**Classification Analysis**

**Identify Big Spenders in Catch the Pink Flamingo**

**Data Preparation**

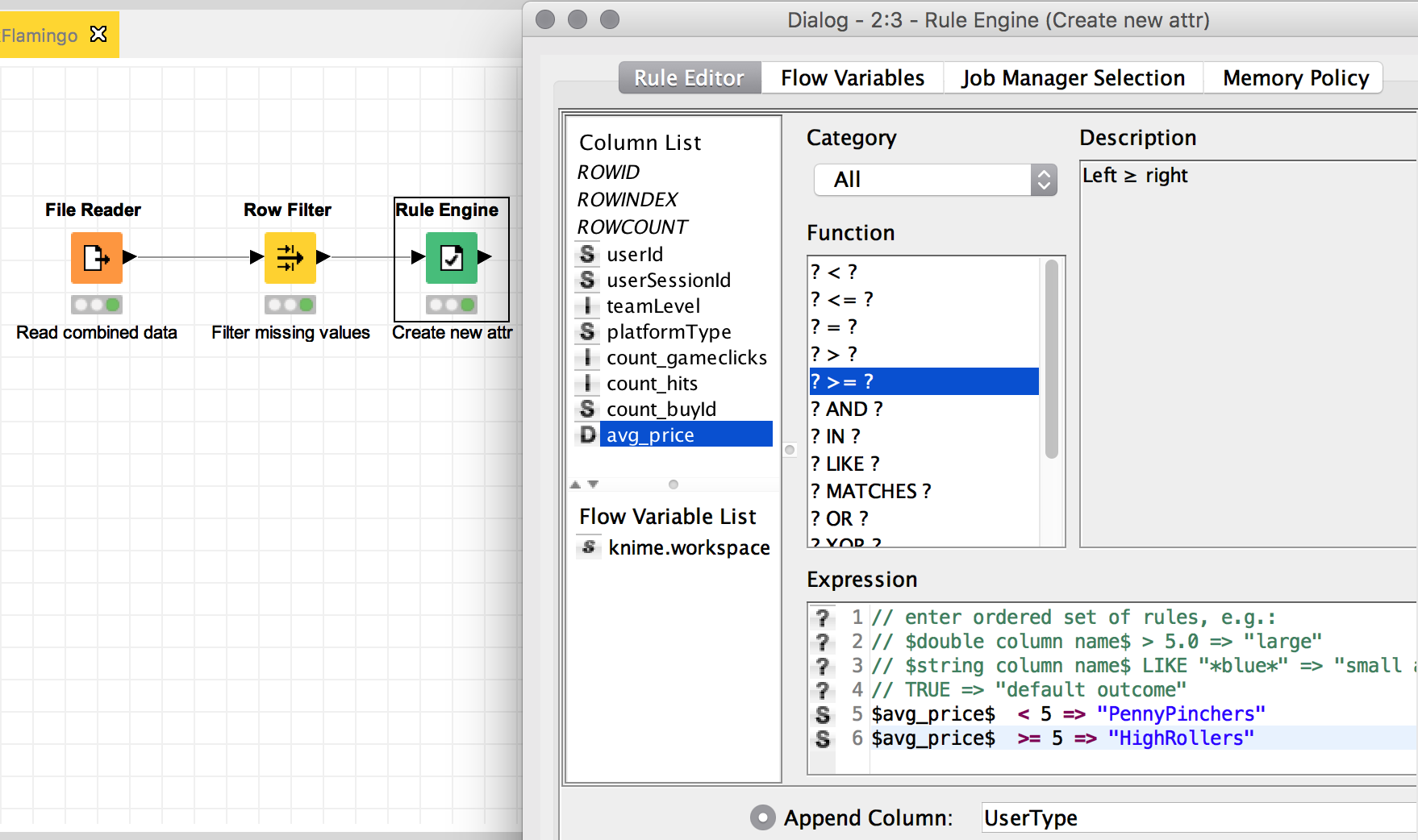
Analysis of combined\_data.csv

Sample Selection

|  |  |
| --- | --- |
| **Item** | **Amount** |
| # of Samples | 4,619 |
| # of Samples with Purchases | 1,411 |

Attribute Creation

A new categorical attribute was created to enable analysis of players as broken into 2 categories (HighRollers and PennyPinchers). A screenshot of the attribute follows:



A Rule Engine node was used for the creation of a new attribute call ‘User Type’, this node takes a rule and match them to each row in the table. If a rule matches, its outcome is added into the ‘UserType’ column. The rule used was:

**IF avg\_price < 5 THEN ‘PennyPinchers’**

**IF avg\_price >= 5 THEN ‘HighRollers’**

The creation of this new categorical attribute was necessary because we need to apply a classification algorithm, a decision tree works over a categorical variable, and we had to transform a quantitative variable into two categories.

Attribute Selection

The following attributes were filtered from the dataset for the following reasons:

|  |  |
| --- | --- |
| **Attribute** | **Rationale for Filtering** |
| userId | It’s an identifier and doesn’t give any information about the user and his likely to purchase big-tickets or not. |
| userSessionId | It’s an identifier and doesn’t give any information about the user and his likely to purchase big-tickets or not. |
| avg\_price | The information is now contained in the new attribute that we created in the step above. |

**Data Partitioning and Modeling**

The data was partitioned into train and test datasets.

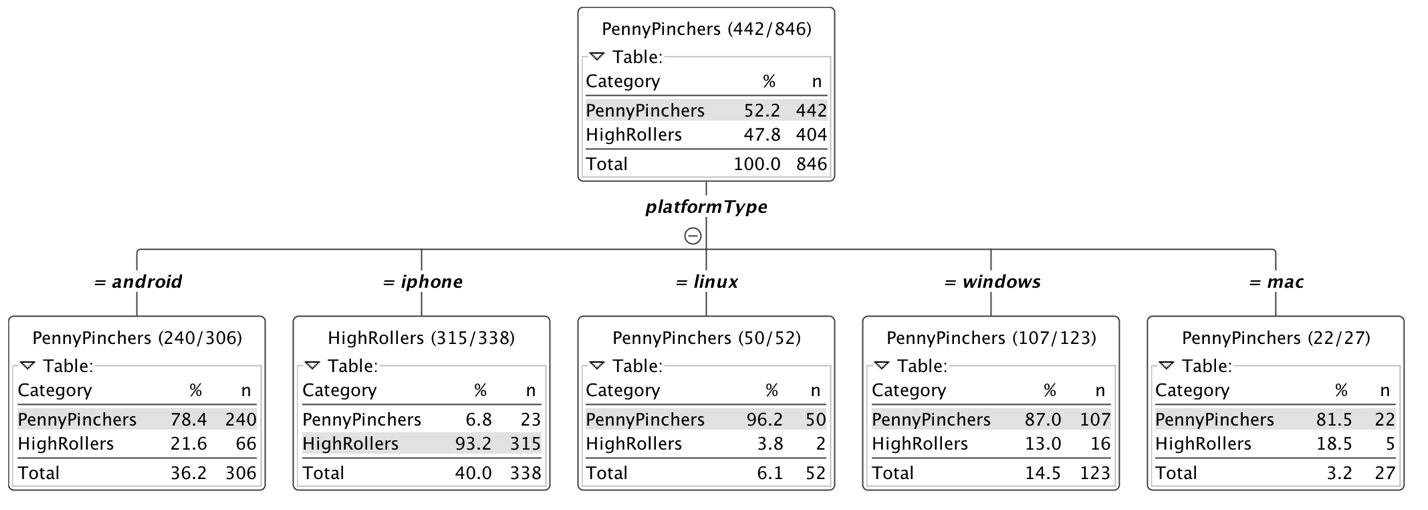
The **train** data set was used to create the decision tree model.

The trained model was then applied to the **test** dataset.

This is important because **we are trying to simulate data (with the train data set) that can occur in a possible future to know what will happen if we commit to the use of a model in making predictions. Then, we use our test data set to measure the accuracy, reliability and credibility of the model.**

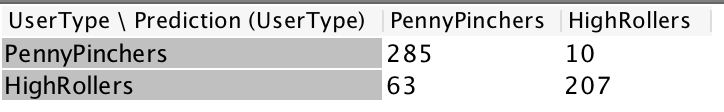
When partitioning the data using sampling, it is important to set the random seed because **we need to be able to reproduce our experiment.**

A screenshot of the resulting decision tree can be seen below:



**Evaluation**

A screenshot of the confusion matrix can be seen below:

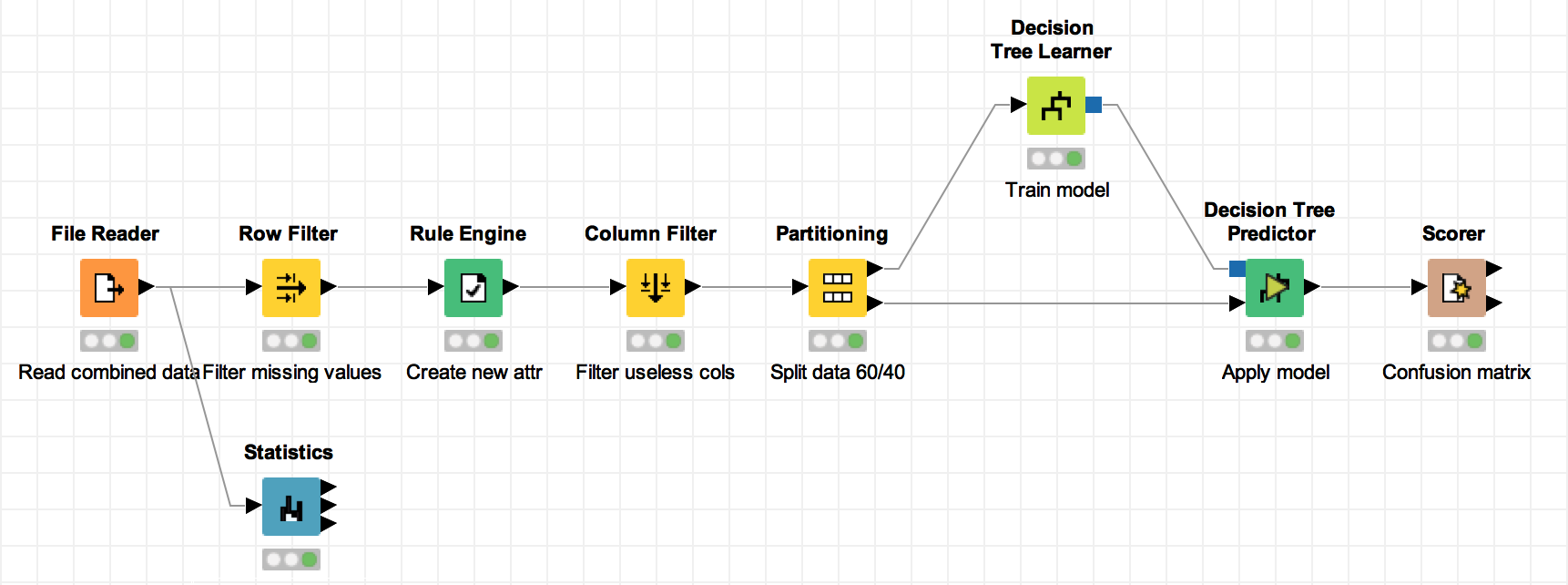


As seen in the screenshot above, the overall accuracy of the model is

* **285 users were correctly identified as PennyPinchers**
* **10 users were wrongly identified as HighRollers**
* **207 users were correctly identified as HighRollers**
* **63 users were wrongly identified as PennyPinchers**

**Analysis Conclusions**

The final KNIME workflow is shown below:



What makes a HighRoller vs. a PennyPincher?

**Based in the classification analysis, we found that platform used by the user, determine strongly the type of consume in the user.**

**In our sample, we found that 40% of the users play Pink Flamingo through iPhone, moreover, 93% of those users are HighRoller users.**

**Linux users present the same characteristic for PennyPincher profile. 96% of Linux users are PennyPicher users, but they only represent 6% of our sample.**

|  |
| --- |
| **Specific Recommendations to Increase Revenue** |
| 1. **Offer more attractive promotions for iPhone users, maybe an aware for n-buys in a month.** |
| 2. **Investigate the reason why users in Linux are not buying items. Realize an analysis for the rest of platform types to determine more strategies for them.** |