**Movie Recommendations using RapidMiner**

1) Importing the data set:

The available dataset are:

1. LINKS.csv: contains mapping of MovieId to IMDBid to TMDBid  
 2. movies.csv: contains movie names, genres  
 3. TAGS.csv: tagged by users and

4. RATINGS.csv: which contains movieId, userId, ratings and timestamp.

The data of interest here is RATINGS.csv. We import it RapidMiner in the following manner:

RapidMiner UI -> View -> Show Panel -> Repository panel ... -> add data

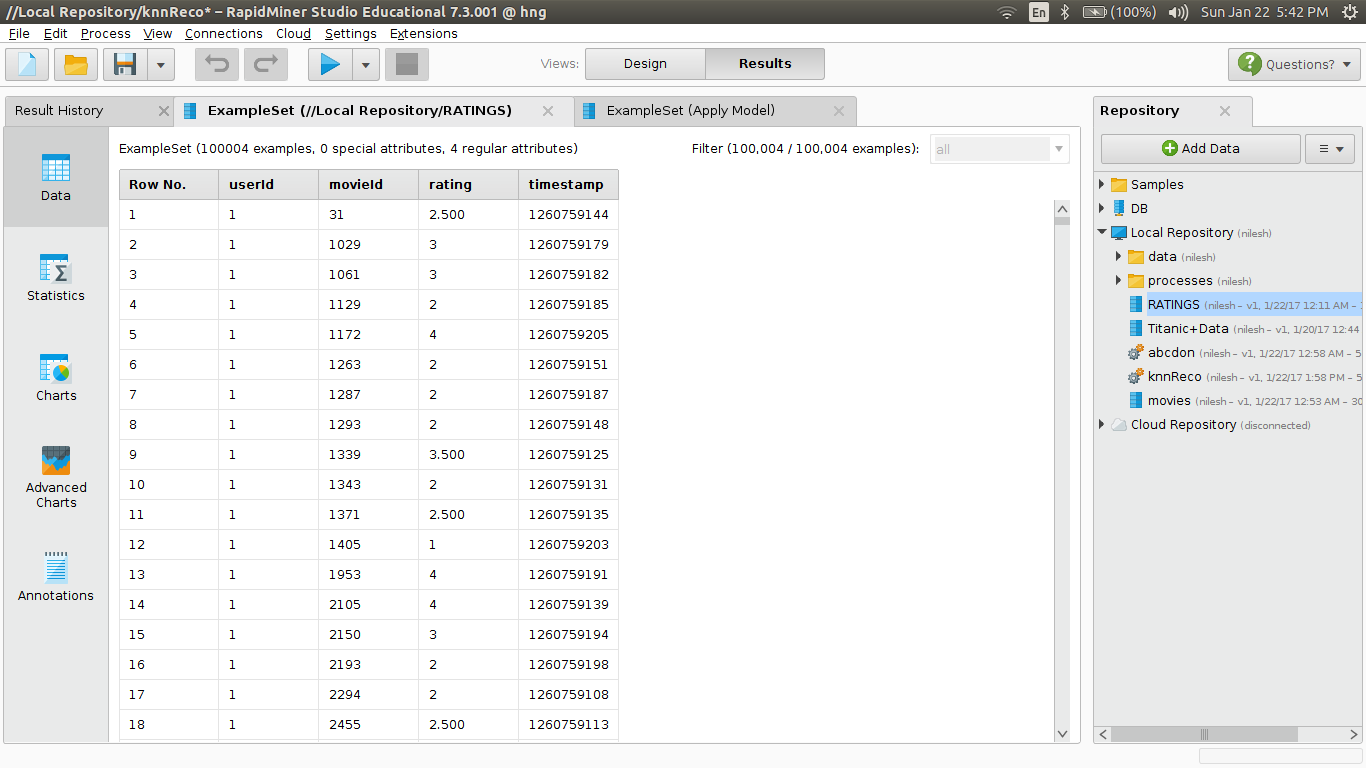


Fig. 1: Input data set: with 100,004 rows and 4 columns.

Now go to Repositories -> Local Repositories -> RATINGS.csv and drag it to Process panel.

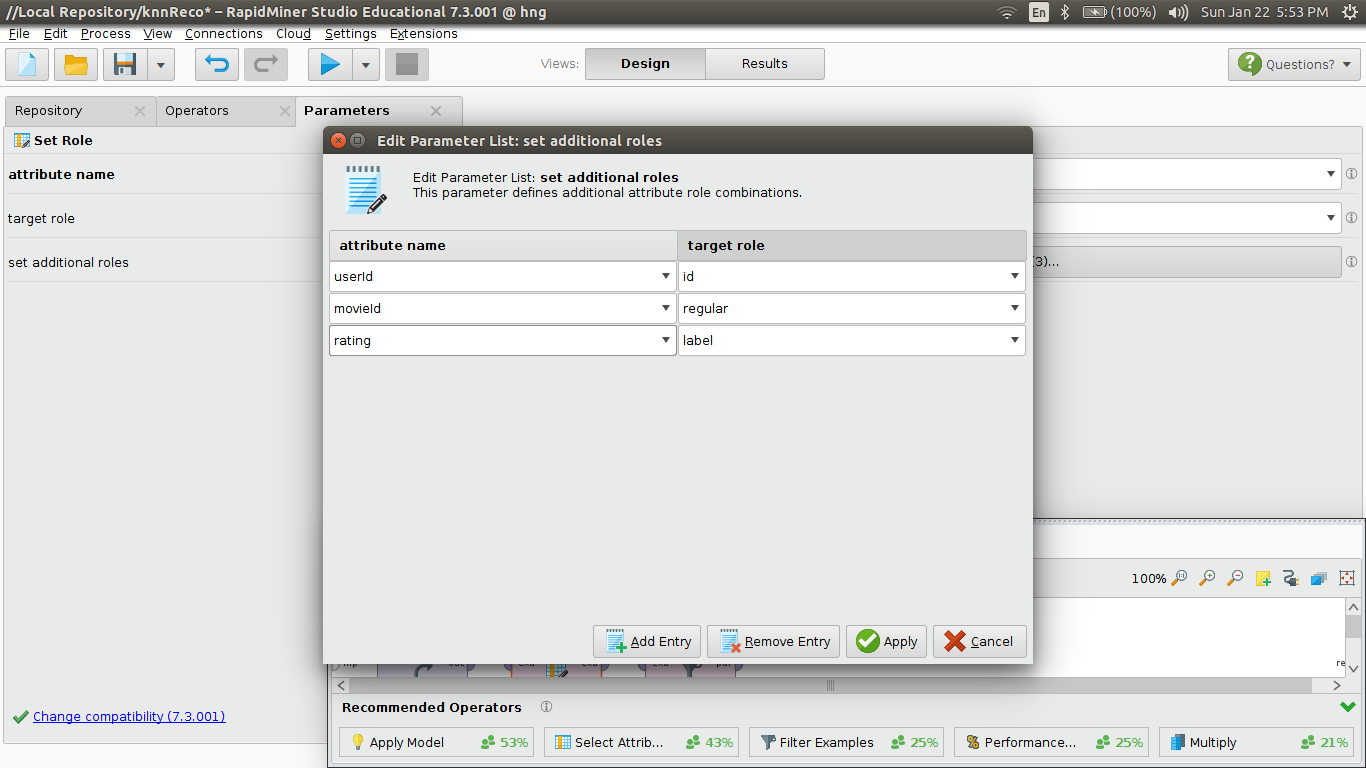
Next, since this is a raw file, we need to set up which columns represents what.

Next, go to Operations panel, drag Set Role operation to Process panel and change the parameters of that operatio in Parameters panel as follows:

Attribute name= movieId

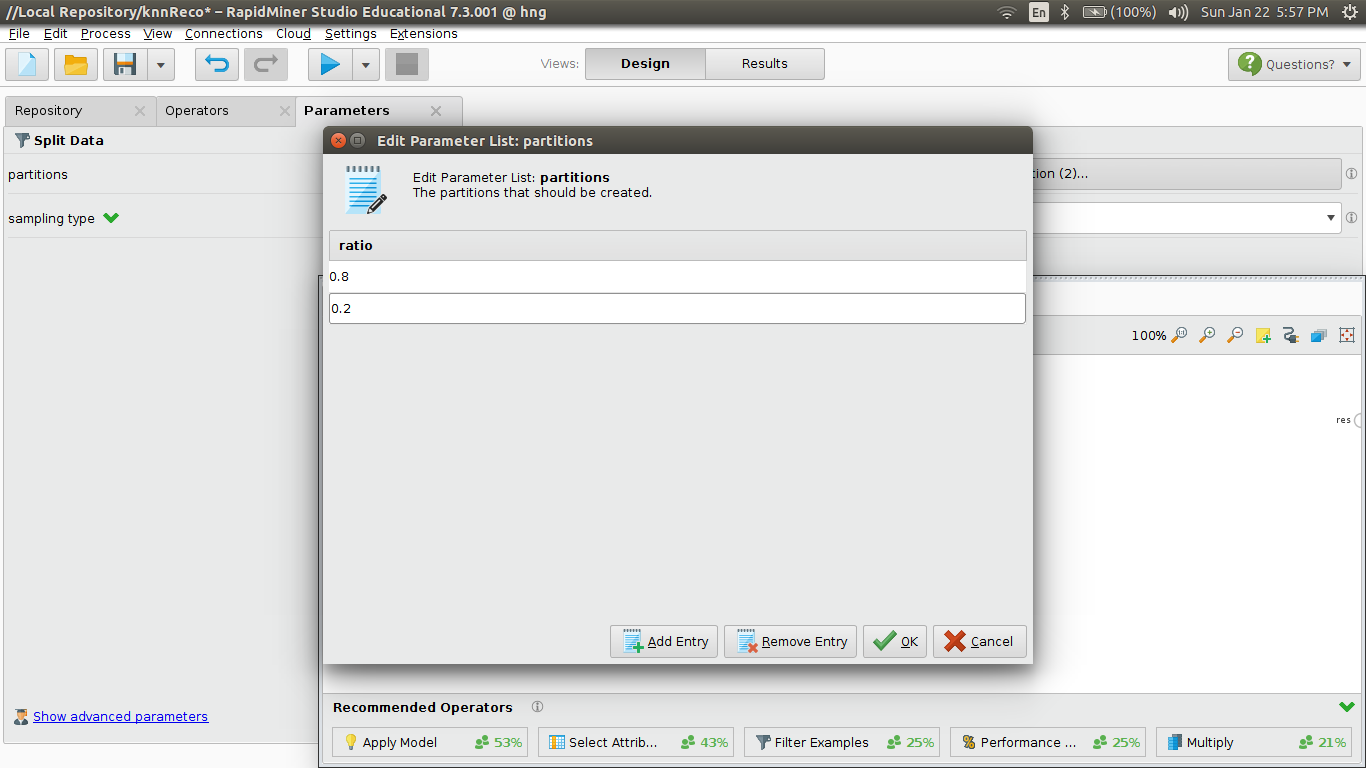
target role = regular

and set additional roles as and clikc apply. Now, the data is in machine readable format.



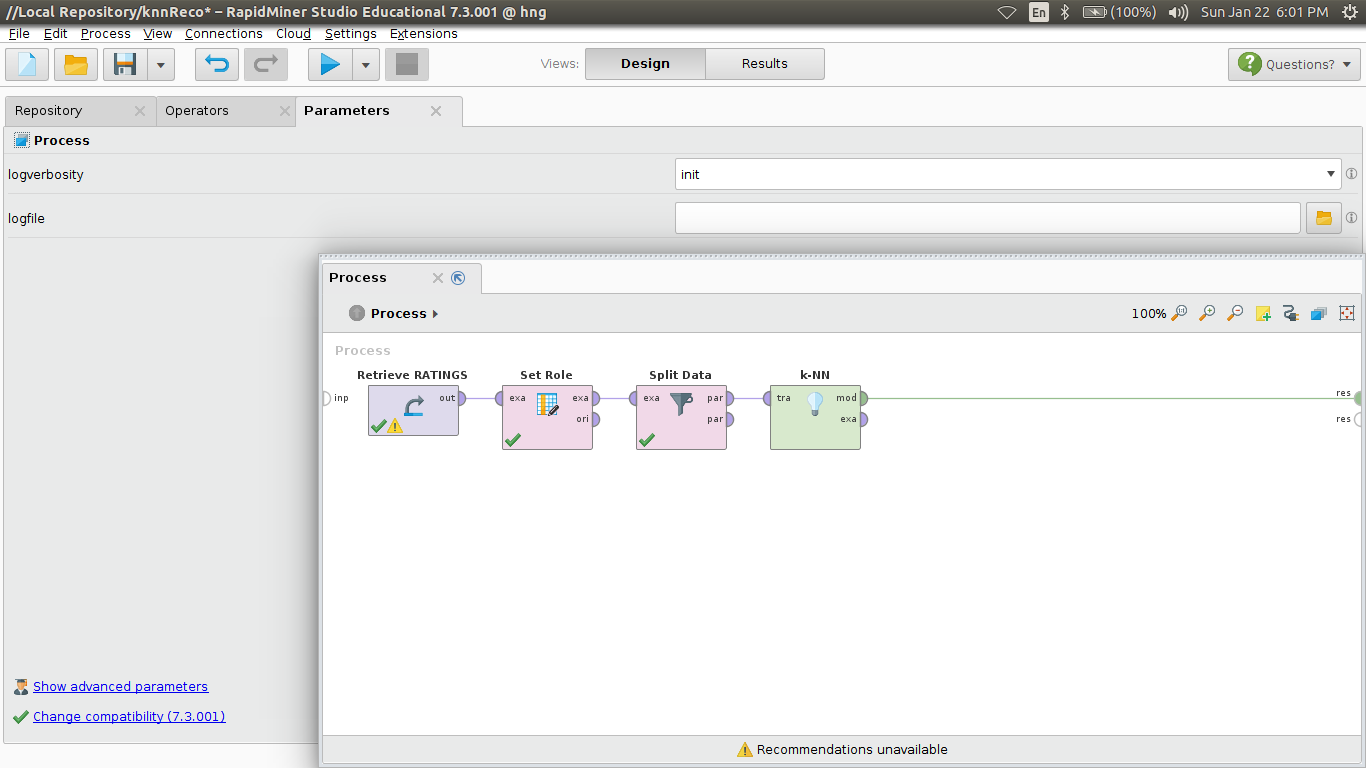
2) Split the data frame into training and testing sets: 80% and 20% resp.

Operators -> Split Data to Process panel. Go to the Parameters panel of this operation and set appropriate shuffling method. Next click on Edit enumeration and select number of partitions with their share; all must sum to 1.



3) Training the model:

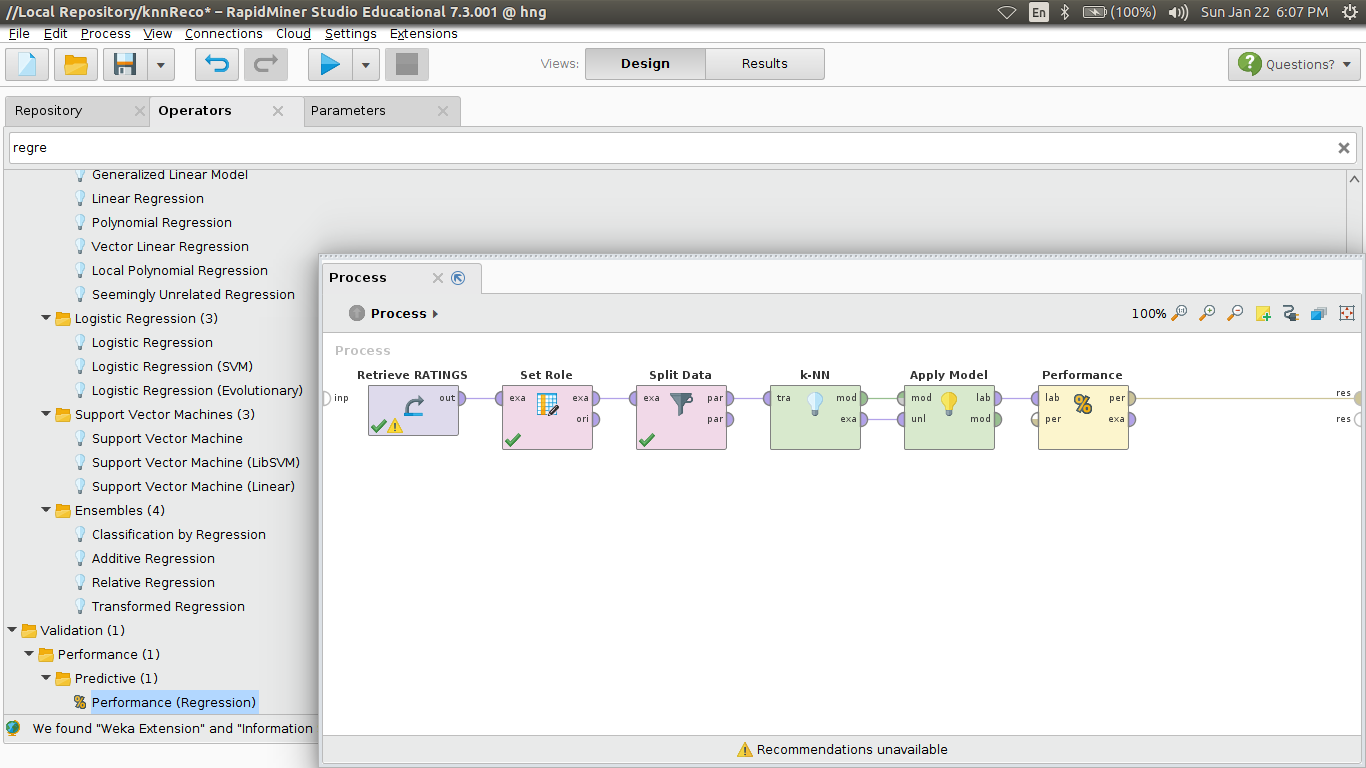
For the illustration purpose of this, let's select the simplest model k-NN and join the training partition node to the train node of k-NN operator. Edit the properties of k-NN in Parameters panel: set k = 10 and appropriate distance/similarity matrics. Now join the model node of the k-NN operation to result node and run the process.



4) Evaluationg model performace:

Now that the model is trained let's evaluate it's performace on the training data itself.

For this, use Apply Model and Performance(Regression, becuase our values are float.) operations and connect the nodes carefully as shown:

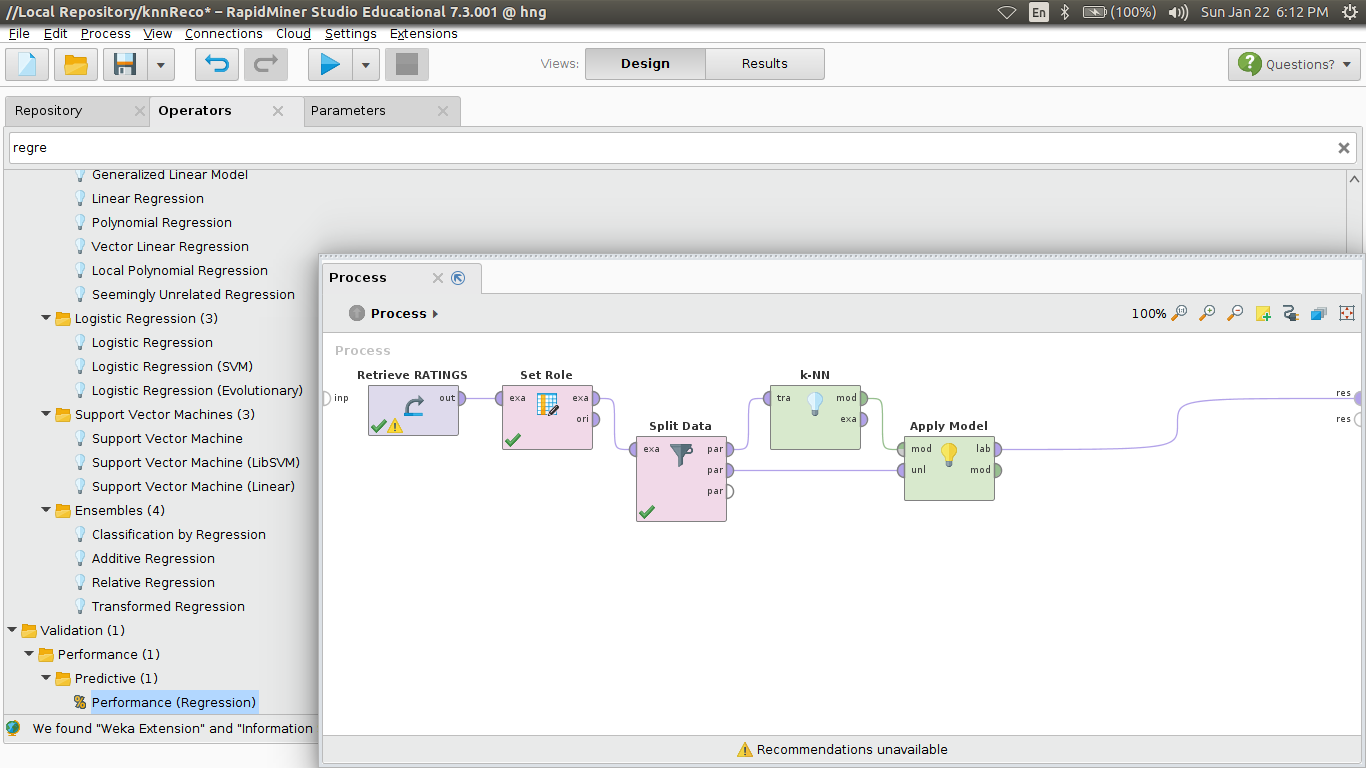


Before running this process, select all the metrics you want to know from the Parameters panel of the Performance operation.

Runnung this process will return the desired metrics for the model performance.

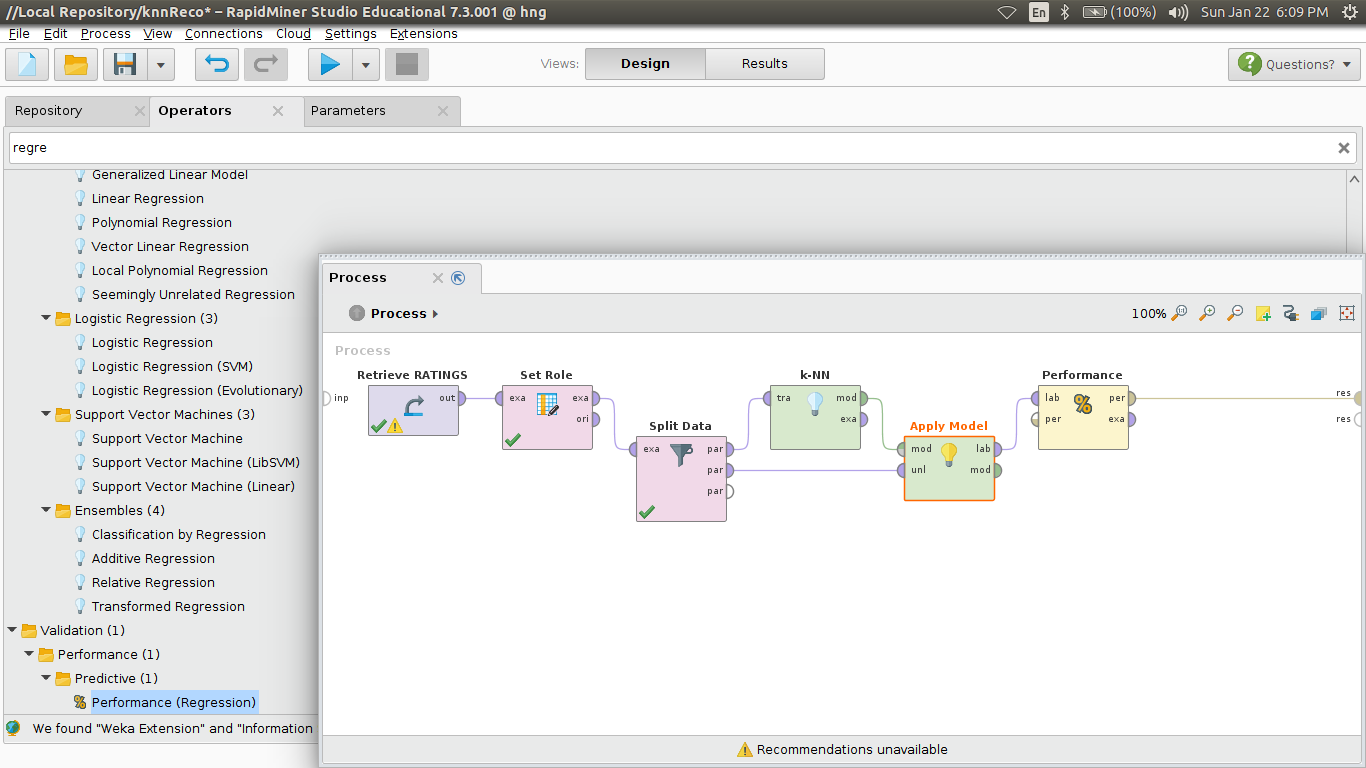
4) Model prediction on unseen data:

In this case, we use the testing data as an unlabelled data for the node of the Apply Model as shown and running the following process will give ratings as prediction of the model:



5) Performance evaluation on unseen data:

Now, to evaluate the performance of model on unseen data, add a Performance (Regression) operation to the Process panel as follows and edit the parameters of the operation, returning the desired metrics of model evaluation such as RMSE.



This model tries to group user according to the movies rated by them. When a new user comes in that cluster, the requested movies are assigned an aggrigated score of the present users.