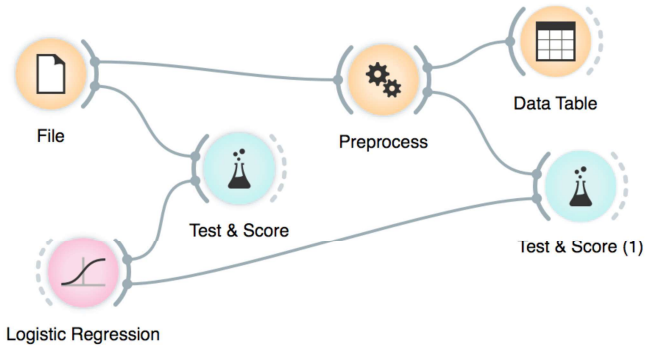
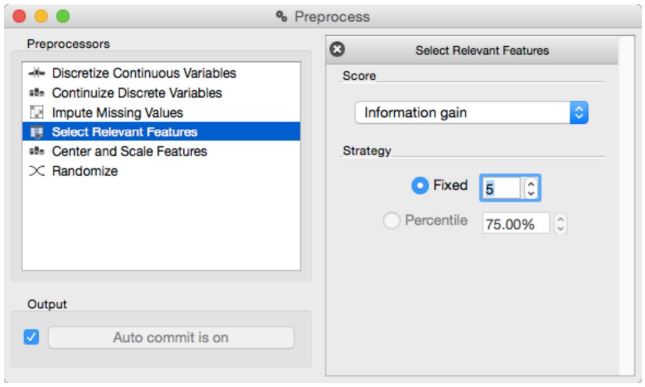
# Lesson 12: A Sneaky Way to Cheat

Gene expression data set we will use was borrowed from Gene Expression Omnibus. There is a special widget in Orange bioinformatics add on that we could use to fetch this and similar data sets. Instead, we will here rely on GEO data set that is preloaded in Orange: geo-gds360. Use File and then "Browse documentation data sets".

Consider a typical gene expression data sets where we have samples in rows and genes expressions in columns. These data sets are usually fat: there are many more genes than samples. Fat data sets are almost typical for systems biology. When samples are labeled with phenotype and our task is phenotype classification, many features (genes) will be irrelevant and most often only a few will be highly correlated with class. So why not simply first select a set of most informative features, and then do the whole analysis? At least cross-validation will then work much faster, as the model inference algorithms will deal with much smaller data tables. Cool. What a nice trick! Let's try it out in the following workflow.




The workflow above uses the data preprocessing widget, which we have configured to select 5 most informative features.




Observe the classification accuracy obtained on the original data set, and on the data set with five best selected features. What is happening? Why?