**Project 2**

**Data Mining for Intrusion Detection**

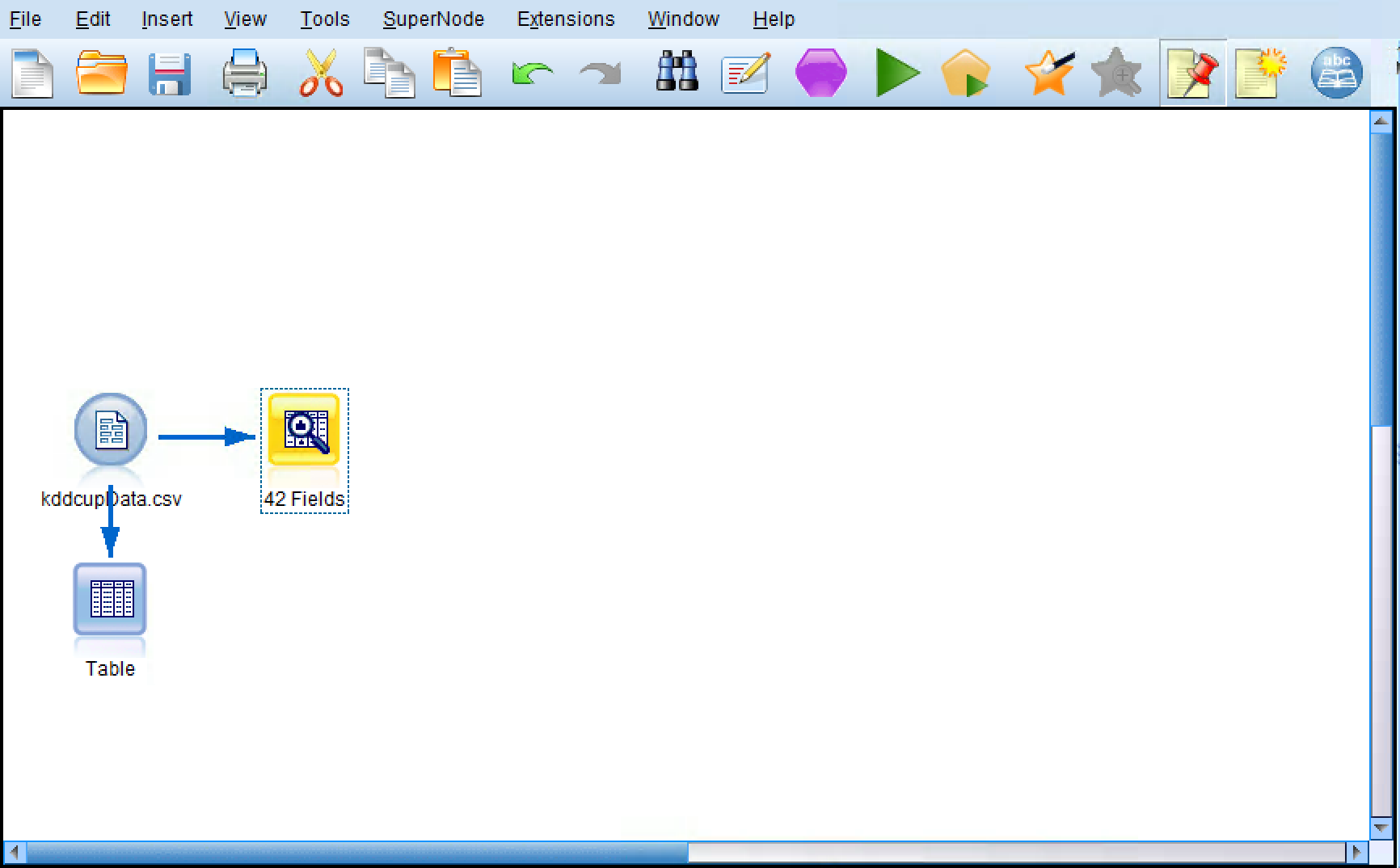
**Executive Summary**

The Data mining is carried with the help of tools and one such tool is a SPSS modeler from IBM. It is a text analytics software application. It allows us to not program lines code instead has statistical and data mining algortihms. It reduces the complexity in the data however huge volume it might be and makes complex predictive models very easy.

Classification models predict categorical class labels; and prediction models predict continuous valued functions. For example, we can build a classification model to categorize bank loan applications as either safe or risky, or a prediction model to predict the expenditures in dollars of potential customers on computer equipment given their income and occupation. There are many classification models that are available: Decision Trees (C5.0), Neural Networks, Bayesian Networks, regression model etc. Each has its own importance and plays a significant role in analyzing the data and predicting the outcomes. The counterpart for Classification is Prediction which is another form of data analysis.

Coming to this project, I have used classification method on the data which finds the intrusion attacks on the computer network from unauthorized third parties or users. There is a huge volume of data which will be sample sized to a lesser number and use the same due to complexity issues. The data set contains 22 attack types which are divided into four categories namely “dos, u2r, r2l and probe”.

First, the data is loaded and the data is connected to the data audit node:

****

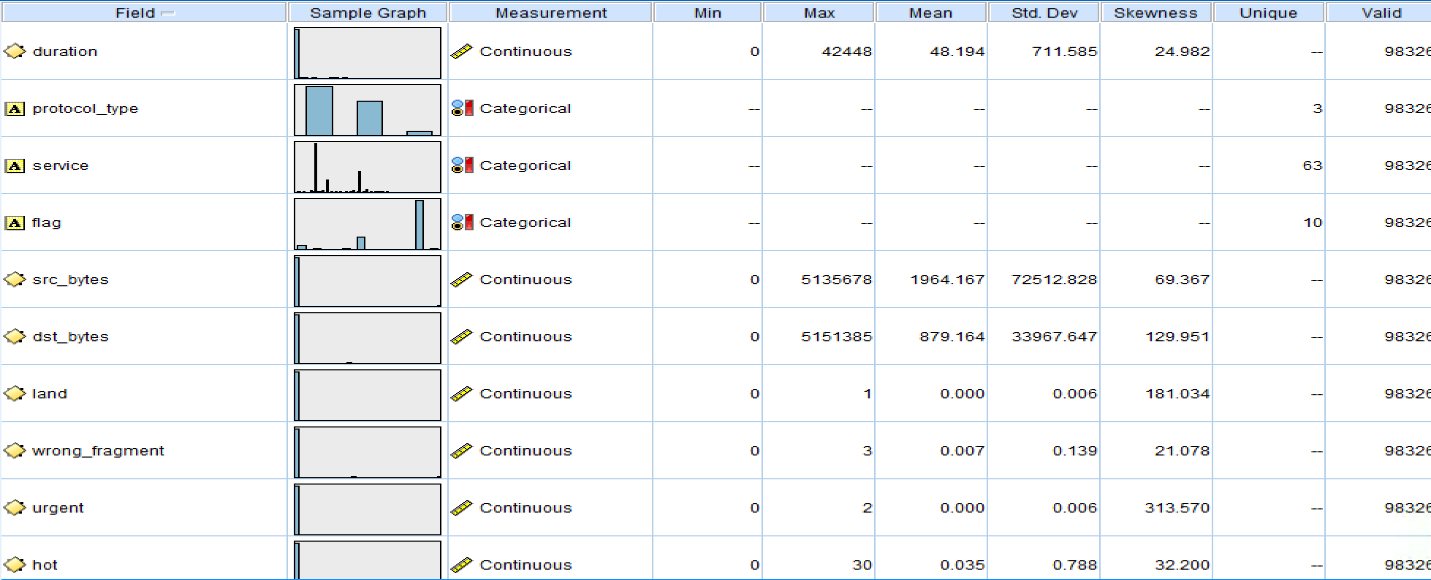
**The data audit node:**

This helps in fetching the details of the data, the statistical data, the minimum, maximum, mean median values are all fetched.

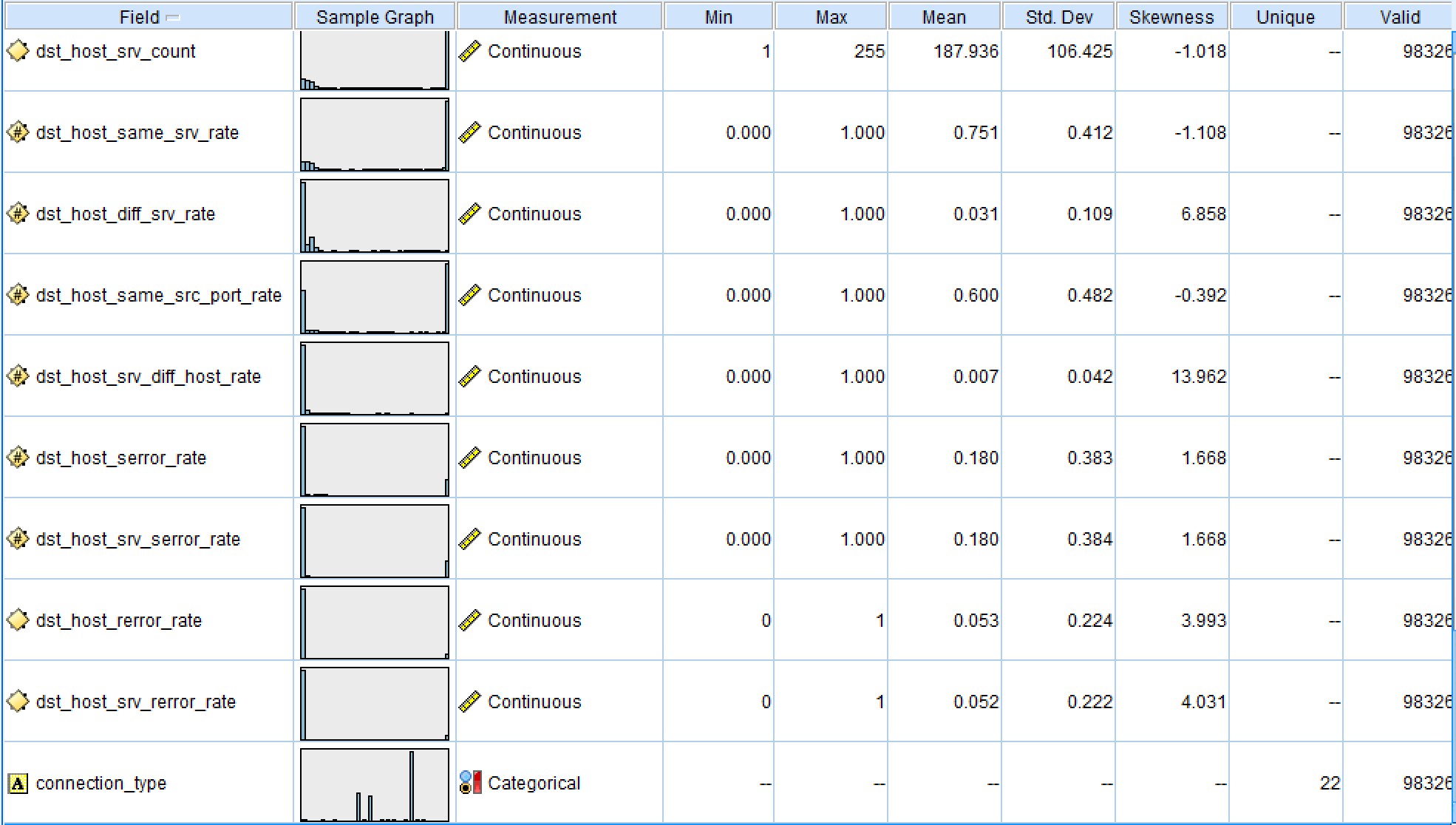
We can see number of outliers present in the data, skewness of the data with the help of data audit.

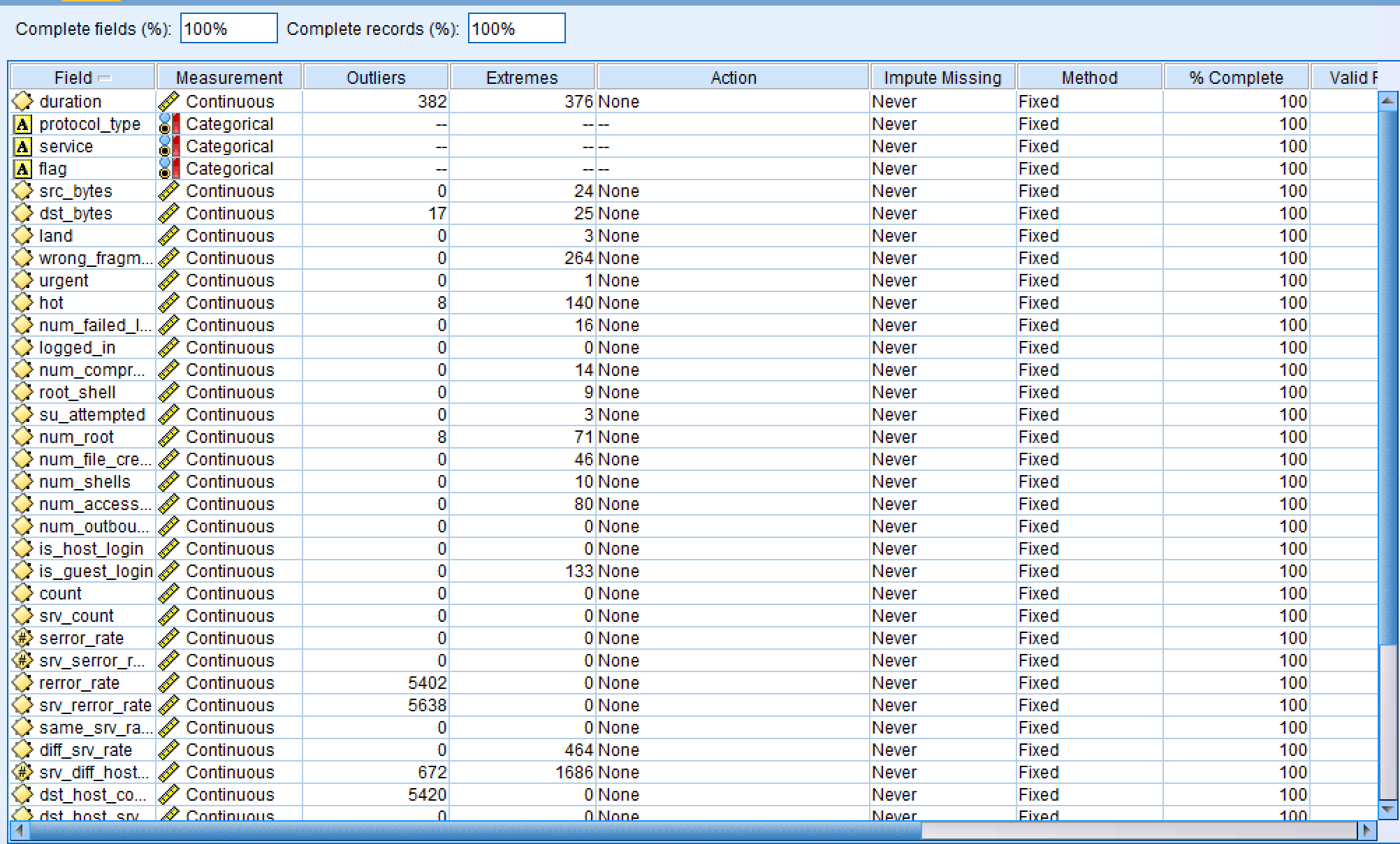
For the given data set there are no missing fields present which means there is 100% data present in all the fields. However that is found out in the data audit node below.

The screenshots below displays detail information about the data set:



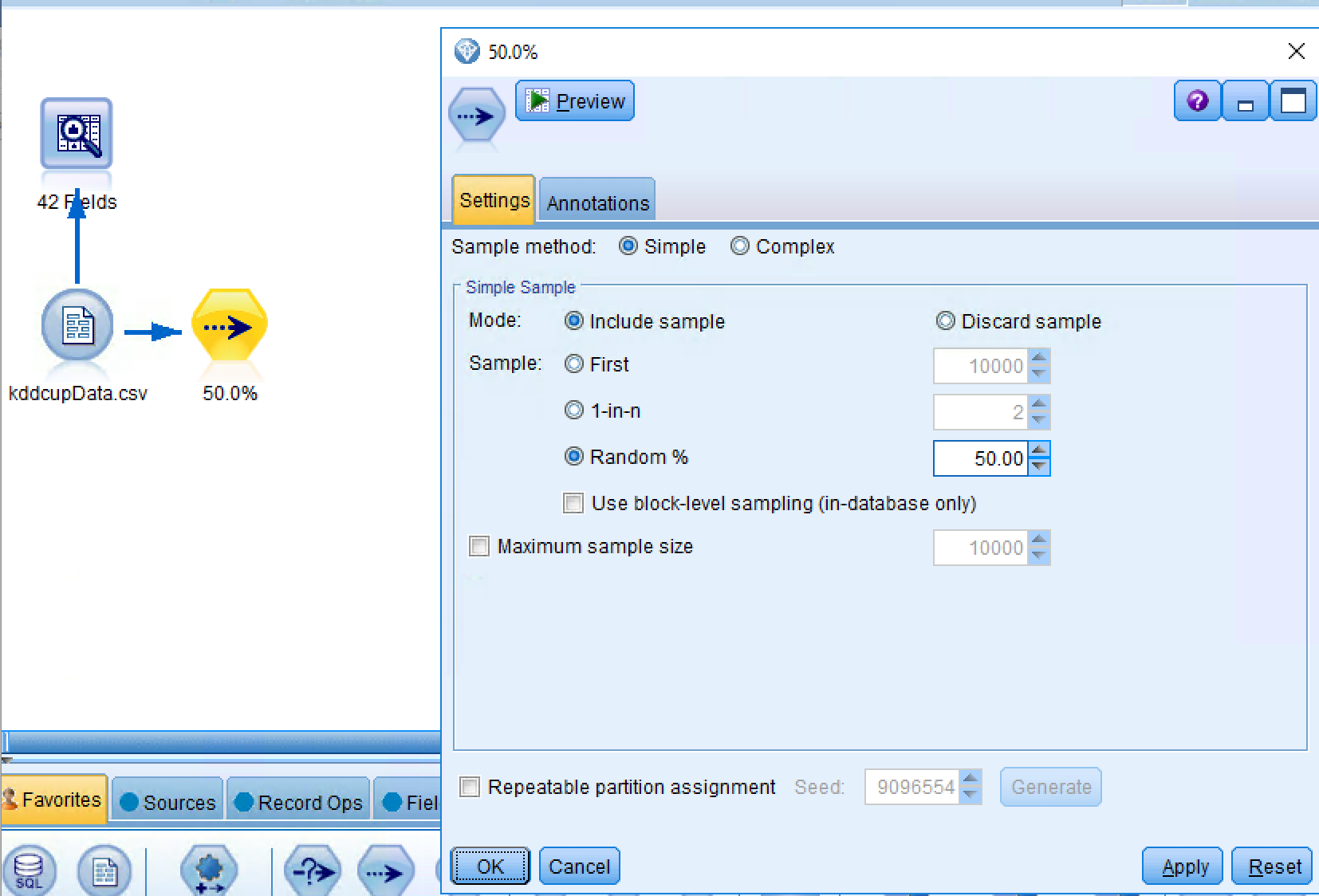






Since, we understood the data set, it is the time to get the sample data from the data set since the size of the data set is very big. To do so, we will use sample node present in Record Operations of the SPSS Modeler. Here I have chosen 50% of the data from the data set for the analysis. The screen shot of this is shown below.

Now we are required to reclassify all the different types of attacks under the name of ATTACKS and the connections which are not attacks i.e. normal connections as Normal. In order to do so in SPSS Modeler we will use reclassify node which is present in field operations. In this node we reclassified all the attacks as per our requirement as discussed above and added this reclasssified data to a new field called reclassified.

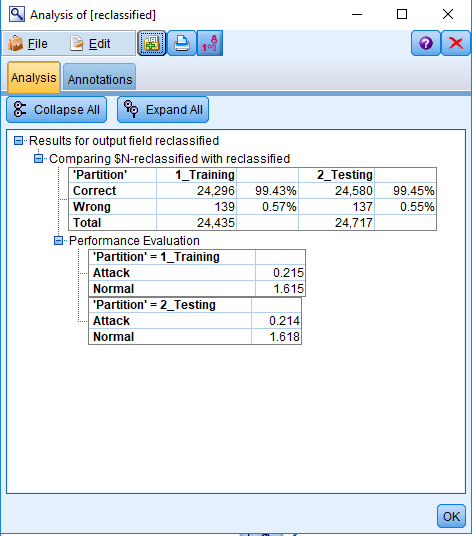
****

Since the data set is very huge so to avoid any biased analysis in any of the model I have normalized the whole data using Auto Data Prep node before performing any type of analysis. I have partitioned data into 50-50%.

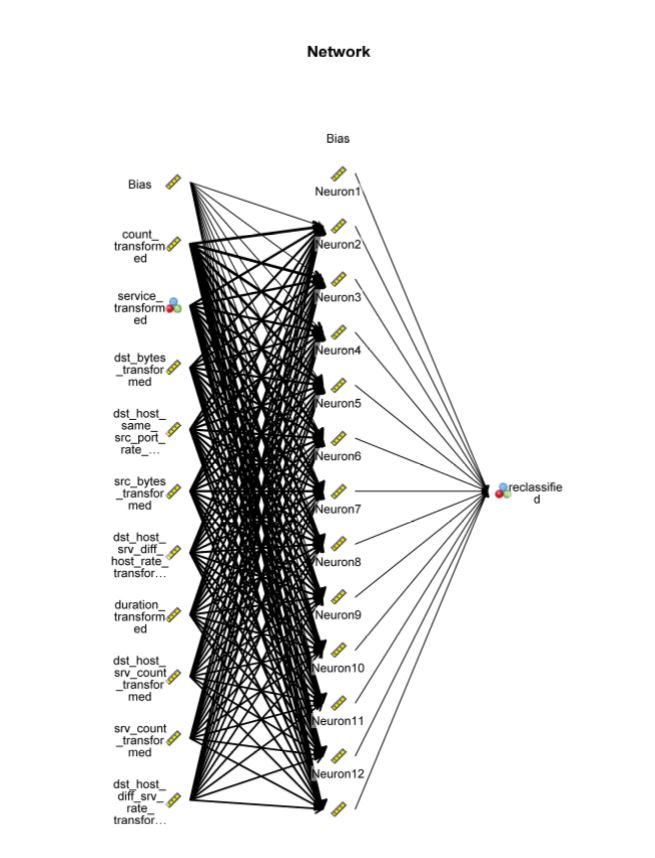
**Neural Network:**

Depending on the configuration of the output, Neural Networks can be used for numerical predictions or classification. Here, we will use Neural networks to classify wether the new incoming connection is Attack or Normal. The following figures shows the full analysis of the Neural network.

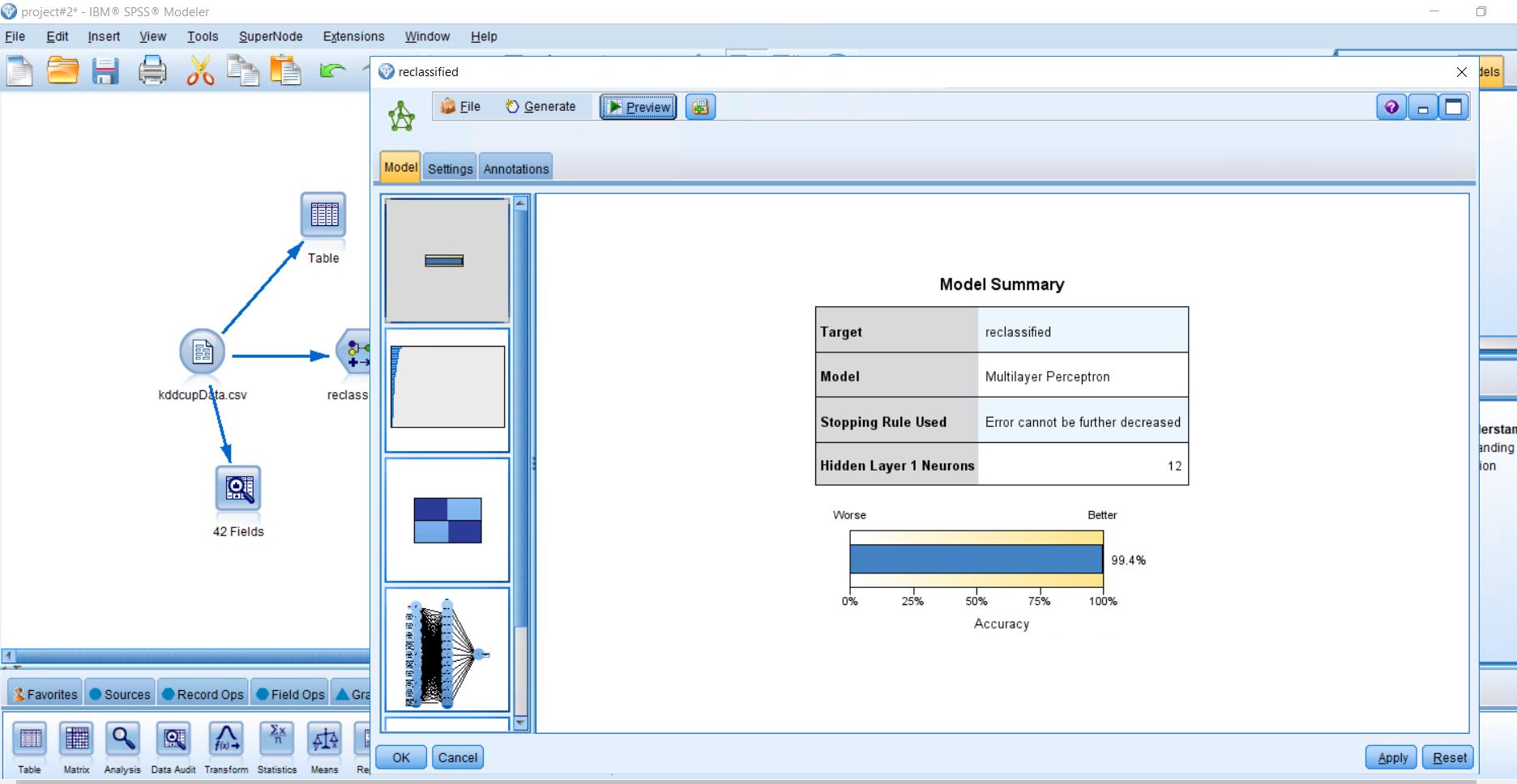
**Analysis:**



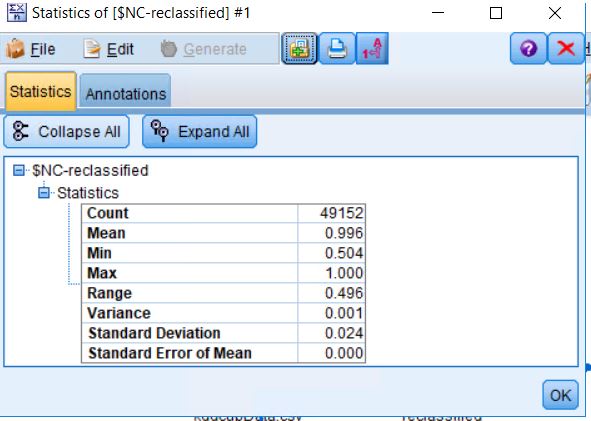
**Graph:**

****

**Neural network output:**

****

**Statistics:**

****

From the above output we can see that the accuracy of predicting the dependent variable of this model is very good.

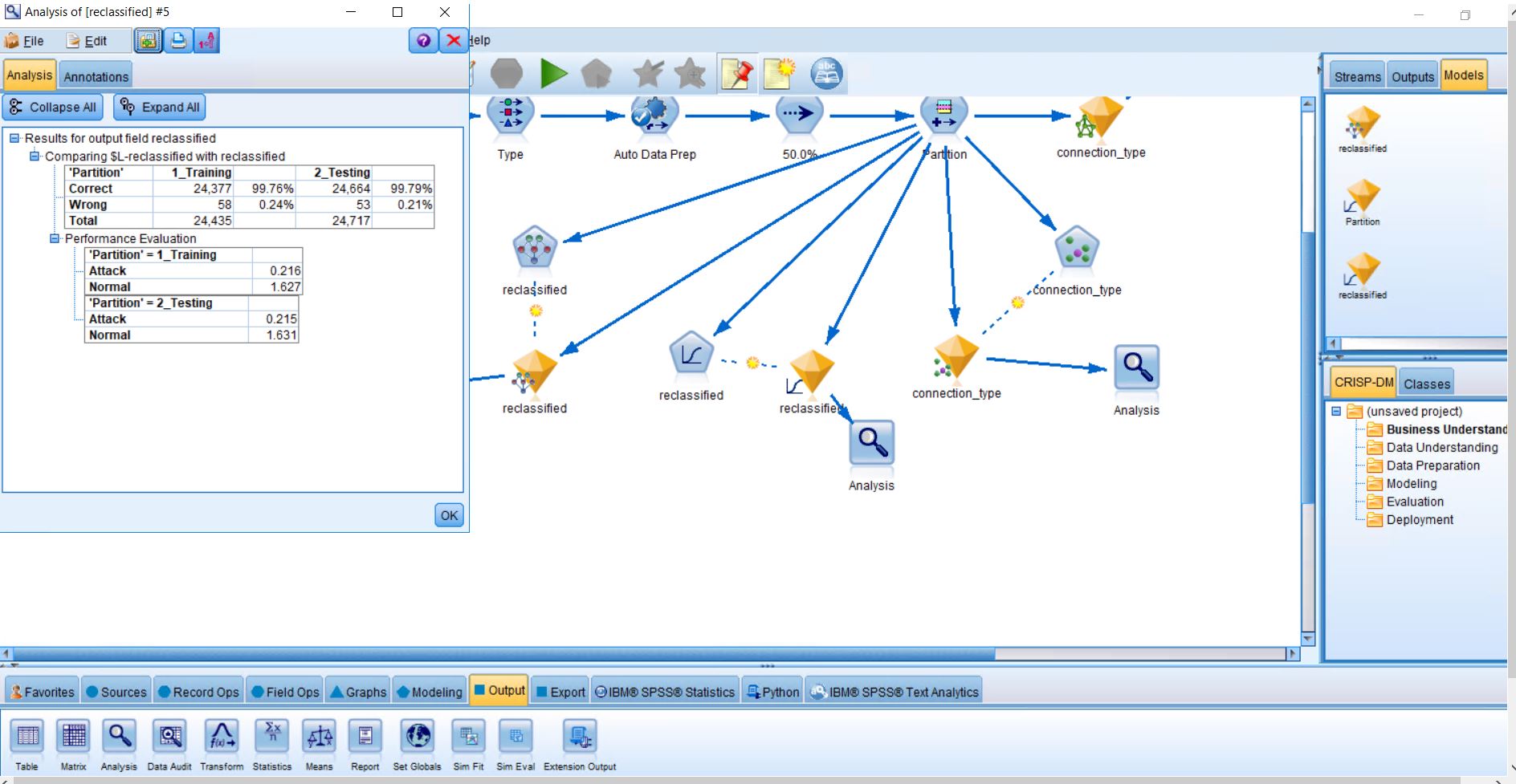
**Logistics:**

Logistic regression is a statistical method for analyzing a dataset in which there are one or more independent variables that determine an outcome. The outcome is measured with a dichotomous variable (in which there are only two possible outcomes) i.e. it only contains data coded as 1 (TRUE, success, pregnant, etc.) or 0 (FALSE, failure, non-pregnant, etc.).

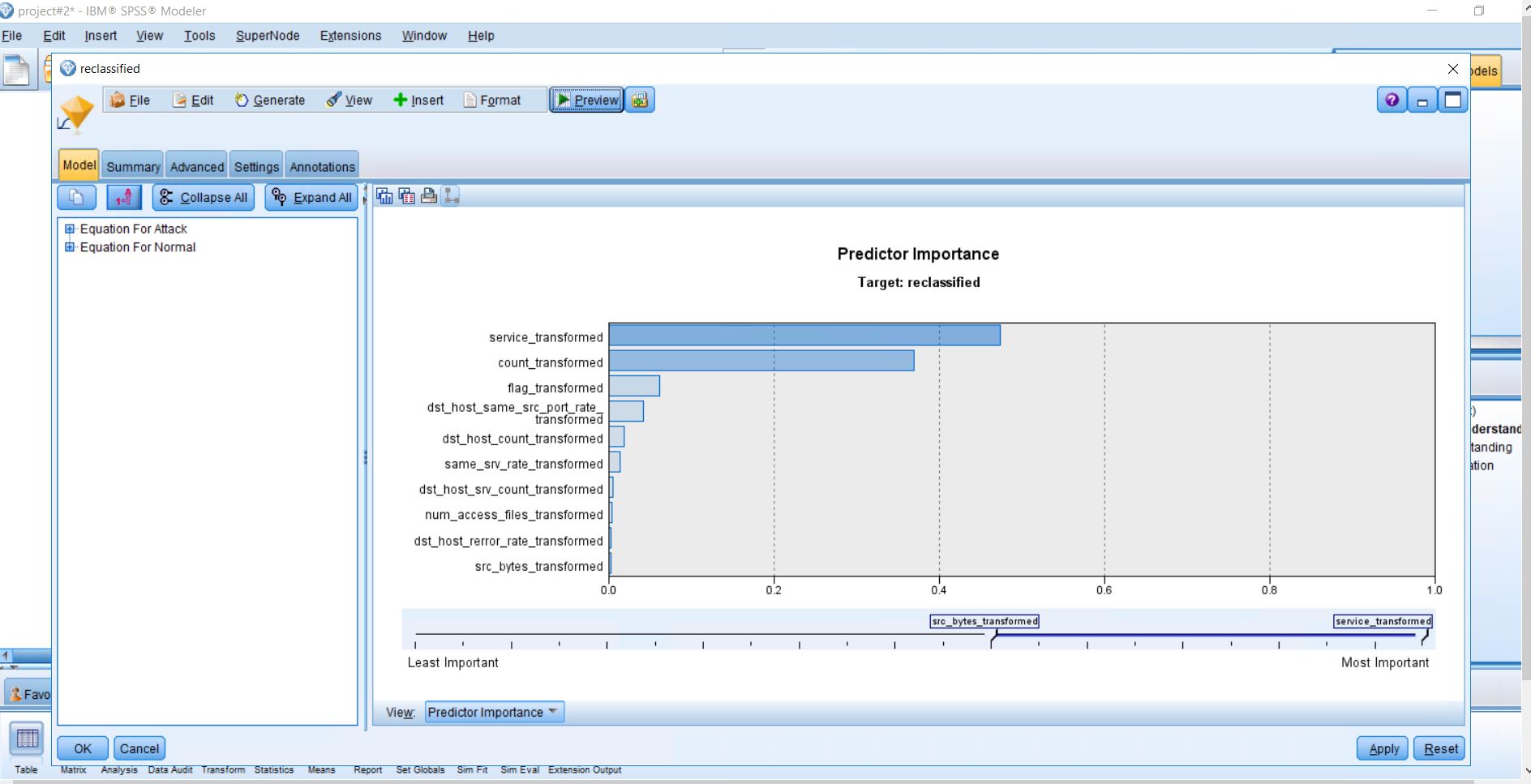
The goal of logistic regression is to find the best fitting (yet biologically reasonable) model to describe the relationship between the dichotomous characteristic of interest (dependent variable = response or outcome variable) and a set of independent (predictor or explanatory) variables. Logistic regression generates the coefficients (and its standard errors and significance levels) of a formula to predict a *logit transformation* of the probability of presence of the characteristic of interest:

The following figures show the full analysis of the logistic analysis model.

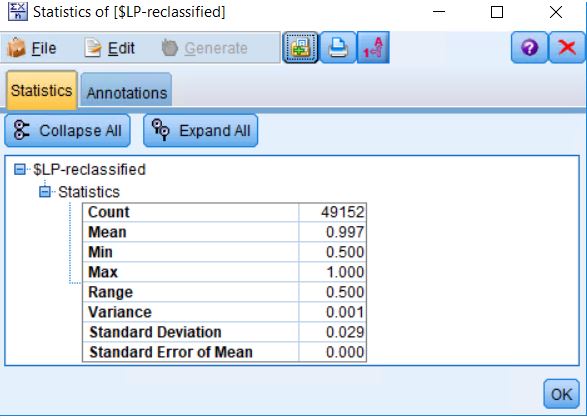
**Analysis:**

****

**Output:**

****

**Statistics:**

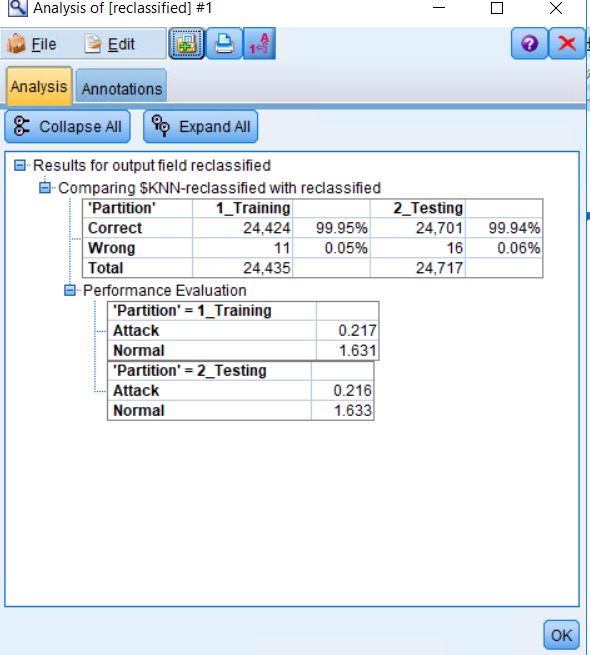
****

**KNN:**

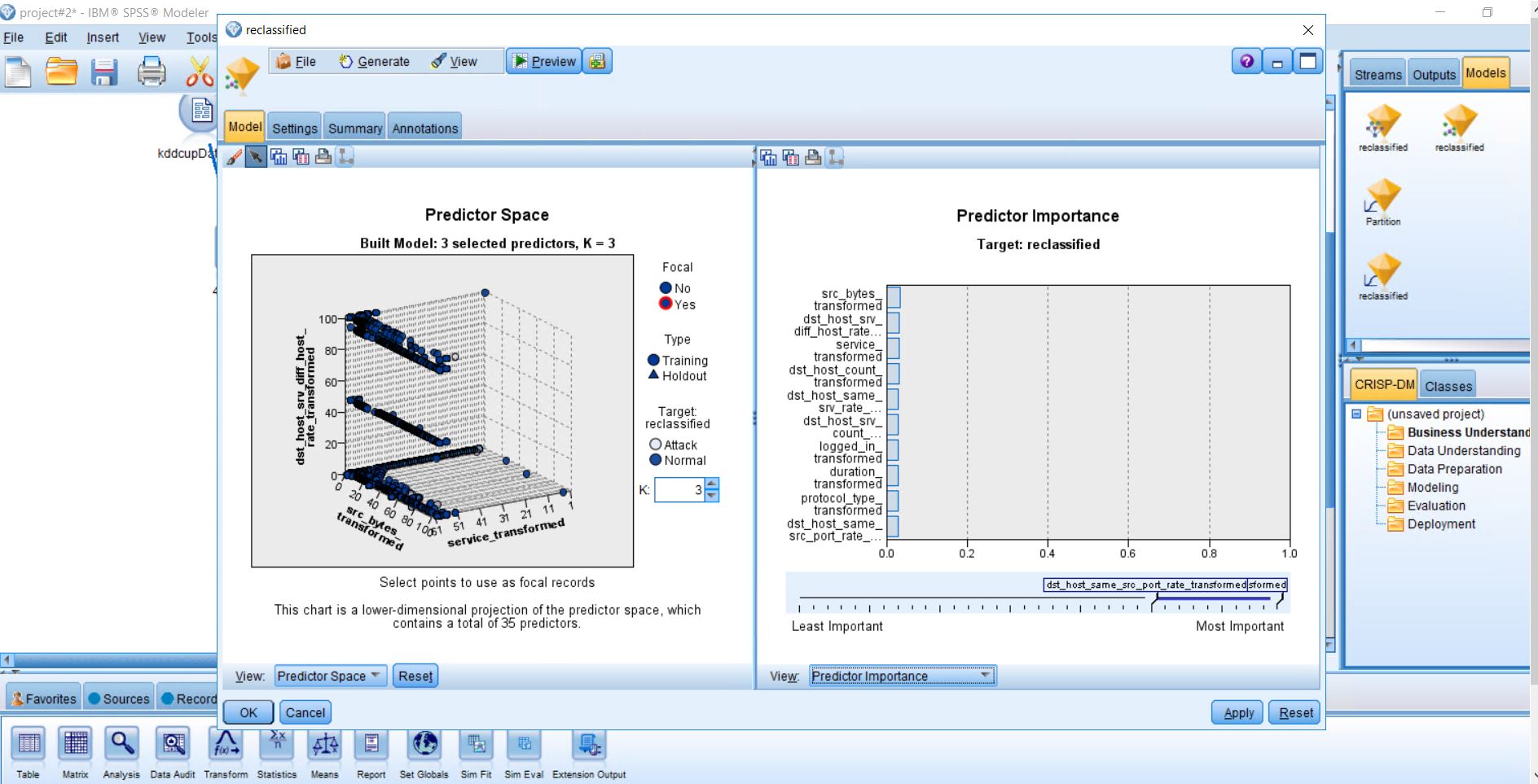
In [pattern recognition](https://en.wikipedia.org/wiki/Pattern_recognition), the *k*-nearest neighbors algorithm (*k*-NN) is a [non-parametric](https://en.wikipedia.org/wiki/Non-parametric_statistics) method used for [classification](https://en.wikipedia.org/wiki/Statistical_classification) and [regression](https://en.wikipedia.org/wiki/Regression_analysis). In both cases, the input consists of the *k* closest training examples in the [feature space](https://en.wikipedia.org/wiki/Feature_space). The output depends on whether *k*-NN is used for classification or regression.

*k*-NN is a type of [instance-based learning](https://en.wikipedia.org/wiki/Instance-based_learning), or [lazy learning](https://en.wikipedia.org/wiki/Lazy_learning), where the function is only approximated locally and all computation is deferred until classification. The *k*-NN algorithm is among the simplest of all [machine learning](https://en.wikipedia.org/wiki/Machine_learning) algorithms.

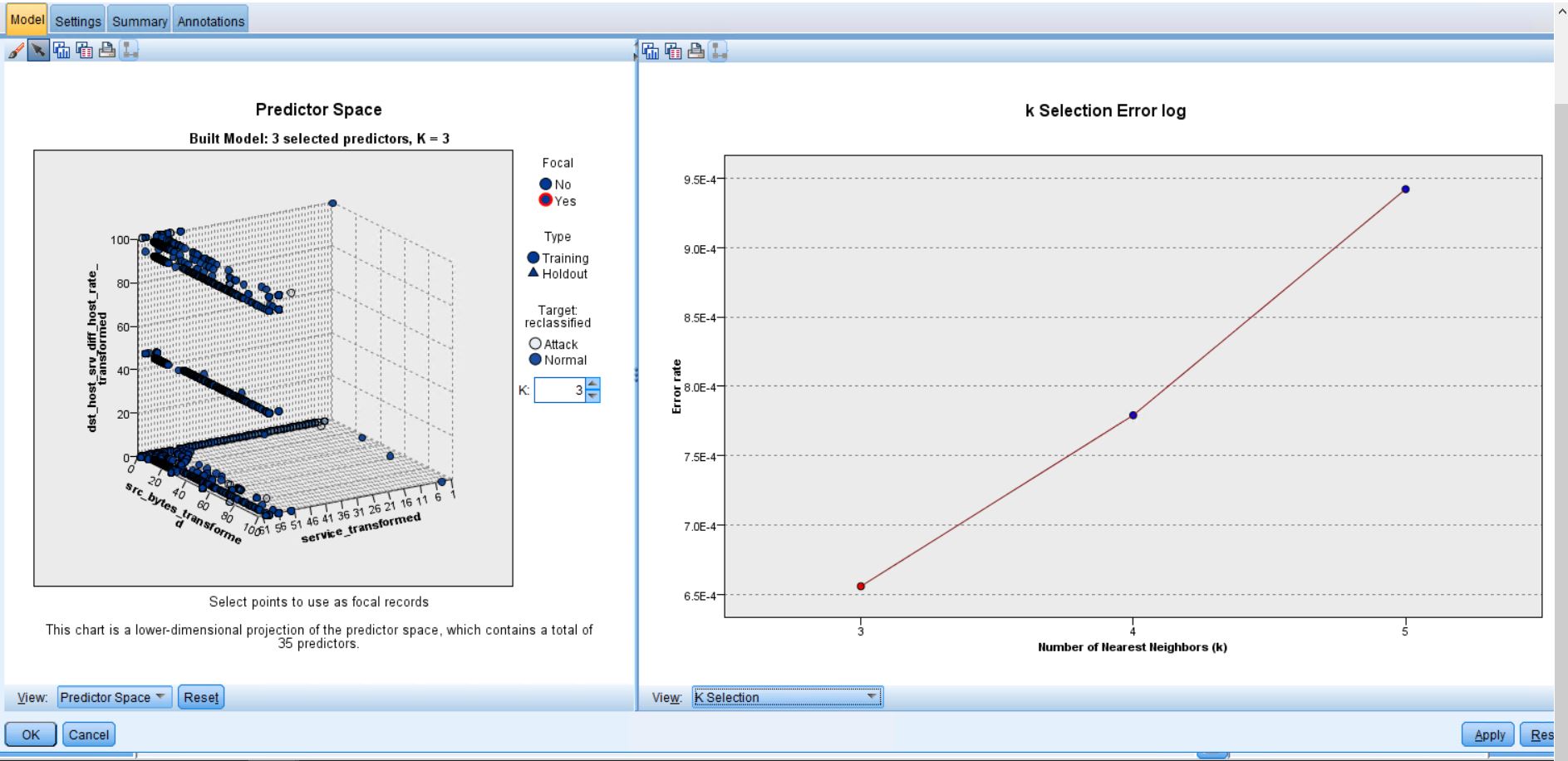
**Analysis:**

****

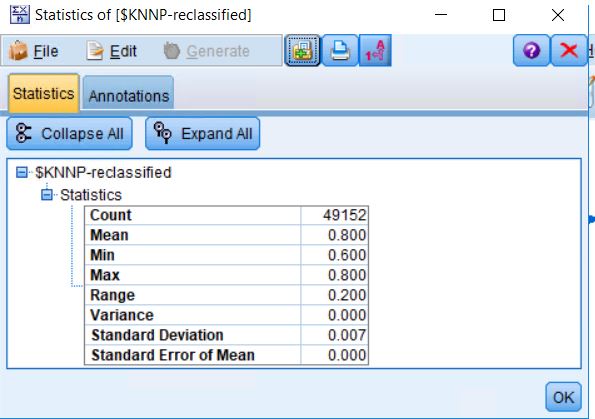
**Predictive Analysis:**

****

**Predicitve Analysis – K selection:**

****

**Statistics:**

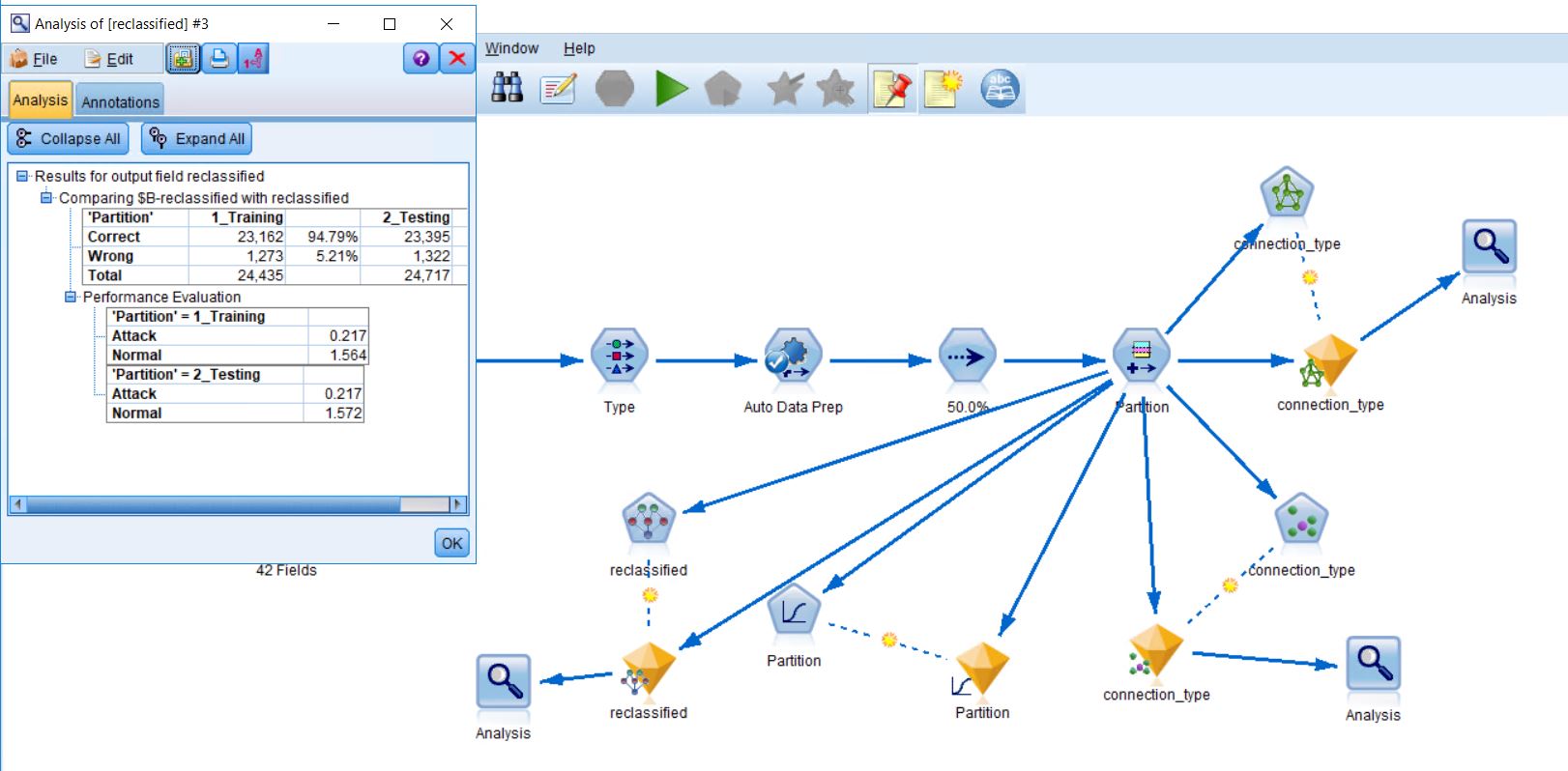
****

**Bayesian:**

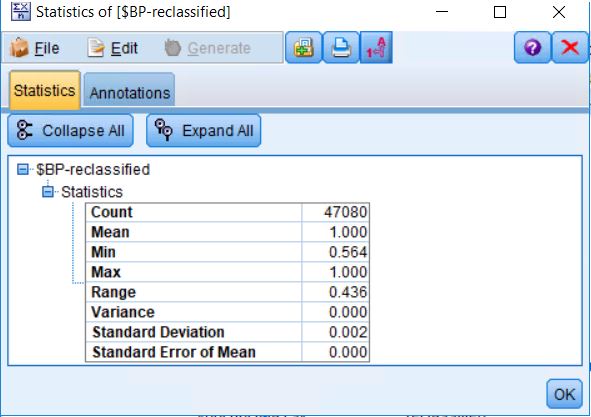
Bayesian statistics is a mathematical procedure that applies probabilities to statistical problems. It provides people the tools to update their beliefs in the evidence of new data.

Bayesian methods provide a rigorous way to include prior information when available compared to hunches or suspicions that cannot be systematically included in classical methods. This method also provides exact inferences without resorting to asymptotic approximations.

**Analysis:**

****

**Statistics:**

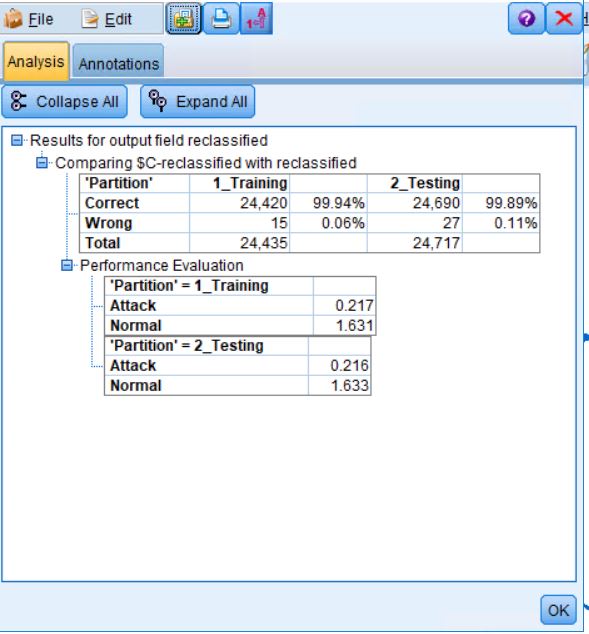
****

**C5.0:**

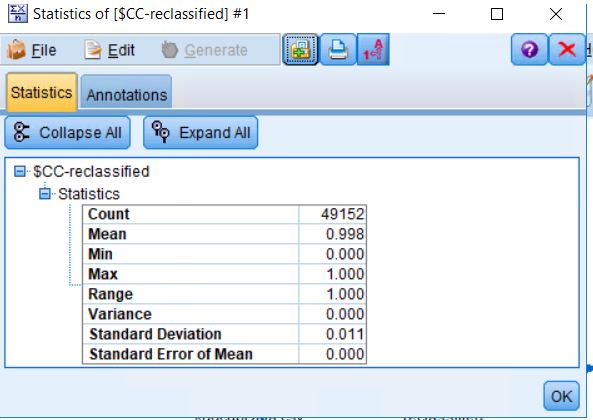
C5.0 tries to extract informative patterns from the data. Below are the Analysis output, Statistical output, and Predictive analysis before 50% and after 50% respectively for all the results.

The C5.0 node can predict only a categorical target. When analyzing data with categorical (nominal or ordinal) fields, the node is more likely to group categories together than versions of C5.0 prior to release 11.0

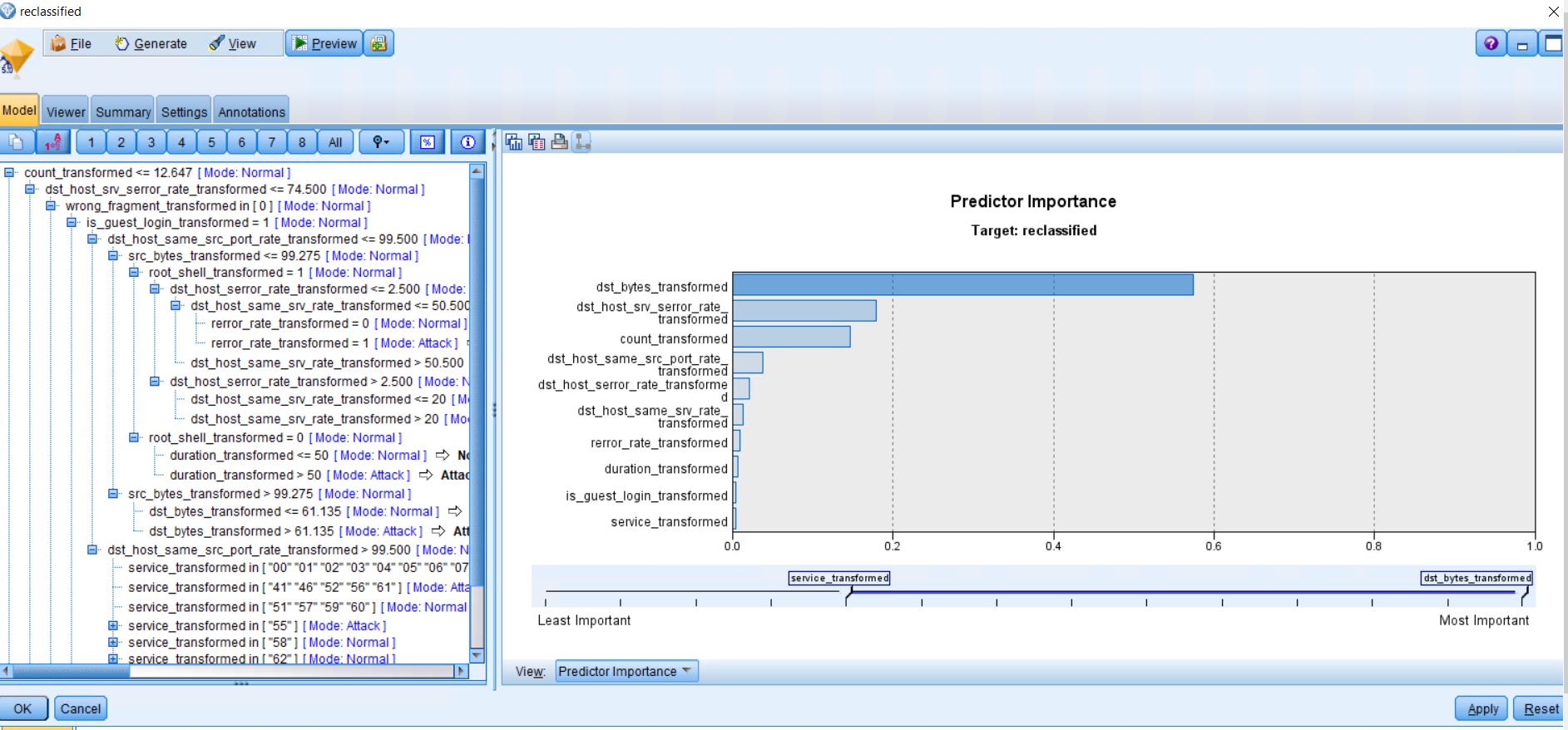
**Analysis:**

****

**Statistics:**

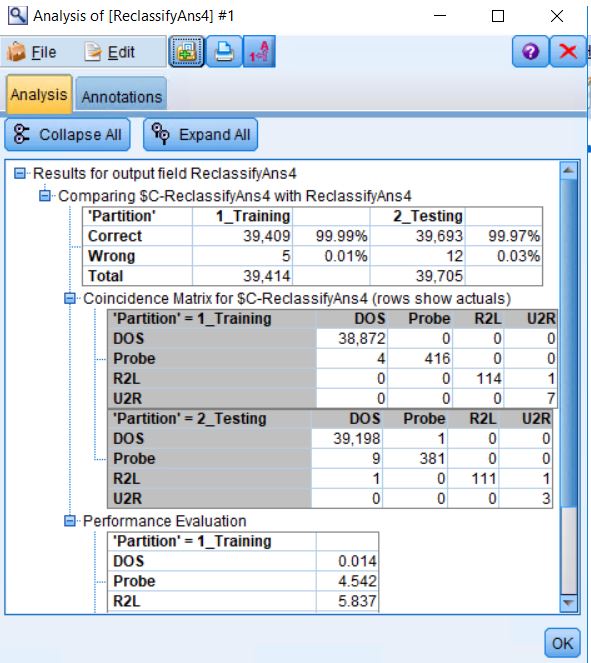
****

**Predictive analysis:**

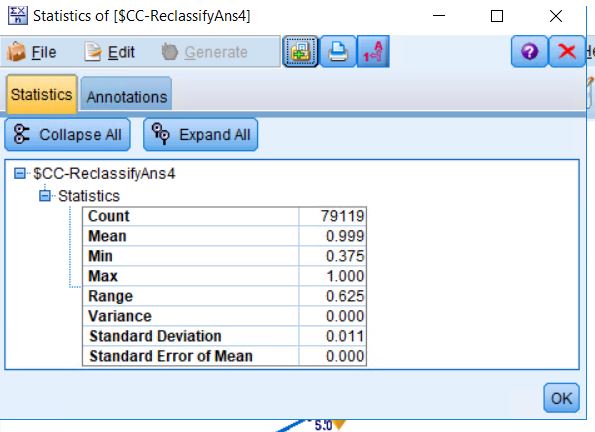
****

I have reclassified it from the original data and connected to the C5.0 again.

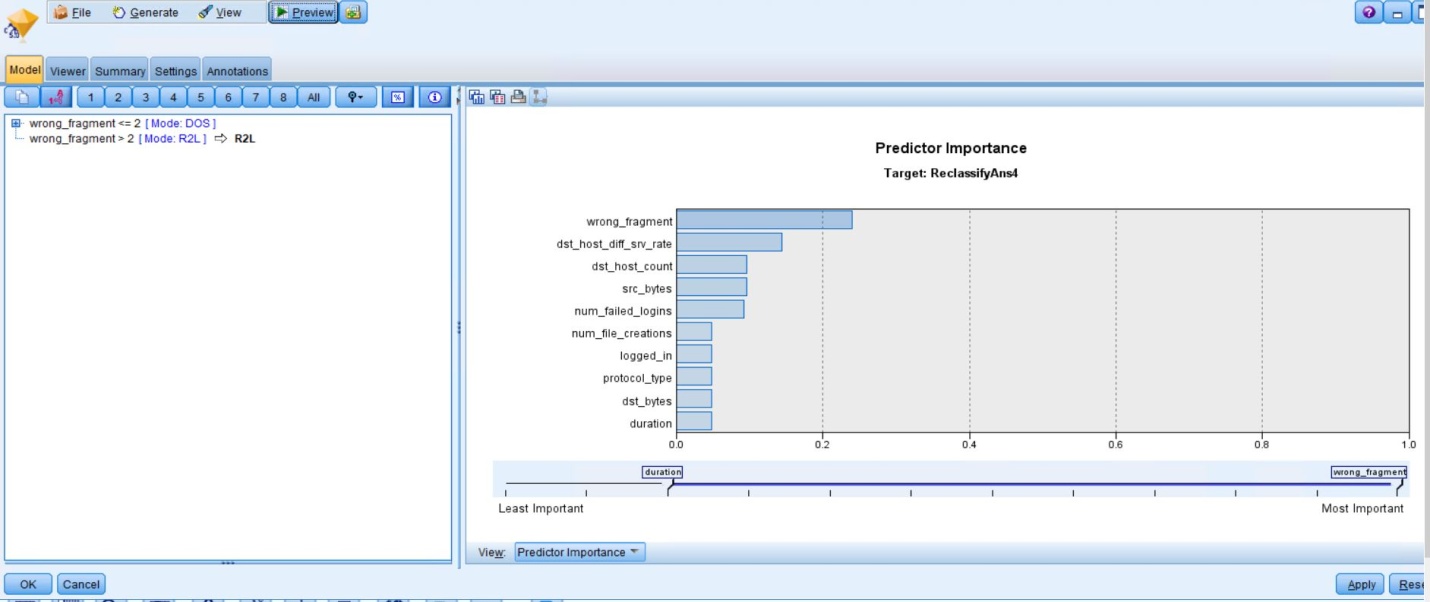
**Analysis before 50 percent:**

****

**Statistics before 50 percent:**

****

**Predictive analysis before 50 percent:**

****

So therefore, as per the predictive analysis graph above, the best one is the wrong\_fragmant.