

CISC 333: Project Report

Speed Dating Matching

Name: Sam Ruscica

Student Number: 10148585

Date: December 1, 2017

Instructor: Professor Skillicorn

Project Outline

The purpose of this project is to demonstrate the knowledge gained from the course material and skills learned throughout the assignments. This will be done by making the most accurate model possible with the given data set. The data is a collection of surveys taken from many different waves of people. Each person gave basic information about themselves and ratings of the other people they met. Most importantly, the individuals also said if they wanted to match with the person they met. By using these attributes, a model will need to predict if the individuals would have matched with each other or not.

Project Approach

The approach for this project will be splitting the project into four main sections; model selection, data manipulation, initial results, and refining. The overall goal will be to surpass the baseline of the dataset. By surpassing the baseline, it will show that the model is an improvement over intuitive guessing. As this is the largest dataset handled in this course, it will be important to make decisions carefully as slight differences can have large results. This design will follow a cyclical approach meaning if the results are unsatisfactory, any step in the design process can be re-visited and adjusted to account for what was learned from the poor results.

Model Selection

Selecting the right model type for this project was essential in saving a lot of time. The time saved would come from not investing time into a non-complementary modeling type for the dataset. The model was selected by looking through the data while keeping the target classes in mind. The data set showed sporadic clustering and that there were two classes, match and don't match. Based on these preliminary findings, it was concluded that several modeling types could be excluded. The first model to be excluded was the decision tree as it is typically only suited for handling more simple divisions between classes. The next model type to be excluded was neural networks as the data set is large and would make runtimes very long. The last two options are the SVM and random forest model. Both appeared to work with the dataset based on initial observation. They were also the top two model types from the assignments which focused on the two-class titanic dataset. A SVM model was selected over the random forest as SVM models are suited for two class data sets by design. Additionally, random forests are an opaque, "black box", and might be regulated in the future.

Cleaning the Data

Manipulating and cleaning up the data is a crucial step in making the model. It has the ability to adjust the attributes to be more suited for the model type as well as save time by removing unwanted data. Due to the large nature of the dataset, cleaning the data was especially important to the accuracy of the model. It appears that as the study was conducted, several attributes were added in as well as attribute ranges were changed. This shows that data mining was an afterthought and will make cleaning the data more difficult. Additionally, data manipulation can be adjusted multiple times throughout the design process for the model. The outlined procedure below is what was planned for the initial model but will be changed once in the refinement stage.

The first step in cleaning the data is identifying and understanding what the different attributes represent. After going through the provided data key file, several attributes were determined to be problematic. To address these attributes, several different correction techniques were used.

The first technique was simply removing the attribute all together. The attributes were either removed because they mimicked the match attribute or because the data was uncorrelated to the match attribute.

The next technique was filling in the missing values throughout the dataset. This was done by replacing the missing values with the mean of the values. This choice may be an issue but will be explored further in the refinement step.

The last correction technique was scaling the data to be in the same range. This was an issue first noticed for waves six through nine as they used a scale from one to ten. The other waves used a system that was one hundred points allocated over the different attributes. The first step in adjusting the small scale was multiplying the values by ten to bring it into the larger range. Then all the values were summed and divided by one hundred, this new ratio was used to divide the values that were multiplied by ten so that it mirrored the one-hundred-point distribution.

Initial Results

In order to evaluate the accuracy of the model, the baseline had to be determined. This was done by first finding the maximum number of potential matches. This ended up being the total number of rows, 8281. Next the number of actual matches was totaled by summing the match attribute. The value of total actual matches was 1365. This give a baseline of 83%, since always guessing "no match" would result in 83% accuracy.

The initial results for this project were surprisingly lower than expected. The linear SVM had an accuracy of 75%, with no data manipulation. Upon trying out different polynomial SVM models, none of them surpassed 78%. The addition of the purposed data manipulation increased the best SVM to 81%. As none of the SVM models were able to match the baseline, a new model needed to be selected. This was chosen instead of modifying the data further since the SVM model might not suit the data.

Refining

Refining the system is done once the initial results are collected and evaluated. The changes made during the refining stage can be small or large depending on the results. Once a change is made, the results can be immediately checked to see if they resulted in an improvement on the overall system. Below are the changes made to each stage.

Model Selection Refining

Changing the model selected can be beneficial in seeing how the current model type compares to a different one. This can be beneficial in two different ways. The first one being, it can be used for confirming the initial choice of model. That is, if the new model used performs worse than the current one, the current model suits the dataset better. The other benefit being, it can be used in determining a new model to use if the current one is unsatisfactory. This can sometimes be hard to determine if lots of data manipulation was done to compliment the older model. This can be somewhat avoided by first seeing how the models compare without any data manipulations.

Due to the poor results of the different SVM models, a new model type needed to be selected. Looking back on the original model selection step, the SVM was chosen over the random forest due to slight possible downsides of the random forest. Since the random forest is the next best option, it was selected as the new model.

Upon changing to the random forest model, the accuracy was far higher being 83.5% with no data manipulation. At this point the model has already slightly surpassed the baseline for the dataset. When the data manipulation was added back in, it improved again to 85%. Since, the new model was far better than the old one and it surpassed the baseline, it will be the new model type for the analysis.

The random forest was then adjusted to try and find the best number of models to use. After an hour of trying different model counts, the results for 100, 500, 1,000, 1,400, 1,500, 1,600, 2,000 were found. They were 84.1%, 84.5%, 85.0%, 85.4%, 86%, 85.6%, and 85.2% respectively. This is an expected outcome as there is a gaussian distribution around the ideal point and the best number of models found was 1,500.

Cleaning the Data Refining

Cleaning the data can be a fantastic way to make improvements to the overall accuracy. Typically, these slight changes to the data manipulation is handled using a greedy attribute technique. This can be handled in a positive or negative way where you either add more attributes to the model or take more away. You would continue doing this until the model showed signs of loosing accuracy.

Another way of refining the data is looking at prior data manipulations and evaluating their effectiveness. This can help confirm what you have done in the past is correct or incorrect. It is very beneficial to notice an issue that data miner has introduced rather than the data set.

Lastly, new data manipulations can be introduced based on the findings from prior results. This could lead to slight or large accuracy changes. Tools like the confusion matrix can be helpful in looking at the kinds of errors occurring. Sometimes certain errors are less harmful than others.

The first change to the current model was using the greedy attribute technique to add in more attributes at random to see if the accuracy improved. The SAT attribute was beneficial but required it to be converted in to a number from a string. It helped increase the accuracy by 0.2%. More attributes were introduced but upon doing so, the accuracy decreased by 0.5% for the 5 different attributes tried. This means that the initial attributes filtered out were chosen correctly. The next change was trying the greedy attribute technique to remove different attributes to see if the accuracy improved. This was more difficult as many attributes are part of wave sets. After trying several different attributes, I found that removing the "Round" attribute helped by 0.2%. I continued to look for more attributes that improved the results when removed but stopped looking after four attempts that decreased the accuracy.

Looking back on prior data manipulations turned out to be very beneficial. The first thing discovered was that there were many missing values in the dataset and the current model had everything filled in using the mean for each attribute. This would cause the overall attribute to fall even closer to the mean since so many values were the mean. To counteract this, a Java Snippet was used to check how many missing values were tied to each person. If the person had more than half the values missing, it was removed from the dataset all together. This change improved the accuracy by a further

0.3%. Upon exploring the issues with missing values, a histogram was used to see the trends in the data. It was noticed that men and women had different mean values for their attributes. To account for this, the men and women were split into two different sections before performing the missing value mean fill. This increased the accuracy of the model by 0.2%.

Results Refined

Once the refining step is complete, a final set of results is produced for evaluating. Overall the random forest model with 1,500 models resulted in an accuracy of 87%. This is 4% above the baseline and 6% above the SVM. Looking at the confusion matrix, it seems the 90% of the errors are caused in a false positive, meaning that it said the people would not match when they actually did.

Problems and Surprises

The task of understanding and manipulating a large dataset was extremely challenging. I found that one of the main issue occurred was simply the time it took to run the model. It was especially difficult during the refining stage where many slight changes had to be made independently and each checked independently. Another challenge was understanding the dataset initially. It was a lot of information to take in but slowly as the model took shape, so did my understanding of the different attributes and their relations to each other.

I was very surprised to see the SVM model not work well on this data. This might have been caused by how narrow the dataset was as SVM models like wide dataset. Apparently, SVM models sometimes just don't fit some datasets and this seems to be that case.

I was also very surprised to see how little my changes made during the refining step effected the accuracy. This might be caused by the fact that I was only changing the attributes I had already changed. If I had effected or combined new attributes it might have given more drastic improvements.

Possible Improvements

There are several things that can be done that have potential to improve the accuracy of the model if there is ample time. The first improvement would be going back and trying to manipulate the attributes that were not correlated to the match attribute. This would be challenging because there is no obvious correlation between them and it might require techniques that were not used in this project. Another improvement would be using the greedy attribute method again, trying to find more attributes or removing more attributes that would increase the accuracy. Due the to number of attributes it was only attempted a few times. This could also be done by looking into the random forest to see which attributes helped make the prediction more than others. Another improvement would be adjusting the number of acceptable missing values per row or changing how the missing rows were filled in beyond by just gender isolated means. It would also be beneficial in trying to reduce the number of false positive results as it would pair people up with would not work well together. The last improvement would be trying to find a transparent model to use. This could be beneficial in the future if the UN passes the "Right to Know" bill.

Conclusions

Overall, this project has stretched my skills and knowledge in the course. The task was to make the most accurate model possible by using the dataset, by using the different attributes given by the individuals in each wave of surveying. The attribute that the model is looking for is match. This shows if the individuals liked each other.

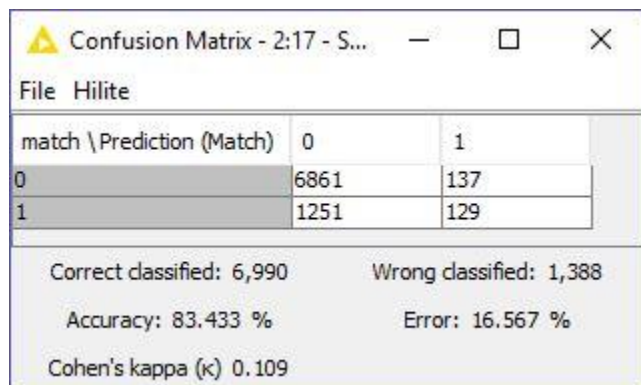
The model making strategy was outlined to be a cyclical four stage system. The first stage would choose the appropriate model to use. The second stage would manipulate the data accordingly. The third stage would evaluate the results of the initial model. The last stage would refine and then re-evaluate the model. If at any point the model produced unsatisfactory results, the model making process would cycle back.

Upon first pass through the model making strategy, the SVM was chosen as it seemed to fit the two class system. Many different data cleaning techniques were used to get rid of bad data, make poor data usable, and deal with missing data. When the model was evaluated, the SVM performed very poorly.

Due to the poor results in the evaluation stage, the refinement stage was extensive. It first chose to replace the SVM model with a random tree. Upon doing so, the new model performed far greater. Next it looked into the refinement of the data manipulation. This made a lot of progress in increasing the accuracy of the model as before being refined it increased the accuracy by 1.5%. This stage improved the accuracy by a further 0.9% after refinement.

The errors that due result from this model tend to be a false negative. This is beneficial for the type model used. This is because this type of data would most likely be used by a dating service of some kind. It would be fine for that service to not match you with someone you would have been good with but it would have been very bad if they match you with someone you would not be good with.

Appendix

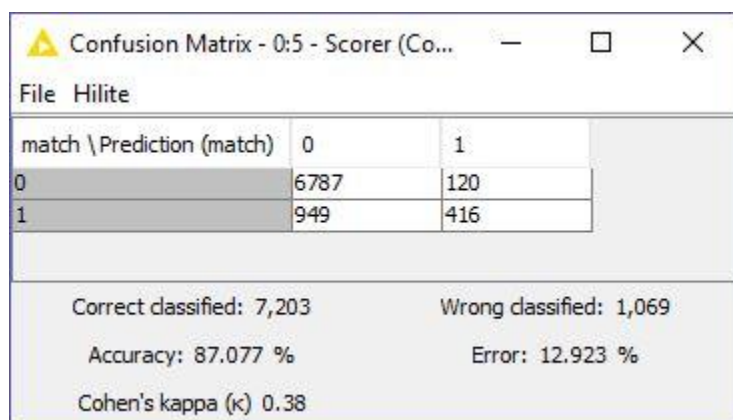


A screenshot of a software window titled "Confusion Matrix - 2:17 - S...". The window has a menu bar with "File" and "Hilite". It contains a table with two columns labeled "0" and "1" under the header "match \ Prediction (Match)". The rows are labeled "0" and "1". Below the table, summary statistics are displayed: "Correct classified: 6,990", "Wrong classified: 1,388", "Accuracy: 83.433 %", "Error: 16.567 %", and "Cohen's kappa (κ) 0.109".

match \ Prediction (Match)	0	1
0	6861	137
1	1251	129

Correct classified: 6,990 Wrong classified: 1,388
Accuracy: 83.433 % Error: 16.567 %
Cohen's kappa (κ) 0.109

Figure 1: Initial Random Forest Test



A screenshot of a software window titled "Confusion Matrix - 0:5 - Scorer (Co...". The window has a menu bar with "File" and "Hilite". It contains a table with two columns labeled "0" and "1" under the header "match \ Prediction (match)". The rows are labeled "0" and "1". Below the table, summary statistics are displayed: "Correct classified: 7,203", "Wrong classified: 1,069", "Accuracy: 87.077 %", "Error: 12.923 %", and "Cohen's kappa (κ) 0.38".

match \ Prediction (match)	0	1
0	6787	120
1	949	416

Correct classified: 7,203 Wrong classified: 1,069
Accuracy: 87.077 % Error: 12.923 %
Cohen's kappa (κ) 0.38

Figure 2: Last Random Forest Test

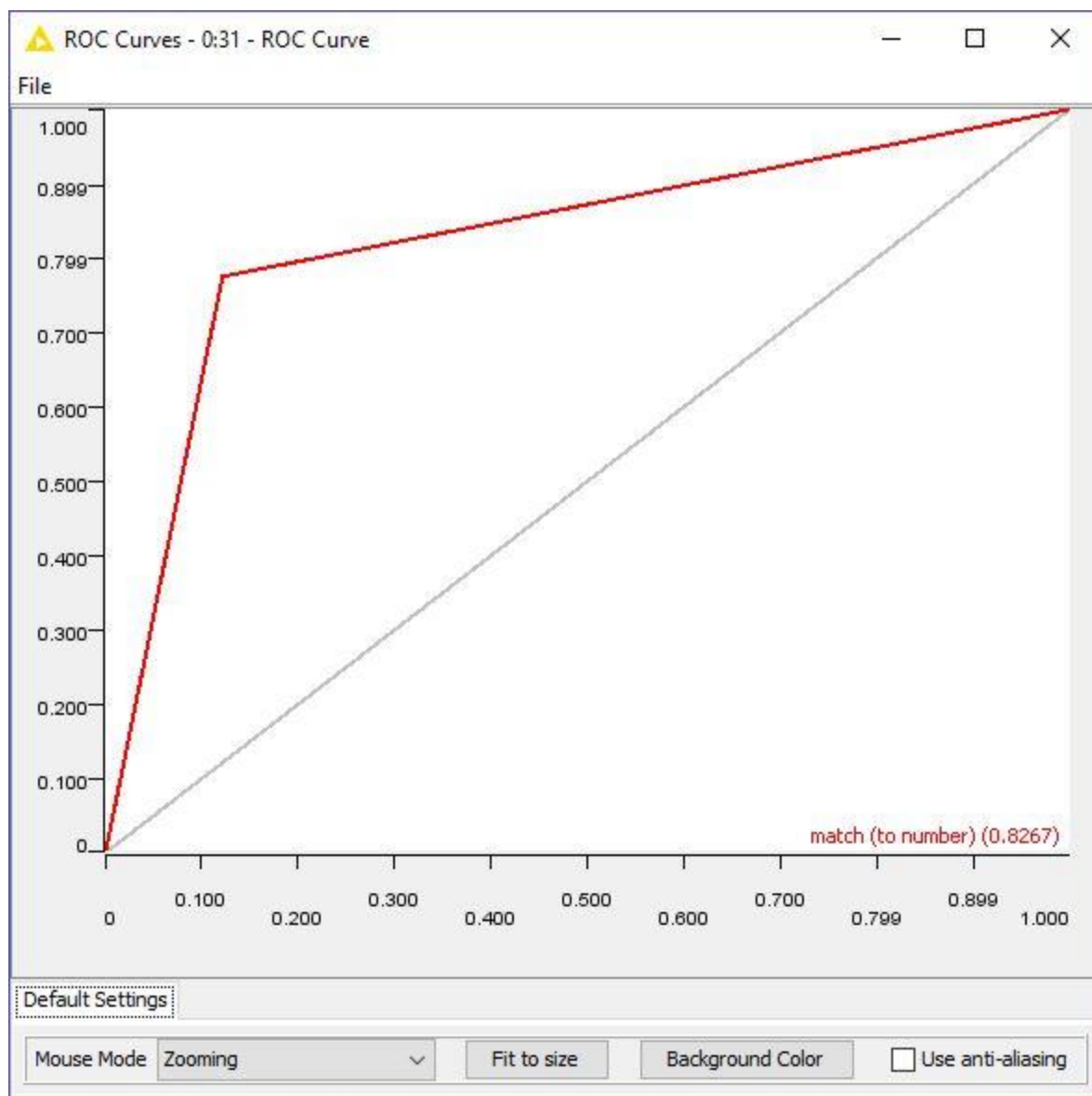


Figure 3: Last ROC Random Forest

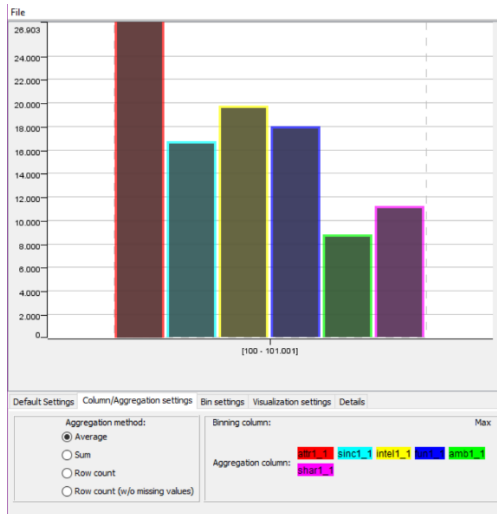


Figure 4: Male Histogram

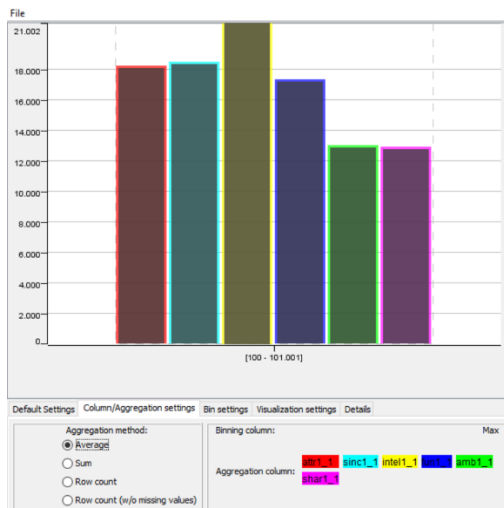


Figure 5: Female Histogram

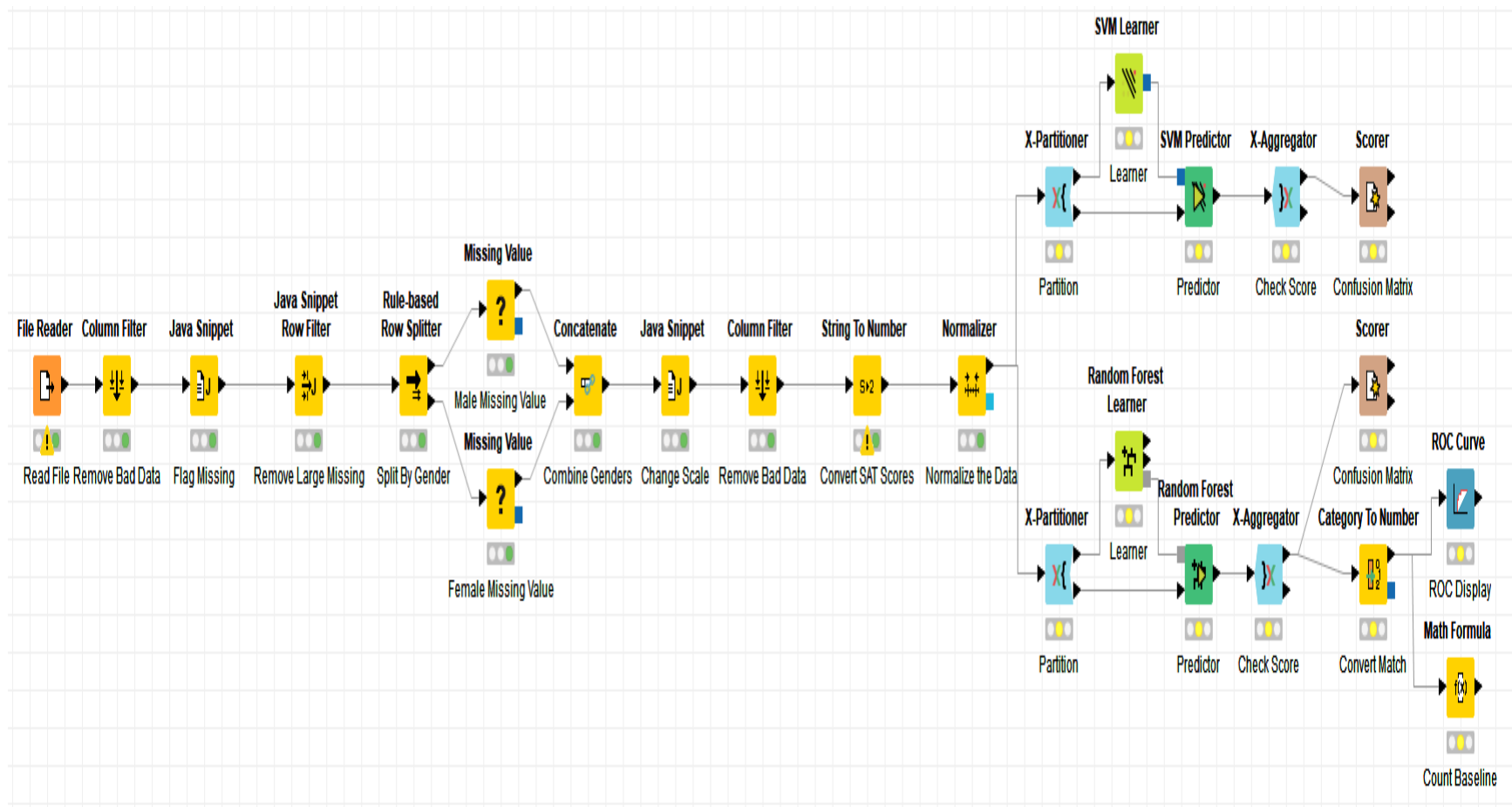


Figure 6: KNIME Nodes Used for Analysis