**Introduction**

**EMY**

**Motivation**

The principal motivation for this analysis is to apply what we have been learning in Machine learning into the analysis of Covid cases, making the predictions of the confirmed cases of Covid during the very beginnings of the pandemic in Europe Region. To know the number of positive cases was primordial since there were many people affected that needed medical services and treatments.

**Business understanding**

To predict confirmed Covid cases, we are going to apply different Machine Learning models, which data was chosen from a Covid repository where we can analyse the beginning of Covid cases in the world, but we decided just to focus in Europe region for making better predictions.

**Business Description**

**Research Question**

Using different training and tests splits in the data, what model could predict better the confirmed cases of Covid in Europe (WHO Region)?

**General goal**

The general goal is to predict how many confirmed Covid cases. This will help European countries to know how the virus is spreading for taking preventive, and medical decisions in the European Region.

**Success criteria/indicators**

We are going to apply 3 different Machine Learning Models in which we are going to compare the R-squared value of each or the Mean Absolut Error to define which is the best model.

**Technologies used**

**Models and machine learning algorithms**

As we are trying to predict a numerical value we are going to apply supervised Models that fit with our data and we decided to use (EMY ESCRIBI LOS MODELOS QUE USASTE POR FIS ) for regression, and ARIMA model as a time series analysis.

**Libraries**

We used different libraries to perform this analysis like Pandas, Seaborn, Matplotlib, NumPy, Standard Scaler, PCA, ARIMA, sm, adfullet, among others.

**Accomplishment Data**

Our Covid Dataset was split through different excel files and we decided to concatenate with the useful information for our project and after that we have 49068 rows and 11 columns in the dataset in which 2 of them are continuous numerical variables, 4 are discrete numerical variables and 5 are categorical variables. We have data collected from February 2020 until July 2020, confirmed, death, recovered cases of Covid and other variables that are available in the Data Dictionary (Appendix 1)

**Source**

The data was chosen from a Kaggle repository found in this link: <https://www.kaggle.com/datasets/imdevskp/corona-virus-report?select=covid_19_clean_complete.csv> (Kaggle, 2020)

**Characterization of the dataset**

**Attributes**

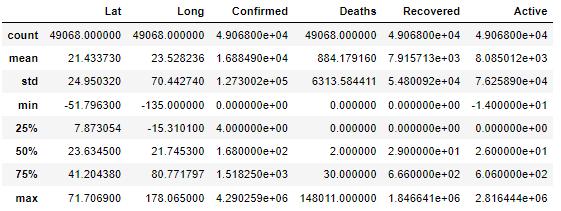
We are going to analyse 11 variables in which “Confirmed” is going to be our target variable to be predicted, and the others are going to be independent variables which will be selected for our analysis.

**Dimensions**

The shape of the Covid dataset to be analysed is 49068 rows and 11 columns.

**Descriptive Statistics**

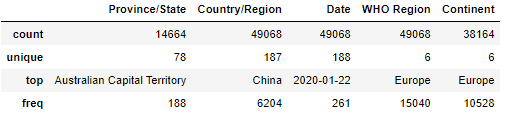
In Figure 1 we are going to see the principal statistics of the numerical variables.

****

*Figure 1 Statistics of the numerical variables in Covid dataset*

In the dataset we can appreciate some details as latitude, and longitude which show us where is every country located, and a collection of confirmed, deaths, recovered, and active cases of Covid.

We are going to analyse the statistics of the categorical variables in Figure 2



*Figure 2 Statistics of the categorical variables in Covid dataset*

In our categorical variables, we can see that we have 187 countries in our dataset, 78 provinces, 188 different dates, 6 WHO Region and 6 different continents

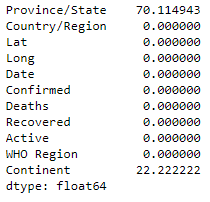
**Data Preparation and Preprocessing**

**Dropping Duplicates**

After dropping duplicates, the number of rows didn't change after dropping duplicates, and this means that our data doesn't have duplicates inside.

**Missing Values**

The NaN values were standardized and we got the next results analyzing missing values:

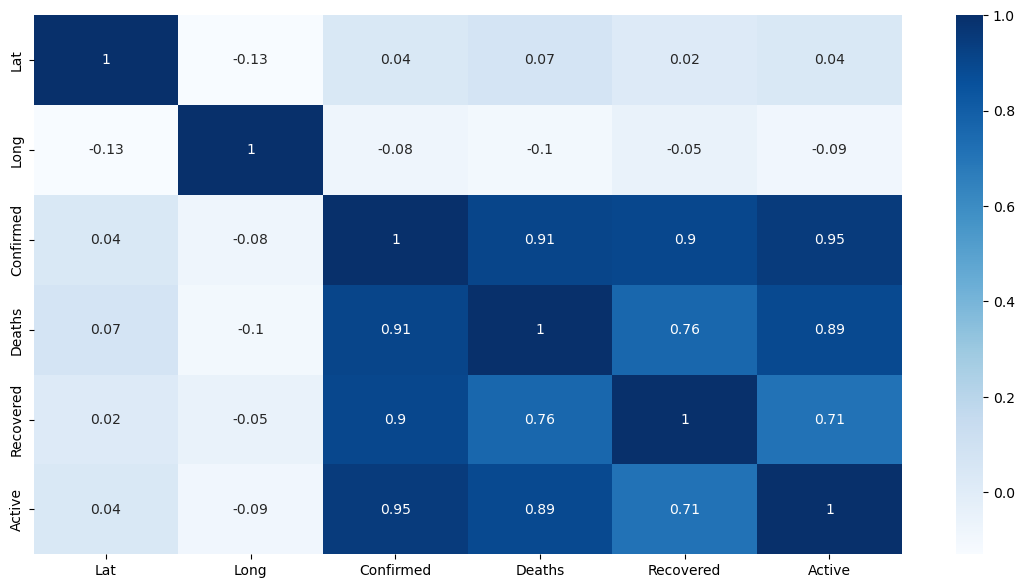


*Figure 3: Percentage of NaN values in each column*

As we can see the variable "Province/state" has around 70% of Null values present, and we decided to drop it since it would create bias and variance leading into bad predictions in our Machine Learning predictions, and in the column “Continent”, we explored that "WHO Region" and “Continent”, we have similar unique values; however, they are not the same for example in Eastern Mediterranean Regions could be included European and non/European countries, that is why we are going to replace the NAN- values as "Unknown".

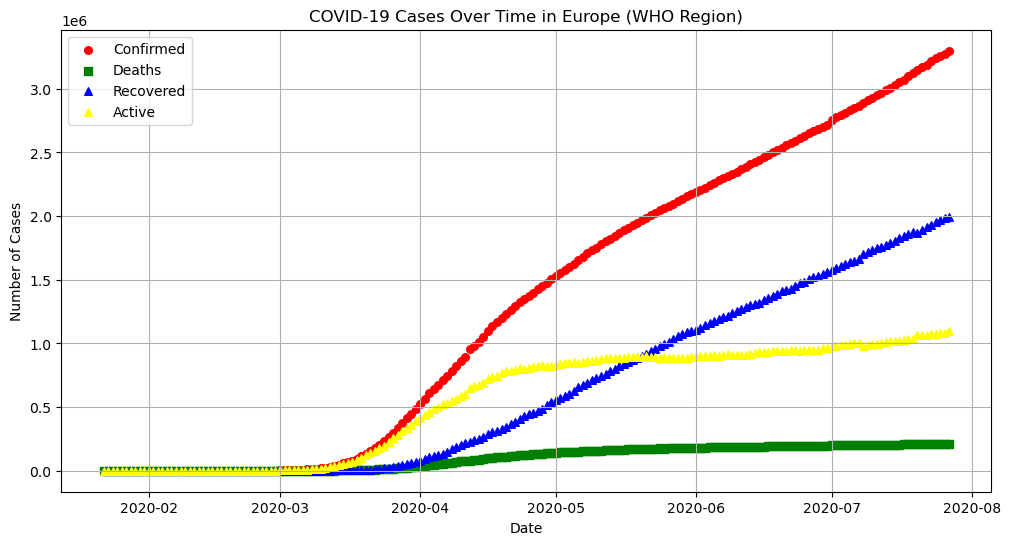
**Data Visualization**

**Correlation:**



*Figure 4: Correlation Matrix between the numerical values of Covid cases*

In Figure 4, in the Correlation Matrix of the numerical values, we can see a strong correlation of the target variable “Confirmed” with “Deaths”, “Recovered” and “Active”, these correlations are really good since they will help us get good Machine Learning models for predictions.



*Figure 5: Visualization of Covid 19 Cases Over Time in Europe (“WHO Region)*

In this Figure 5, we can see a positive tendency trend in “Confirmed” Covid cases, but also, we can see how was the distribution of the Covid cases Over time in Europe according to our data time provided. We can see also that until march the cases were still around 0. Besides, we can appreciate how was the tendency of Deaths, Recovered, and Active cases in Europe Region.

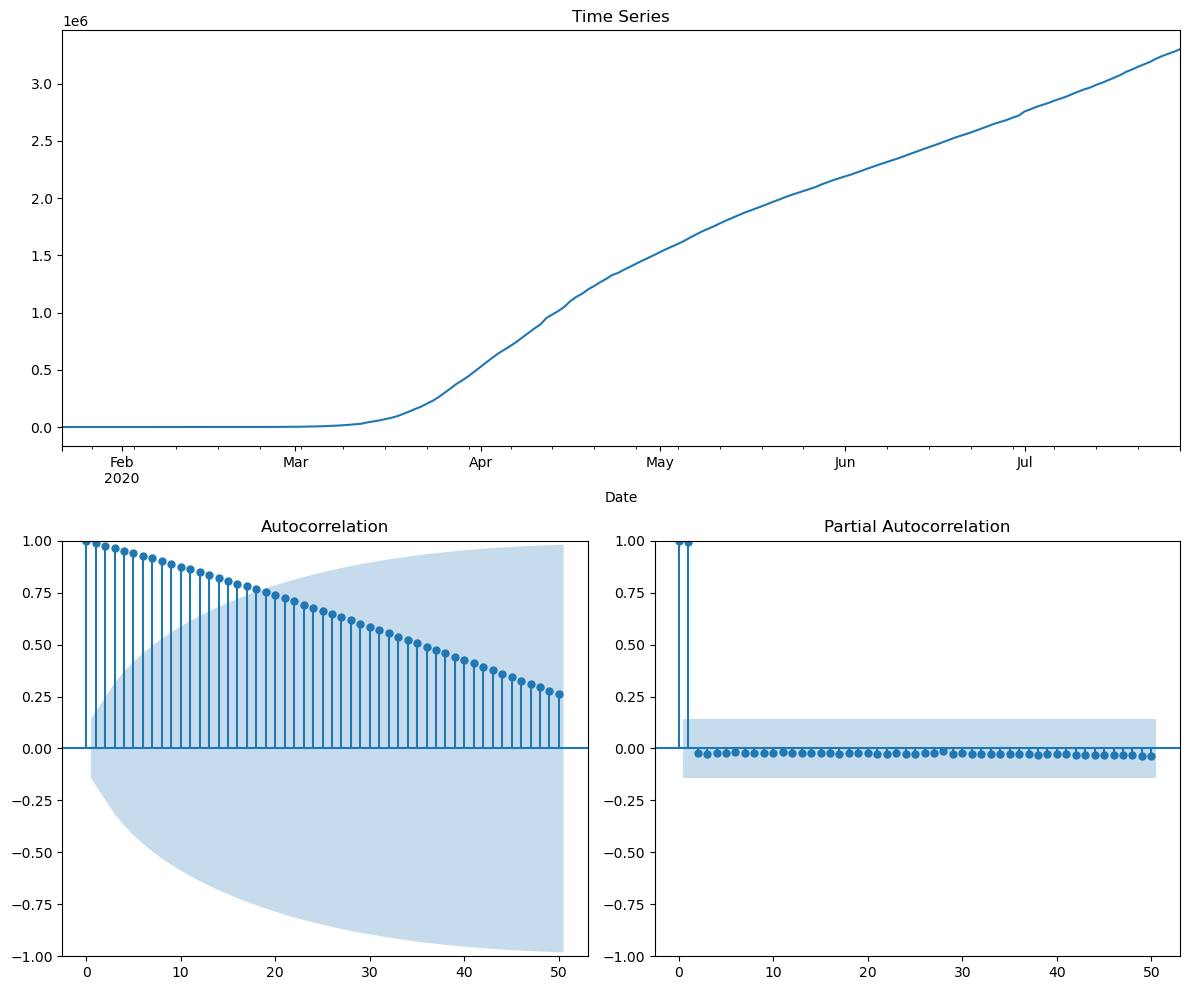
**Normalizing the data**

Fo the time series model we didn’t normalize the data since we have only one column to analyze. However in KNeighbord Regressor and Random Forest Regressor, we used all the columns of the dataset and we tried to normalize with Standard Scaler, MinMax Scaler and Robust Scaler, but our machine learning results give around 100% of accuracy and that is why we decided not to scale the data for the following Machine Learning Models.

**Machine Learning Models**

**A ) Arima Time Series Model:**

The first Machine Learning model analysed is ARIMA model in which we applied techniques learned in class to make predictions over time, as our data is very reduced of just 7 months, we decided not to apply SARIMAX since Covid cases didn’t have too much peaks during this time, following a continuous increasing tendency.



*Figure 6: Visualization of Covid 19 Cases Over Time in Europe (“WHO Region)*

We did the previous plot (Figure 6) to see if our data is shown as stationary, but we can see that our data is not stationary since it has an increasing trend in the first visualization where we see the months vs confirmed Covid cases, we can also see a decreasing autocorrelation and analyzing the partial autocorrelation, we can see around the first 2 lags are around 1 in the y-axes, but then we can see the following lags in negative values close to zero and we can see that there is a relationship present there.



*Figure 7: Results of the mean and variance values of the first 110 rows and afterwards*

We can see that the variance is changing between the first 110 rows and afterwards and we can say that the data is not stationary, and to confirm this using a mathematical method, we will use Dickey Fuller Test.

For the Dickey Fuller test, we stated the next Hypothesis to compare that is the null hypothesis (H0), and the alternative one (HA):

H0: The time series is non-stationary

HA: The time series is non-stationary

The significance level stablished by default is 0.05



*Figure 8: ADF and p-value results applying Dickey Fuller Test*

Applying Dickey Fuller Test, we can see that the p value is more than 0.05; so, we reject the null hypothesis, and we can say that our dataset is non-stationary.

**Splitting the data**

Before applying techniques to transform the data stationary, we did 3 different splits for the analysis.

**Dimensionality Reduction**

**Feature Engineering**

**Models**

**Results**

**Conclusion**

**Appendix**

**Appendix 1: Data Dictionary**

**Appendix 2: CRISP-DM - Part 1**

**Appendix 3: CRISP-DM - Part 2**

**Team Collaborations**

**Mijail Blanco’s Collaboration**

**Emily Herbas Collaborations**

**References**

[www.kaggle.com](http://www.kaggle.com). (2020). COVID-19 Dataset. [online] Available at: <https://www.kaggle.com/datasets/imdevskp/corona-virus-report?select=covid_19_clean_complete.csv>.