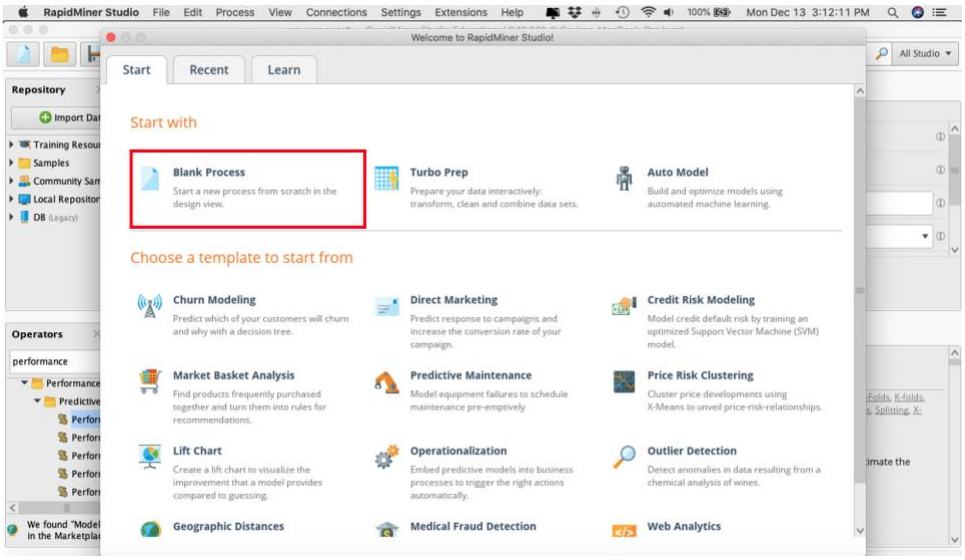
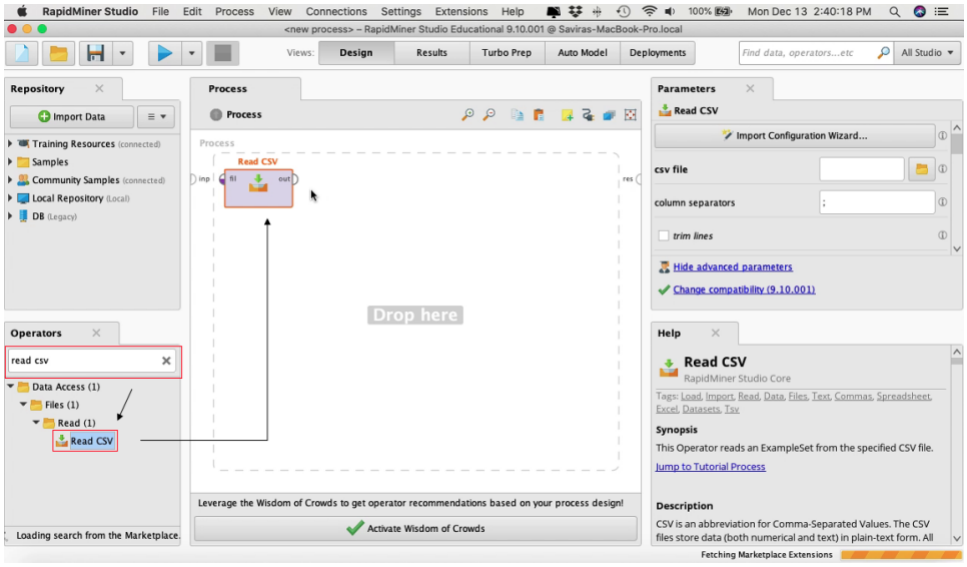
Data Processing using K-Nearest Neighbor (KNN) in RapidMiner Studio

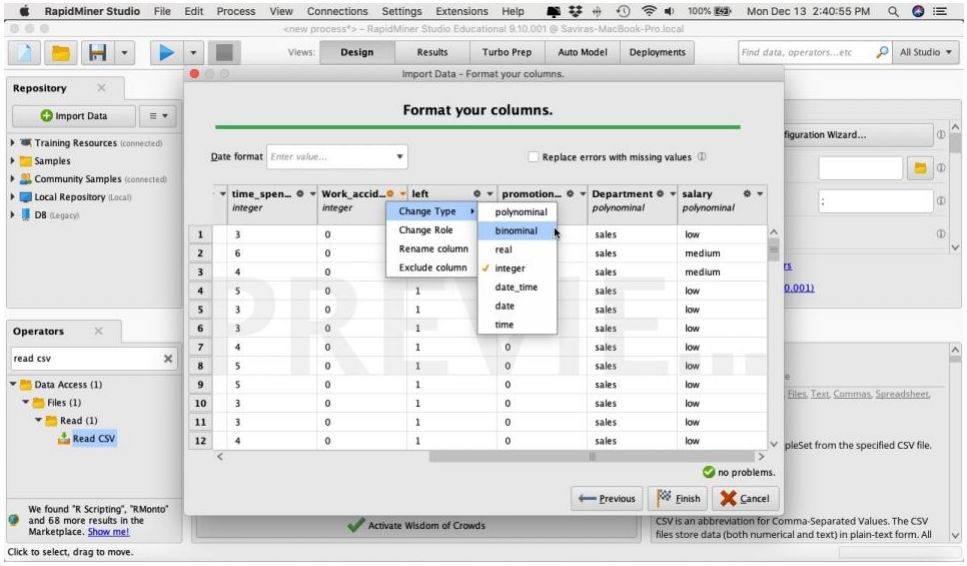
In this study, RapidMiner Studio Version 9.10 was used. To open a new worksheet, first select New Process on the File menu, then select Blank Process.



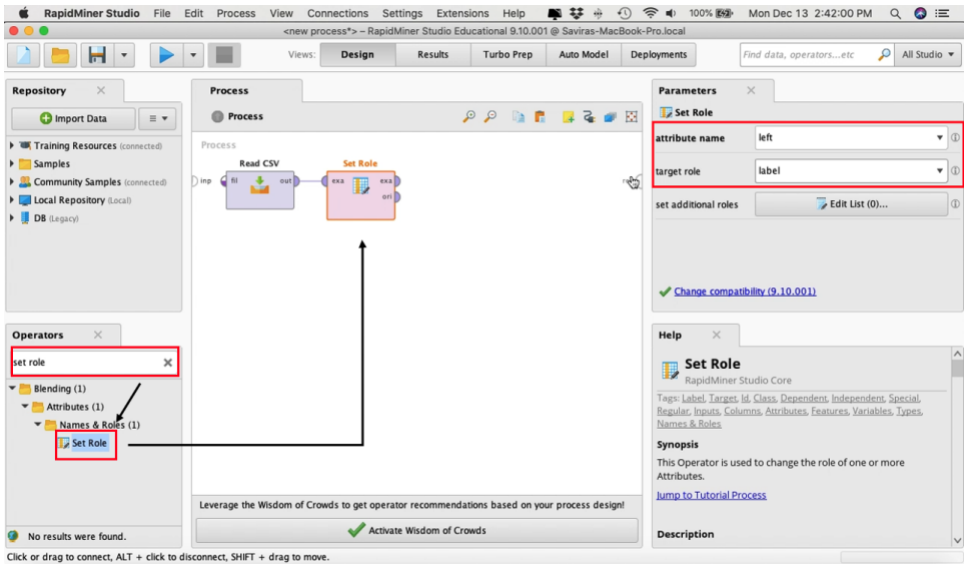
On the operator menu on the lower left, look for an operator named Read CSV, then drag it into the blank worksheet named Process.



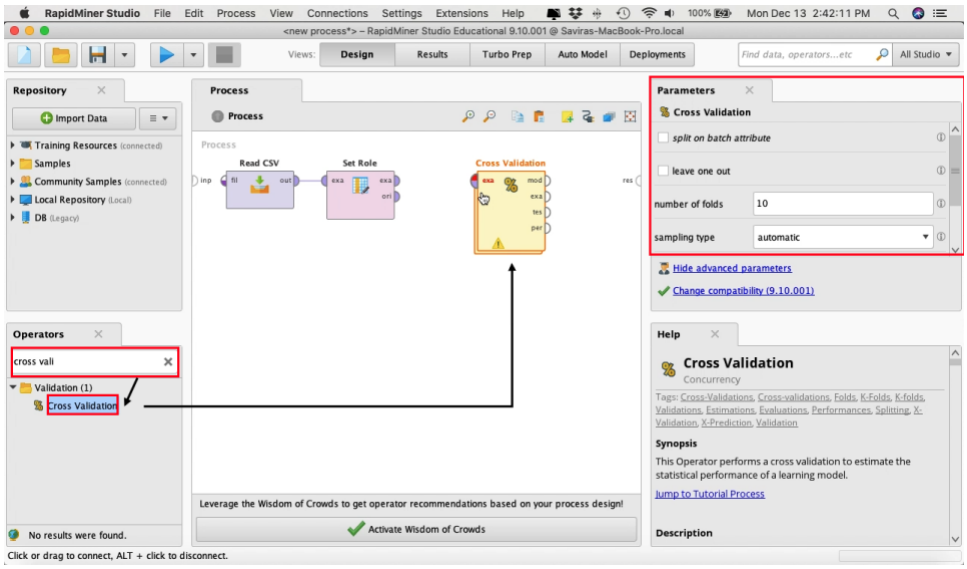
Click the Read CSV operator to open the dataset. In this data, *the work_accident, left, and promotion_last_5years* are detected as numeric variables. To change it to binary variables, Change Type command is called, then the binomial attribute is selected.



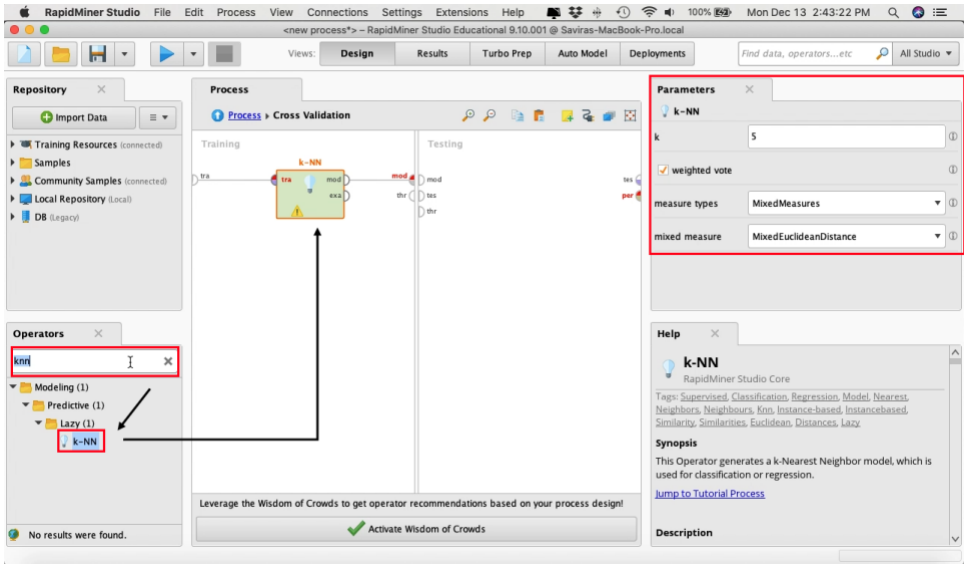
Next, in the operator column, look for an operator named Set Role and drag it into the worksheet. The point on the output of Read CSV is then connected to the example point of Set Role. Set Role is used to define the target response variable. In the parameter column on the top right, *left* is selected as the target variable we choose.



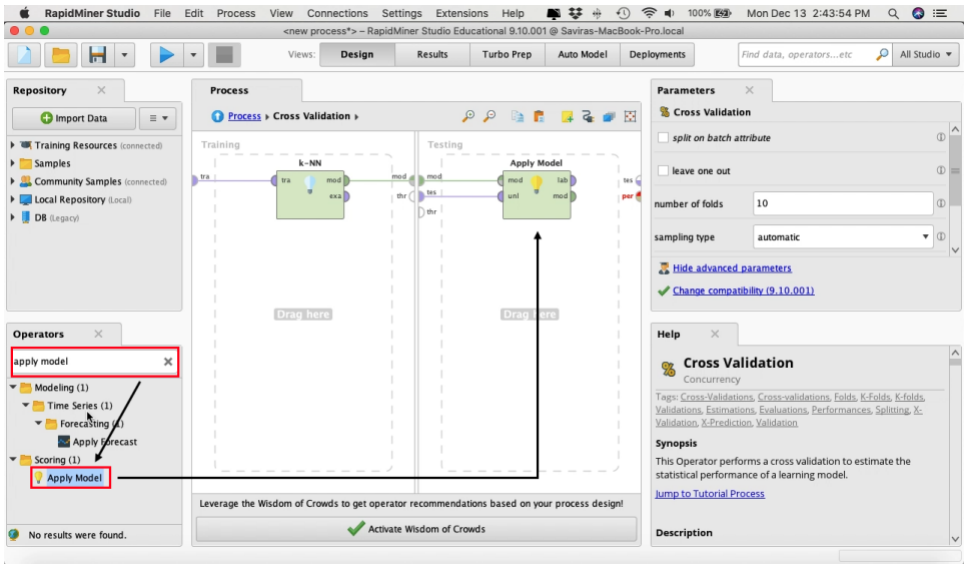
In this study, the partition of training data and testing data was carried out using cross validation. On the operator menu, look for Cross-validation, then drag it into the worksheet. In the parameter tab on the top right, there is a setting to decide the number of folds to be used. This time the default number of folds is used, which is ten folds.



In Cross Validation menu, there are two tabs, training and testing. In training tab, drag the KNN operator which can be found inside the operator column. In the parameter tab on the top right, there are several parameters that need to be changed, namely the number of *k*, type of combination function, type of distance calculation, and distance calculation method. In this study, the default setting of RapidMiner was used, namely five *k* pieces, weighted voting as combination function, calculating distances with MixedMeasures to accommodate numerical and categorical variables, and Euclidean distance method to calculate distances.

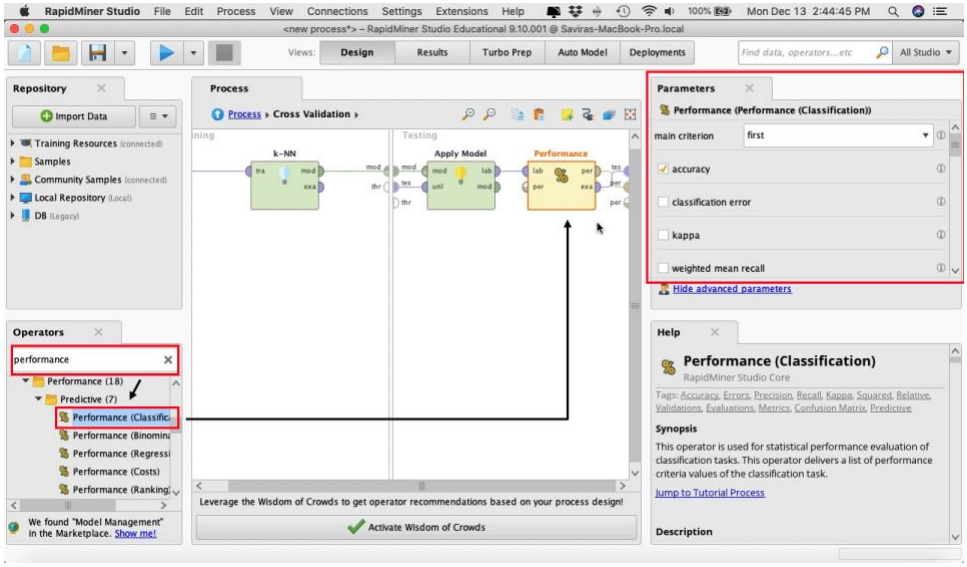


To apply the model obtained from training process to testing data, Apply Model operator is used. The point from KNN is connected to the point in the Apply Model to tell which model will be used in predicting testing data. The testing data point is connected to the unlabelled data point in Apply Model to indicate the data as testing data.

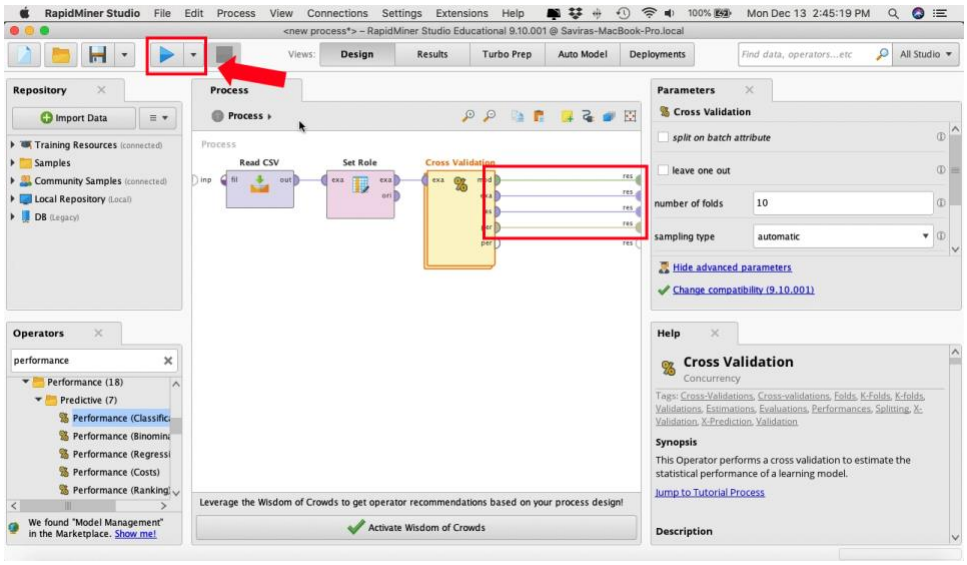


To assess the performance of KNN model, Performance (Classification) operator is used. The accuracy option was chosen as we want to find out the accuracy of the model in predicting employee turnover. The two *labelled data* points in Apply Model and Performance are connected.

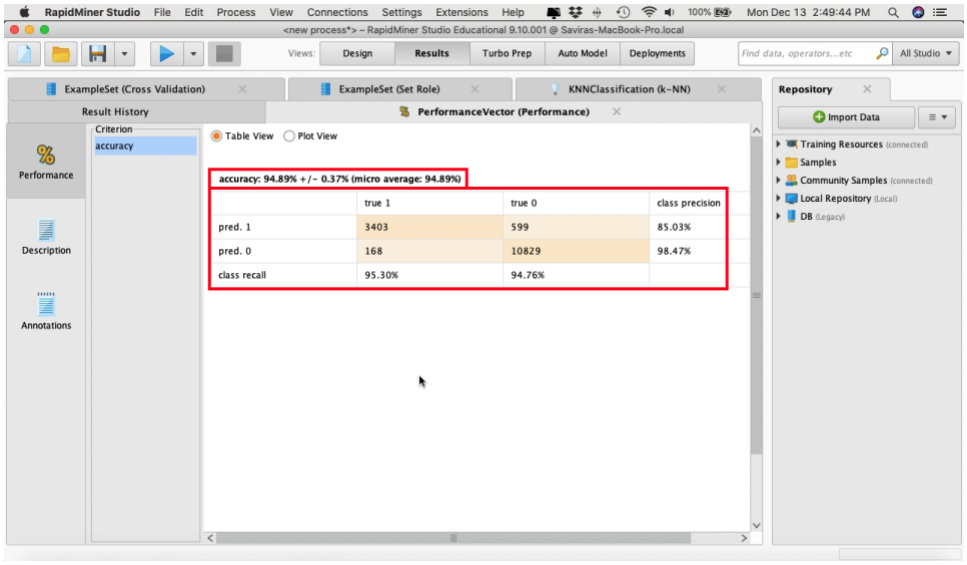
Then the performance point in Performance is connected to the performance point at the end of Cross Validation. The performance example point in Performance operator is connected to the testing data point at the end of Cross Validation which will be compared to determine the accuracy of the model in making predictions.



Once the process in Cross Validation window is complete, the model can be run immediately. The model, example, testing, and performance nodes are each connected to the result endpoint. Make sure each operator is connected to another operator for the model to run.



When the process is complete, the output will appear as a confusion matrix as shown below.



Result

The output generated by RapidMiner Studio can be seen below. The accuracy of the model in predicting employee turnover is 94.89%. This means that the model's ability to predict employee turnover is very good.

accuracy: 94.89% +/- 0.37% (micro average: 94.89%)

	true 1	true 0	class precision
pred. 1	3403	599	85.03%
pred. 0	168	10829	98.47%
class recall	95.30%	94.76%	