entropy (and “a bias towards selecting attributes with a large number of values”)
  - **gain\_ratio:** “adjusts the information gain for each attribute to allow the breadth and uniformity of the attribute values” (default: **gain\_ratio**)
  - **gini\_index:** “a measure of impurity of an ExampleSet” and splits on an attribute which “gives a reduction in the average gini index of the resulting subsets,” so resulting subsets are less dispersed and less variant (gini is a measure of inequality)
  - **accuracy:** attributes selected “for split that maximizes the accuracy of the whole Tree”
- **Maximal depth:** the size of the decision tree in layers (default: **20**)

# SETTING PARAMETERS FOR DECISION TREES (CONT.)

## Post-pruning (induced decision tree performance, accuracy, sensitivity)

- After a decision tree has been induced, post-pruning can be brought into play by assessing the tree's ability to differentiate classifications with accuracy.
- Confidence level for the "pessimistic error calculation of pruning" ("Decision Tree (concurrency)") (a measure to ensure that the generated rules hone pretty closely to the data, without expansive error bars)(default: yes, and 0.25)

# SETTING PARAMETERS FOR DECISION TREES (CONT.)

- **Pre-pruning** (the stopping criteria for leaf node generation...as in when should the decision tree stop growing?):
  - **Minimal gain:** the threshold gain that has to be achieved at a node before it is split (default 0.1)
  - **Minimal leaf side:** the minimal number of examples in a leaf's subset (parent node size) (default: 2)
  - **Minimal size for split:** the root node is = "to the total number of examples in the ExampleSet," and thereafter, only the nodes  $\geq$  minimum size for the split can be split (default: 4)
  - **Pre-pruning alternatives:** available methods for additional attempts to split leaf nodes in a decision tree if other pre-pruning parameters limit splitting (if measures indicate that splitting will not necessarily add "to the discriminative power of the entire tree") (default: 3)

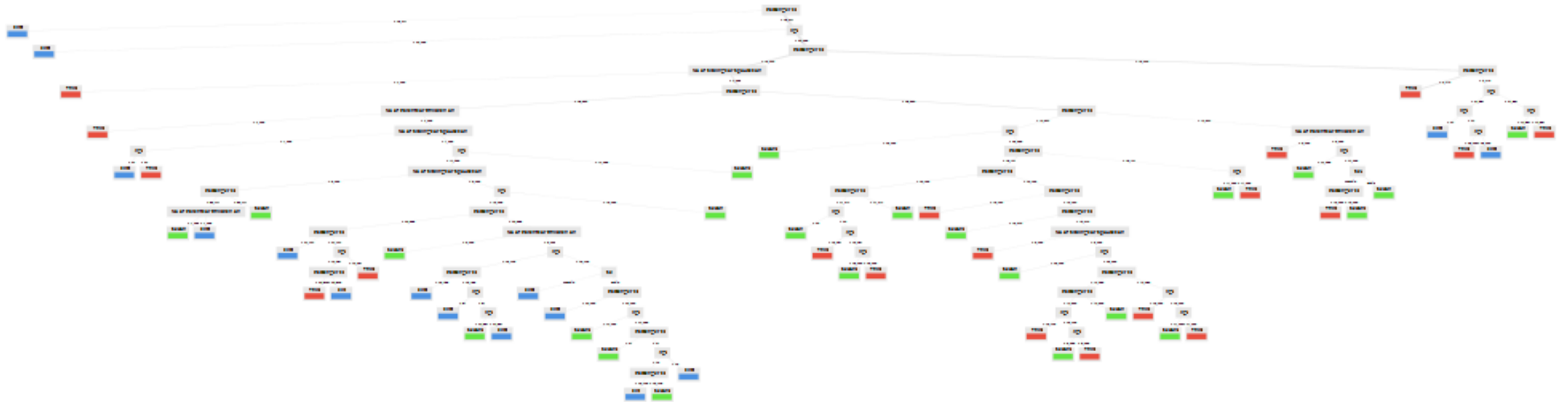
# SETTING PARAMETERS FOR DECISION TREES (CONT.)

- If all objects belong to one class or category, in a particular leaf node, then tree growth stops at that particular node. (There is nothing else to divide / split.)
- The point of a decision tree is to identify the best split attribute (variable) and the optimal split point (the numerical or dummy information in the edges connecting the leaves).
- A risk with decision trees (if the parameters are set up a certain way) is that they overfit to the training data with so much specificity that the trees and rules cannot apply effectively to similar data; in other words, they do not generalize in an applicable way.
  - Smaller and simpler decision trees are thought to be more generalizable / applicable to broader ranges of similar data.



# SETTING PARAMETERS FOR DECISION TREES (CONT.)

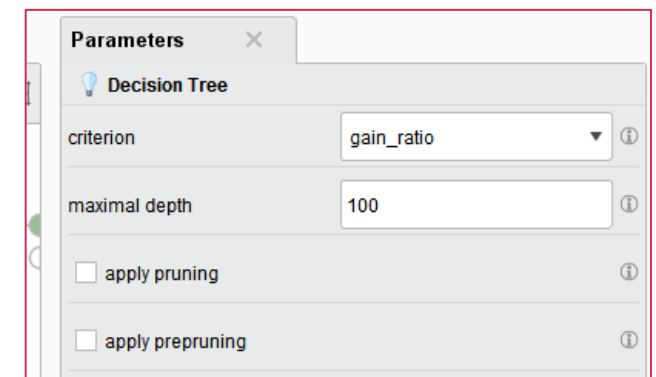
- More training data (up to a certain point) enhances the rule-making and discriminant capability of decision trees.
- **Caveat:** Even with  $N = \text{all}$ , decision trees will not have full predictivity because of the nature of data, the inherent noise in data, the limits to the decision tree model, data cleaning and handling, the thin separations between some categories, and other factors.



# AN UNPRUNED DECISION TREE

From: unlabeled Titanic dataset in RapidMiner Studio, "Passenger Fare" focus; no pre-pruning (runs parallel to tree generation / growth), no post-pruning

More nuanced interaction effects between variables (but also a visual sense that the tree may have run amuck a little)



# UNDERSTANDING MODEL ACCURACY

- It is possible to run statistical tests against a decision tree to see how well its induced rules apply to non-training data.
- Usually, a dataset is split 70:30 with the 70% used for training and the 30% used for testing. The test is for both the following:
  - “sensitivity” (“true positive rate” / recall / “probability of detection”), and
  - “specificity” (“true negative rate”)
- How much noise or uncertainty is there in the classification assessment? ([“Sensitivity and specificity,”](#) Aug. 11, 2017)
- The error rate is the total number of misclassified points (whether Type 1 or Type 2 errors), divided by the total number of data points.
  - The accuracy rate is one minus the error rate.

# UNDERSTANDING MODEL ACCURACY (CONT.)

- Decision tree models are in competition with other means of finding patterns in data (like neural networks, like discriminant analysis, and others).
- The nature of datasets (and how they were cleaned) may affect how efficient particular models are.
  - Some methods may be sensitive to noise and complexity.
- There is a value to having parsimonious models or those that are pared-down, spare, and simple enough to apply to other contexts. The power of decision trees is not in minute details...but the broad-scale co-variates and broad splits, when it comes to predictivity. [In contrast, a descriptive use of decision trees may be better un-pruned, so the various data nuances may be extracted. However, the human analyst(s) then have to be comfortable with complexity.]

# 3. DATA STRUCTURES





# 4. MODEL ACCURACY AND DECISION TREE VALIDATION

Type 1 and Type 2 Errors



# 2X2 CONTINGENCY TABLES

(IN A BINARY CLASSIFICATION CONTEXT)

Table of Error Types		Null Hypothesis Is...	
		Actually True / Only Chance at Play	Actually False / More than Chance at Play
Decision about Null Hypothesis	<b>Reject Null Hypothesis</b> (There is more than chance acting on this data.)	<b>Type 1 Error</b> False Positive +- Incorrect decision in identifying the null hypothesis as true (when it is actually false)	Correct Inference True Positive ++ Correct decision in identifying the null hypothesis as false and correctly rejecting the null hypothesis
	<b>Do Not Reject the Null Hypothesis</b> (There is nothing more than chance acting on this data.)	Correct Inference True Negative -- Correct decision in identifying the null hypothesis as true and not rejecting the null hypothesis	<b>Type 2 Error</b> False Negative -+ Incorrect decision in identifying the null hypothesis as false (when it is actually true)



# CLASSIFICATION ERRORS / CONFUSION MATRIX

Classification Errors		Actual Classification	
		Does Not Belong to Class	Belongs to Class
Decision about Classification	Assess that Something Belongs to Target Class Decision	Type 1 Error False Positive +-	Correct Inference True Positive ++
	Assess that Something Does Not Belong to Target Class Decision	Correct Inference True Negative --	Type 2 Error False Negative -+

# RECEIVER OPERATING CHARACTERISTIC (ROC CURVE)

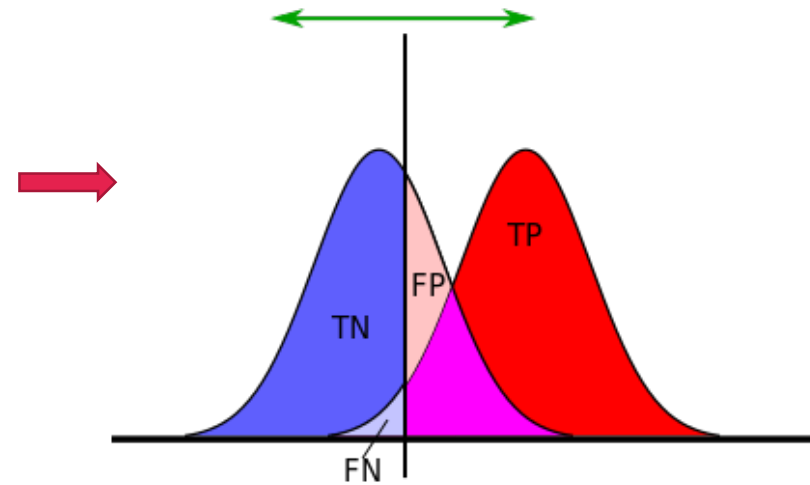
TP = true positive  
FP = false positive  
FN = false negative  
TN = true negative

50

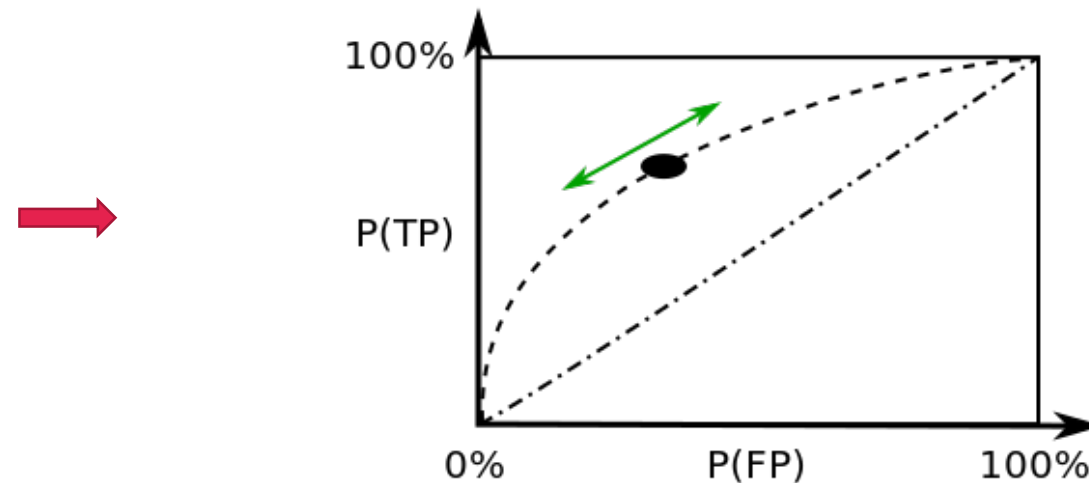
Vertical line or criterion value (which moves horizontally) determines sensitivity of the instrument for detection. What is to the right of the line is seen as positive, and what is to the left is seen as negative. The point is to minimize classification errors.

The ROC Curve visualization shows the true positive rate (y-axis, indication of sensitivity) plotted against the false positive rate (x-axis) at different cut-off points of the assessment. A perfect test has an ROC curve that reaches the upper left corner, which suggests 100% sensitivity, and 100% specificity, without errors. A perfect test would have no overlap between the two distributions.

The diagonal shows what would be the test sensitivity if it were only going by chance (as with a "random classifier"). Above the diagonal are good classifiers, and below the diagonal are poor ones (with lots of false positives).



TP	FP
FN	TN



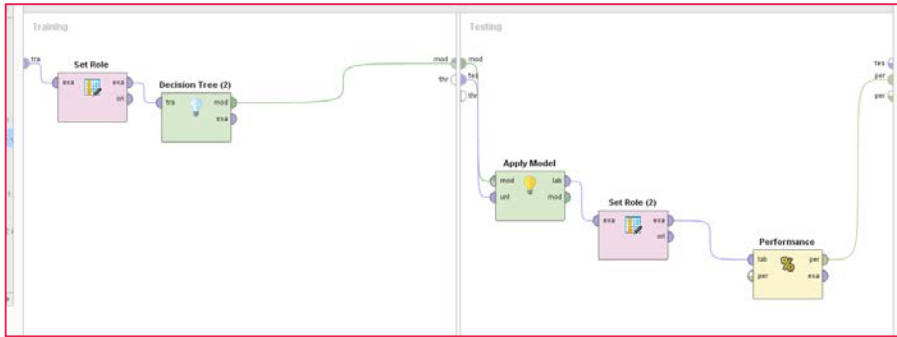
# A MULTI-CLASS CONFUSION MATRIX

	PREDICTED					
ACTUAL	A	B	C	D	E	
A	$TP_A$	EAB	EAC	EAD	EAE	
B	EBA	$TP_B$	EBC	EBD	EBE	
C	ECA	ECB	$TP_C$	ECD	ECE	
D	EDA	EDB	EDC	$TP_D$	EDE	
E	EEA	EEB	EEC	EED	$TP_E$	

The true positives run along the diagonal. Anything off-diagonal are the errors. The "EBA" should be read as the error value where the true "B" is misread as "A". (The draft table above was recreated from [Dr. Nouredin Sadawi's "Evaluating Classifiers: Confusion Matrix for Multiple Classes,"](#) Aug. 30, 2014)

# ARRIVING AT “GENERALIZABLE” DECISION TREE MODELS

- The trick is to avoid “overfitting” a model to the training data. If a model is “overfit,” it is too closely specified to the data from which it was originate and so may tell you a lot about patterns in that data but does not transfer or generalize to independent datasets.
  - Model “underfitting” occurs when a model does not sufficiently capture or describe the underlying training data.
- A “cross-validation” tests how well the model would perform against independent data.
- The independent data against which a model is tested may be excerpted from the original data and set aside as test data.



# RUNNING A CROSS-VALIDATION ON THE DECISION TREE MODEL

1. On the process pane, delete the decision tree. In Validation -> Performance -> Predictive, select "Performance (Classification)".
2. Within the Cross-Validation tool, put the Decision Tree with the Set Role labels in the Training pane (left)...and in the Testing pane, put Apply Model -> Set Role -> Performance. Define the column with the target classification.
3. In the Training pane, the links between the "ports" should be training -> example -> example -> training -> model -> model ports
4. In the Testing pane, the links between the ports should originate from two sources: model and testing | | -> model and unlabeled data | | -> labeled to example set input -> example set -> labeled -> performance -> performance ports

This cross-validation was run on 10 folds (subsets), and the data was sampled automatically. With such a large dataset, this cross-validation could have been run with fewer than 10 folds. Parallel execution was enabled. (All were default settings).



**Repository**

+ Add Data

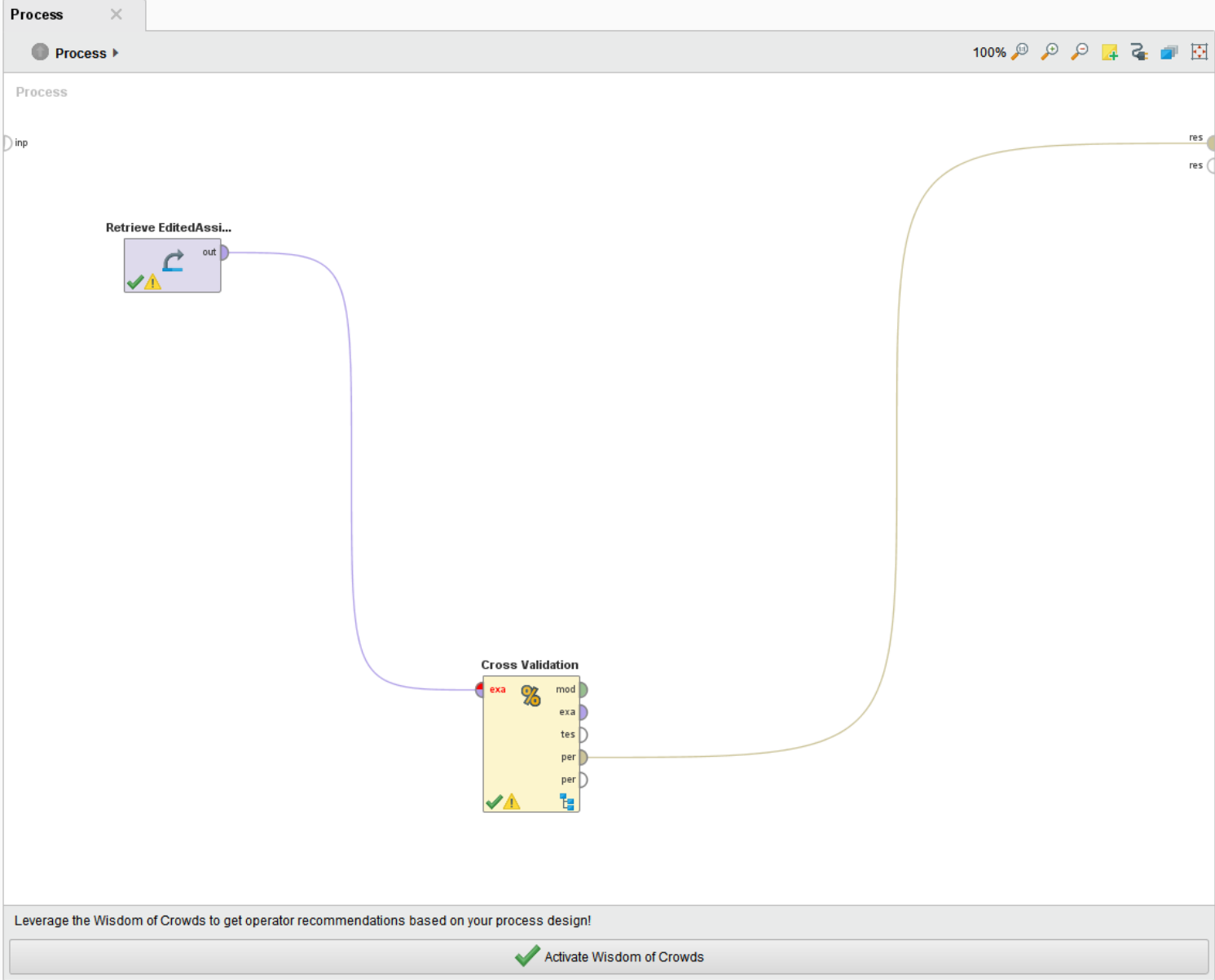
- Samples
- DB
- Local Repository (shalin)
  - data (shalin)
  - processes (shalin)
    - Assignment\_dim-00000-7f909300 (shalin - v1, 8/14/17 12:51)
    - EditedAssignmentDataforDecisionTreeRendering (shalin - v1, 8/14/17 12:51)
    - FauxDatasheet1 (shalin - v1, 7/31/17 1:15 PM - 4 kB)
    - FauxishData (shalin - v1, 7/31/17 5:39 PM - 2 kB)
    - RandomForestRunonAssignmentData (shalin - v1, 8/15/17 1:15 PM - 2 kB)
    - RunonNutritionData (shalin - v1, 7/31/17 12:19 PM - 2 kB)
    - TitanicSetandRandomForest (shalin - v1, 8/15/17 2:25 PM - 2 kB)
- Cloud Repository (disconnected)

**Operators**

performance

- Validation (19)
  - Performance (17)
    - Predictive (7)
      - Performance (Classification)
      - Performance (Binominal Classification)
      - Performance (Regression)
      - Performance (Costs)
      - Performance (Ranking)
      - Performance (Support Vector Count)
      - Performance (Attribute Count)
    - Segmentation (4)
      - Cluster Count Performance
      - Cluster Distance Performance
      - Cluster Density Performance

We found "Model Management" in the Marketplace. [Show me!](#)



**Parameters**

Process

logverbosity: init

logfile: [ ]

resultfile: [ ]

random seed: 2001

send mail: never

encoding: SYSTEM

[Hide advanced parameters](#)

[Change compatibility \(7.5.000\)](#)

**Help**

Process

RapidMiner Studio Core

**Synopsis**

The root operator which is the outer most operator of every process.

**Description**

Each process must contain exactly one operator of this class, and it must be the root operator of the process. This operator provides a set of parameters that are of global relevance to the process like logging and initialization parameters of the random number generator.

<new process\*> - RapidMiner Studio Free 7.5.000 @ HL-ITAC-LTSHALI

File Edit Process View Connections Cloud Settings Extensions

Views: Design Results

Need help?

**Repository**

+ Add Data

- Samples
- DB
- Local Repository (shalin)
  - data (shalin)
    - processes (shalin)
      - Assignment\_dim-00000-7f909300 (shalin - v1, 8/14/17 12:51)
      - EditedAssignmentDataforDecisionTreeRendering (shalin - v1, 8/14/17 12:51)
      - FauxDatashet1 (shalin - v1, 7/31/17 1:15 PM - 4 kB)
      - FauxishData (shalin - v1, 7/31/17 5:39 PM - 2 kB)
      - RandomForestRunonAssignmentData (shalin - v1, 8/15/17 1:15 PM - 2 kB)
      - RunonNutritionData (shalin - v1, 7/31/17 12:19 PM - 2 kB)
      - TitanicSetandRandomForest (shalin - v1, 8/15/17 2:25 PM - 2 kB)
  - Cloud Repository (disconnected)

**Process**

Process > Cross Validation

100%

Training

Testing

Drag here

Drag here

**Operators**

performance

- Validation (19)
  - Performance (17)
    - Predictive (7)
      - Performance (Classification)
      - Performance (Binominal Classification)
      - Performance (Regression)
      - Performance (Costs)
      - Performance (Ranking)
      - Performance (Support Vector Count)
      - Performance (Attribute Count)
    - Segmentation (4)
      - Cluster Count Performance
      - Cluster Distance Performance
      - Cluster Density Performance

We found "Model Management" in the Marketplace. [Show me!](#)

Activate Wisdom of Crowds

**Parameters**

Performance (Performance (Classification))

main criterion first

- accuracy
- classification error
- kappa
- weighted mean recall
- weighted mean precision
- spearman rho
- kendall tau
- absolute error
- relative error
- relative error lenient

[Hide advanced parameters](#)

**Help**

**Performance (Classification)**  
RapidMiner Studio Core

Tags: Accuracy, Errors, Precision, Recall, Kappa, Squared, Relative, Validations, Evaluations, Metrics, Predictive

**Synopsis**

This operator is used for statistical performance evaluation of classification tasks. This operator delivers a list of performance criteria values of the classification task.

[Jump to Tutorial Process](#)

**Description**

This operator should be used for performance evaluation.



Result History PerformanceVector (Performance)

Criterion  
accuracy

Performance

Description

Annotations

Table View Plot View

accuracy: 78.31% +/- 1.09% (mikro: 78.31%)

	true published	true deleted	true unpublished	class precision
pred. published	7831	1505	664	78.31%
pred. deleted	0	0	0	0.00%
pred. unpublished	0	0	0	0.00%
class recall	100.00%	0.00%	0.00%	

78.31% accurate with +/- 1.09%

So applied to similar data (generalizability), this decision tree model is about 78% accurate in its classifications, particularly when it comes to identifying the published category of assignments.

Result History PerformanceVector (Performance)

Performance

**PerformanceVector**

PerformanceVector:  
accuracy: 78.31% +/- 1.09% (mikro: 78.31%)

ConfusionMatrix:

True:	published	deleted	unpublished
published:	7831	1505	664
deleted:	0	0	0
unpublished:	0	0	0

Description

Annotations

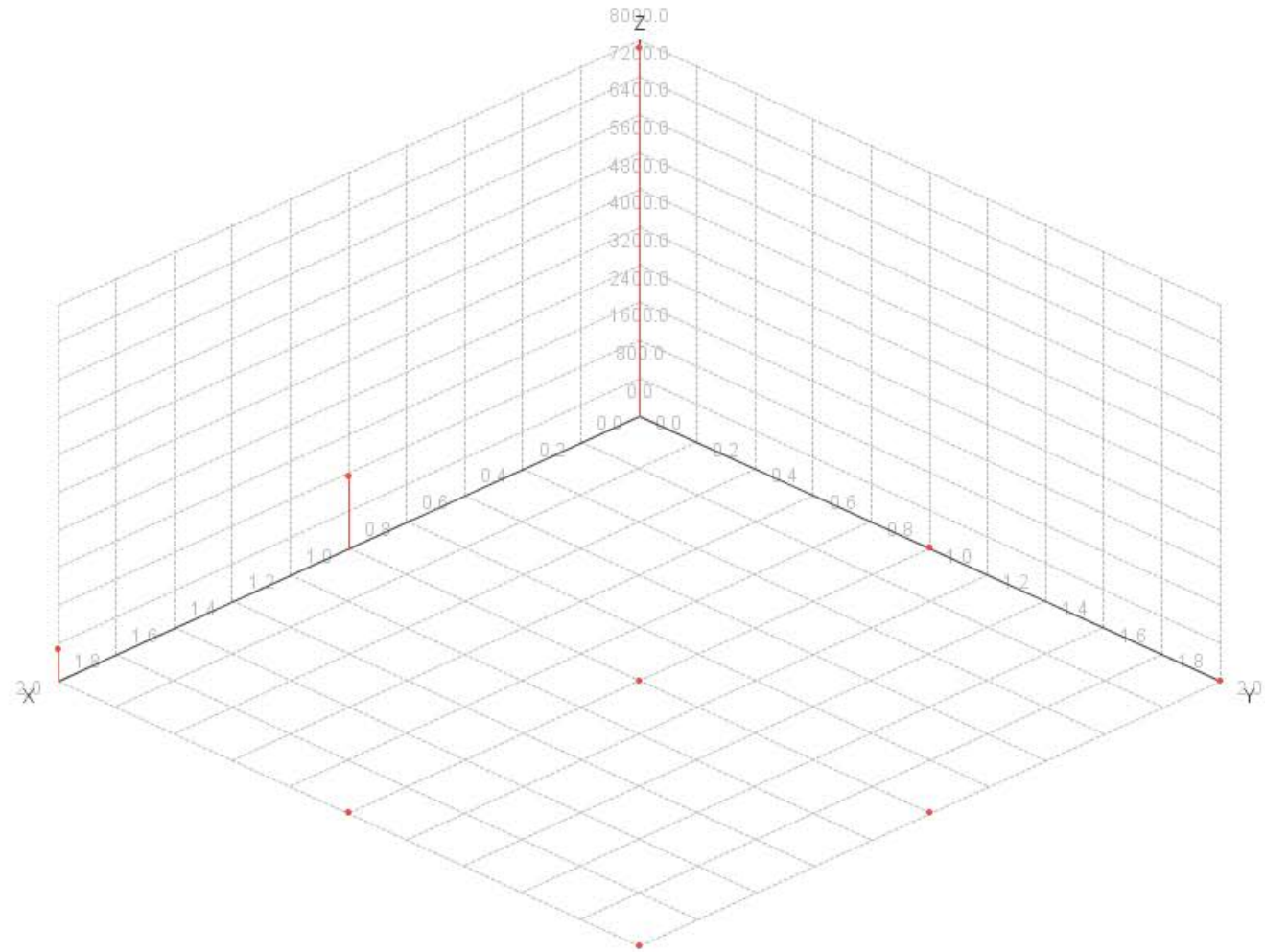
Repository

Add Data

- Samples
- DB
- Local Repository (shalin)
  - data (shalin)
  - processes (shalin)
    - Assignment\_dim-00000-7f909300 (shalin - v1, 8/14/17 12:51 PM)
    - EditedAssignmentDataforDecisionTreeRendering (shalin - v1, 8/14/17 12:51 PM)
    - FauxDatashet1 (shalin - v1, 7/31/17 1:15 PM - 4 kB)
    - FauxishData (shalin - v1, 7/31/17 5:39 PM - 2 kB)
    - RandomForestRunonAssignmentData (shalin - v1, 8/15/17 1:12 PM - 2 kB)
    - RunonNutritionData (shalin - v1, 7/31/17 12:19 PM - 2 kB)
    - TitanicSetandRandomForest (shalin - v1, 8/15/17 2:25 PM - 2 kB)
- Cloud Repository (disconnected)



Confusion Matrix (x: true class, y: pred. class, z: counters)



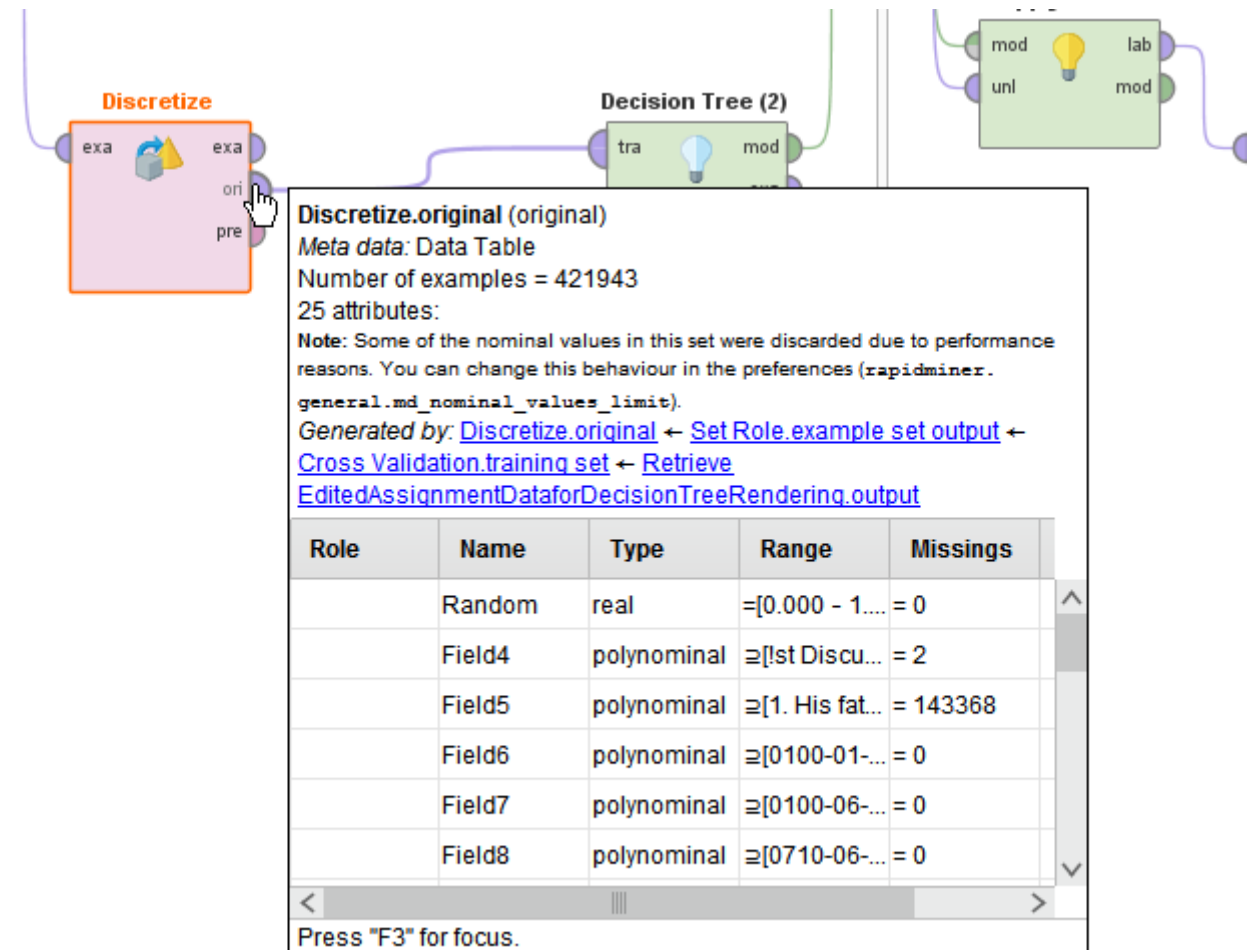
# POST-RUN INTERNAL MODEL VALIDATION (TO THE DATA)

Next steps in the model validation may be to change up the parameters of the decision tree model to see what may increase its performance.

One attempt to improve involved [discretizing](#) by frequency over the original data.

Other attempts were made to discretize on other dimensions as well.

These did not result in any change in performance accurate.



**Discretize**

exa exa  
ori  
pre

**Decision Tree (2)**

tra mod

mod lab  
unl mod

**Discretize.original (original)**  
 Meta data: Data Table  
 Number of examples = 421943  
 25 attributes:  
 Note: Some of the nominal values in this set were discarded due to performance reasons. You can change this behaviour in the preferences (`rapidminer.general.md_nominal_values_limit`).  
 Generated by: [Discretize.original](#) ← [Set Role.example set output](#) ← [Cross Validation.training set](#) ← [Retrieve](#) [EditedAssignmentDataforDecisionTreeRendering.output](#)

Role	Name	Type	Range	Missings
	Random	real	=[0.000 - 1....	= 0
	Field4	polynomial	≥[!st Discu...	= 2
	Field5	polynomial	≥[1. His fat...	= 143368
	Field6	polynomial	≥[0100-01-...	= 0
	Field7	polynomial	≥[0100-06-...	= 0
	Field8	polynomial	≥[0710-06-...	= 0

Press "F3" for focus.

# POST-RUN INTERNAL MODEL VALIDATION (TO THE DATA) (CONT.)

- In this particular case, the model classification accuracy is high for one classification of assignments but not so much on two other classifications. This is an issue (which arises in part because the training data is overwhelmingly of one type).
- There are likely other issues as well, but exploring these will be beyond the purview of this live presentation and digital leave-behind slideshow.

# EXPLORING AND HYPOTHESIZING FROM DECISION TREES

- The resultant decision trees are the basis for starting analyses...not the end point.
  - Each of the levels and splits of the decision trees have meaning (potentially). The levels closer to the root node are more meaningful in a broad-scale way than those by the leaves because the processing is iterative, and each level affects subsequent ones.
  - Each of the branches (and the related values:  $>$ ,  $<$ ,  $\geq$ ,  $\leq$ ) have meaning (potentially).
- Researchers will be hypothesizing about what the decision trees mean. They will need to articulate what the resulting decision trees mean (both informationally and action-wise). [Decision trees are not used in isolation but in a particular context.]
- It is important to know the data intimately: what it contains, how it was acquired, how it was sourced (provenance), how it was cleaned, and so on.

# EXPLORING AND HYPOTHESIZING FROM DECISION TREES (CONT.)

- Surprises are good...but it is important to troubleshoot apparent “nonsense” results...
- It helps to know the theoretical background of the issue at hand, so theoretical frameworks and reasoning may be brought into play.

# 5. RAPIDMINER STUDIO AND DRAWING DECISION TREES; PARAMETERS; SEQUENTIAL WALK-THROUGHS

(a free delimited educational version of the software  
for non-commercial research contexts)



# DEMONSTRATION

<new process> - RapidMiner Studio Free 7.5.000 @ HL-ITAC-LTSHALI

File Edit Process View Connections Cloud Settings Extensions

Views: Design Results

Need help?

**Repository**

data

- Deals (v1)
- Deals-Testset (v1)
- Golf (v1)
- Golf-Testset (v1)
- Iris (v1)
- Labor-Negotiations (v1)
- Market-Data (v1)
- Polynomial (v1)
- Products (v1)
- Purchases (v1)
- Ripley-Set (v1)
- Sonar (v1)
- Titanic (v1)
- Titanic Training (v1)
- Titanic Unlabeled (v1)

**Operators**

Search for Operators

- Data Access (47)
- Blending (77)
- Cleansing (26)
- Modeling (129)
  - Predictive (61)
    - Lazy (2)
    - Bayesian (2)
    - Trees (9)
      - Decision Tree
      - Random Forest
      - Gradient Boosted Trees
      - CHAID
      - ID3
      - Decision Stump

**Process**

Process

Retrieve Market-Data

Retrieve Weighting

Retrieve Titanic Unla...

Retrieve Ripley-Set

Retrieve Deals

Retrieve Labor-Neg...

Set Role

Decision Tree

**Parameters**

Decision Tree

criteria: gain\_ratio

maximal depth: 100

apply pruning:

confidence: 0.25

apply prepruning:

minimal gain: 0.0

minimal leaf size: 2

minimal size for split: 4

number of prepruning alternati...: 100

Hide advanced parameters

**Help**

Decision Tree

Concurrency

Tags: Supervised, Classification, Model, Id3, I4.8, I4.8, C4.5, C4.5, C5.0, C5.0, Cart, Chaid, Trees

**Synopsis**

Generates a Decision Tree for classification of both nominal and numerical data.

[Jump to Tutorial Process](#)

**Description**

A decision tree is a tree-like graph or model. It is more like an inverted tree because it has its root at the top and it

Leverage the Wisdom of Crowds to get operator recommendations based on your process design!

Activate Wisdom of Crowds

# 6. EDUCATIONAL DATA WITH DEFINED OUTCOMES AND CANDIDATE COVARIATES

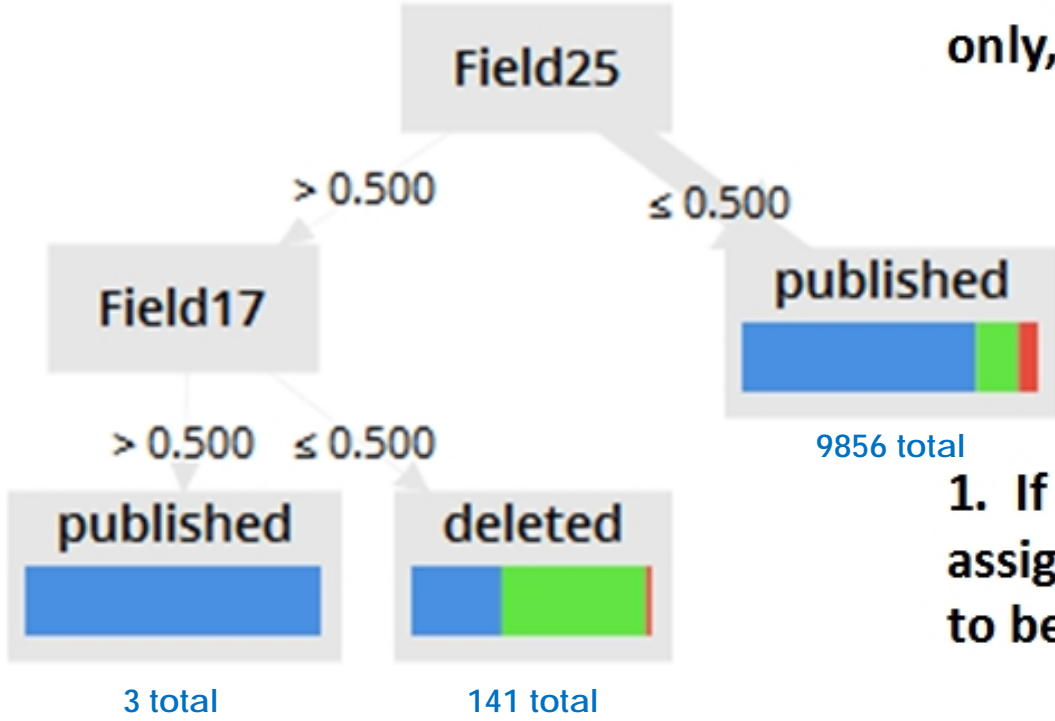




# FROM ONLINE LEARNING DATA: ONE DECISION TREE FROM A RANDOM FOREST MODEL (NEXT SLIDE)

- Note that the particular decision tree (one of 10 iterations from the random forest approach) gives a sense of frequencies.
- Assignment muting and peer review are surfaced as relevant co-variates affecting the outcome states of the assignments: published, unpublished, or deleted.
- The other variables are not sufficiently influential on the outcome variables.
- The following decision tree is one of 10 exported in the random forest process.

Note that Fields 25 and 17 are both binary fields only, with 1 (true) and 0 (false) options.



1. If an assignment is not muted (the majority of assignments are unmuted), it is much more likely to be published...but not in every case...

2. If assignments are muted and require peer reviews, there is a greater likelihood of the assignment being in a published vs. unpublished state.

If assignments are muted but do not require peer reviews, these tend to be deleted.

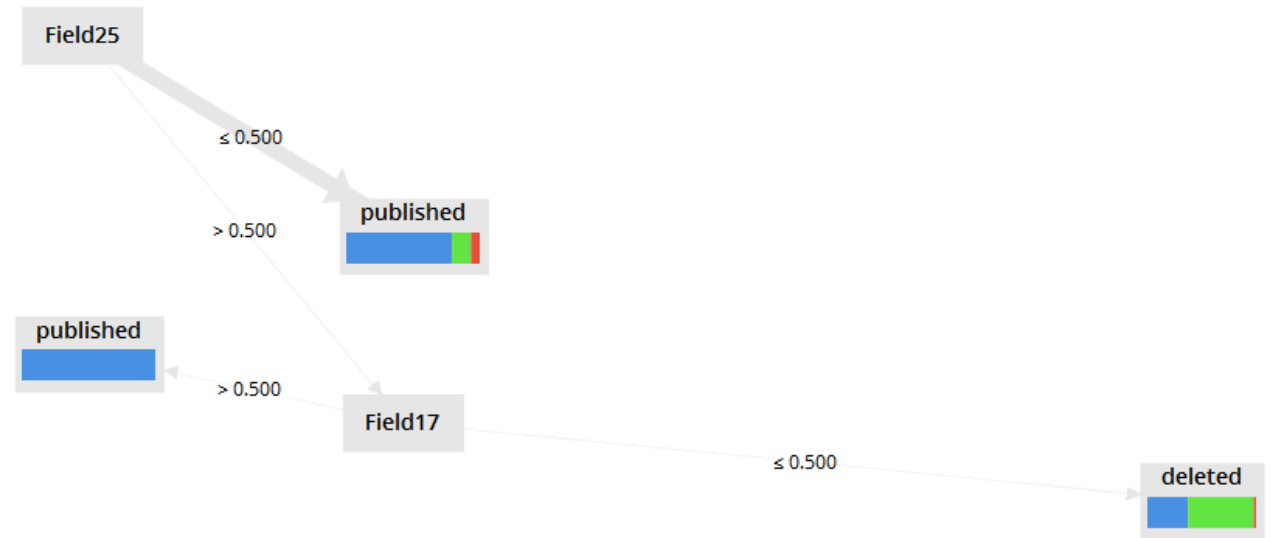
```
Tree
Field25 > 0.500
|   Field17 > 0.500: published {published=3, deleted=0, unpublished=0}
|   Field17 ≤ 0.500: deleted {published=53, deleted=85, unpublished=3}
Field25 ≤ 0.500: published {published=7803, deleted=1409, unpublished=644}
```

Note that the largest numbers of exemplars are the unmuted, and these are represented at the right branch: 9856 total.

# DIFFERENT SPATIAL REPRESENTATIONS OF DECISION TREES

The underlying data, particularly its density or sparsity, will affect how the respective data visualizations look.

- Tree
- Tree (Tight)
- Radial
- Balloon
- ISOM
- KKLayout
- FRLayou
- Circle
- Spring



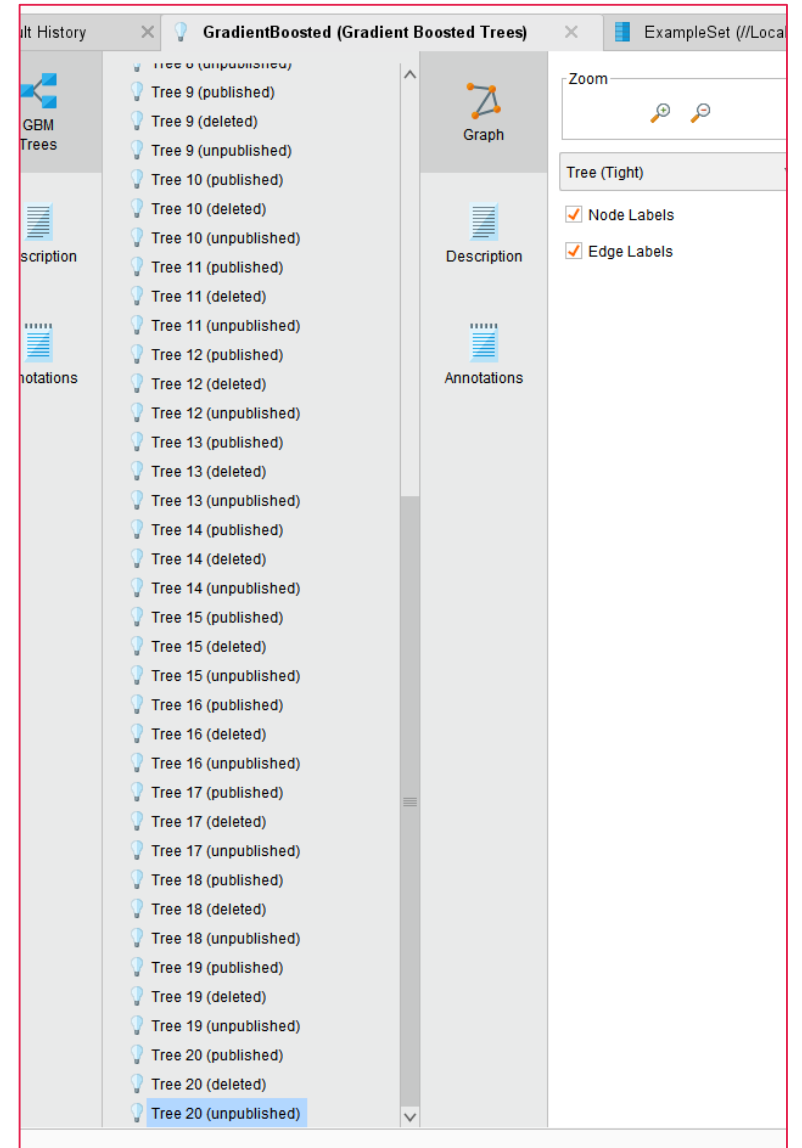
spring layout

# GRADIENT-BOOSTED DECISION TREE

Gradient-boosted decision trees are a refinement on basic decision trees because they learn from the residual errors between actual data and the predicted data, in order to ensure closer predictivity modeling, by iterating over the data. The final tree should be more sensitive to the underlying data than the earlier trees.

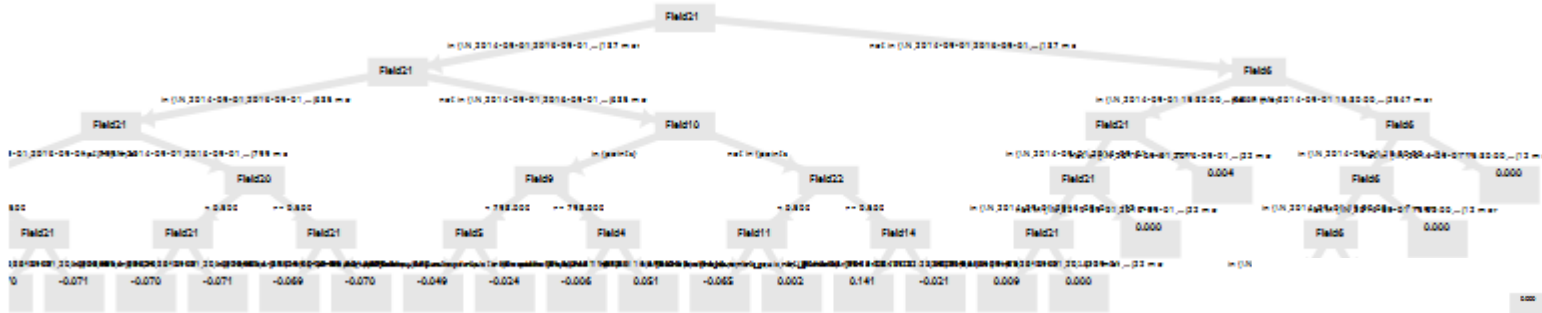
Selected parameters:

- 20 trees deep
- Maximal depth: 5
- Minimum rows: 10
- Minimum split improvement: 0
- Number of bins: 20
- Learning rate: 0.1
- Sample rate: 1
- Distribution: Auto
- Early stopping: 0
- Maximum runtime seconds: 0

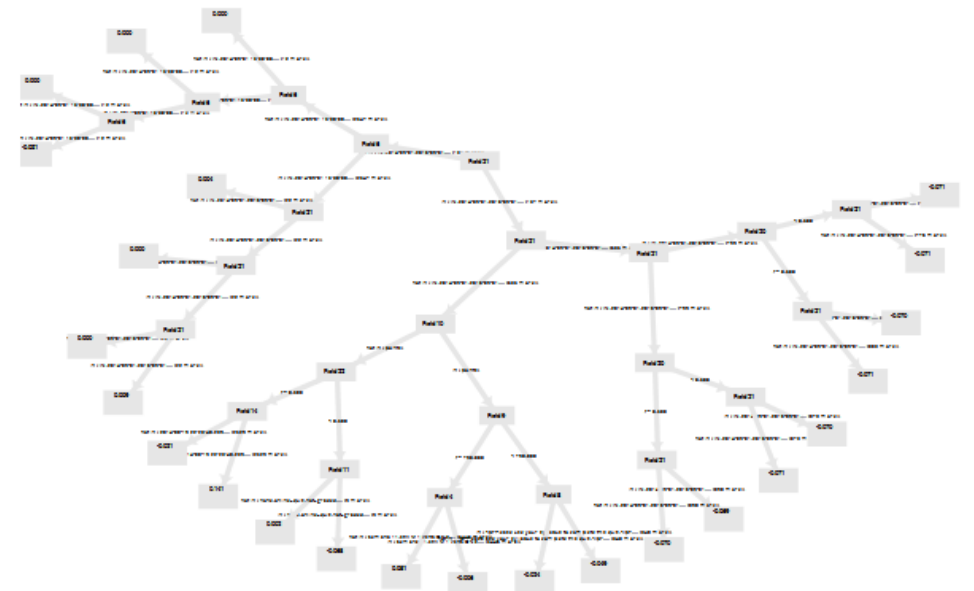


# GRADIENT-BOOSTED DECISION TREE (ZOOMED-OUT VIEWS)

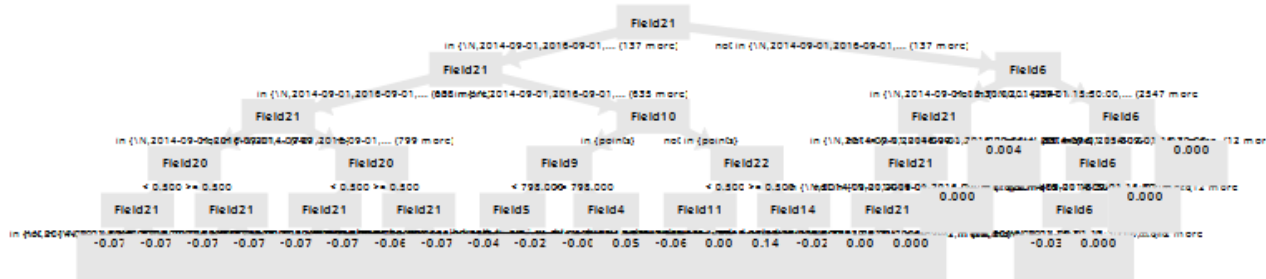
tree



radial



tree (tight)

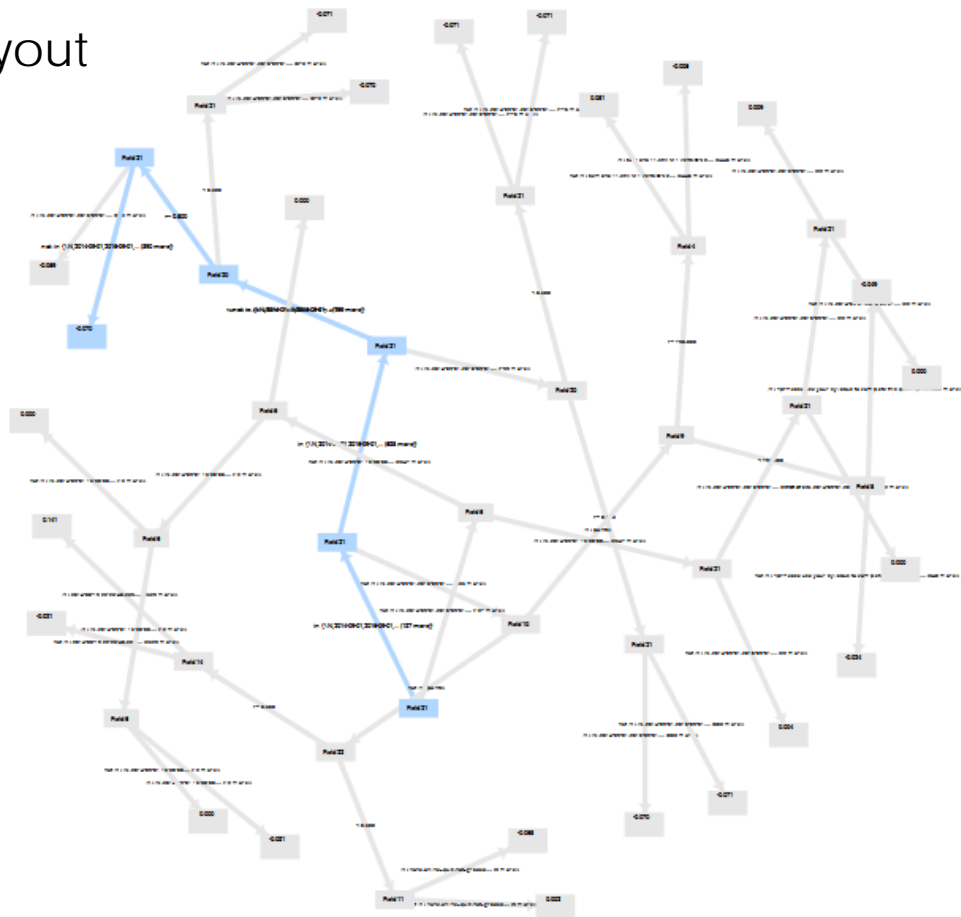




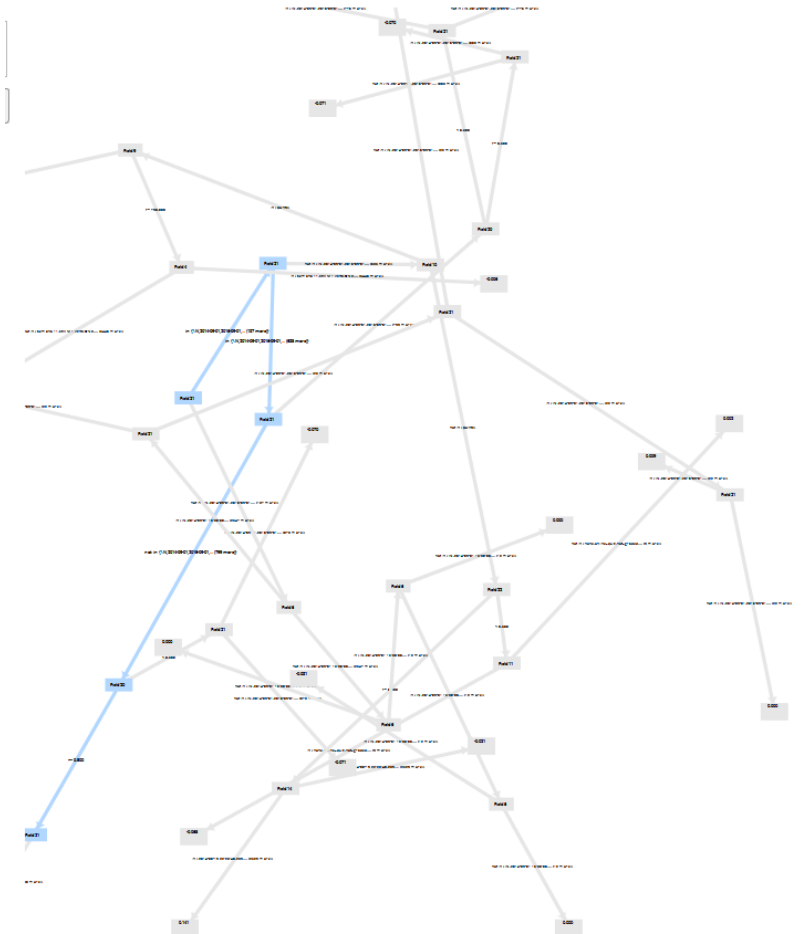
# GRADIENT-BOOSTED DECISION TREE

(ZOOMED-OUT VIEWS) (CONT.)

KKLayout

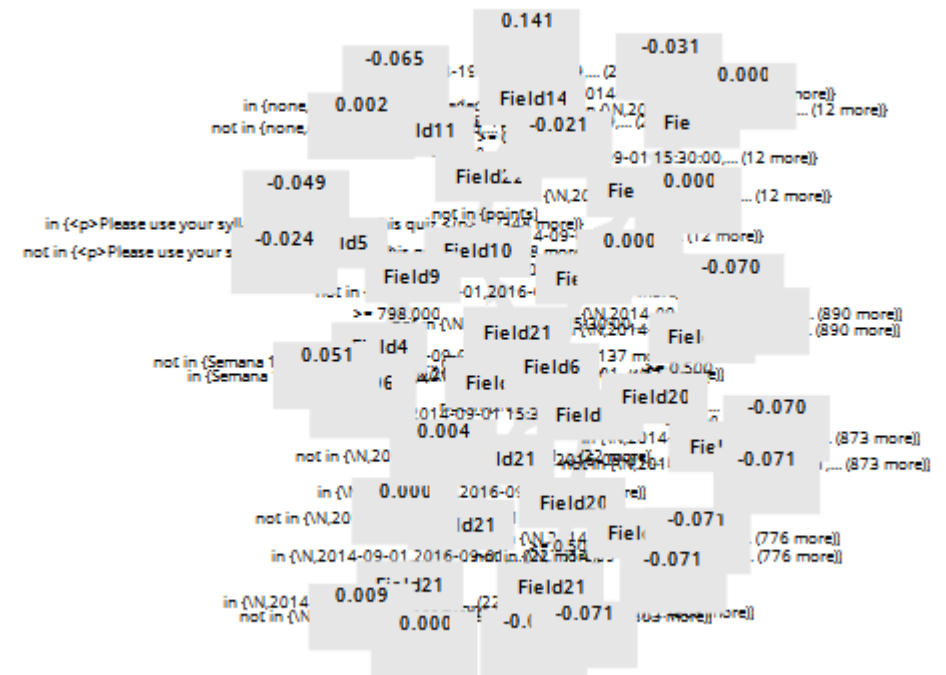
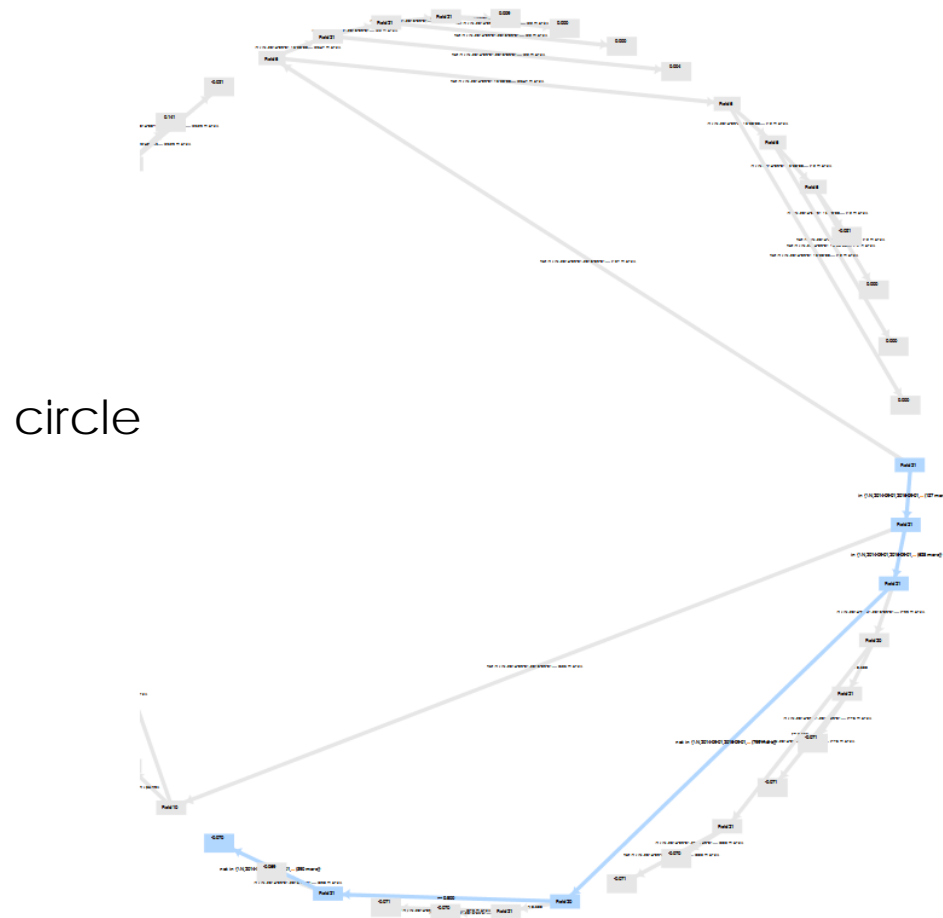


FRLayout



# GRADIENT-BOOSTED DECISION TREE

(ZOOMED-OUT VIEWS) (CONT.)





# HOW DECISION TREES MAY BE USED IN EDUCATION AND LEARNING

- Understanding frequency distributions by class for particular populations (how likely a population is to break out into particular categories, based on the data sample or a full set)
  - Predicting the class membership of a single record based on a decision tree
  - Predicting the class membership frequency of a group set based on a decision tree (with varying levels of accuracy)
- Understanding key attributes / variables / co-variates in affecting classification

# HOW DECISION TREES MAY BE USED IN EDUCATION AND LEARNING (CONT.)

- Applying informed decision analysis in understanding particular critical covariates for classification and using that information to affect decision-making and action-taking
  - Identifying a particular critical assignments that learners have to pass in order to pass a class
  - Identifying particular behaviors / features of learners that predict success in a particular learning sequence

# SOME SAMPLE ASKABLE QUESTIONS

- What are critical assignments and activities for learners to be placed in certain grade categories? Or in a binary pass / fail (P/F) classification?
  - How can instructors and GTAs prepare learners to do better for high-value assignments and activities?
  - What are ways to mitigate for gaps in preparedness?
- What are necessary features (such as do-or-die activities) for learners who select into a particular domain field vs. another? Or who end up in various percentile categories of earnings a year out after graduation? Or who end up in particular region-based relocations after graduation?
  - Out of a set of possible influencers / contributory variables, which are the most potent in predicting target class labels? (And what of this data can surprise us?)

# PRE-DATA EXPLORATION HYPOTHESIZING

- Explore the data, and in selecting the class outcomes grouping, hypothesize about possible associations that one might find in inducing a decision tree.
- Document the hypotheses.
- Revisit these once the decision trees and random forests have been created.
- Revise the hypotheses with the new information.
- It is important to capture thinking prior in order to be able to observe how the thinking changed with new data and new data visualizations.
- Post-hoc theorizing is valuable, too, but in a way that may be overly swayed by the available outcomes.

# INTERNAL VALIDITY

- Internal validity suggests the following:
  - The constructs are cleanly theorized. The concepts are accurately and precisely defined.
  - The variables that may be indicators of the constructs are clearly defined.
  - The research is clean.
  - The data collection is sufficient.
  - The predictive analytics models applied are correct to the data.
  - The data is processed cleanly.
  - The algorithms run correctly.
  - The applied logics are correct.
- Models can work cleanly with internal validity and still have poor external validity.

# EXTERNAL VALIDITY

- To be “externally valid,” a model has to achieve the following: It has to...
  - model the studied aspects of the real world accurately (usually both backwards and forwards in time, historically and predictively...and in the present...descriptively)
  - offer predictive insights about the world with a degree of accuracy
  - inform understandings
  - inform decision-making
  - be transferable or generalizable to other contexts
- In other words, a model may function wonderfully in a way that is separated from the world, but that is not particularly helpful in any applied way in the world.

# EXTERNAL VALIDITY (CONT.)

- What can confound a decision tree analysis (in terms of external validity)?
  - There may be other contributing variables to the target outcomes...that are not represented in the dataset. That relevant data may not have been collected. The available information is a limiting factor.
  - The quality of the collected data may be an issue. If the quality is suspect, the insights will not necessarily apply.
  - The theorizing may be off. The construct that the variables are supposed to represent may be poorly conceptualized and poorly defined, and the variables themselves may not be indicators of that construct.
  - The in-world phenomenon may itself be elusive to observe, elusive to measure (or to collect on). (The in-world phenomenon may not exist.)
  - A decision tree analysis approach may be the wrong type of analysis to find the relevant variables that contribute to the particular target outcomes / classifications.
  - The outcome classifications themselves may be poorly conceptualized and poorly defined.
  - And others...

# EXTERNAL VALIDITY (CONT.)

- Assuming that the dataset captures authentic and sufficient data...in a thorough way...and the decision tree was properly induced from the data...and the tree itself was properly interpreted...the next step might be to explore whether the insights from the decision tree are externally valid:
  - Do the insights from the decision tree reflect the real-world? (Do the data patterns identified inform accurately on real-world phenomena?)
- And further, will the interventions designed from the decision tree rules change ground truth in the world?
  - Or are the suggestive data patterns identified actually “anti-patterns,” solutions that look good initially but result in failure and unintended negative outcomes in the real world?



# EXTERNAL VALIDITY (CONT.)

- One other point: The world is constantly changing, so a model by definition may only be good for a time in the past, present, and future...and then has to be updated for applicability.

# 7. SUMMARY



# A FEW POINTS

- Decision trees are a method of machine learning that identifies patterns in quantitative and categorical / nominal data. The data can be heterogeneous.
- In general, the data can be large and analyzed at-scale.
- Decision trees capture macro-level patterns in the data, not more nuanced analytics (which would require other methods and some human close-reading).
- The machine-learning process involves the induction of rules from the training data (variables that inform the classification of records) that identify data patterns but also enable predictive analytics of non-training (test) data.

## A FEW POINTS (CONT.)

- Various techniques may be applied to strengthen the decision tree analysis method: data cleaning, data pre-processing, decision tree parameter-setting, and others.
- Decision trees are more effective with some types of data than others. The model is never perfect even with full data sets because of inherent noise in data.
- Machine learning is about finding patterns in data that may be invisible to the unaided human eye...and to theorizing...and other methods at getting at truth.

## A FEW POINTS (CONT.)

- Model accuracy may be ascertained in RapidMiner Studio using various validation methods. Here, a cross-validation was run with output on classification accuracy performance.
- There are methods to apply to the decision tree sequence to enhance the decision tree classification performance.
- The decision tree machine learning method may be applied to educational and online learning data (as long as the data has a target classification and attributes / variables in the right data format or accurately convertible to the right data format).

# CONCLUSION AND CONTACT

- Dr. Shalin Hai-Jew, Instructional Designer
  - iTAC, Kansas State University
  - 212 Hale Library
  - 785-532-5262
  - [shalin@k-state.edu](mailto:shalin@k-state.edu)
- **Notes:** The presenter is an early user of machine-learning-based decision trees. The presenter has no tie to the software maker of RapidMiner Studio.
- **Thanks!** Thanks to the event organizers for accepting this presentation.

# NOTES ABOUT RAPIDMINER'S SUPPORT FOR EDUCATION

- **Free Educational Licenses:** RapidMiner offers free licenses for academic usage to “students, professors and researchers” through [their educational license program](#). The RapidMiner Academia team offers support for users of this license. These generally run on three-year cycles.
- **RapidMiner Community Features:**
  - There are forums for users of RapidMiner Studio.
  - The company offers commercial trainings and certification for “Certified RapidMiner Analyst” standing.
  - In the tool, there is a “wisdom of crowds” aspect that provides “advice” based on how users solve similar challenges that the user has arrived at.