**ORANGE**

*Datasets:*

We have used two datasets in the Orange to predict and have better insights for COVID-19 data-

1. HDI (Provided by Dataset feature of Orange)
2. COVID-19 dataset which has infected cases listed according to country, continent and dates

*Extra Features:*

We merged both the datasets to have some extra information for each country from HDI source in COVID-19 dataset like –

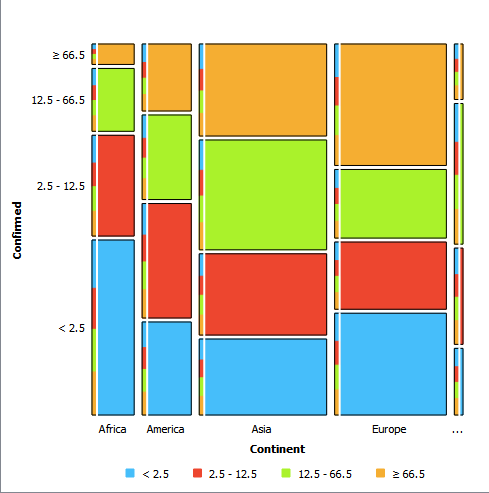
1. Physicians per 10,000 people
2. Population median range
3. Total unemployment

*Target:* Confirmed

*Validation:* Population > 0

*Visualizations:*

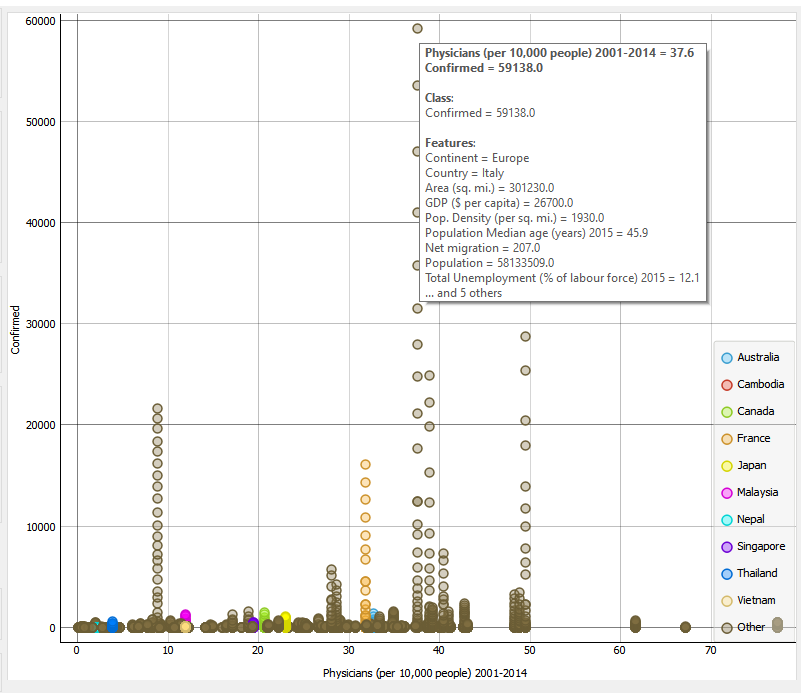
Mosaic Display



*- Europe has highest number of cases when compared to other continents (orange portion indicating percentage >66.5 is bigger for Europe)*

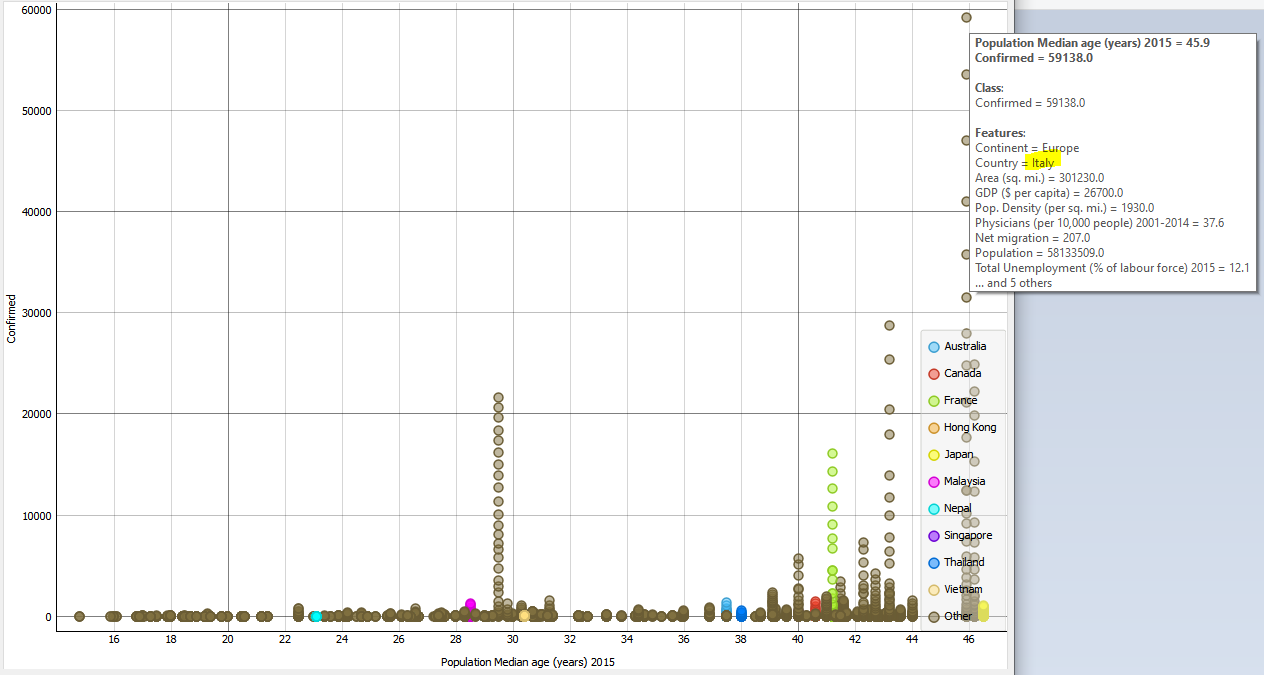
Scatter plots:

1. X-axis -> Physicians per 10,000 people , Y-axis -> Confirmed cases, Color -> Country



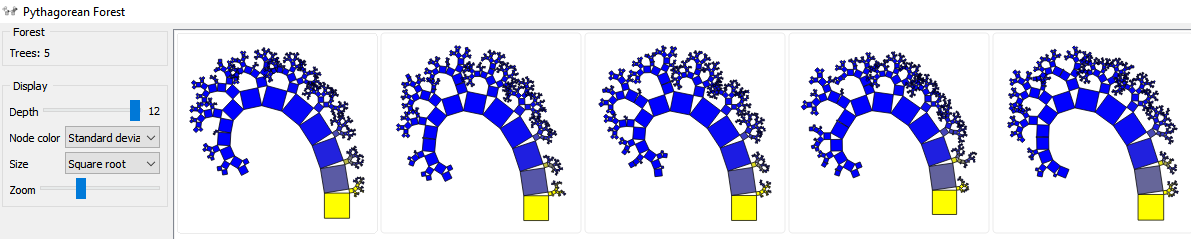
* *For Italy, the number of confirmed cases is highest whereas number of physicians available per 10,000 people are just 37.*

1. X-axis -> Population median age , Y-axis -> Confirmed cases, Color -> Country



*Italy has more population of people with age group > 45 years. Since, the number of confirmed cases are more in Italy, it can be assumed easily from the above graph that the people of age group more than 45 are infected more.*

*Pythagorean Forest*



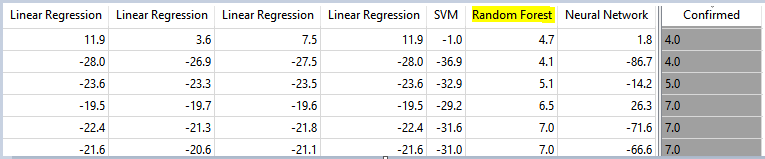
*Algorithms used for prediction:*

1. *Linear Regression*

*Linear Regression with 4 different flavours-*

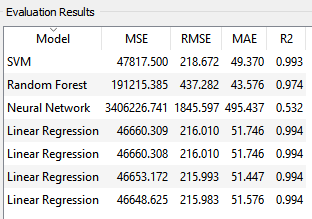
1. *No regularization*
2. *Lasso*
3. *Ridge*
4. *Elastic net*
5. Random Forest
6. SVR
7. Neural network

Prediction Results-



It can be concluded from the prediction results that the results of “**Random Forest**” are best

Evaluation results-



Even though RMSE is high for Random Forest when compared to other models, we have the best prediction results. Since the parameters for Random Forest can be easily modified and used in python, we chose “Random Forest” algorithm for our further analysis.