# Regression models with COVID-19 datasets

Linear Regression, Decision Tree Regression and Gradient-boosted tree regression are modeled and evaluated with COVID-19 datasets from data repository by Johns Hopkins CSSE. Jupyter notebook was used for analysis since It was possible to import PySpark MLlib into Jupyter notebook as below.

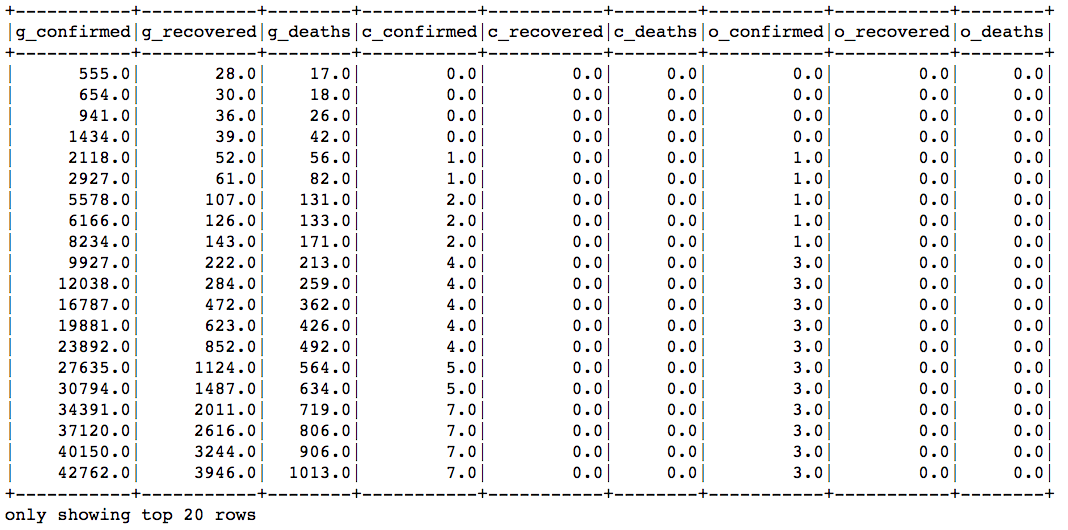
*import pyspark*

*from pyspark.sql import SparkSession*

*spark = SparkSession.builder.appName('Ops').getOrCreate()*

## Data Preparation and Preprocessing

In order to predict the fatality rate of Ontario, Three CSV files of confirmed, recovered and deaths for Global time series are used. The final version of Spark DataFrame created for regression modeling contains 9 columns of confirmed cases, recovered cases and deaths cases in three groups of global, Canada and Ontario as below.



## Linear Regression

The vector of 8 columns is used as input columns for Linear Regression model and the number of Ontario deaths is label columns for the model. After fitting, the vector of coefficients can be seen below.



The intercept of model is very close to zero as 0.16802774778269086 and RMSE(root mean squared error) is around 1 as 0.9810006656488205.

The most of elements in the vector of coefficients are very close to zero except one column, so that most of columns are not related to the fatality rate of Ontario. Only one column shows 0.7166 and that columns is the fatality rate of Canada. So, it is safe to say that the death rate of Canada is linearly related to that of Ontario and the 72% of Canadian deaths are the deaths of Ontarians.

## Decision Tree Regression and Gradient-boosted tree regression

Two of decision tree and Gradient-boosted tree regressions are modeled and evaluated. Both of models show the same RMSE of lower than 5s. This is relatively low and it can be regarded as an accurate model for predictions. However, it can be caused by highly related data of Canadian deaths column.