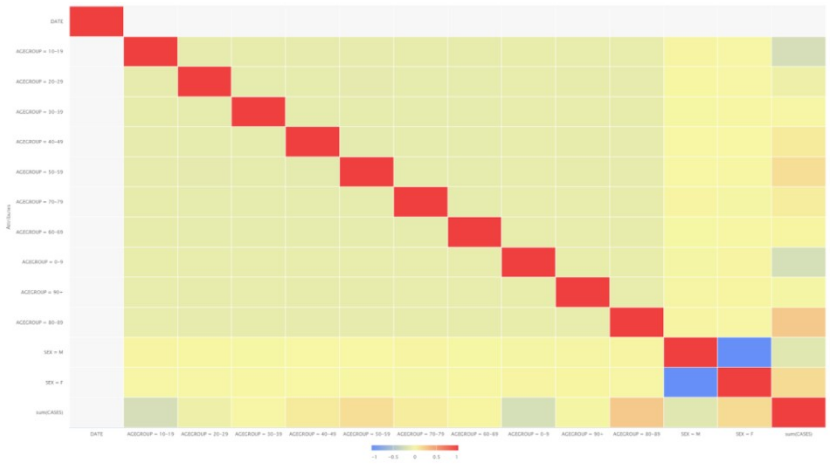
Zhou Menghui r0768060
Tran Tong Khanh Linh r0648586
Praveen Kumar Pratheek r0771488

Rapid Miner Assignment - Covid19BE

Data Analysis

1. When will the peak be reached?
   - This cannot be identified accurately because in the training set the number of cases increases from day to day.
   - Therefore a model in which there is an increase and then a decrease cannot be created and therefore a peak cannot be identified.
2. Has the curve flattened after the social distancing?
   - In the graph we can see that the number of cases has increased exponentially througout the given time frame, even after the implementation od social distancing measures.
   - However a case could be made that the number of cases might have risen even more rapidly if social distancing measures had not been imposed.
3. What should be the decision on lockdown, extend or not to extend
   - It can be seen from the April 15th figures that the number of cases was still increasing exponentially so at that time the lockdown should not have been stopped.
   - However now (May 30th) the predicted number of cases is much higher than the actual number of cases and therefore the lockdown seems to have made a huge impact and the danger of the coronavirus seems to have been nullified to a huge extent and the lockdown can now be stopped.

o Impact of age and sex group in the number of cases

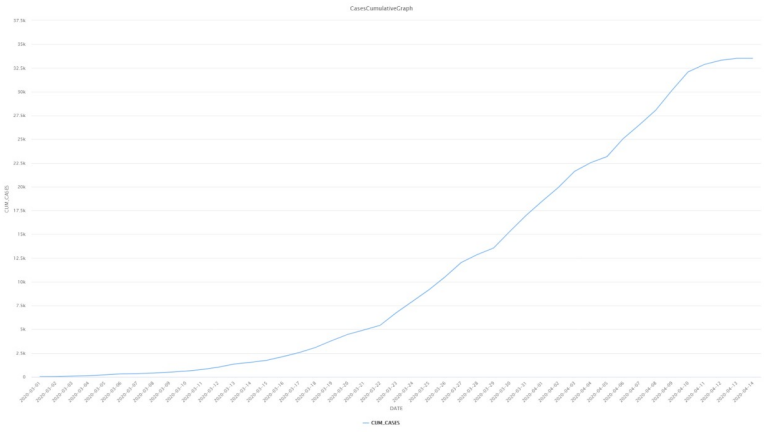


- From the above and below diagrams derived by using the correlation matrix operator, we can see that there is a slight positive correlation between the number of cases and deaths and the 85+ age group and a slight negative correlation between the number of cases and deaths and the age group 0-24.

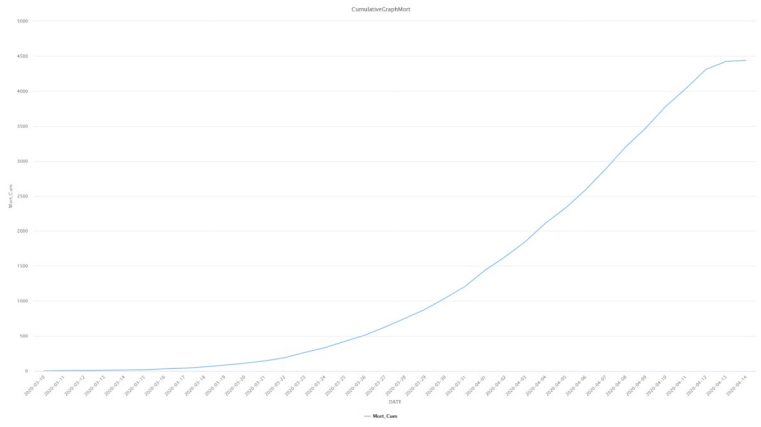
- The sex does not seem to have any meaningful correlation with the number of cases or deaths.

o Correlation among the different attributes

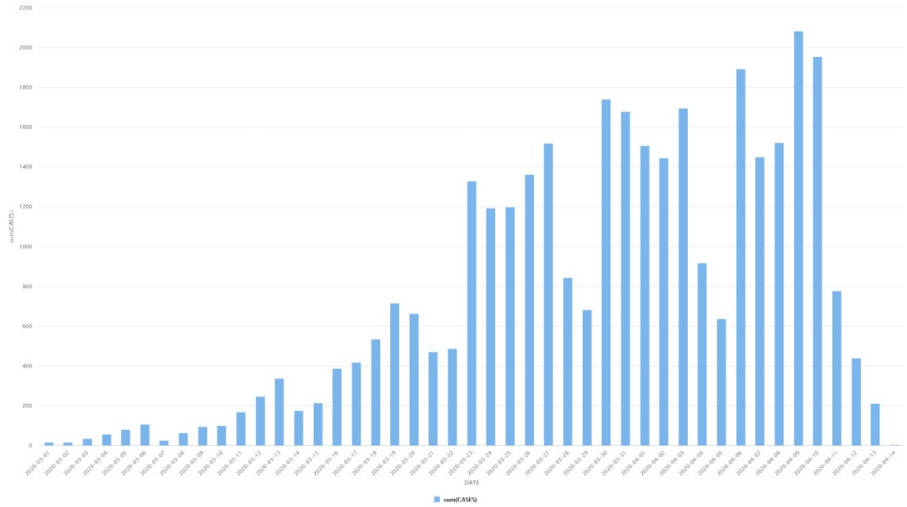
o Calculate the cumulative case number: slightly above32.5k cases from the diagram below

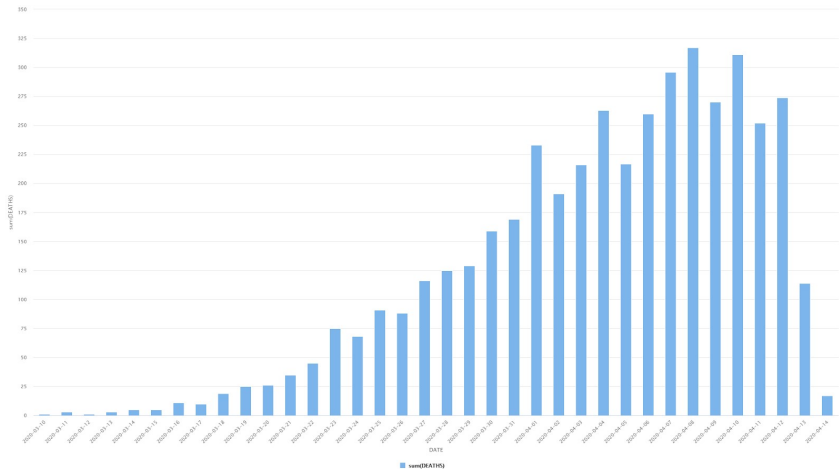


o Calculate the cumulative death number: slightly below 4500 cases as shown below.

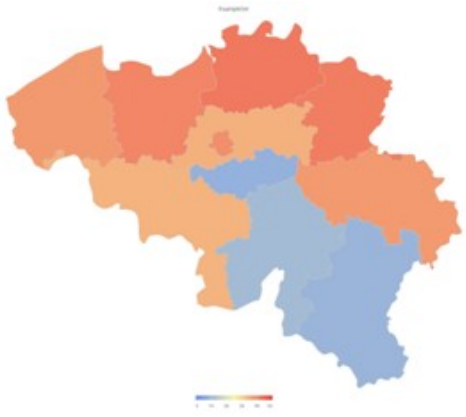


o Calculate the case progression and mortality rate:





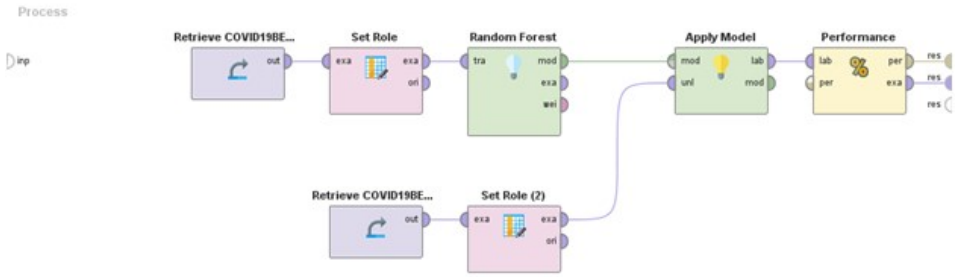
Visualization



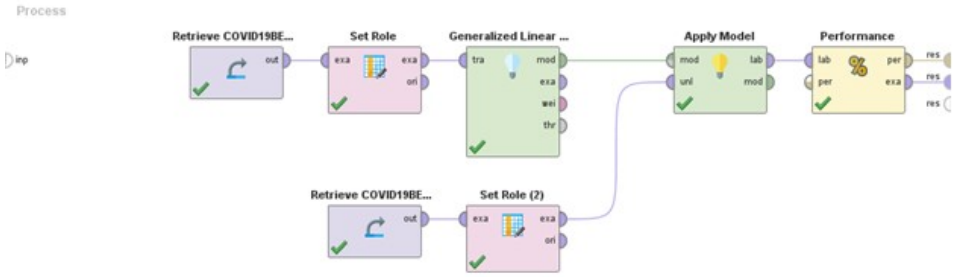
- the number of cases is the highest in East Flanders, Limburg and Antwerp and is less in the southern regions of Namur, Luxembourg and Walloon Brabant.

Description and Evaluation of the RapidMiner models

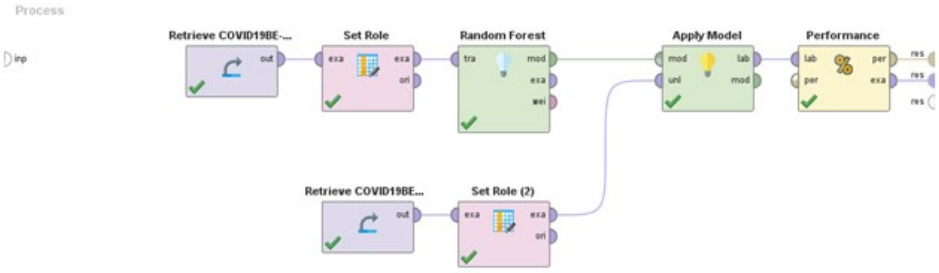
- We can see from the below RapidMiner models that the RandomForest model works the best for prediction of the number of cases and the number of deaths while using the validation set provided as it is considered the best model overall in the industry.
- The Decision Tree model is the second best model for predicting the number of deaths again as expected but does not perform accurately when predicting the number of cases. We therefore used the Generalised Linear Model algorithm for the number of cases as it is also well regarded in the industry for regression purposes.
- The root mean squared error was used to evaluate the models as it is the most well used metric. The rms values decrease even further when the training set itself is also used for validation.
- For creating the datasets we deleted all the rows where data was not present and did not consider imputing any values because the commonly used imputing techniques like average, median etc cannot accurately guess the number of cases.
- We can see that the rms values for the models is not extremely good. This is because the amount of data used for training is not high. Another reason could be that the models used i.e. Random Forest, Decision Trees and Generalised Linear Models are not adapted to be used for time series data.



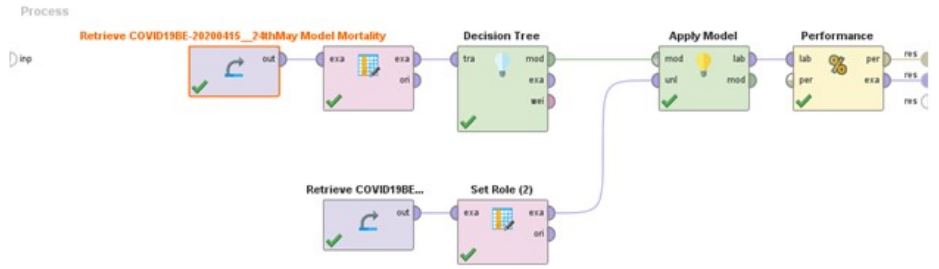
root_mean_squared_error: 72.365 +/- 0.000



root_mean_squared_error: 102.516 +/- 0.000



root_mean_squared_error: 39.469 +/- 0.000



root_mean_squared_error: 59.951 +/- 0.000

Predicted numbers

- we used the ARIMA algorithm which is a time series prediction algorithm.
- In a prediction from 15 April 2020, the country will reach **36419** cases on 20 Apr 2020 (actual number is 41,834), **47153** cases on 05 May 2020, **86164** cases on 05 Jun 2020.