

Getting Started With Machine Learning

Minitab and SPM
presented by ADDITIVE

Sample datas available call +49-6172-5905-30 to
discuss your application

Machine Learning

What is machine learning

A machine learning algorithm “teaches” a computer to recognize patterns using available data.

Data is usually split into a training set and a test set:

- ▶ Training (or learn) data is used to create the model.
- ▶ Test data is used to assess model performance.

Popular machine learning techniques include tools you are likely already familiar with such as regression and logistic regression. Tree-based techniques, such as CART and TreeNet, are also frequently used for machine learning.

When to use machine learning

Machine learning techniques can be supervised or unsupervised:

- ▶ Use supervised machine learning techniques when you have a response, or target, variable (Y) and multiple predictors (X's). Supervised machine learning tools include Regression, Binary Logistic Regression, CART, Random Forests, TreeNet, and MARS.
- ▶ Use unsupervised machine learning techniques when no response exists and you want to find groupings in your data. Unsupervised machine learning tools include Principal Components Analysis and Cluster Analysis.

Why use machine learning

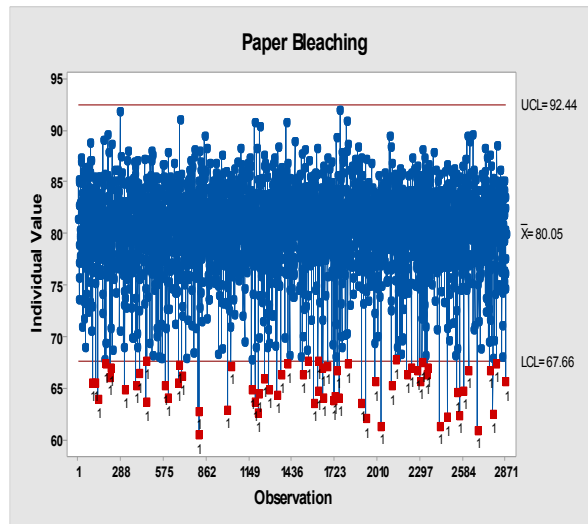
Use machine learning techniques to answer questions such as:

- ▶ What is the root cause of process defects?
- ▶ What factors contribute to excess variation in my process?
- ▶ What is the predicted cost of manufacturing a custom order?
- ▶ What conditions indicate machine degradation?

Sample: Pulp Bleaching Process

Problem

A paper manufacturer needs to use current process data to determine which factors are contributing to excessive variation in a pulp bleaching process. The process is very non-stable, creating an unacceptable defect rate.



Data collection

Process sensors collect data in real time. The results are saved in a CSV file for offline analysis.

Tools

- ▶ Graphs
- ▶ CART

Data set

PulpBrightness.csv

Variable	Description
Production Rate	Predictor
Discharge pH	Predictor
Caustic	Predictor
MgSO4	Predictor
H2O2	Predictor
O2	Predictor
Conductivity	Predictor
Unbleached	Predictor
Brightness	Continuous Response (Target)
Low	Categorical Response (Target) – Indicates when Brightness falls below the lower control limit

Importing data

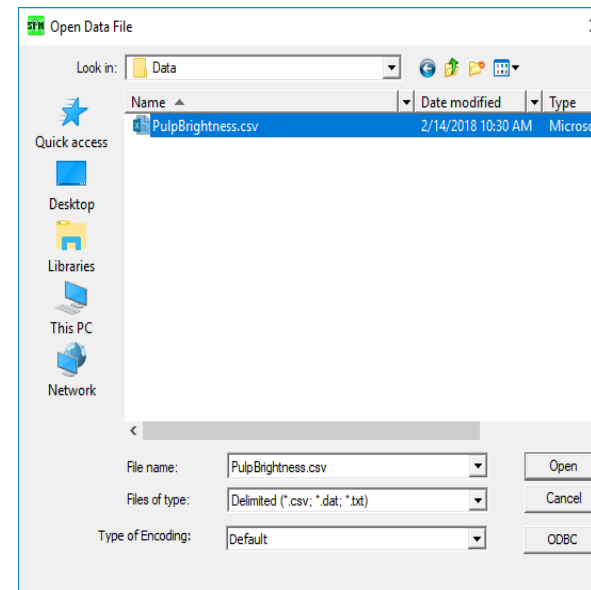
Importing Data

You can import data from many different file types or connect directly to your database via ODBC. To use ODBC, you may need to install ODBC drivers on your system.

The data for this example is in a CSV file.

Open Data File

1. Choose **File > Open > Data File**.
2. From **Files of type**, choose **Delimited (*.csv; *.dat; *.txt)**.
3. Select PulpBrightness.csv as shown below.



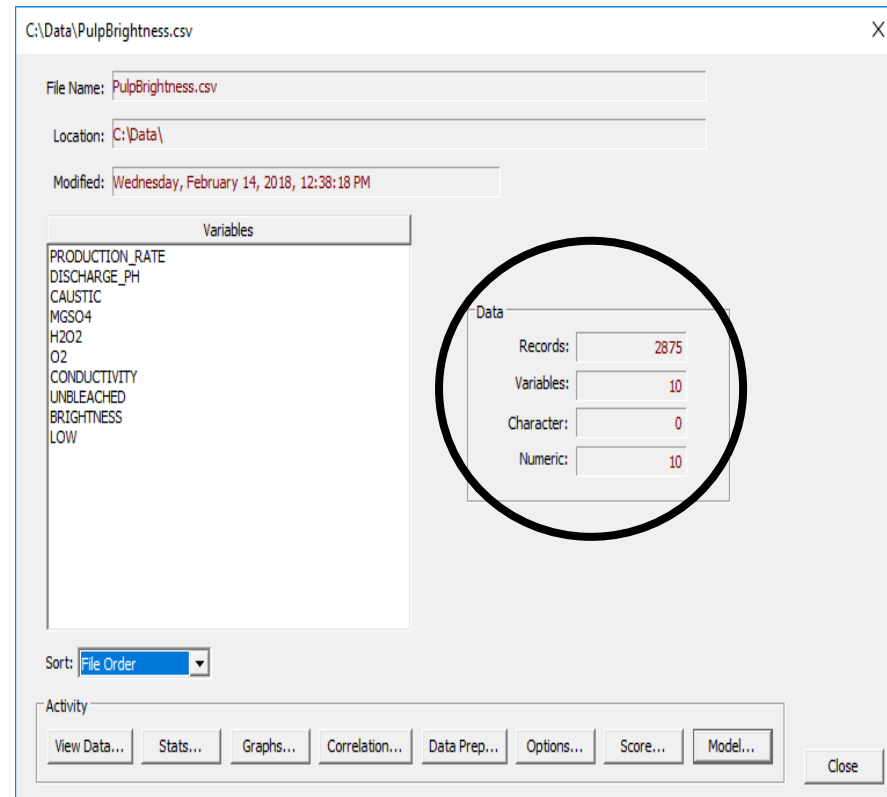
4. Click **Open**.

Importing data

By default, you will see an alphabetical list of variables in the data source. To put these variables in the order they appeared in the original file, choose **File Order** from the **Sort** drop-down.

Note that this data set contains 2875 rows and 10 variables. All variables are numeric in this data set because the contents of each variable, or column, is a number. Alternatively, a character variable is a variable that contains at least one cell with characters or text.

You can view your data, summary statistics, and graphs. For this example, we will begin by viewing some simple graphs.



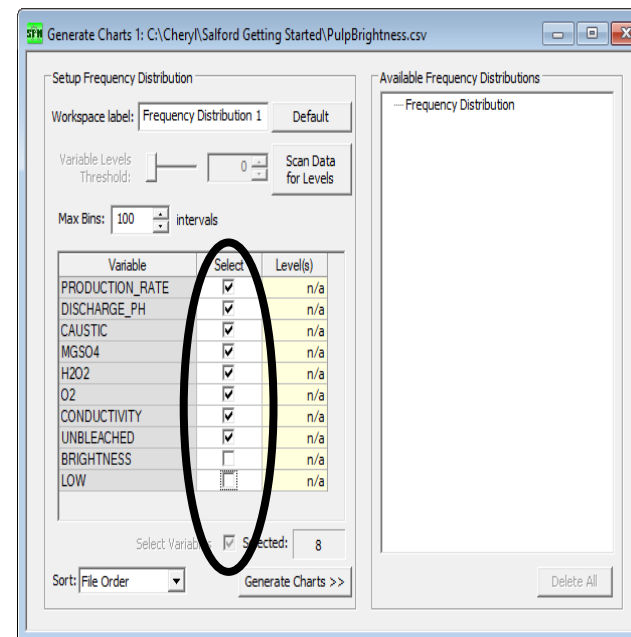
Visualizing data

Graphs are an important preliminary step in data analysis. They enable you to examine the data and to identify patterns, relationships, and potential problems.

Because a relatively small number of predictor variables exist, select all of them to view in a graph. For now, we will focus on only the predictor variables, not the target, or response variables.

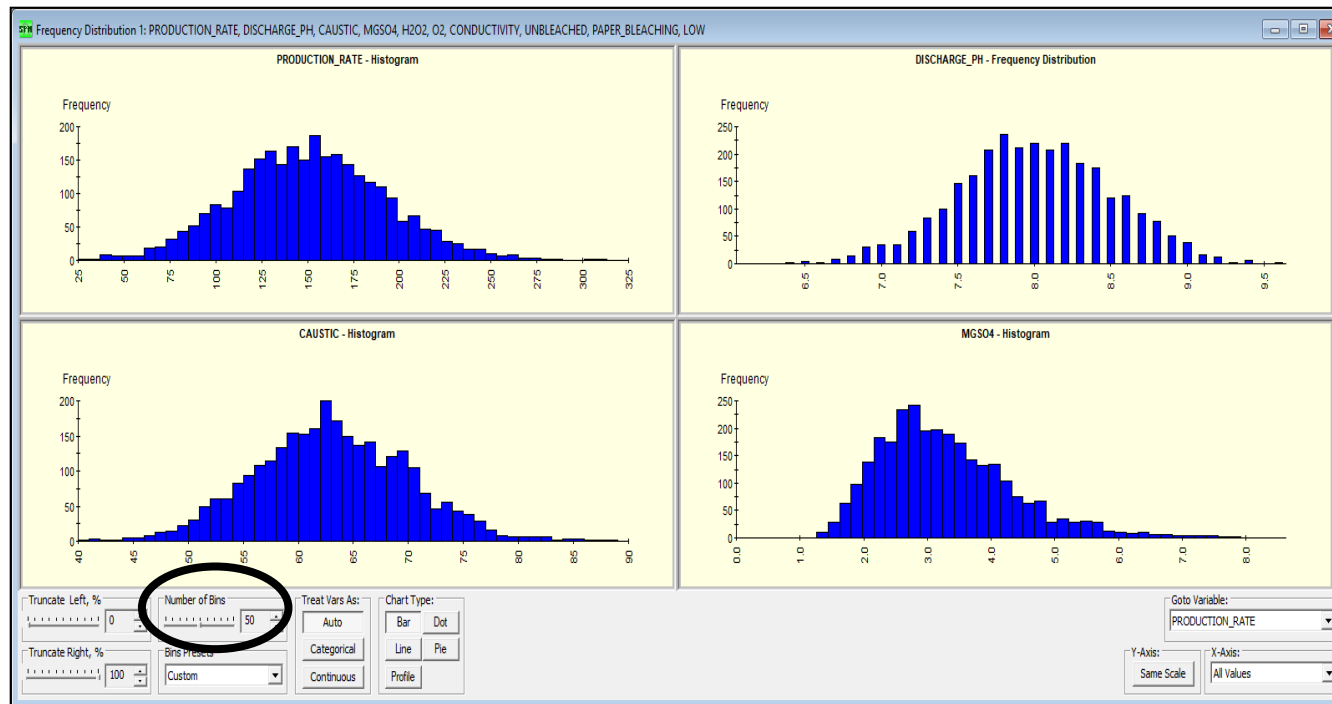
Graphs

1. Click the **Graphs** button.
2. From **Sort**, choose **File Order**.
3. Click the **Select** heading to select all the variables, check **Select Variables**, then uncheck **BRIGHTNESS** and **LOW** as shown below.



4. Click **Generate Charts**.

Visualizing data



The predictors in this data set all happen to be continuous. Histograms are very useful for looking at the shape, center, and spread of continuous variables.

Use the controls at the bottom of the Graph window to interact with these graphs. For example, by increasing the **Number of Bins** using the slider at the bottom of the window, some additional patterns emerge.

Notice that:

- ▶ Production rate, discharge pH, and caustic appear to be roughly bell-shaped, while MGSO4 is right-skewed.
- ▶ Because more rounding occurs with the pH measurement, the discharge pH appears slightly more discrete than the other predictor variables.

Visualizing data

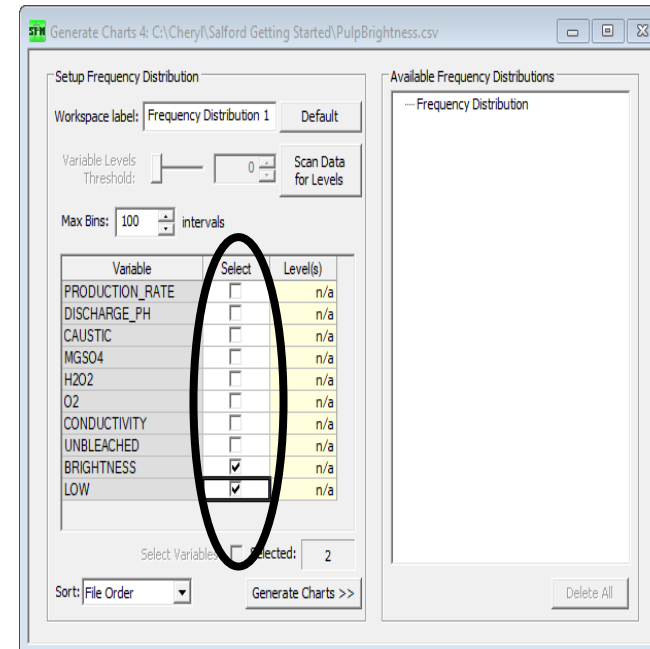
Data scientists often refer to a response variable as a target variable. Return to the dialog box to view the two target variables:

- ▶ Brightness is a continuous target that represents the actual brightness value.
- ▶ Low is an indicator variable that identifies when the brightness measure falls below the lower control limit.

We will consider both target variables because they provide two ways of looking at this problem. LOW allows us to focus on what predictor variables are most associated with the brightness measure falling below its expected range. We can then follow up this investigation by looking at the brightness measure itself to see the relationships between the key predictor variables and brightness more closely.

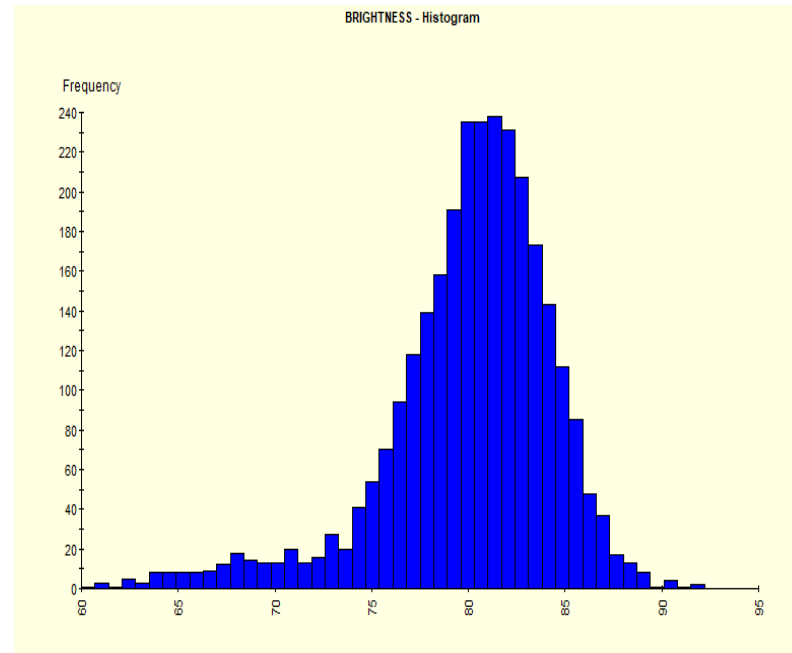
Graphs

1. Choose **Explore > Frequency Distribution**.
2. Check **BRIGHTNESS** and **LOW** as shown below.



3. Click **Generate Charts**.

Visualizing data

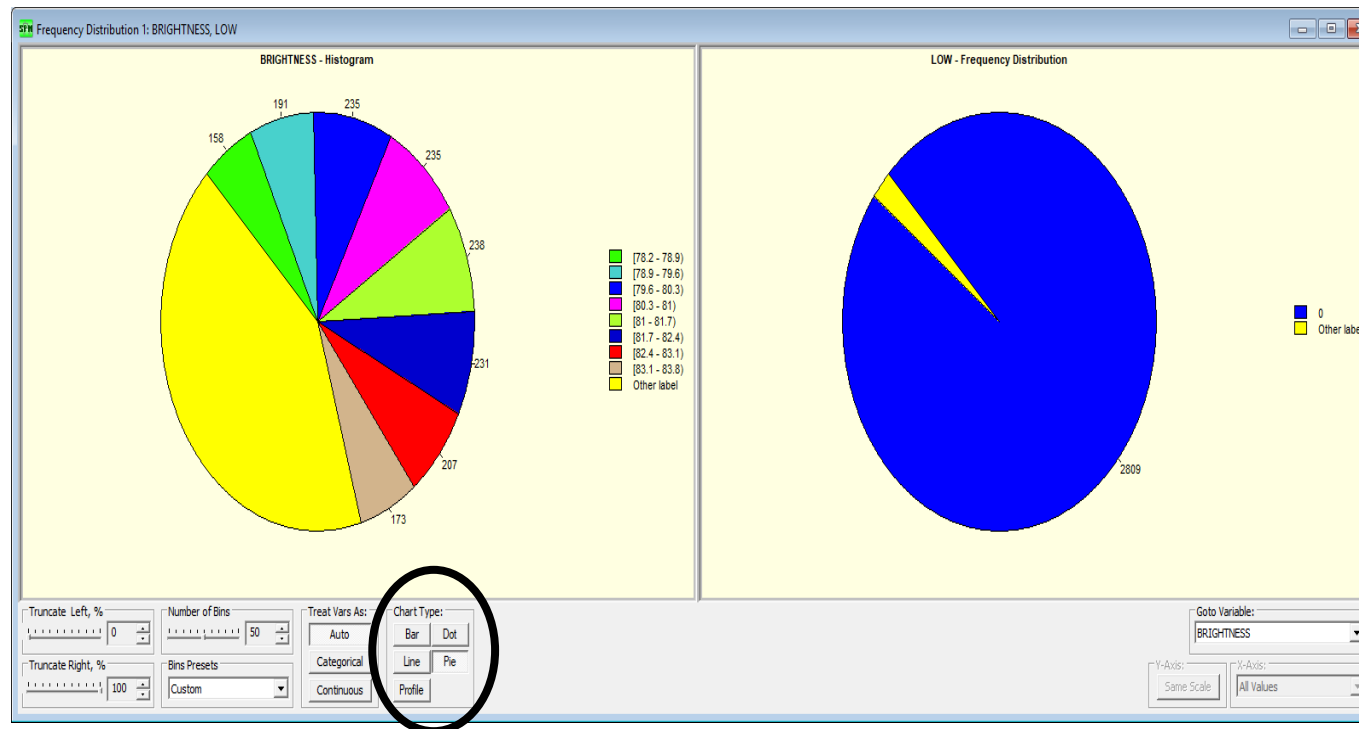


Brightness is a continuous target. The default **Chart Type** of **Bar** makes sense for this variable. The resulting bar graph for a continuous variable will be a histogram.

Once again, you can change the **Number of Bins** to adjust the histogram visualization. In the graph above, the number of bins was increased to 50.

Notice the long left tail in the brightness variable histogram. These are the points where the process data fall below the lower control limit.

Visualizing data



To switch from histograms and bar charts to pie charts, click **Pie** under **Chart Type**.

Low is an indicator variable that identifies when the Brightness measure falls below the lower control limit. Because low is categorical, it is useful to view it in a Pie Chart. Notice that only a small proportion of the cases fall below the lower specification limit. These values fall in the “Other label” category because they are the least frequent in occurrence.

Next, let's use CART to see which predictor variables contribute the most to the brightness value falling below the lower control limit.

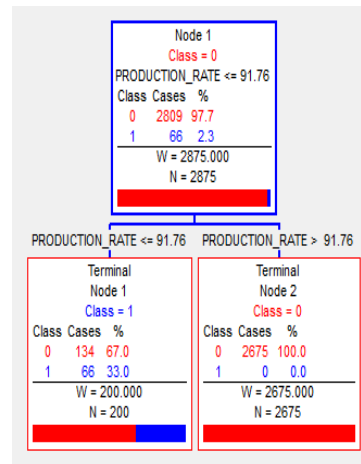
CART

What is CART

CART is an acronym for **C**lassification and **R**egression **T**rees. CART is a decision tree algorithm that works by creating a set of yes/no rules that splits the response or target variable into partitions based on the predictor settings.

The splits on predictor settings are performed sequentially, one predictor at a time, in order of their influence on the target variable. The first split is on the predictor that has the strongest effect on separating the target variable into two groups.

Each subsequent split further refines the target variable into more precise groupings. This partitioning of the data optimizes the accuracy in prediction of the target variable given new settings of the predictors. When the partitioning is complete, the resulting prediction is the target variable mean (continuous data) or event probability (categorical data) in the final grouping or node for the combination of predictor settings.



When to use CART

Use CART to:

- ▶ Identify the most important predictors of a response or target
- ▶ Discover combinations of predictor settings that are most likely to lead to a specific outcome
- ▶ Visualize your findings
- ▶ Create business rules that are easy to understand, use and apply to your process in real time

Why use CART

CART answers questions such as:

- ▶ What is the root cause of the defects in my process?
- ▶ What rules should I put in place to monitor my process that identify when to perform corrective action?

Fitting a CART model

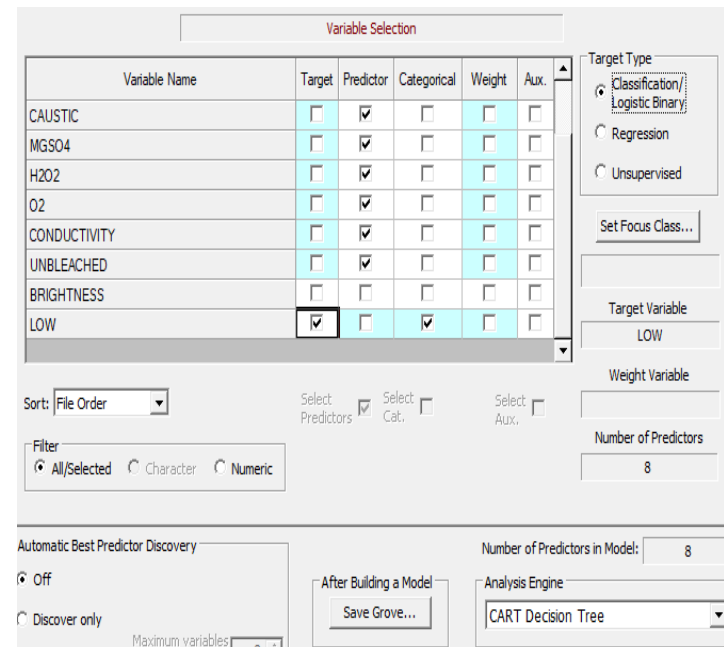
Recall that the LOW variable represents observations falling below the lower control limit on a control chart. These observations represent when the brightness measure fell below its expected range. Identifying the root cause associated with these out-of-control points allows us to determine the sources of this excessive variation in the process.

LOW is a categorical target variable, so the **Target Type** is **Classification/Logistic Binary**. When you select LOW as your target, SPM will automatically check the **Categorical** box for this variable.

Because BRIGHTNESS is a response, not a predictor, you will leave this unchecked for this first model.

CART Model Setup

1. Choose **Model > Construct Model**.
2. From **Sort**, choose **File Order**.
3. From **Target Type**, choose **Classification/Logistic Binary**.
4. Check **Low** in the **Target** column.
5. Check the 8 predictors in the **Predictor** column.
6. From **Analysis Engine**, choose **Cart Decision Tree**.



The screenshot shows the 'Variable Selection' dialog box in Minitab. The main table lists variables with checkboxes for Target, Predictor, Categorical, Weight, and Aux. The 'LOW' variable is selected as the target, and all other variables are selected as predictors. The 'Target Type' is set to 'Classification/Logistic Binary'. The 'Sort' is set to 'File Order'. The 'Number of Predictors' is 8. The 'Analysis Engine' is set to 'CART Decision Tree'.

Variable Name	Target	Predictor	Categorical	Weight	Aux.
CAUSTIC	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
MGS04	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
H2O2	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
O2	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
CONDUCTIVITY	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
UNBLEACHED	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
BRIGHTNESS	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
LOW	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Target Type: Classification/Logistic Binary, Regression, Unsupervised

Target Variable: LOW

Weight Variable: (empty)

Number of Predictors: 8

Sort: File Order

Filter: All/Selected, Character, Numeric

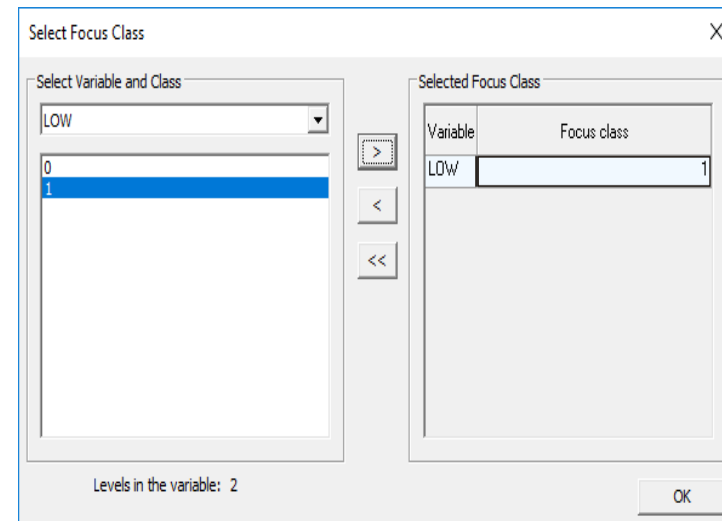
Analysis Engine: CART Decision Tree

Fitting a CART model

The target variable, Low, is an indicator variable where 1 indicates the observation was outside of statistical control. Because we are looking for the root cause of a point being outside of statistical control, 1 is the focus class for this target variable.

CART Model Setup (continued)

7. Click **Set Focus Class**.
8. Select **1** on the left side of the dialog box, then click the right arrow to assign the value **1** as the focus class.



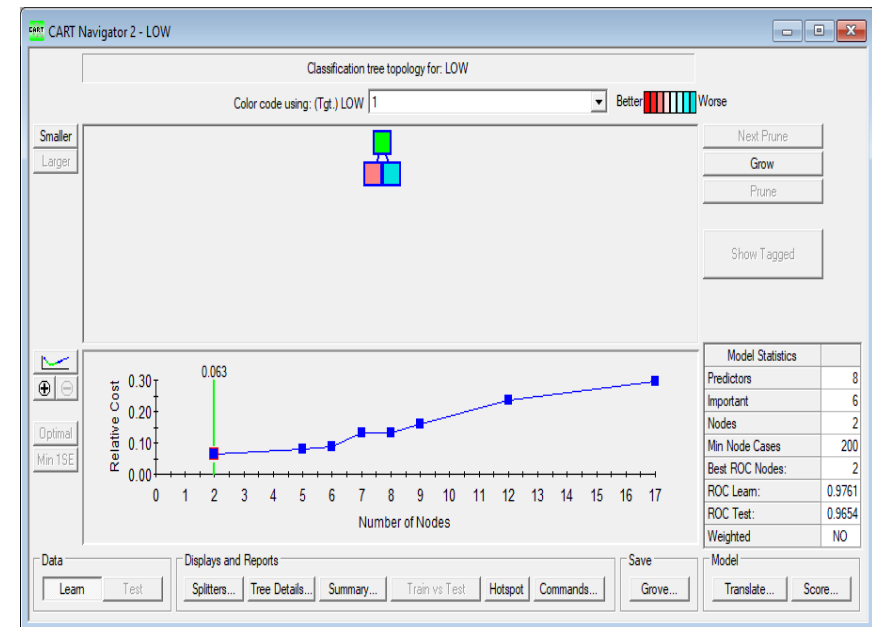
9. Click **OK**.
10. Click **Start**.

Fitting a CART model

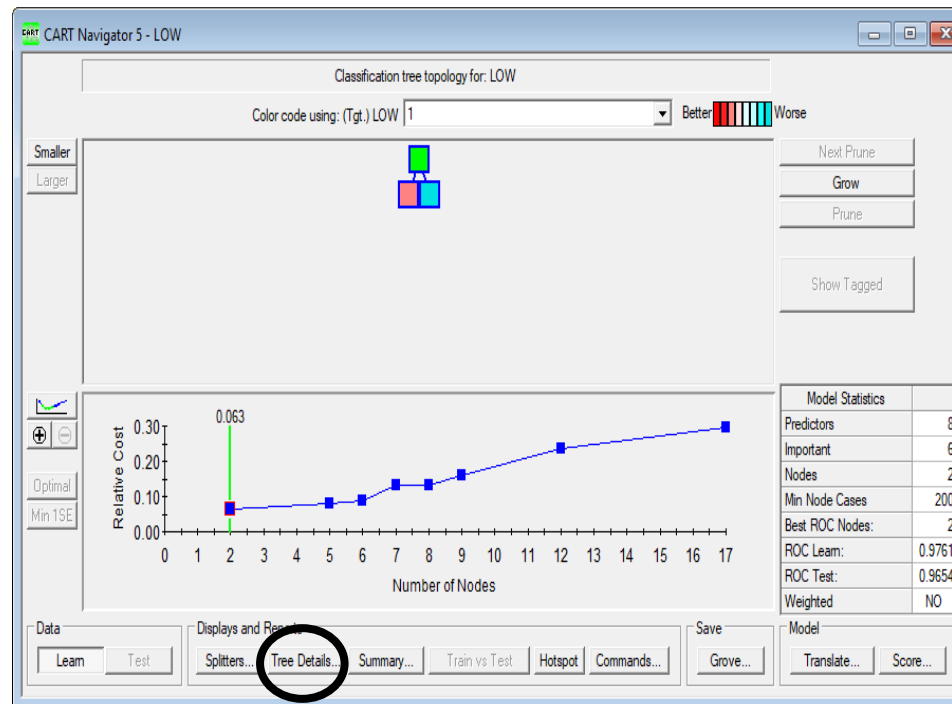
Recall that machine learning techniques involve splitting the data into learn and test sets where the test set is used to evaluate the model's predictive ability. By default, the CART model uses 10-fold cross validation. With 10-fold cross validation, the observations are randomly split into 10 groups or folds. To estimate the error in the prediction, 10 iterations are run where at each iteration, 9 of the 10 folds are used to create a CART decision tree and the remaining fold is held out as the test set to assess the model accuracy.

SPM compares several CART models, each containing a different number of nodes. The nodes represent the number of groups that the CART method uses to partition the data. The simplest possible CART model would contain two nodes, where the data are split into two groups based on a single value of a single predictor. The most complicated possible CART model would contain up to N nodes where each unique observation would fall in its own node. Putting each observation into its own group or node would provide a perfect fit for the learn data, but would not likely predict well for new observations. Cross validation using a learn and a test set allows us to balance the complexity of the model with the ability of the model to predict for new observations.

The “optimal” model is the model that minimizes the relative cost, or error rate of the tree with regard to predicting the value of the target for the holdout data. Here, the 2-node tree has the lowest relative cost.



Viewing the tree



Under **Displays and Reports**, click **Tree details**.

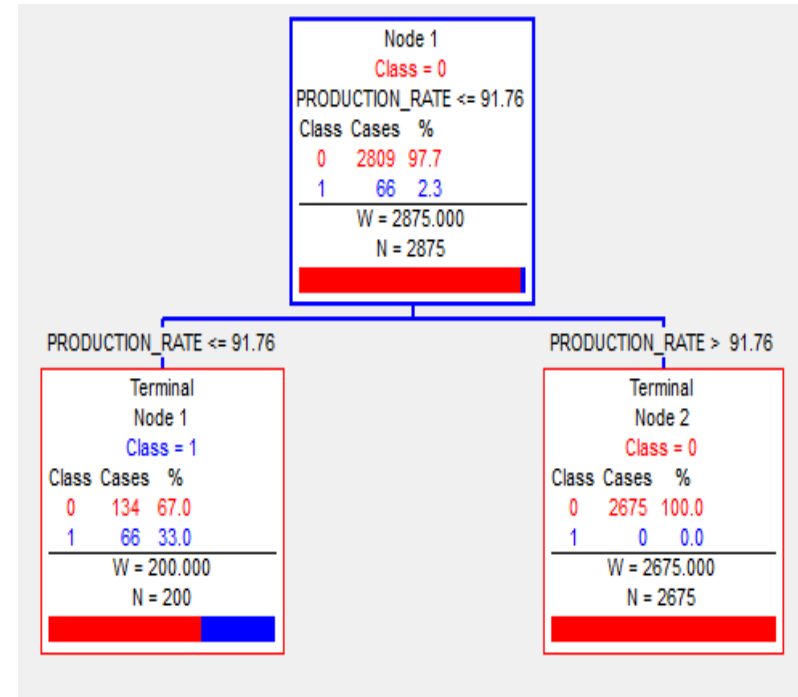
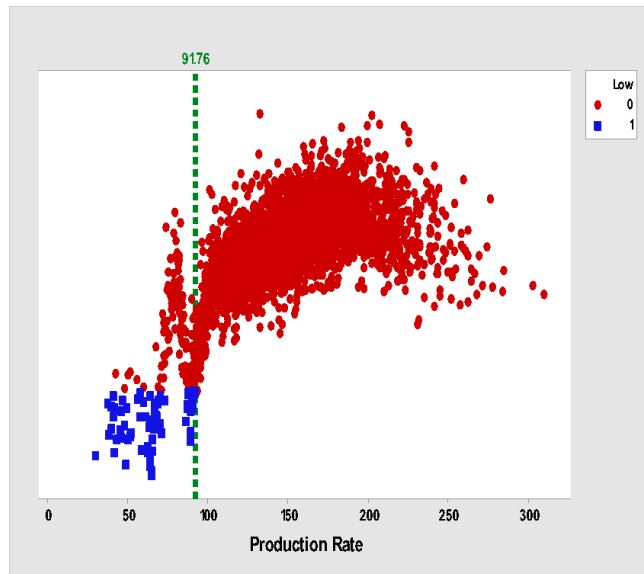
The simple 2-node tree splits the data by Production Rate. All of the observations that are unusually low (outside of statistical control) had a production rate less than or equal to 91.76. So, our first rule is:

If production rate ≤ 91.76 , then the estimated probability of the process being out of control is relatively high (33%). If production rate > 91.76 , then the process is likely in statistical control.

Viewing the tree

The Minitab graph below explains why this rule works. The CART model finds the vertical line corresponding to production rate that best separates the Low = 0 from the Low = 1 group.

Recall that 10-fold cross validation was used to select the model with the best accuracy in prediction. The resulting 2-node model shown here, uses all of the data.

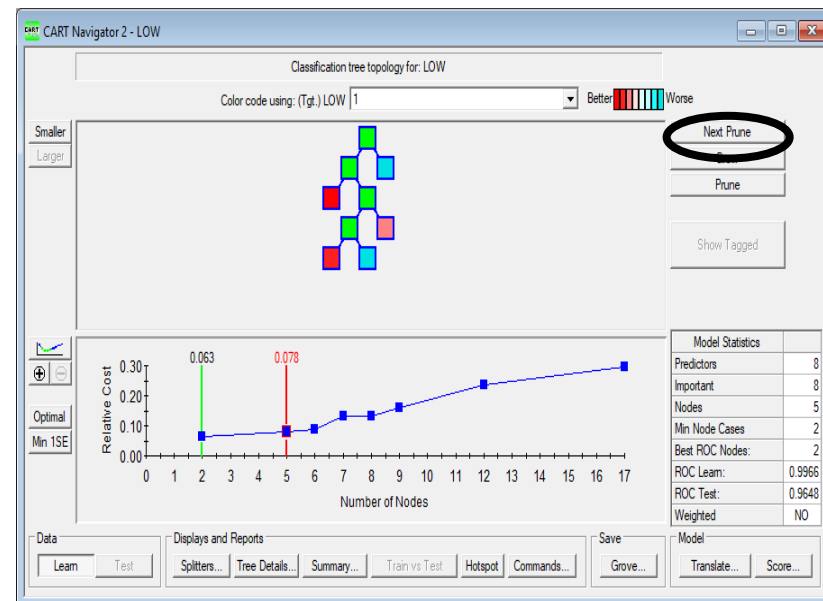


Growing the CART model

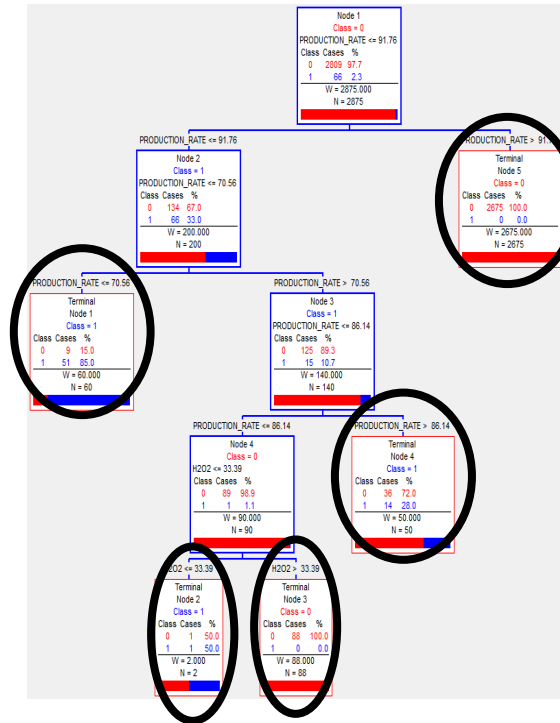
For these data, the model with the smallest error in prediction is the simple, 2-node model that splits on production rate. However, you can grow a larger tree for further investigation. Here, we will investigate the more complicated 5-node model.

CART Model Grow

1. Return to the **CART Navigator** window. If you can't find it, choose **Window > Cart Navigator – LOW**.
2. Click **Grow**.
3. Click **Tree Details** if the tree isn't currently visible. You may want to arrange your **CART Navigator** and **Tree Details** windows so that both are visible together on your screen.



Growing the CART model



For these data, the model with the smallest error in prediction is the simple, 2-node model that splits on production rate. However, you can grow a larger tree for further investigation. The five terminal nodes are circled above.

If you break down production rate even further, there is an 85% chance that the process will be out of control if the production rate is less than 70.56. Also, notice that the next new variable that CART splits is H2O2. Here, H2O2 only matters when the production rate falls between 70.56 and 86.14. This is represented in the node circled in the bottom left corner. But notice that there are only 2 observations in this node. This is likely why the more complicated 5-node model did not predict as well as the simpler 2-node model.

ADDITIVE your partner

This is the point where we stop our small introduction. Sample datas and further steps in this scenario are available.

- ▶ Contact our Team...
- ▶ E-Mail: spm@additive-net.de
- ▶ Phone: +49- 6172-5905-30
- ▶ Web: www.additive-net.de/spm

ADDITIVE Soft- und
Hardware für Technik und
Wissenschaft GmbH
Max-Planck-Straße 22b
D-61381 Friedrichsdorf