Data Mining

SPSS Clementine 12.0

7. C5.0 Algorithm

Spring 2010

Instructor: Dr. Masoud Yaghini

Outline

- Overview
- Deriving a New Field
- Building a Model
- C5.0 Node
- Browsing the Model
- Using an Analysis Node
- References

Overview

Overview

- The C5.0 node builds either a decision tree or a rule set.
- The model works by splitting the sample based on the field that provides the maximum information gain at each level.
- The target field must be categorical. Multiple splits into more than two subgroups are allowed.

Drug Treatments

- For this section, imagine that you are a medical researcher compiling data for a study.
- You have collected data about a set of patients, all of whom suffered from the same illness.
- During their course of treatment, each patient responded to one of five medications.
- Part of your job is to use data mining to find out which drug might be appropriate for a future patient with the same illness.
- This example uses the data file named DRUG1n.

Drug Treatments

- The data fields used in the demo are:

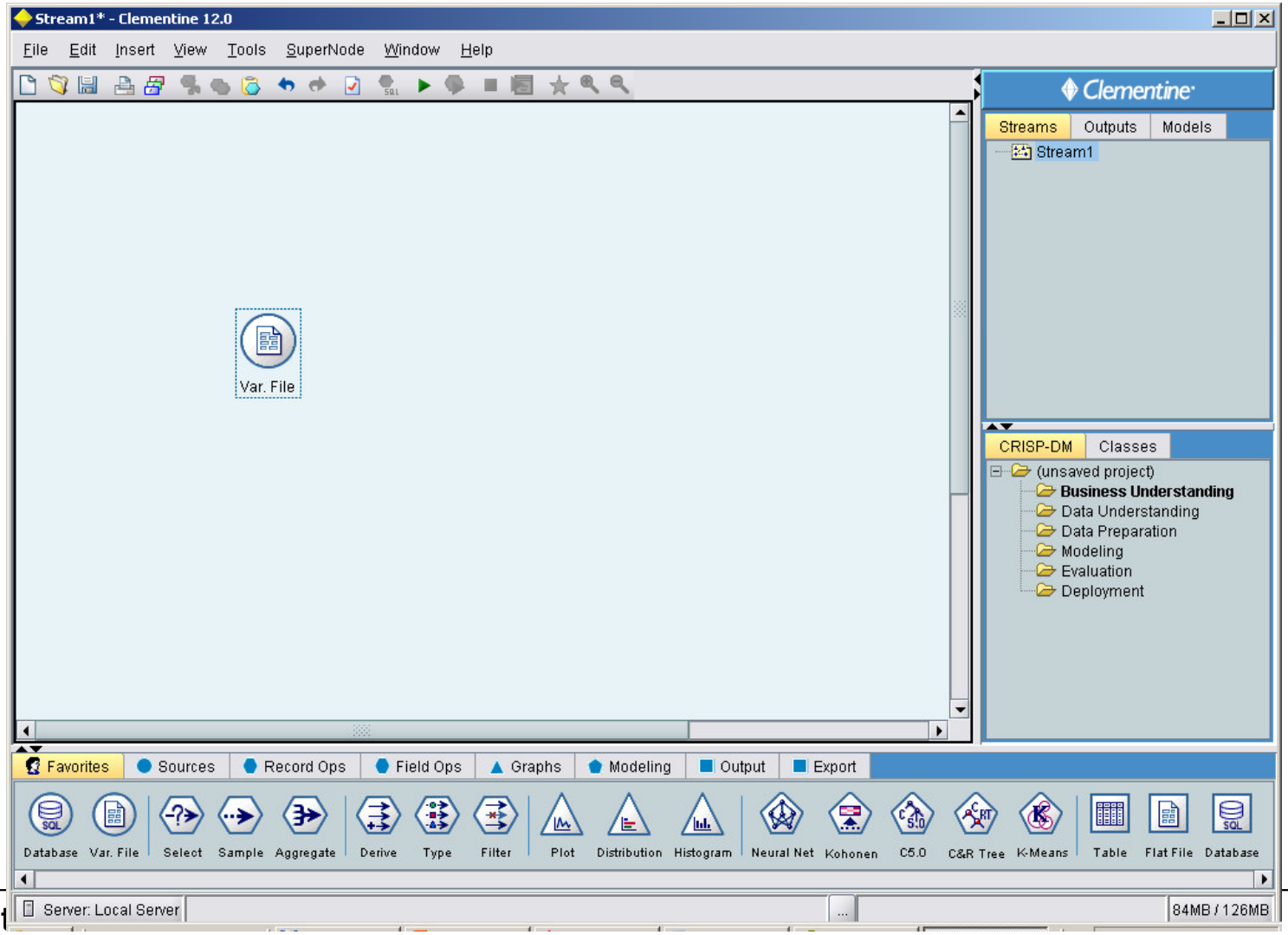
Data field	Description
Age	(Number)
Sex	*M* or *F*
BP	Blood pressure: *HIGH*, *NORMAL*, or *LOW*
Cholesterol	Blood cholesterol: *NORMAL* or *HIGH*
Na	Blood sodium concentration
K	Blood potassium concentration
Drug	Prescription drug to which a patient responded

Building Data Stream

- You can read in delimited text data using a Variable File node.
- You can add a Variable File node from Sources tab
- Next, double-click the newly placed node to open its dialog box.

Reading in Text Data

- Adding a Variable File node



Reading in Text Data

- Variable File dialog box

Var. File

Refresh

C:\Program Files\SPSSInc\Clementine12.0\Demos\DRUG1n

File: C:\Program Files\SPSSInc\Clementine12.0\Demos\DRUG1n

```
Age,Sex,BP,Cholesterol,Na,K,Drug
23,F,HIGH,HIGH,0.792535,0.031258,drugY
47,M,LOW,HIGH,0.739309,0.056468,drugC
47,M,LOW,HIGH,0.697269,0.068944,drugC
```

Read field names from file | Specify number of fields 1

Skip header characters: 0 | EOL comment characters:

Strip lead and trail spaces: None Left Right Both

Invalid characters: Discard Replace with

Encoding: Stream default | Decimal symbol: Stream default

Delimiters

Space Comma Tab

Newline Other

Non-printing characters

Allow multiple blank delimiters

Lines to scan for type: 50

Quotes

Single quotes: Discard

Double quotes: Discard

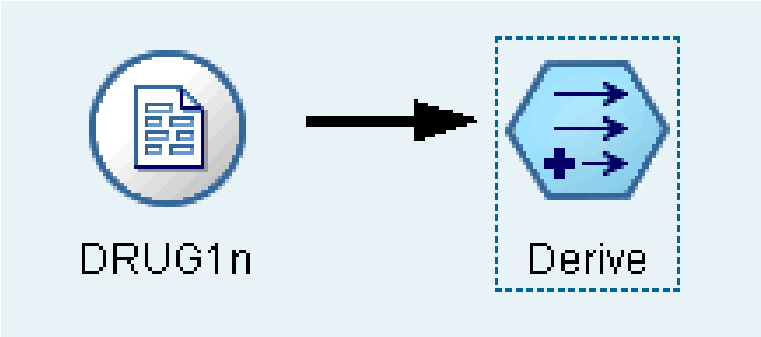
File Data Filter Types Annotations

OK Cancel Apply Reset

Deriving a New Field

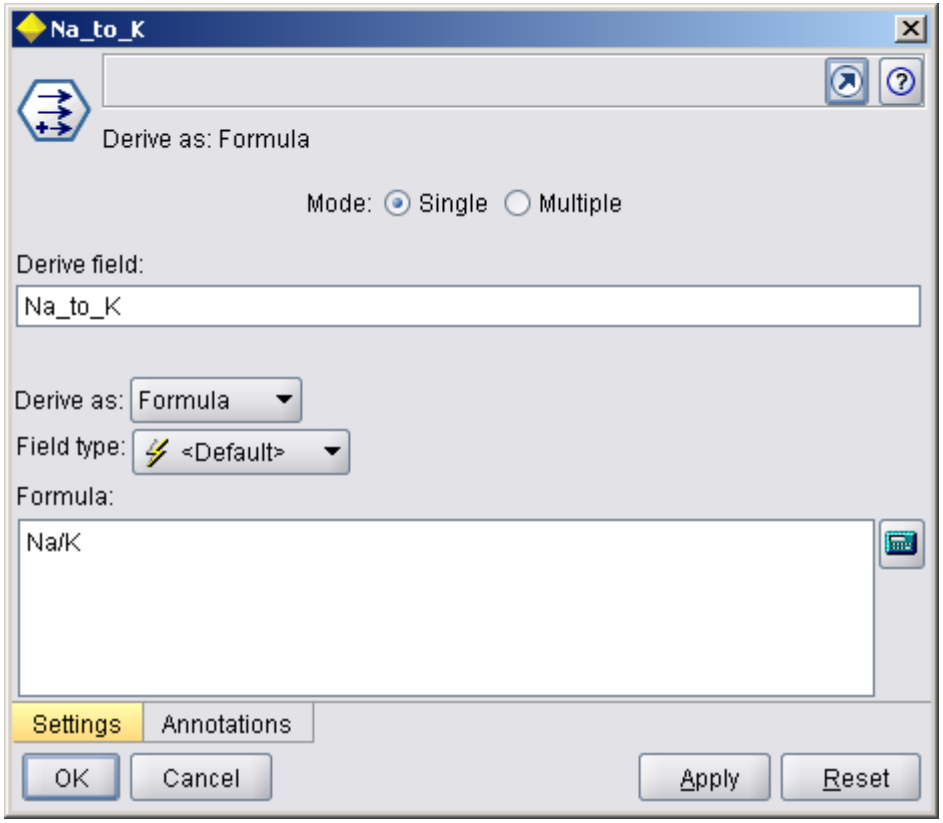
Deriving a New Field

- Since the ratio of sodium to potassium seems to predict when to use drug *Y,* you can derive a field that contains the value of this ratio for each record.
- This field might be useful later when you build a model to predict when to use each of the five drugs.
- Add a Derive node into the stream, then double-click the node to edit it.



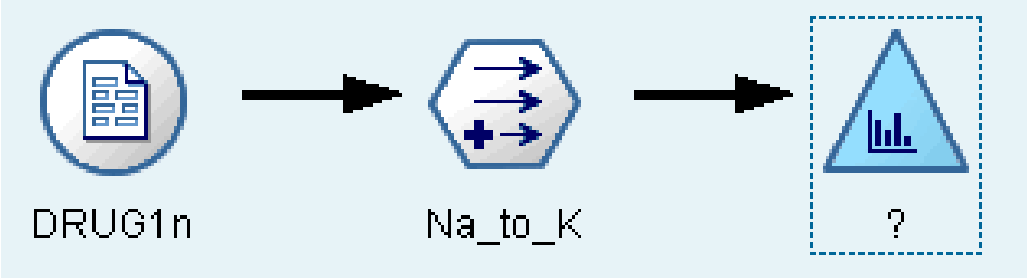
Deriving a New Field

- Name the new field *Na_to_K*.
- Since you obtain the new field by dividing the sodium value by the potassium value, enter Na/K for the formula.



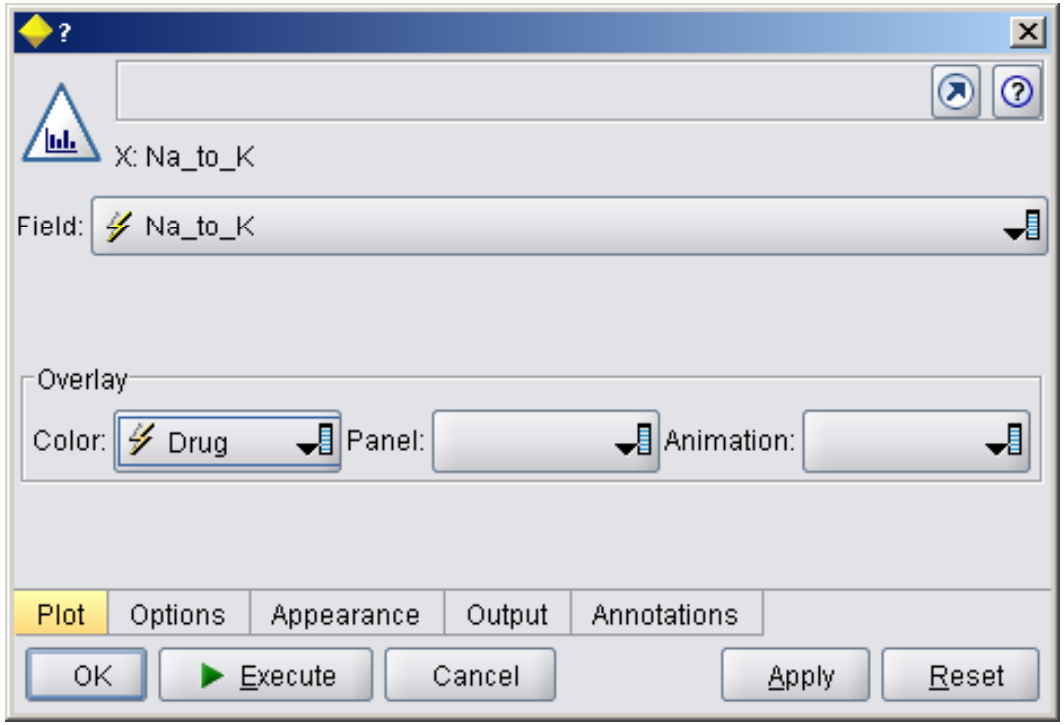
Deriving a New Field

- You can check the distribution of your new field by attaching a Histogram node to the Derive node.



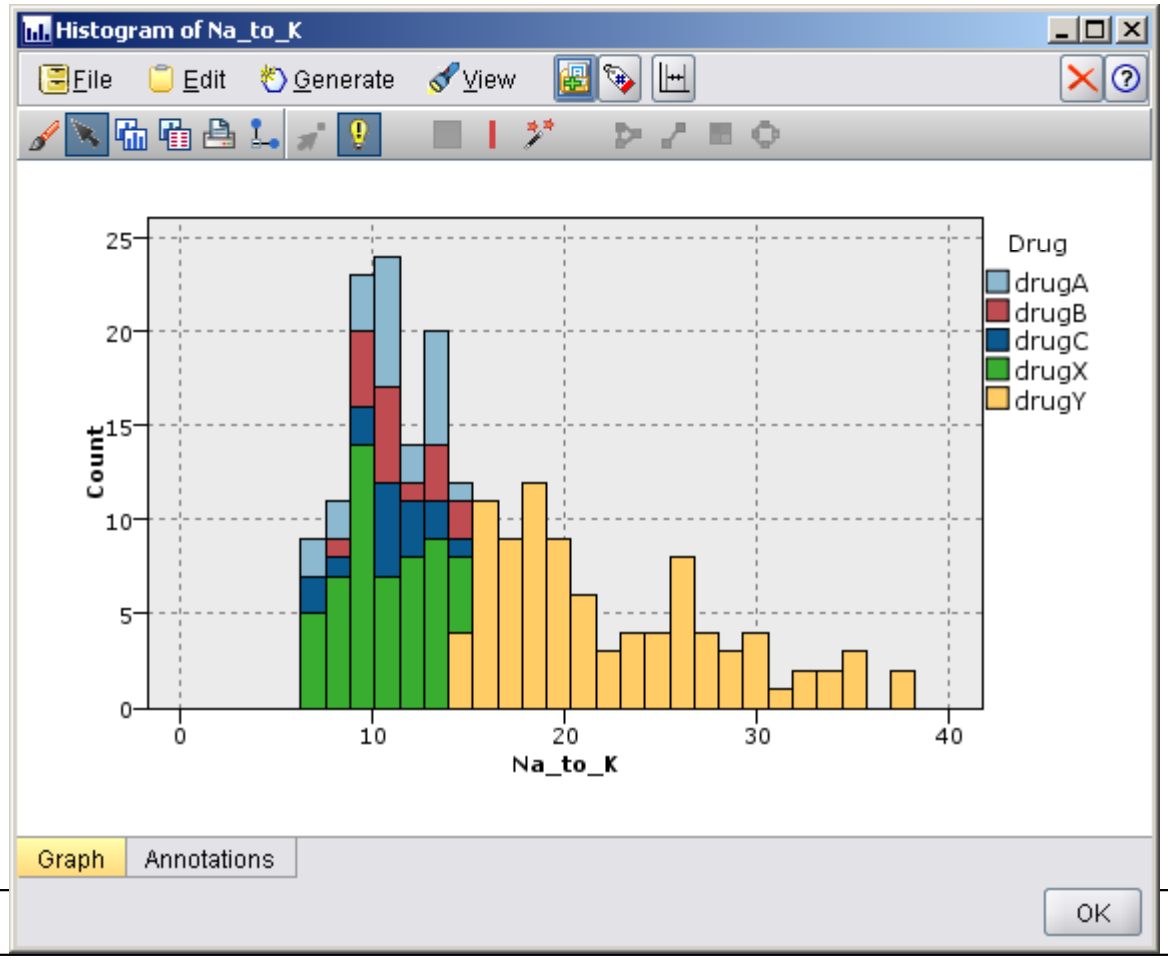
Deriving a New Field

- In the Histogram node dialog box, specify *Na_to_K* as the field to be plotted and *Drug* as the overlay field.



Deriving a New Field

- Execute the stream. Based on the display, you can conclude that when the *Na_to_K* value is about 15 or above, drug *Y* is the drug of choice.



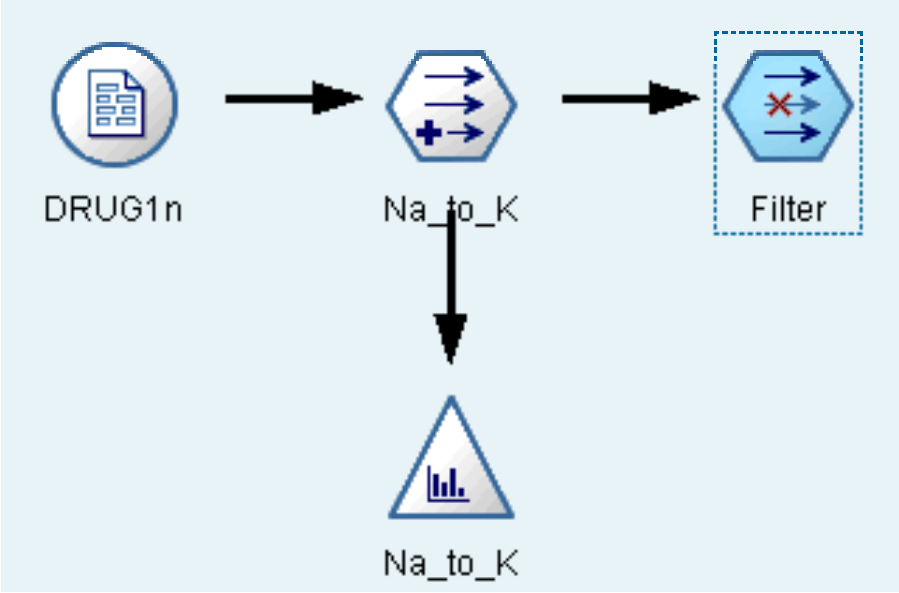
Building a Model

Building a Model

- By exploring and manipulating the data, you have been able to form some hypotheses.
- The ratio of sodium to potassium in the blood seems to affect the choice of drug, as does blood pressure.
- But you cannot fully explain all of the relationships yet.
- This is where modeling will likely provide some answers.
- In this case, you will use try to fit the data using a rule-building model, C5.0.

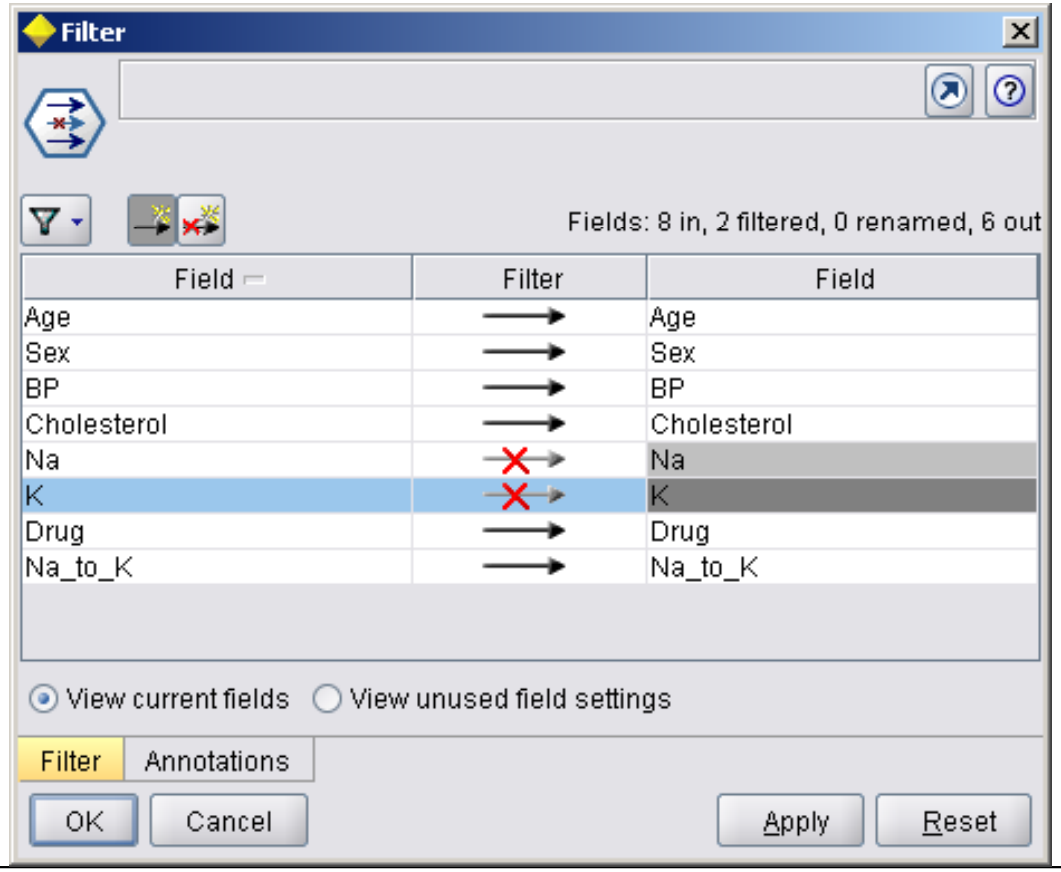
Building a Model

- Since you are using a derived field, *Na_to_K*, you can filter out the original fields, Na and K, so that they are not used twice in the modeling algorithm.
- You can do this using a Filter node.



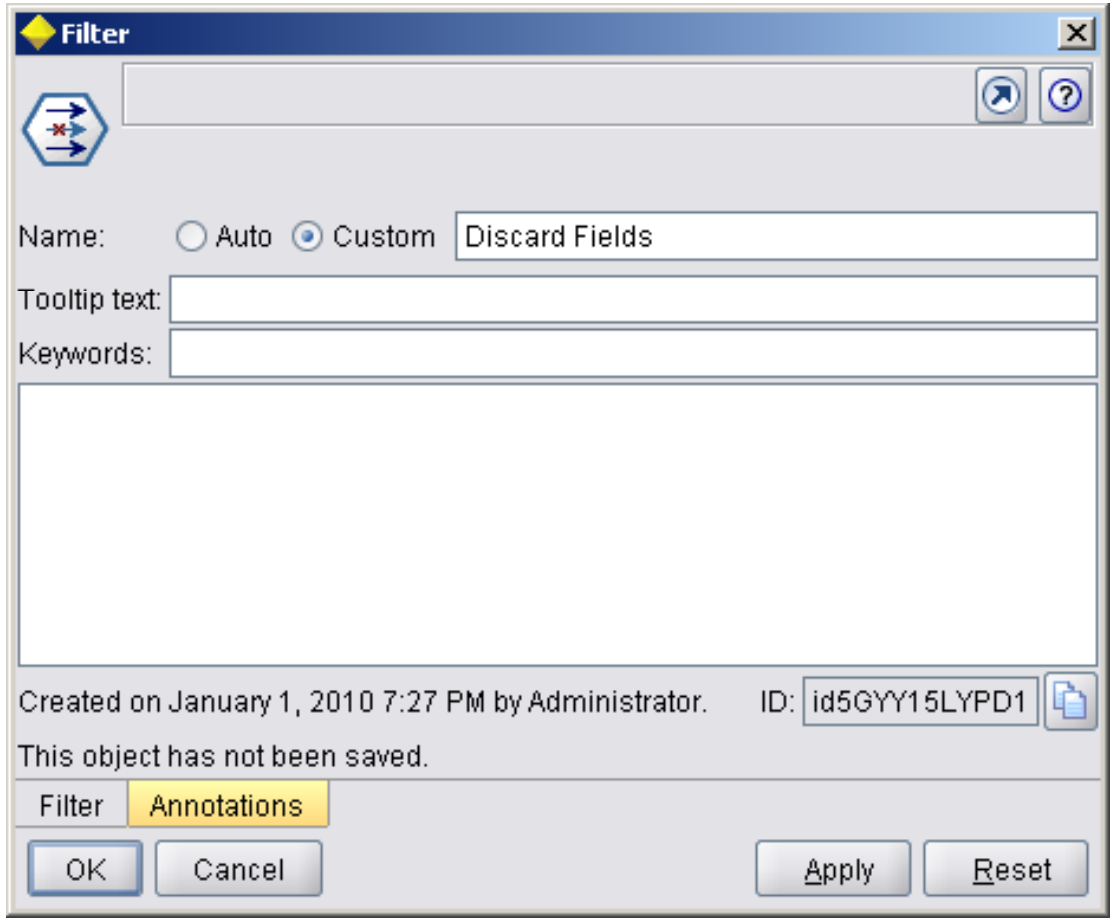
Building a Model

- On the Filter tab, click the arrows next to *Na* and *K*.
- Red Xs appear over the arrows to indicate that the fields are now filtered out.



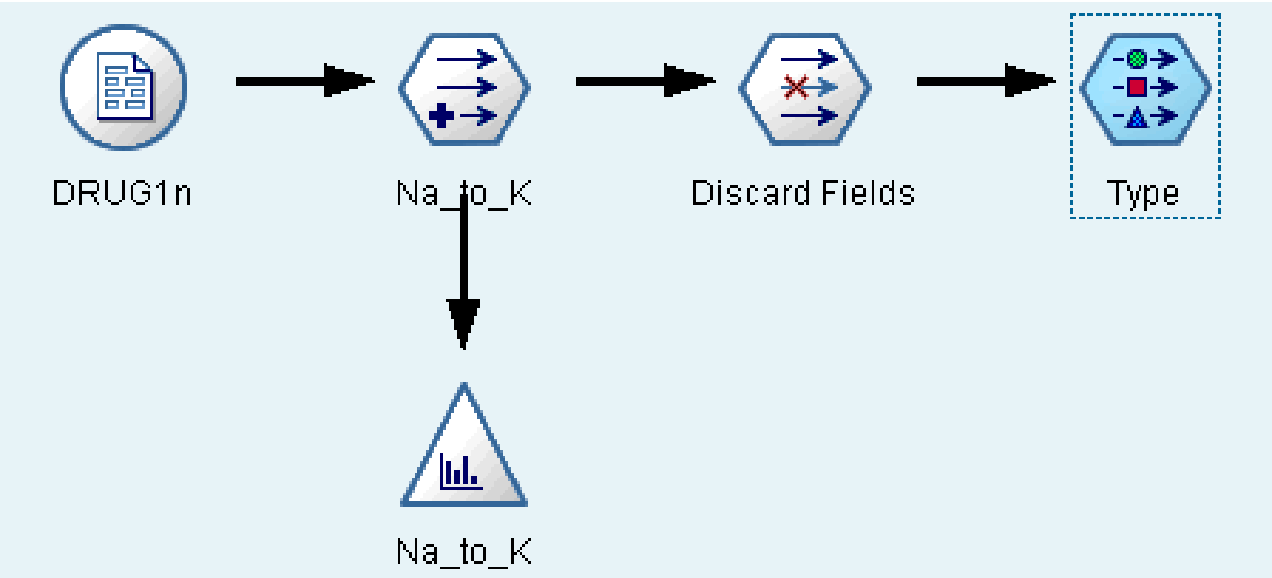
Building a Model

- Change the default name of Filter node to Discard Fields



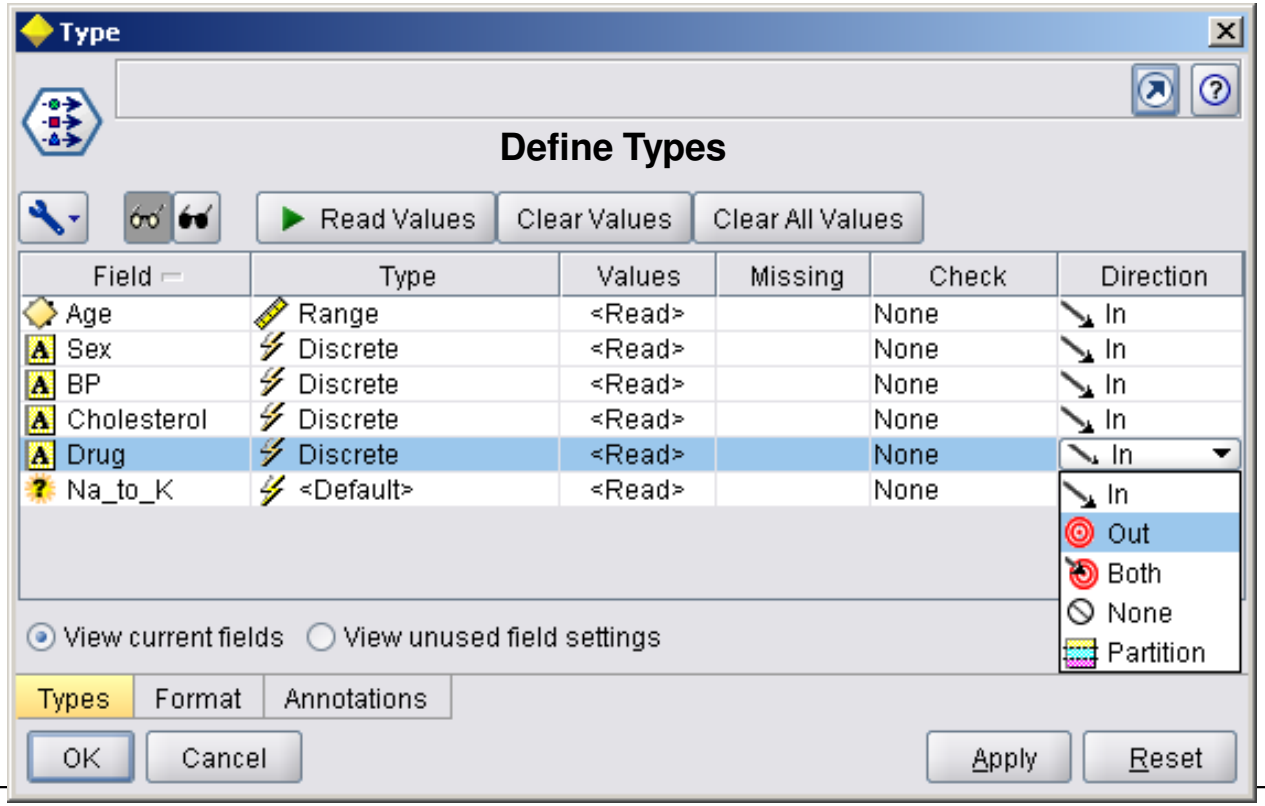
Building a Model

- Attach a Type node connected to the Filter node.
- The Type node allows you to indicate the types of fields that you are using and how they are used to predict the outcomes.



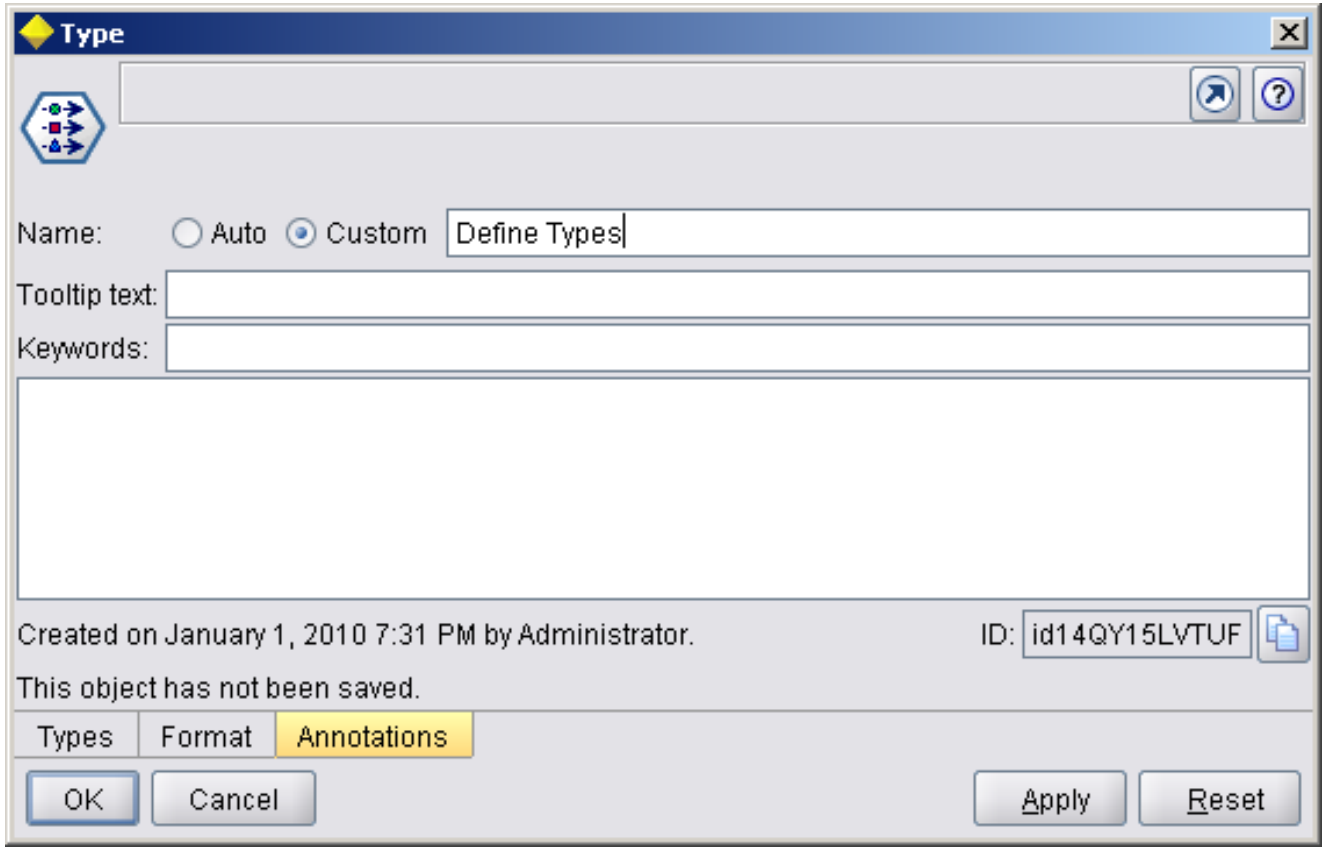
Building a Model

- On the Types tab, set the direction for the Drug field to Out, indicating that Drug is the field you want to predict.
- Leave the direction for the other fields set to In so they will be used as predictors.



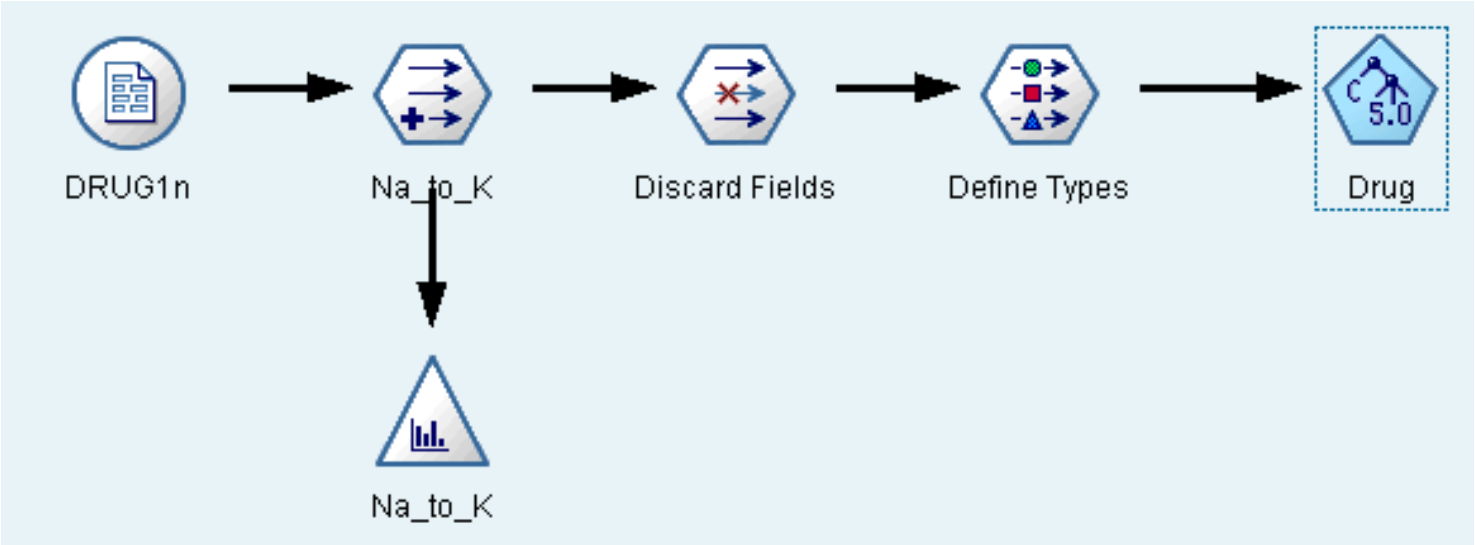
Building a Model

- Change the default name of Type node to Define Types



Building a Model

- To estimate the model, place a C5.0 node in the workspace and attach it to the end of the stream as shown.
- Then click the green Execute button to execute the stream.



C5.0 Node

C5.0 Node

- This node uses the C5.0 algorithm to build either a **decision tree** or a **rule set**.
- A C5.0 model works by splitting the sample based on the field that provides the maximum **information gain.**
- Each subsample defined by the first split is then split again, usually based on a different field, and the process repeats until the subsamples cannot be split any further.
- Finally, the lowest-level splits are reexamined, and those that do not contribute significantly to the value of the model are removed or **pruned.**

C5.0 Node

- The C5.0 node can predict only a categorical target (Set or Ordered Set fields).
- **Requirements**
  - To train a C5.0 model, there must be one categorical (i.e., Set or Ordered Set) Out field, and one or more In fields of any type.
  - Fields set to Both or None are ignored.
  - Fields used in the model must have their types fully instantiated (their type can not be string or type less).

C5.0 Node

- **Strengths**
  - C5.0 models are quite robust in the presence of problems such as missing data and large numbers of input fields.
  - They usually do not require long training times to estimate.
  - In addition, C5.0 models tend to be easier to understand than some other model types, since the rules derived from the model have a very straightforward interpretation.
  - C5.0 also offers the powerful **boosting method** to increase accuracy of classification.

C5.0 Node

- **C5.0 Node Model Options**

Drug

Model name: Auto Custom

Use partitioned data

Output type: Decision tree Rule set

Group symbolics

Use boosting Number of trials: 10

Cross-validate Number of folds: 10

Mode: Simple Expert

Favor: Accuracy Generality

Expected noise (%): 0

Fields Model Costs Analyze Annotations

OK Execute Cancel Apply Reset

C5.0 Node Model Options

- Model name:
  - Specify the name of the model to be produced.
  - Auto
    - With this option selected, the model name will be generated automatically, based on the target field name(s). This is the default.
  - Custom
    - Select this option to specify your own name for the model nugget that will be created by this node.
- Output type
  - Specify here whether you want the resulting model nugget to be a Decision tree or a Rule set.

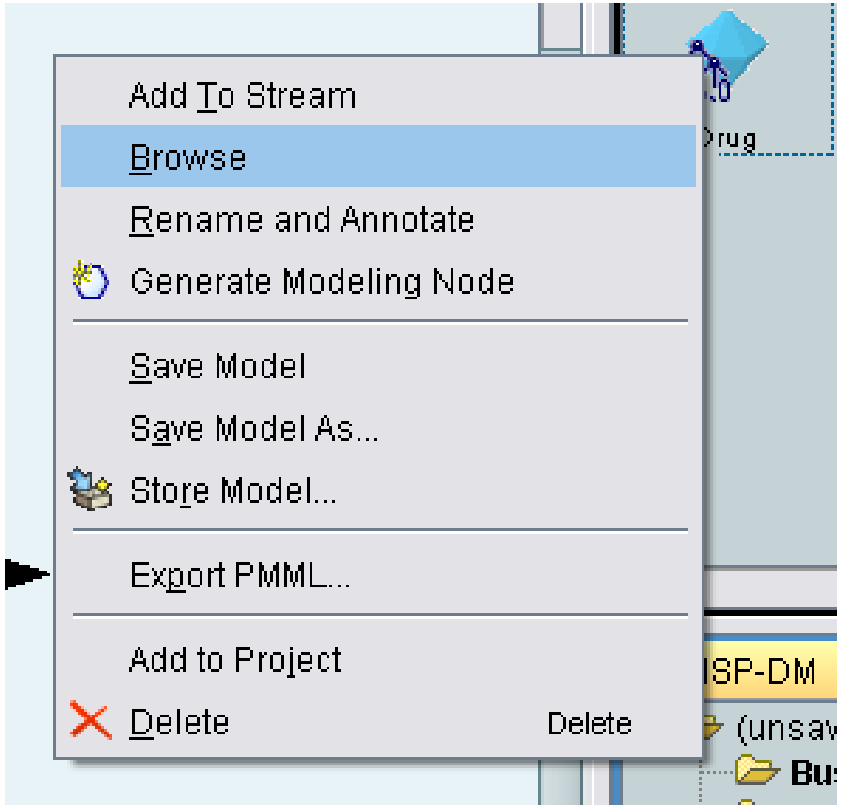
C5.0 Node Model Options

- Group symbolics
  - If this option is selected, C5.0 will attempt to combine symbolic values that have similar patterns with respect to the output field.
  - If this option is not selected, C5.0 will create a child node for every value of the symbolic field used to split the parent node.
  - For example,
    - if C5.0 splits on a COLOR field (with values RED, GREEN, and BLUE), it will create a three-way split by default.
    - However, if this option is selected, and the records where COLOR = RED are very similar to records where COLOR = BLUE, it will create a two-way split, with the GREENs in one group and the BLUEs and REDs together in the other.

Browsing the Model

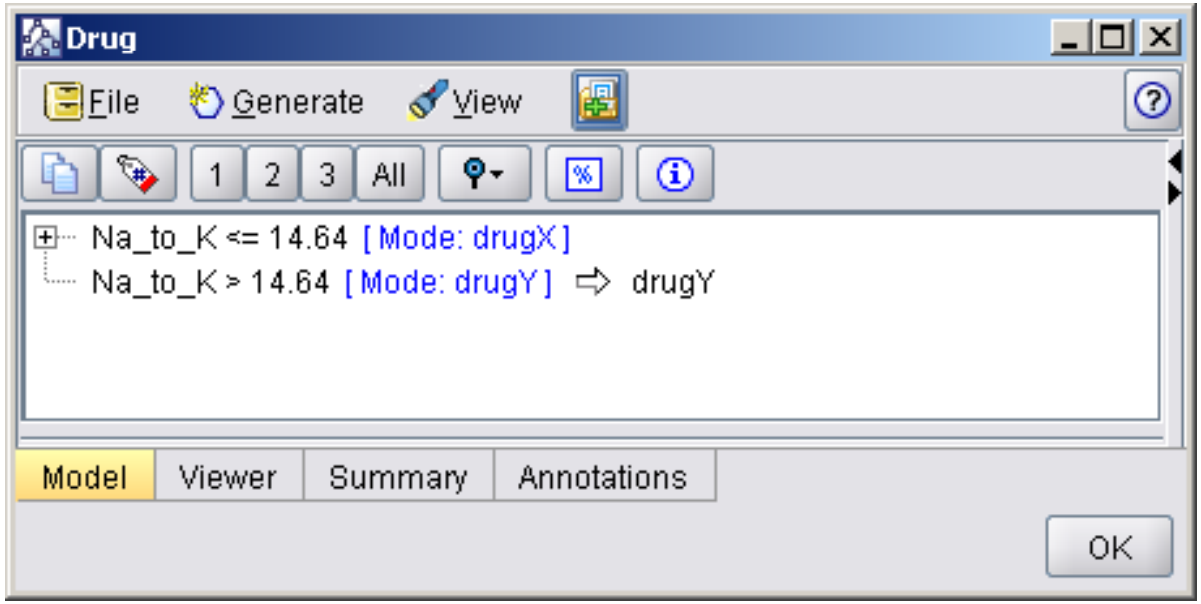
Browsing the Model

- When the C5.0 node is executed, the generated model node is added to the Models tab. To browse the model, right-click the icon and choose Browse from the context menu.



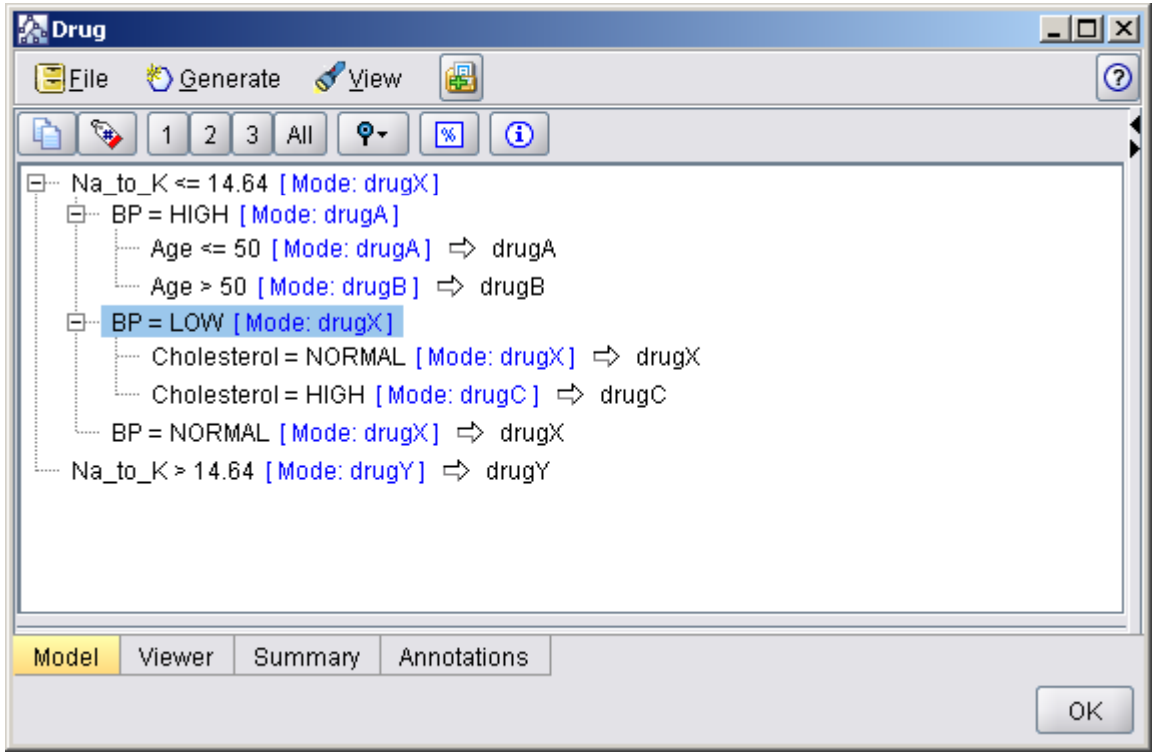
Browsing the Model

- The Rule browser displays the set of rules generated by the C5.0 node in a decision tree format.



Browsing the Model

- Initially, the tree is collapsed. To expand it, click the All button to show all levels.

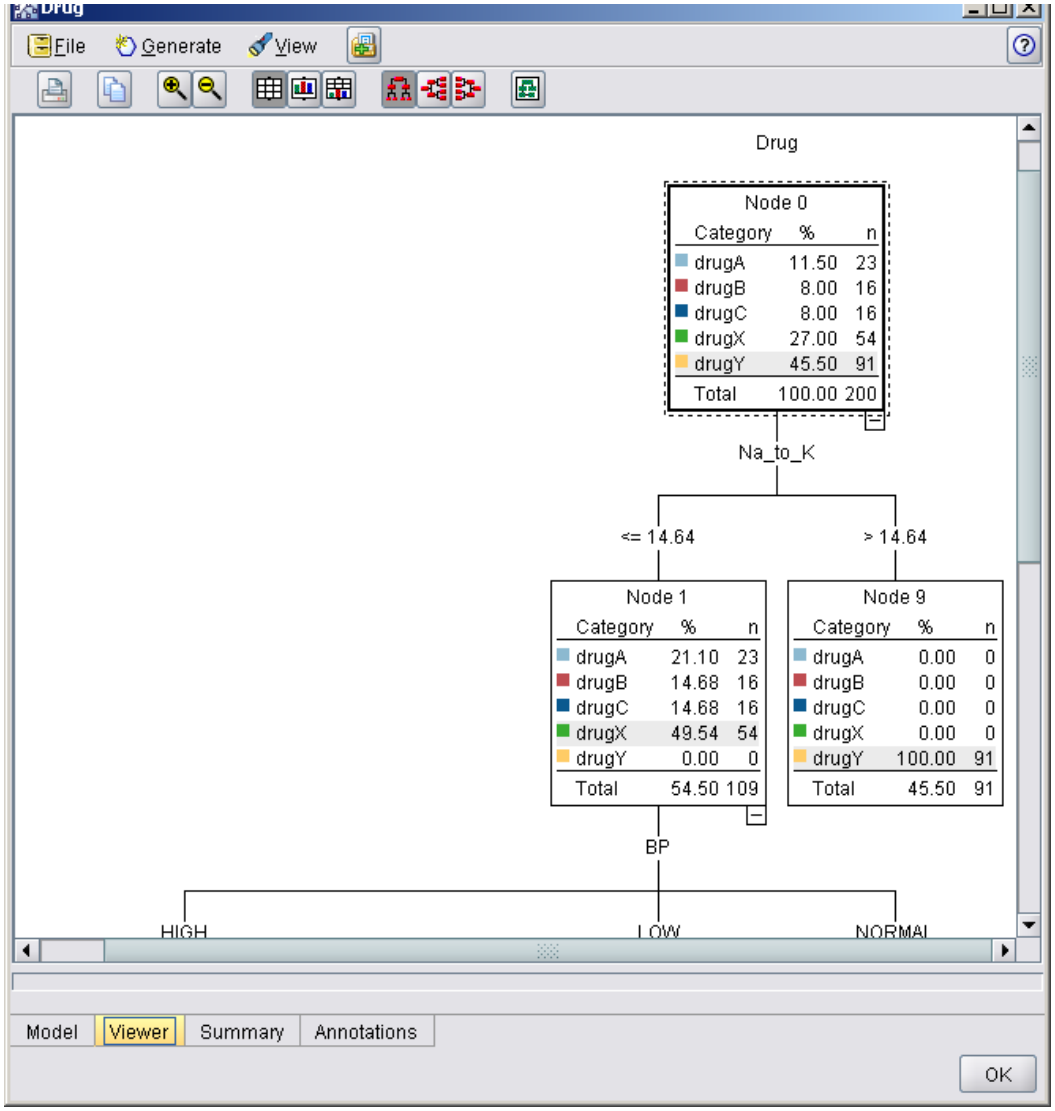


Browsing the Model

- Now you can see the missing pieces of the puzzle. For people with an *Na-to-K* ratio less than 14.642 and high blood pressure, age determines the choice of drug.
- For people with low blood pressure, cholesterol level seems to be the best predictor.

Browsing the Model

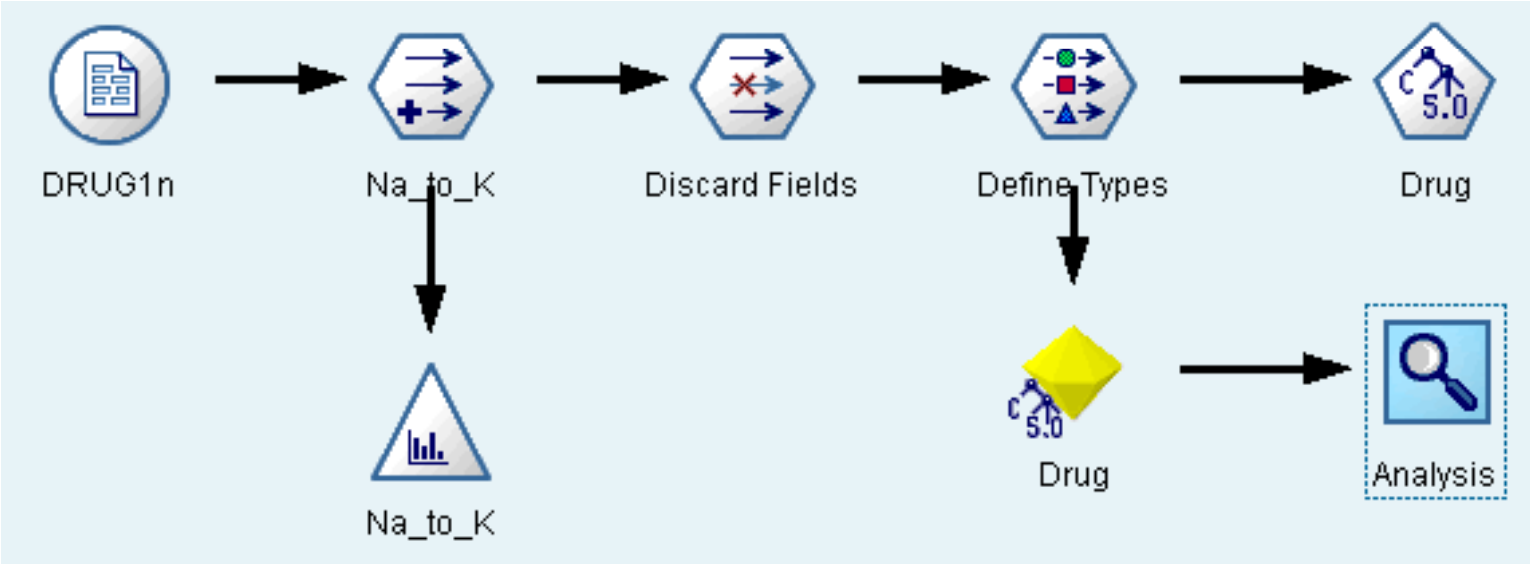
- The same decision tree can be viewed in a more sophisticated graphical format by clicking the Viewer tab.
- Here, you can see more easily the number of cases for each blood pressure category, as well as the percentage of cases.



Using an Analysis Node

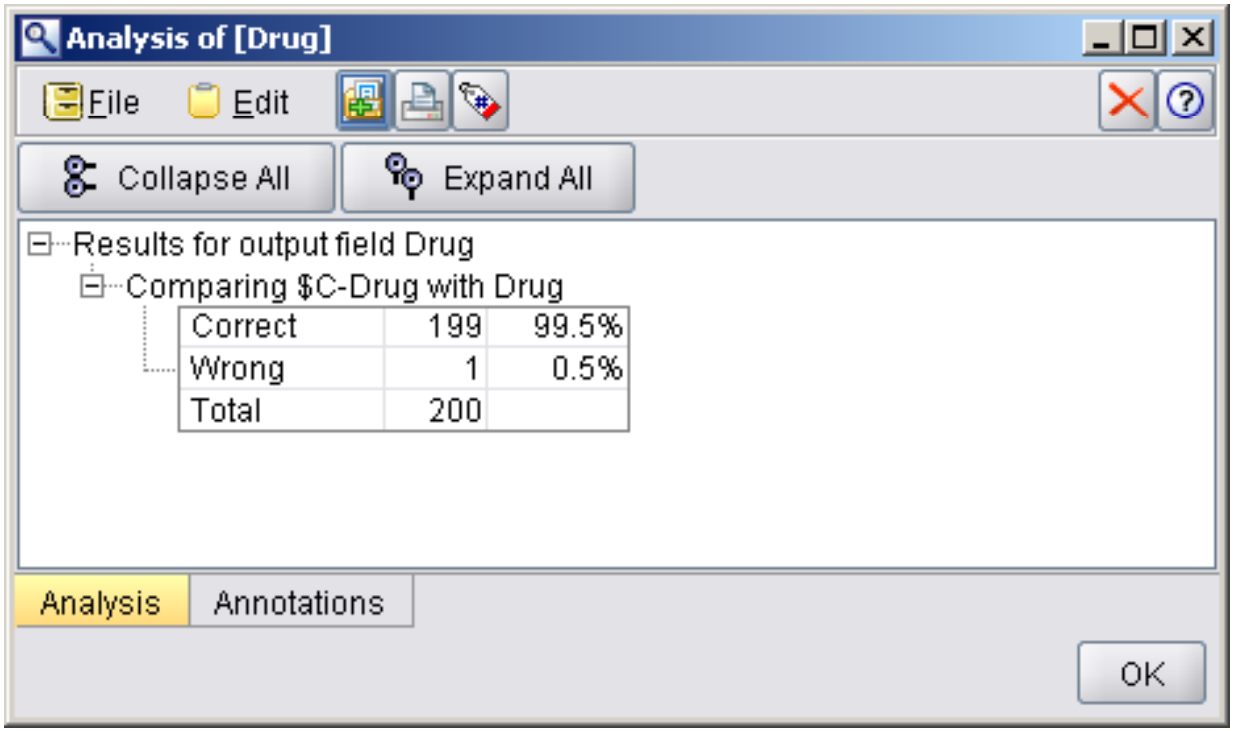
Using an Analysis Node

- You can assess the accuracy of the model using an Analysis node.
- First, attach the C5.0 model to the stream, and then attach an Analysis node.



Using an Analysis Node

- The Analysis node output shows that with this artificial dataset, the model correctly predicted the choice of drug for almost every record in the dataset.



Using an Analysis Node

- With a real dataset you are unlikely to see 100% accuracy, but you can use the Analysis node to help determine whether the model is acceptably accurate for your particular application.

References

References

- **Clementine® 12.0 Clementine Applications Guide,** 2007. (Chapter 8)
- **Clementine® 12.0 Clementine Modeling Nodes,** 2007. (Chapter 6)

The end