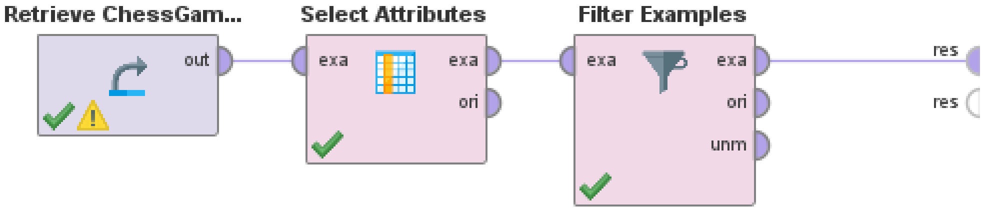
Data Mining Assignment Learning diary

Author: Javier Cordero Luna

ChessGamesOpeningBlackWhiteVersusVictory dataset:

Processes:

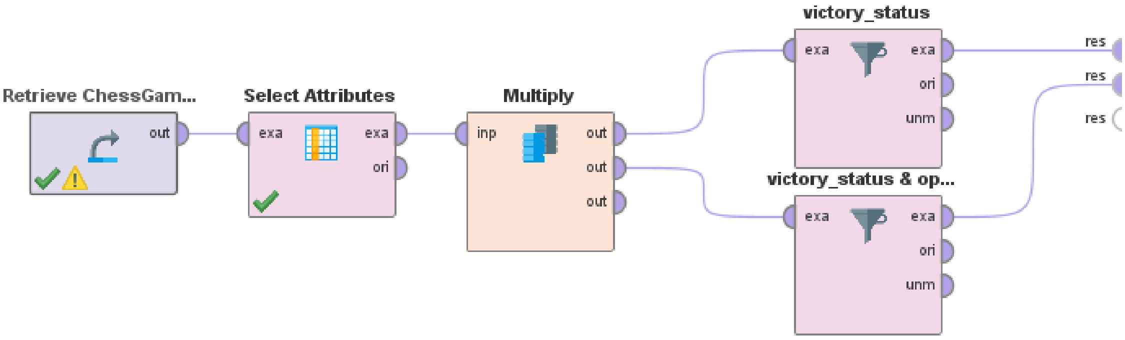
1. Retrieve data from the dataset.
2. Select the attributes: id, opening_name, turns, victory_status, and winner.
3. Filter by victory_status where it equals "mate."



Questions to be answered:

- How many moves, on average, are needed to win?
  - On average, it took about 65 moves to achieve victory by checkmate.
- Who wins more, white or black? What is the percentage difference?
  - In terms of the number of victories, we can clearly say that white has more wins, although the difference with black is not too large (52.87% white wins vs. 47.13% black wins).
- What is the most winning opening? Is it played more by white or black?
  - The winning opening is Van't Kruijs opening, with over 150 victories, compared to the second most winning opening, the Sicilian Defense.
  - Although the general percentage of victories is dominated by white, the pieces that play this opening the most are black, with more than 78% of executions.

These were the processes used to reach these conclusions:



In this case, and to answer the last question, I needed to apply an additional filter to obtain the data with the necessary clarity.

To answer the questions posed, I needed to have "id," "turns," "victory_status," and "opening_name" available.

Finally, to clean up the workspace, I included everything in a subprocess to make it more understandable.

By completing this first task with this dataset, I have a better understanding of how to retrieve and process the data to obtain what I need. It is true that I used the Rapidminer tutorial to make the learning curve less steep. With regard to the dataset, I have come to understand the correlation between openings and the color of the pieces.

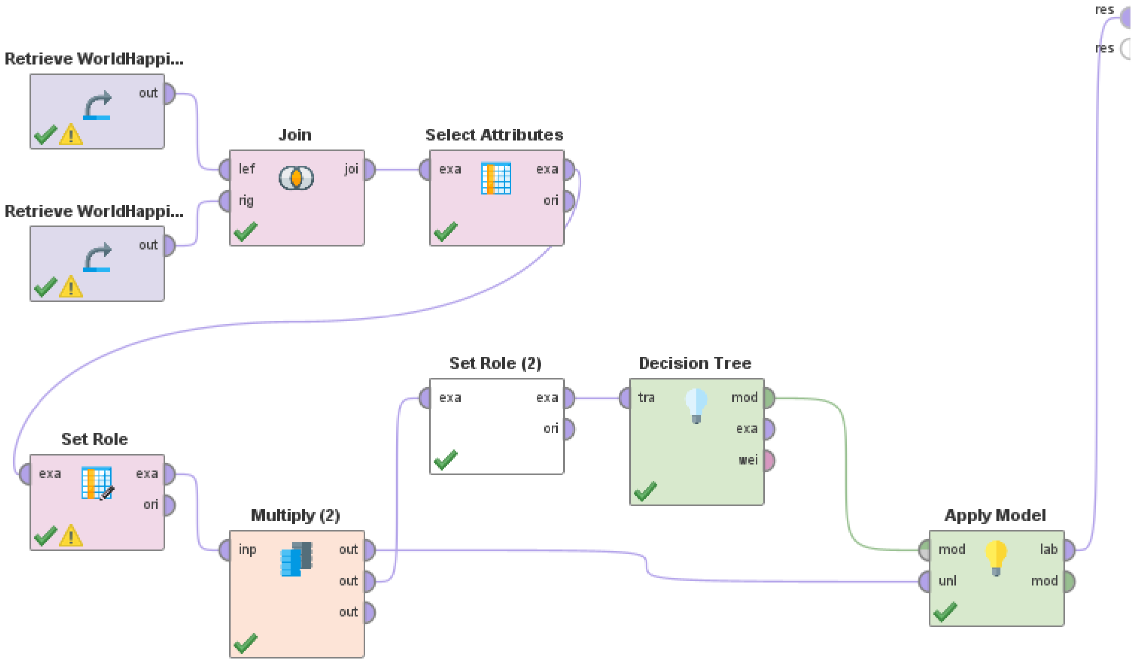
tripadvisor_HotelReviewRatingPrediction dataset:

I used the AutoModel tool included in RapidMiner to predict the rating for the hotel. After executing different models, we can conclude that the best rating could be a 5.0 (Support Vector Machine simulation), and the others are between 3.949 and 4.366, according to the different models run.

World Happiness dataset

Processes:

1. Retrieve data from both datasets.
2. Join them by country or region.
3. Select the attributes needed to conduct the study.
4. Set the roles so that a decision tree model can be run.
5. Run the decision tree model and apply the model.



As we can see, Finland continues to be the country with the best global score. In addition, at the global level, we see that intermediate scores prevail, not so much the extremes, resembling a normal Gaussian bell curve.

On the other hand, the distribution of Healthy Life Expectancy is not the same, as it is clearly skewed to the right, as is Freedom to Make Life Choices.