Prediction and Forecast of Global Covid-19 Cases

Introduction

The development of machine learning models provides pattern detection and prediction from data. Applying machine learning algorithms is of the utmost importance to be able to get useful insights into enormous amounts of complex data, that otherwise would be extremely hard if not impossible to discover, allowing people to make important decisions.

This work aims to develop machine learning models using data from Kaggle.com to predict the cumulative number of Covid-19 confirmed cases, the number of resulting fatalities across the world and forecast for future days. Some important graphs have been made on how the trend is moving and which countries were the most affected.

Exploratory Data Analysis

The datasets used have important data about the number of Covid related fatalities and confirmed cases for each country for each day from 23rd January 2020 until 10th of June 2020.

We did visualization such as histograms, pie charts, kernel density estimate (KDE), to help visualize the distribution of each numerical feature.

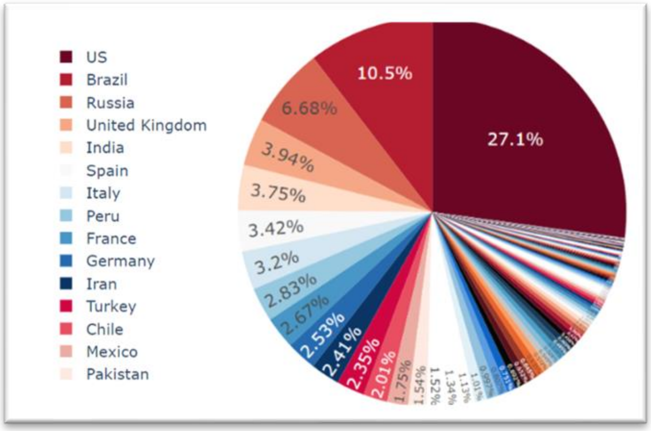


Figure 1 - Pie-chart for the percentage of global confirmed cases

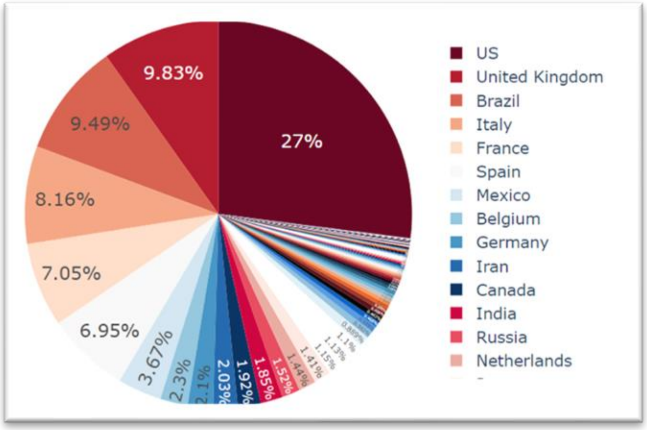


Figure 2 - Pie-chart of the percentage of global fatalities

We can observe from Figure 1, the country that had the highest percentages of confirmed cases was the United States (US) with 27.1% of all confirmed cases of Covid cases, followed by Brazil with 10.5% of confirmed cases and Russia with 6.68%.

While for the percentage of fatalities, in Figure 2, the US has the highest percentage with 27% of the global fatalities, next comes the United Kingdom with 9.83% and Brazil with 9.49%.

The US is the country with the largest population, so that could explain the high percentage of space it takes in the pie-charts above.

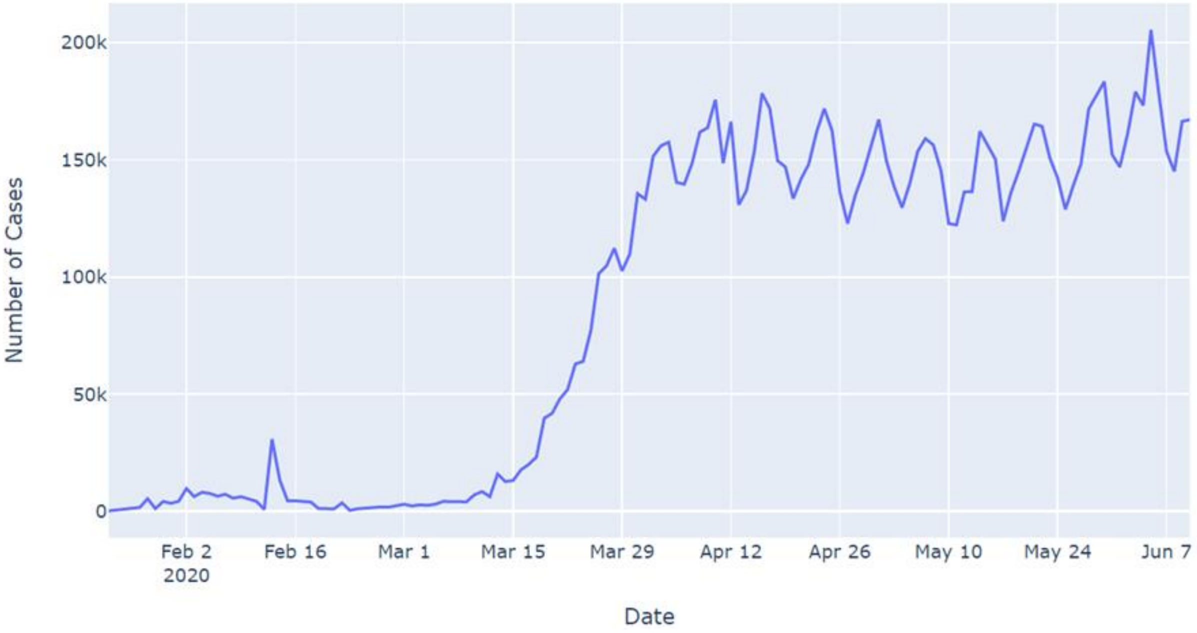


Figure 1 - Global number of confirmed Covid cases over the first 5 weeks

From the graph above we see an exponential growth of the confirmed cases of Covid, between March 12th and April 4th. After this exponential growth, there is a constant of peaks every ~3 days, this raises the need to smooth out the data to get a more precise prediction.

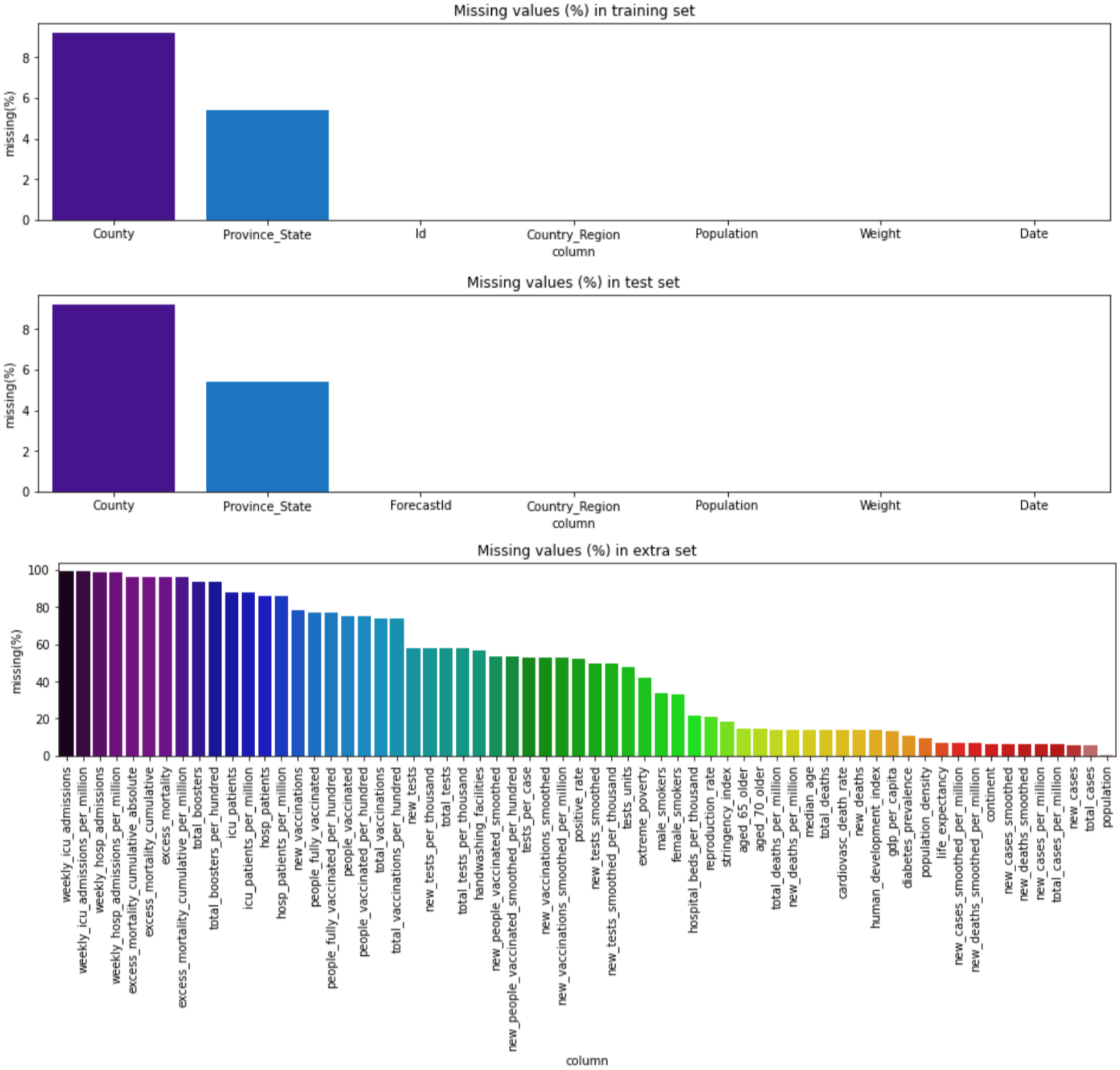


Figure 2 - Missing values from the test, train, and extra datasets

We found some missing values in County and Province_State variables in the train and test dataset, and in our extra dataset we found a much higher percentage of missing values, where some features had more than 95% of missing values; therefore, we have decided not to include the extra data in our ML model after consideration.

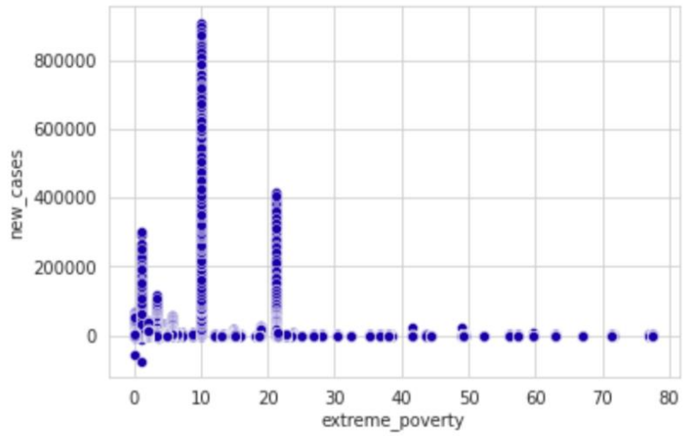


Figure 3 - Relation between new cases and extreme poverty

From the graph above (Figure 5) we can see that with increased poverty there is a smaller number of cases. There is a possibility that in some countries due to lack of medical amenities and inflated cost of health care, people are not getting tested, and cases are not being reported.

Data Preparation

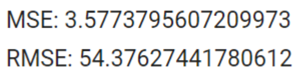
Data preparation is a crucial step in the development of a ML prediction model because it is when we decide which columns are dropped and which columns stay so that we can enhance the predictability of our model. Province_state, County and ID features were dropped because they do not have any value for our prediction model.

The rolling mean is commonly used with timeseries data to reduce short term fluctuations and highlight long term fluctuations. In our dataset, we have added rolling means of 5, 10, 15 and 30 days to smooth out the fluctuations found during the EDA.

Comparing ML Models

We tested 6 ML models to evaluate which would be better at predicting the confirmed cases for the future. After fitting our data to each model and getting the predictions, we used metrics such as RMSE, MAE and R2 to see which model performed better.

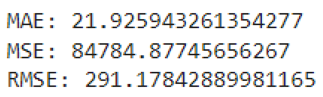
1. **Extreme Gradient Boosting** - this model combines weak learners that are regression trees, and each of them maps an input to one of its leaves that contains a continuous score.



```
MSE: 3.5773795607209973
RMSE: 54.37627441780612
```

Figure 4 - Extreme Gradient Boosting Results

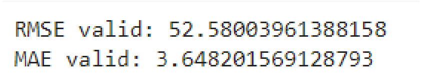
2. **Decision Tree Regressor** - This algorithm applies regression in the form of a tree to make decisions about the data to then be able to make predictions.



```
MAE: 21.925943261354277
MSE: 84784.87745656267
RMSE: 291.17842889981165
```

Figure 5 - Decision Tree Regressor Results

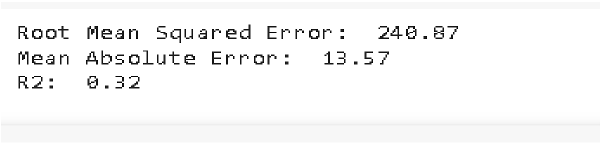
3. **Random Forest Regressor** - This model is a supervised learning algorithm that uses an ensemble learning method for regression. It acts as an algorithm estimator that combines the result of many decision trees and then decides which works best and outputs are optimal.



```
RMSE valid: 52.58003961388158
MAE valid: 3.648201569128793
```

Figure 6 - Random Forest Regressor Results

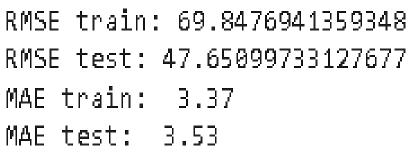
4. **Light GBM** - This model uses a gradient boosting framework, and it is based on a decision tree to increase the efficiency of the model. It is used for ranking and classification.



```
Root Mean Squared Error:  240.87
Mean Absolute Error:  13.57
R2:  0.32
```

Figure 7 - Light GBM Results

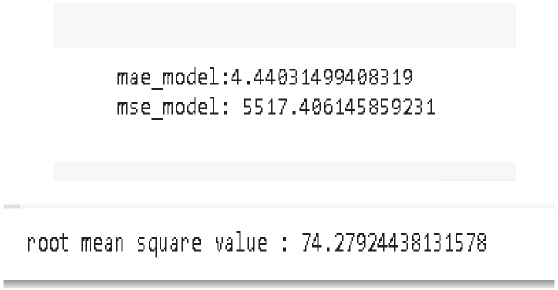
5. **Linear Regression** - It is supervised by a machine learning model. This model finds the best fit linear line between independent and dependent variables.



```
RMSE train: 69.8476941359348
RMSE test: 47.65099733127677
MAE train:  3.37
MAE test:  3.53
```

Figure 8 - Linear Regression Results

6. **Gradient Boost** - Generates multiple weak learners in a sequential way that the present learner is always more effective than the previous one and combines their prediction to make an accurate prediction.



```
mae_model:4.44031499408319
mse_model: 5517.406145859231
```

```
root mean square value : 74.27924438131578
```

Figure 11 - Gradient Boost Results

Conclusion / Best Model

RMSE stands for Root Mean Squared Error and MAE stands for Mean Absolute Error and these measurements show us how far the values predicted from the machine learning models are from the true values, so the lower the MAE and RMSE the better at predicting the models are, and we can say our model has higher accuracy in the prediction.

We concluded that the best model to predict our data is the **Random Forest Regression** algorithm, since it is the model that shows to have the lowest values of RMSE (52.58) and MAE (3.64), so it is the model that predicts future data more accurately.