Predicting Carbon Dioxide Emissions until 2030 using Conventional Forecast Techniques and Machine Learning Models

Pedro Albuquerque

Dissertation written under the supervision of professor Nicolò Bertani

Dissertation submitted in partial fulfilment of requirements for the MSc in Business Analytics, at the Católica Lisbon School of Business and Economics, 5th April 2023.

Abstrato

Esta dissertação estuda as emissões globais de dióxido de carbono e estima previsões até 2030 usando técnicas de previsão convencionais e métodos de aprendizagem automática. Os dados sobre as emissões por tipo de combustível e fontes renováveis serão analisados para todos os continentes e para Portugal. Verificamos que o modelo de aprendizagem automática Random Forest produz previsões mais precisas do que o ARIMA convencional. As previsões feitas mostram que os níveis previstos de emissões de dióxido de carbono nos vários continentes variam significativamente. As emissões devem aumentar na Ásia, América do Sul e África, enquanto diminuem na Europa, América do Norte, Austrália e Portugal. No entanto, nenhum dos continentes está no caminho certo para alcançar a meta de redução de 55% estabelecida pelo Acordo de Paris. América do Norte e Europa estão progredindo, mas ainda estão 20% abaixo dos valores esperados.

Palavras-Chave: Acordo de Paris, Dióxido de Carbono, ARIMA , Machine Learning

Título: Previsão das emissões de dióxido de carbono até 2030 usando técnicas de previsão convencionais e modelos de Machine Learning

Autor: Pedro Manuel de Oliveira Albuquerque

Abstract

This dissertation investigates global carbon dioxide emissions and provides forecasts through 2030 using both conventional forecasting techniques and machine learning methods. Data on emissions by fuel type and renewable sources will be analysed for all continents and Portugal. We find that the Random Forest machine learning model produces more accurate forecasts than the conventional ARIMA. Our findings show that the predicted levels of carbon dioxide emissions on the various continents vary significantly. Emissions are predicted to rise in Asia, South America, and Africa while falling in Europe, North America, Australia, and Portugal. However, none of the continents are on track to reach the 55% reduction target set by the Paris Agreement. North America and Europe are making progress, but still 20% away from the expected values.

Keywords: Carbon Dioxide; Paris Agreement; ARIMA; Machine Learning.

Title: Predicting Carbon Dioxide Emissions until 2030 using Conventional Forecast Techniques and Machine Learning Models

Author: Pedro Manuel de Oliveira Albuquerque

Contents

Table of Contents

List of Figures

Chapter 1

Introduction

Carbon dioxide emissions from human activity are a major contributor to climate change, with negative impacts on both the environment and human health, CO2 is the most significant and long-lasting contributor to greenhouse gases(GHGs). In recent years, there has been a growing need to reduce these emissions to mitigate the effects of climate change. In this study, we aim to investigate the carbon dioxide emissions in all continents and make a forecast until 2030 using machine learning and forecast techniques.

To achieve this goal, we will collect and analyse data on carbon dioxide emissions in all continents and Portugal, including data on emissions by type of fuel and renewable energy sources. In addition, we will use both machine learning and traditional forecast techniques, such as the ARIMA model, Random Forest and Neural Networks, to make predictions about future emissions.

The results of this study will provide valuable insights into the current state of carbon dioxide emissions in the world in the next ten years, and perceive how the different continents will reach or if they will stay far from the values agreed in the Paris Agreement.

We found out that there are significant differences in the predicted levels of carbon dioxide emissions among the different continents. It was found that Asia, South America, and Africa will experience an increase in emissions, while Europe, North America, and Oceania will see a decrease.

One limitation of the study is the ongoing conflict in Ukraine, which could lead to an increase in the use of coal and other high-emission fuels, potentially affecting the accuracy of the forecast. Overall, it can be concluded that while some regions are on track to decrease carbon dioxide emissions, none of the continents are on track to reach the 55% reduction target set by the Paris Agreement. The machine learning model, Random Forest, performed better for majority of the predictions compared to the ARIMA technique. Further research is necessary to understand the factors driving these trends and to develop strategies for reducing emissions in all regions. Also, the results of this study complement the newly released 6th IPCC report, where a forecast of the carbon dioxide emissions is presented based on the implemented policies globally.

Chapter 2

Literature Review

Context

Carbon Dioxide Laws & Agreements

As indicated in Chapter One, the carbon dioxide (CO2) emissions are one of the most important concerns worldwide, as climate change is a global emergency that goes beyond national borders, especially in the European Union and North America, where both entities are trying for several years to implement new measures and laws to reduce the carbon dioxide (CO2) emissions and the reduction of the greenhouse gases.

Global leaders from more than 100 nations came to an agreement on 12 December 2015 to track climate change and establish some actions and timeframes to lessen its harmful effects.

The Paris Agreement has established a long-term goal for all the countries, with the primary goal being to reduce greenhouse gases to limit a temperature increase of 1,5°C (United Nations, Paris Agreement 2015). These countries will meet every five years to track the commitments made by all of the countries. They must submit a Nationally Determined Contribution (an updated national climate action plan). (United Nations, United Nations 2016)

On 14 July 2021, European Union adopted a series of legislative proposals, to try to achieve climate neutrality by 2050, with an intermediate deadline by 2030, where the objective is a decrease of 55% in the greenhouse gases (Comission 2021). This value is going to be used in future analyses and forecasts of this manuscript as an indicator to see if the European Union is going to achieve it.

Also in European Union we have the Renewable Energy Directive: This is a directive that was adopted by the EU in 2009. The Renewable Energy Directive sets targets for the use of renewable energy in the EU, and requires member states to implement policies and measures to promote the use of renewable energy.

Regarding the United States of America, we have the Clean Air Act, a federal law that regulates air pollution, including carbon emissions from power plants and other sources. The Clean Air Act requires the Environmental Protection Agency (EPA) to set standards for air pollution and to regulate the emission of greenhouse gases, such as carbon dioxide. (Agency 2022)

Also in North America, but for Canada the Clean Energy Act was passed in 2020. This act aims to reduce greenhouse gas emissions and increase the use of clean energy in Canada. It establishes targets for reducing emissions, and requires the federal government to implement policies and measures to meet these targets.
In conclusion, the Paris Agreement and various legislative proposals and directives in place in Europe and North America to promote the use of renewable energy and regulate greenhouse gas emissions from various sources are significant steps taken by the international community to reduce emissions and combat climate change.

Types of Fossil Energies

Carbon dioxide is created all around the planet and released into the atmosphere every second. This production results from the burning of several fuel types, each of which emits varying levels of carbon dioxide (CO_2).
The most common types of fossil fuels that are used and will be taken into account in this study are:

Coal: This fossil fuel is a sedimentary rock that is black or brownish-black in color and is made of the remains of extinct plants that existed millions of years ago. Coal is essentially made up of carbon and is the most abundant of the fossil fuels.

Oil: Also known as petroleum, is a liquid fossil fuel found beneath the Earth's surface. It was created from the buried remnants of aquatic animals that perished millions of years ago and were covered by sediment and rock.

Natural gas: Is a gaseous fossil fuel found in deposits deep beneath the Earth's surface. It is mostly made of methane and is created from the same organic components as oil.

All three of these fossil fuels are non-renewable, which means that if we keep using them at our current rate, they will eventually run out. Some types of non-renewable energies are not fossil, some examples of these energies are Nuclear energy and Hydrogen. Even not being fossil fuels, these energy sources still have environmental impacts and are not considered sustainable in the long-term.

The amount of CO2 emitted per unit of energy output will be evaluated to compare emissions among fuels. The fuel type that emits the most pounds of carbon dioxide per million British thermal units (Btu) is coal, and the fuel type with the lowest value is natural gas. (Administration, U.S. Energy Information Administration 2022)

Renewable Energies

Renewable energies are sources of energy that are naturally replenished and are not depleted when used. These types of energy are considered to be more sustainable and environmentally friendly than non-renewable energy sources, such as fossil fuels, which are finite and contribute to climate change when burned.

Some common types of renewable energy include:

Solar energy: This is energy that is generated from the sun's rays, either through the use of solar panels or through concentrated solar power systems. Solar energy can generate electricity, heat water, and power buildings and homes. In Portugal, this type of energy represents 2% of the energy.

Wind energy: This is energy that is generated by the movement of wind, which is converted into electricity using wind turbines. Wind energy is a clean and renewable electricity source widely used in many countries. Wind is the most used source of renewable energy in Portugal, with a representation in the production of energy of 29,1% of the energy in Portugal

Hydroelectric energy: This is energy that is generated from the movement of water, either through the use of dams or through the use of tidal power. Hydroelectric energy is a clean and renewable source of electricity that is widely used in many countries. Hydroelectric energy is one of the most used renewable energies in Portugal, with a representation in the production of energy of 22,6%.

Biomass energy: This is energy that is generated from organic materials, such as wood, agricultural waste, and biogas. Biomass energy can be used to generate electricity, heat buildings and homes, and power transportation.

Geothermal energy: This is energy that is generated from the heat of the earth's core, which can be harnessed to generate electricity or to heat buildings and homes.

Renewable energies are important because they can help reduce our reliance on non-renewable energy sources, which contribute to climate change and have negative environmental impacts. They also provide a source of energy that is not subject to the price fluctuations and supply disruptions that can be associated with non-renewable energy sources. (Administration, U.S. Energy Information Administration 2022)

Methods

Forecasts

In this next section, we are going to describe which were the methods used in this dissertation to forecast the carbon dioxide emissions in all the six continents and Portugal.
Understanding what constitutes forecast and why it is necessary in various situations is essential to comprehending one of the key goals of this thesis.
Making predictions with the greatest degree of accuracy possible is a crucial technique frequently used as a decision-making tool. This involves taking into account all of the data that is available, including past and present data as well as knowledge of any upcoming events that may have an impact on the forecast's outcome.
There are many different methods that can be used to make forecasts, including statistical techniques, machine learning algorithms, and expert judgment. The accuracy of a forecast depends on the quality of the data used to make the prediction, as well as the assumptions and methodologies used in the forecasting process.
Below, statistical models and machine learning models used, are going to be explained in detail, to give a better insight of them, how they work and how they were used to be helpful in this project.

Naïve

A naive model is a simple model that is used as a baseline for comparison with more complex models. It is called "naive" because it makes unsophisticated assumptions about the data, generating process assumptions, that are unlikely to be true in real-world situations.
For example, in the context of classification tasks, a naive model might assume that all the features in the data are independent of each other, and that the class probabilities are independent of the feature values. This is known as the naive Bayes assumption. A naive Bayes classifier is a simple probabilistic classifier based on this assumption.
Naive models are often used as a baseline to compare the performance of more complex models. They can provide a good starting point for building more sophisticated models, and can serve as a reference point to determine whether the added complexity of a more complex model is warranted. (Rish 2001)

Some examples of Naïve methos are described below, in this dissertation, the naïve method used is Last Value Method.

Last Value Method only presupposes that the following value in the time series will be the same as the most recent value observed. For instance, the forecast for the following period would be 50 tons of CO2 emissions,if the most recent observed value in the time series was 50 tons of CO2 emissions.

Simple Average Method: This method assumes that the time series' future values will be equal to the historical data's average.

Seasonal Naive Method: This approach makes the naive assumption that the subsequent value in the time series will match the previous value from the same season.

ARIMA

The time series forecasting technique known as ARIMA(p,d,q) models, or "Auto Regressive Integrated Moving Average," is based on describing the autocorrelations presented in data, as

well as its own lags and lagged forecast errors. Where p is the order of the autoregressive part (AR), d is the degree of first differencing involved, and q is the order of the moving average (MA) part. (Athanasopoulos 2018)

$$y'_t = c + \phi_1 y'_{t-1} + \cdots + \phi_p y'_{t-p} + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q} + \varepsilon_t,$$

Holt-Winters

A well-liked statistical technique for forecasting time series data with trend and seasonal components is the Holt-Winters forecasting model, sometimes referred to as triple exponential smoothing. The level (lt), trend (bt), and seasonality (st) of a time series are all estimated by the Holt-Winters model using three smoothing equations. These are the equations:

This equation approximates the level of the time series at each point in space and time. The level represents the series' starting point without taking seasonality or trends into account.

Equation for the trend: This equation calculates the time series' trend, overall size and direction of change over time, at each point in the data.

The seasonal equation calculates the time series' seasonal component at each point in time.

By extending the level, trend, and seasonal components into the future, the model employs these equations to produce forecasts. Combining the projections generates a final forecast for each moment in time.

For time series data with trend and seasonality, the Holt-Winters model is especially helpful. Even in cases when the data is non-linear or exhibits abrupt behavioural changes, it can nevertheless produce accurate projections. It might not work for time series data with erratic patterns or significant volatility.

The additive variation, shown by the following formula, is typically used when seasonal variations are roughly constant across the series.

$$\hat{y}_{t+h|t} = \ell_t + hb_t + s_{t+h-m(k+1)}$$

When seasonality varies proportionate to the level of the series, the multiplicative technique is chosen. The formula for the seasonality variable, which is reported as a percentage and used to alter the series by dividing through the seasonal component. (Athanasopoulos 2018)

$$\hat{y}_{t+h|t} = (\ell_t + hb_t)s_{t+h-m(k+1)}$$

Machine Learning

It is the ability of systems to use algorithms and statistical models to learn from training data that is relevant to a given problem in order to automate the construction of analytical models and complete related activities. (Janiesch 2021)

The first step in a Machine Learning model starts with the preparation process, which is responsible for cleaning, pre-process, and format the input data so that the machine learning algorithm can use it. This could involve cleaning the data of useless or missing information, scaling or normalizing the features, and transforming the information into numerical or categorical values.

Making a machine learning model comes after the data has been prepared. The problem that needs to be solved and the kind of data that are accessible will determine which model is used. For instance, neural networks are frequently used for picture data, while random forest and gradient boosting models are frequently utilized for text data.

The machine learning model is trained on the prepared data by tuning its parameters to reduce the discrepancy between the projected output and the actual output. The model's parameters are modified iteratively during this process until the result for the training data can be properly predicted.

Once trained, the model can be used to new data to generate predictions or judgments. To achieve this, new data is sent through the trained model, which produces an output based on the patterns discovered in the training data. The quality of the model and the quantity of training data determine how accurate the predictions or judgments will be.

Numerous industries use machine learning, including research, engineering, healthcare, finance, and business. Machine learning models that produce precise predictions can shed light on complex data sets, aid in scientific research, and guide the development of commercial plans.

Random Forest

A random forest is an ensemble machine learning algorithm that is used for classification and regression. It is a type of decision tree algorithm that creates a set of decision trees from randomly selected subsets of the training data. Then it aggregates the predictions of each tree to make a final prediction.

The main advantage of using a random forest is that it can handle large amounts of data and can make predictions accurately, even when the data is noisy or has missing values. Each tree in the forest is trained on a random subset of the data, and the final prediction is made by averaging the predictions of all the trees. This helps to reduce overfitting, which is a common problem in decision tree algorithms. (Breiman 2001)

Random forests are often used in applications such as image and speech recognition, recommendation systems, and fraud detection. They are also often used as a benchmark for comparing the performance of other machine learning algorithms. To do this algorithm we used *RandomForestClassifier*, a Python library that helps to build Random Forest models.

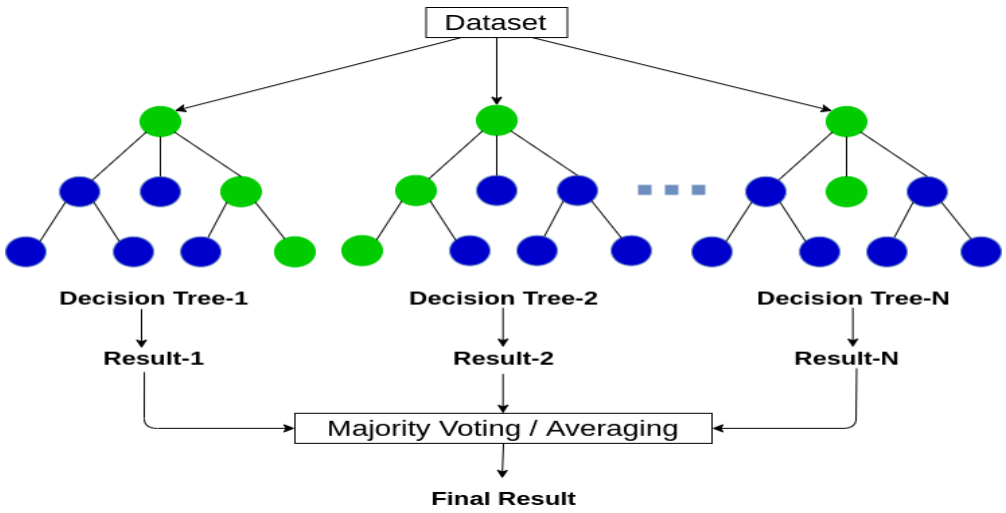

Figure 1 - Random Forest Model

Gradient Boosting

One of the ensemble technique variations that uses numerous weak models combined for greater overall performance is known as gradient boosting. This model involves three main elements.

The loss function, which measures how well the model performs on the training data, is the first component of gradient boosting. The type of problem being solved determines the loss function to be used, for regression we should use the squared error but for classification it should be use a logarithmic loss, the goal of this loss function is to be optimized.

The weak learner, a straightforward model that produces predictions based on training data, is the next component. Regression trees are frequently used as weak learners to correct the residuals in the predictions. Regression trees are decision trees that generate actual values rather than just binary values for splits.

The final ensemble model is created by combining the output of the regression tree with the output of other trees. Then, a new tree is added to the ensemble at each iteration of the algorithm in order to fix the mistakes caused by the earlier trees and build a strong model to minimize the loss function, in this dissertation we are going to use XGBoost that is an optimized gradient boosting library present in Python.

Gradient descent is used to estimate a set of parameters, such as the weights in a neural network or the coefficients in a regression equation. The weights are changed to reduce inaccuracy after calculating loss or error.

Decision trees are used in place of parameters. We add a tree to the model to help reduce the loss before beginning the gradient descent technique. We achieve this by parameterizing the tree, altering its parameters, and then moving forward by lowering the residual loss. (Brownlee 2016)

One of the main advantages of gradient boosting is that it can handle a large number of features and is not prone to overfitting. It is also robust to noise and can handle missing values in the data. However, it can be computationally expensive to train, as it involves training a large number of weak models. When compared to Random Forest model, these two Machine Learning techniques are similar, both ensemble learning techniques.

The key differences between them are the type of error minimized, Random Forest minimizes the variance error while gradient boosting minimizes the residual error, the training speed and the approach to building the ensemble.

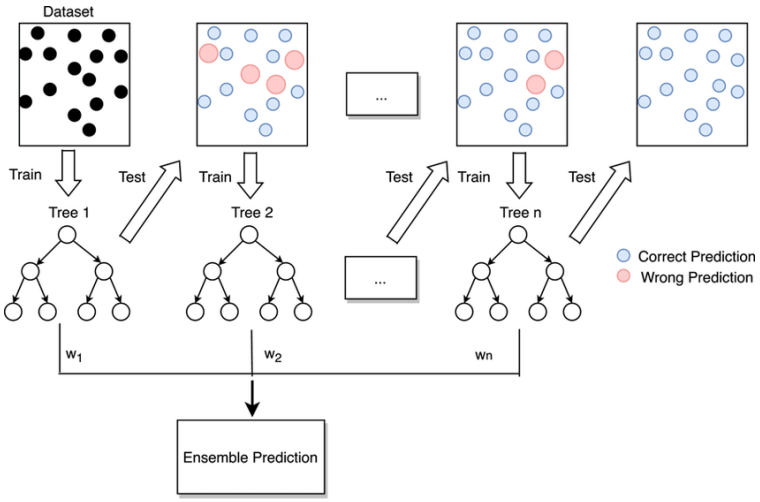

Figure 2 - Gradient Boosting Model

Neural Networks

A machine learning method known as a neural network is based on the structure and operation of the human brain. It is made up of a group of interconnected processing nodes known as neurons that cooperate to identify patterns in data.

Because they can learn complicated correlations between inputs and outputs and make predictions based on historical data, neural networks are particularly well-suited for forecasting. A neural network is often trained on a set of historical data in forecasting applications to discover patterns and correlations between variables. After training, a neural network can predict future values using fresh input data.

Another application of neural networks is time-series forecasting, which aims to estimate future values in a time series. In this instance, a sequence of historical data points is used to train a neural network, which then learns to predict the next value using past information. Since anticipating future values is important, this sort of forecasting is frequently utilized in the financial markets, weather forecasting, and other applications. In our case we will use the algorithm to forecast carbon dioxide emissions, by using the Python library *TensorFlow* to develop it.

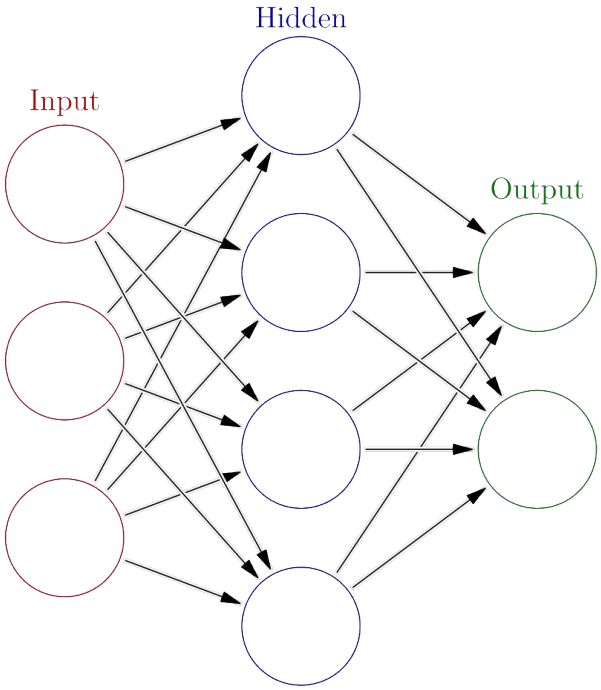

Figure 3 - Neural Network Model

Error Metrics

The way to choose the best model in both parts is going to evaluate them regarding the errors that each model has, choosing the one that has the lower error. The two techniques used are MSR(Mean Square Residuals) and RMSE(Root Mean Squared Error)

MSR, described by the formula below, also known as mean squared error of regression (MSE), is a measure of the difference between predicted and true values in a regression model. It is calculated as the average squared difference between the predicted values and the true values.

$$MSR = \frac{SSR}{n} = \frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{n}$$

RMSE, described by the formula below, is a way to know how well our model fits the data by telling us the average distance between the actual and predicted values. It is commonly used as a loss function in machine learning and optimization problems. RMSE is defined as the square root of the mean squared error (MSE), which is the average squared difference between the predicted values and the true values.

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{y}_j)^2}$$

Chapter 3

Methodology

I will concentrate on outlining all the procedures and approaches used to process all the data in this chapter. Beginning with the preparation and analysis of the data and ending with the application of forecasting and machine learning techniques, where the outcomes may be understood and explained.

Describing problem & Data

As discussed before, this study's objective is to analyse and understand the behavior of Carbon Dioxide emissions and Renewable Energies in all the continents and Portugal.

The first step was to find trustful and meaningful data, for that I resorted to Our World in Data (Oxford 2022), a website developed by a research team that is based at the University of Oxford that have data and focuses on large worldwide problems.

Five datasets were extracted, three regarding Carbon Dioxide annual emissions per country, per region and per type of fuel and the other two about Renewable Energies, one that presents per country how much energy(in percentage) comes from Renewables and the other one, a more detailed one that have detailed information about how much Eletricity is produced in TWh, for each country and by type of Renewable Energy method(Wind, Hydro, Solar and Bioenergy), all of the datasets present data for all the countries of the last fifty years.

For some countries, for example United Kingdom, the Carbon Dioxide emissions are presented since 1855.

The next stage was to comprehend each dataset in greater detail. To do this, the initial step was to focus primarily on the years after 1949, as it was only after that year that data for all nations worldwide became available. The following year, 2020, was also eliminated from the dataset because it was the year that Covid-19 began. Due to the pandemic and all the safety precautions taken, this year's data is not typical compared to years past; therefore, if we used data from 2020, our forecast would be biased and severely impacted by a highly unusual circumstance.

A descriptive analysis was then performed on two of the five datasets, first to determine which nations produced the most, the least renewable energy and carbon dioxide emissions, also the population for continent was added to the dataset and a scatter plot was designed to perceive the relation between CO2 Annual Emissions and Population for each continent. The results are shown in the three tables below, which include the Top 10 and Bottom 10 nations for 2019, the final year of our observation.

In Figure 4, we can see the scatter plot, that represent Annual CO2 Emissions vs Population in 2019, we can clearly see that Asia is by far the main carbon dioxide contributor, but is also by far the continent with the largest population with almost a value of 5 billion people. For the rest of the observations, we can see that Africa has a lot more population than North America and Europe, but produces only 20% of the carbon dioxide emissions of these two continents.

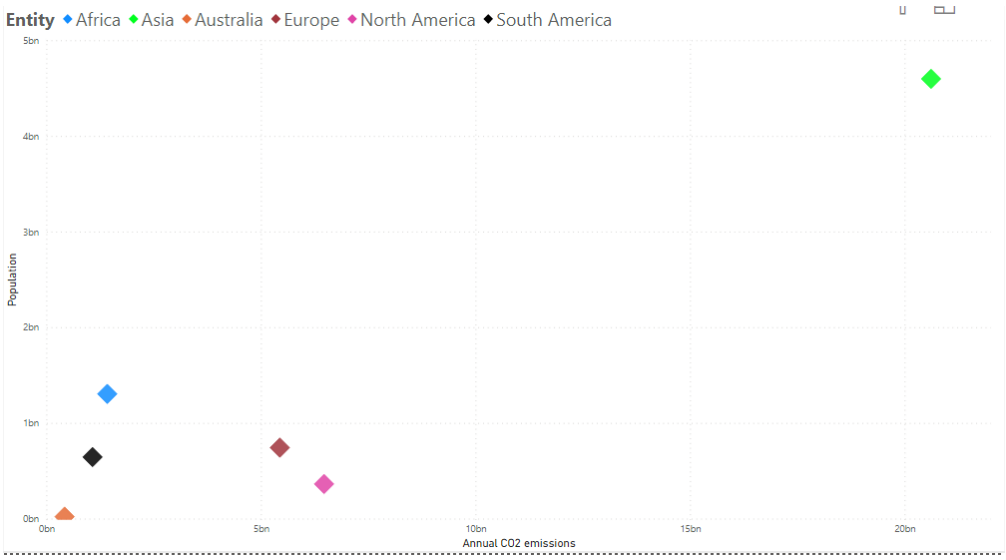

Figure 4 - Scatter Plot, Annual CO2 Emissions vs Population in year 2019

The table 1 shows the top 10 countries that emit more CO2 per year, where 70% of them are located in Asia, being China the leading country with almost the double of the emissions than United States that appear in second place. For the remaining 30% we found two European countries(Germany and Russia) and one North American country.

Entity	Code	Year	Annual CO2 emissions
China	CHN	2019	10489988555
United States	USA	2019	5255816207
India	IND	2019	2625968148
Russia	RUS	2019	1679449327
Japan	JPN	2019	1105929335
Iran	IRN	2019	733365847
Germany	DEU	2019	711427809
Indonesia	IDN	2019	660593985
South Korea	KOR	2019	648024558
Saudi Arabia	SAU	2019	622412749

Table 1 - Top 10 Countries that produce more CO2 emission(tons)s in 2019

Regarding the bottom ten countries that emit Carbon Dioxide emissions, we can observe in Table 2, as expected, that the countries of this list are third world countries.

Entity	Code	Year	Annual CO2 emissions
Tuvalu	TUV	2019	7553
Saint Helena	SHN	2019	10729
Niue	NIU	2019	11330
Wallis and Futuna	WLF	2019	26437
Montserrat	MSR	2019	28610
Nauru	NRU	2019	56651
Saint Pierre and Miquelon	SPM	2019	57220
Kiribati	KIR	2019	67981
Cook Islands	COK	2019	79312
Sao Tome and Principe	STP	2019	126263

Table 2 - Bottom 10 Countries that produce more CO2 emissions in 2019

Following a preliminary analysis, attention turned to emissions for each continent. Based on the analysis of the first table, it was anticipated that Asia will have the highest value of carbon dioxide emissions in 2019. North America and Europe are the ones that come next in the list with very similar values. Africa and South America have almost the same value and with the less emissions come Australia. In Table 3 we are showing the values for each continent.

Entity	Year	Annual CO2 emissions
Africa	2019	1408478929
Asia	2019	20608592701
Australia	2019	414516833
Europe	2019	5430238970
North America	2019	6460726238
South America	2019	1065509648

Table 3 - 2019 Annual CO2 Emissions for each Continent

In Figure 5, we have a different perspective of the values presented above, where we can see that Asia is responsible for 58% of the worldwide carbon dioxide emissions, and Europe and North America have similar percentage values of 18% and 15% respectively. The other continents combined produce less than 10% of the world emissions.

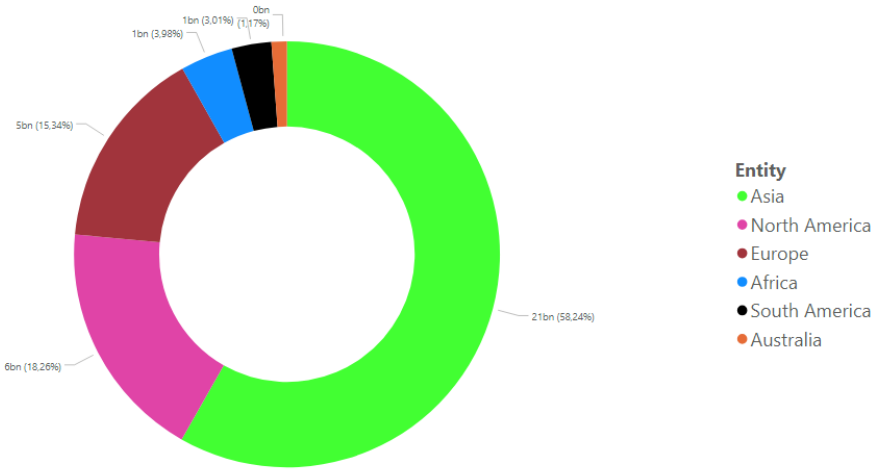

Figure 5 - Pie Chart representing the worldwide CO2 Emissions in 2019

Now the objective is plotting into graphics the values of emissions since 1949, to understand through visualizing the data, how these continents are behaving through the time, and if the agreements and laws implemented had effect in the decreasing of the carbon dioxide emissions.

Europe

Regarding Europe, one of the countries that produce more CO2 in the history, we can see a massive decrease of around 35% from 1990 until now. There is a factor that explains the decrease of 20% in the emissions in the decade of 90 to 00's, that was the collapse of the economies of the Soviet Union and countries in Eastern Europe. What these countries have in common, is that they were heavily based on coal, which is the type of fuel that produces more carbon dioxide per unit than all the others. (Administration, U.S. Energy Information Administration 2022)

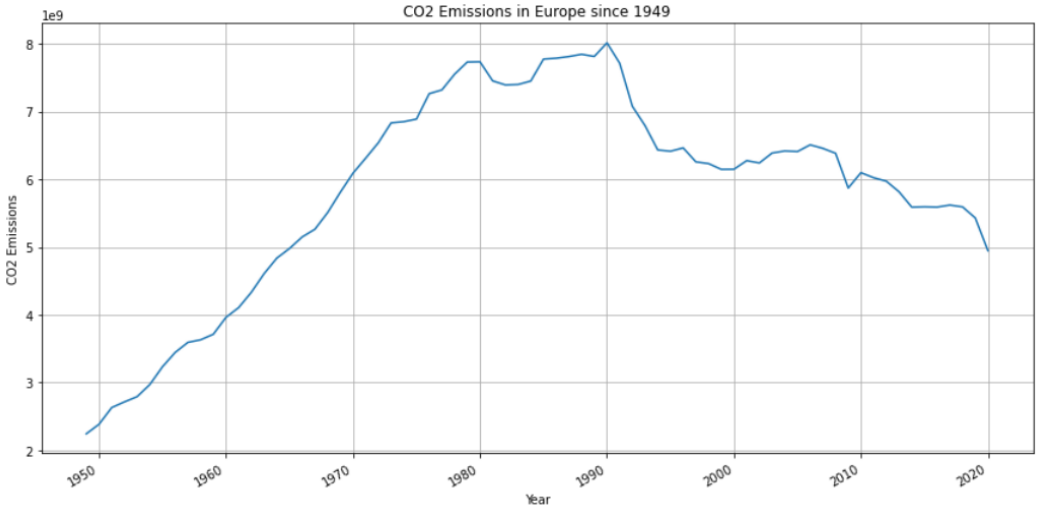

Figure 6 - CO2 Emissions in Europe since 1949

As said above, and we can confirm in Figure 7, in the year 1990 the carbon emissions coming from coal start do decreasing a lot, which leaded to a decrease of emissions overall in Europe.

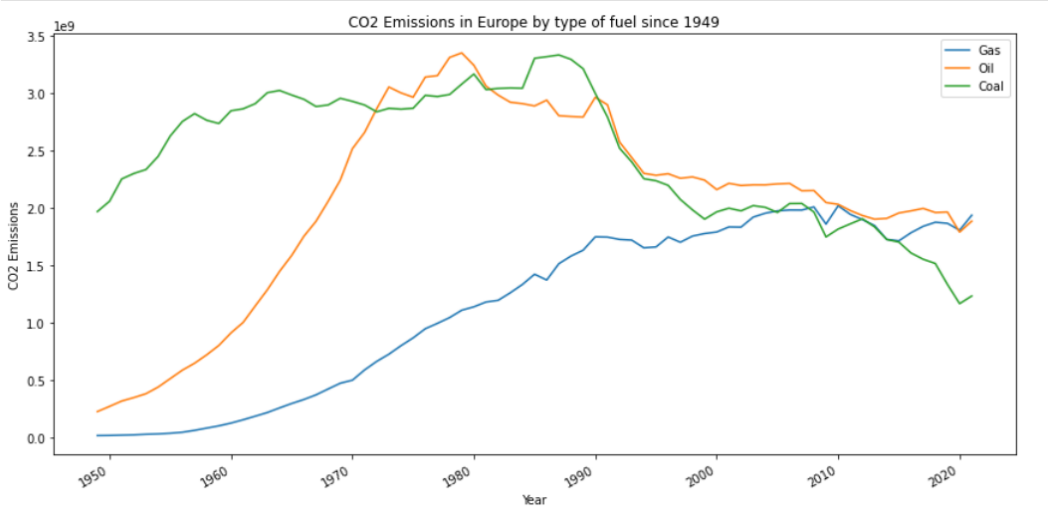

Figure 7 - CO2 Emissions in Europe by type of Fuel since 1949

When looking to the Renewable Energies, we can see than in Europe the main font of Energy coming from Renewable energies source is by far Hydro. However, we can also see that energy coming from wind sources have increase more than the double in the past 10 years. Solar and Other sources of Renewable Energies produce almost the same amount of energy in TWh.

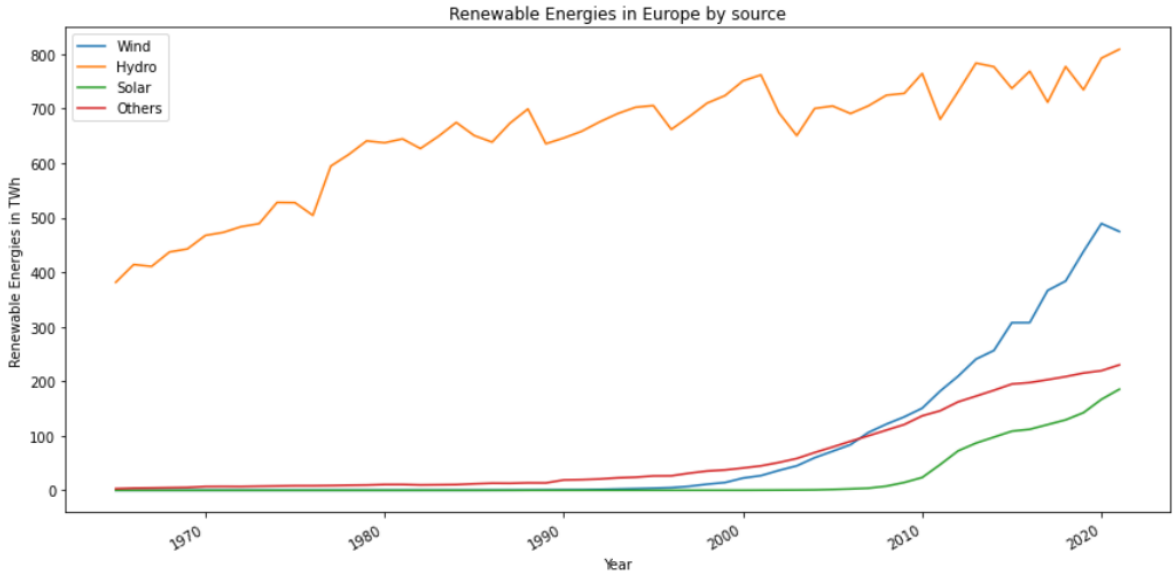

Figure 8- Renewable Energy Production in Europe since 1965 by source

North America

Looking for North America CO2 emissions since 1949 into the present we can observe an exponential increase until 2005, after this year the emissions start to decline until the present by around 20%, this decrease can be explained by the increase and deployment of renewable energies and the substitution of natural gas in the industry. (Kristina Mohlin 2019)

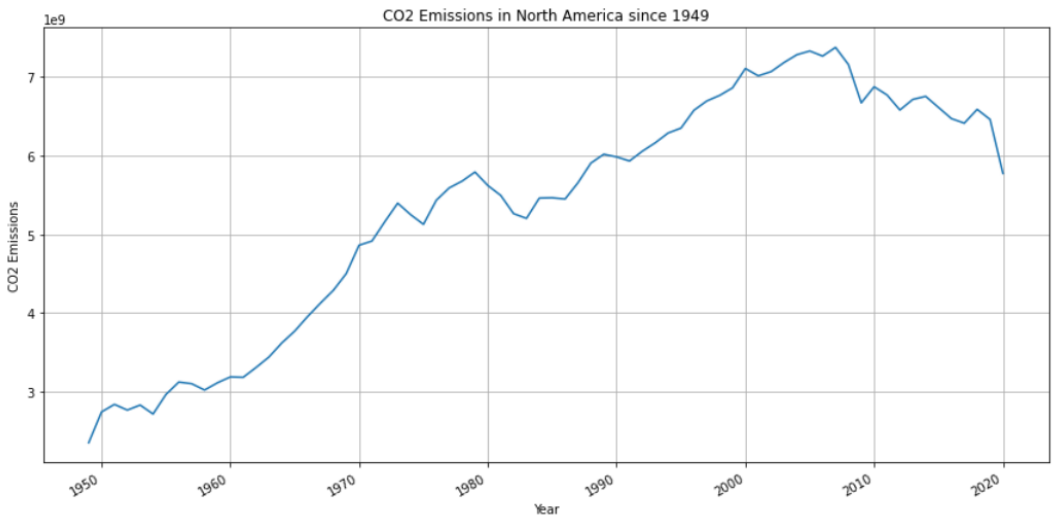

Figure 9 - CO2 Emissions in North America since 1949

When analyzing the Figure 10 of Renewable Energy Production in North America, we can see that after 2005, the year that the Carbon Dioxide emissions start to decrease significantly, the Renewable Energy Production almost doubled, increasing from 800 TWh to 1400 TWh.
This increase, is due to the huge growth of renewable energies productions coming from wind, the values coming from this source, have increase 400% in the last decade, we can also see that solar source have start becoming to produce values that will affect this continent beneficial.

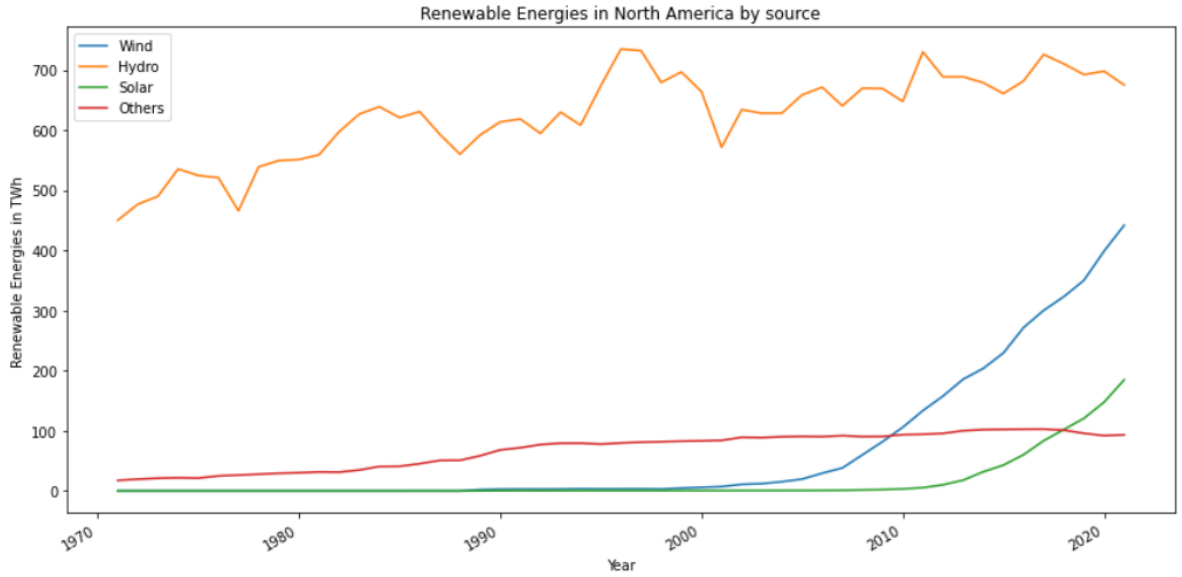

Figure 10 - Renewable Energy Production in North America since 1971 by source

Looking to North America we can observe the same that happened years before in Europe, with a decrease of 50% in coal emissions in the last decade, which led to the decrease of overall emissions.

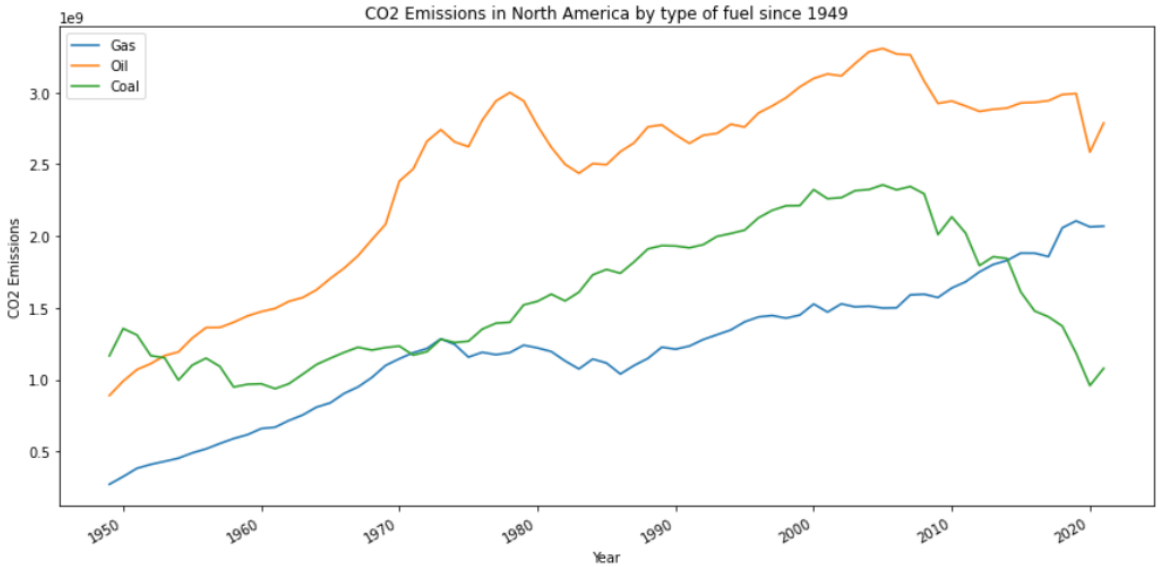

Figure 11 - CO2 Emissions in North America by type of fuel since 1949

South America

Regarding all the countries in South America, the same carbon dioxide growths happens from 1949 until 2013, where there is a turning point and the emissions start to decline into today, this is mainly due to the growth and use of renewable energies that are substituting the fossil fuels, responsible for the CO2 emissions.

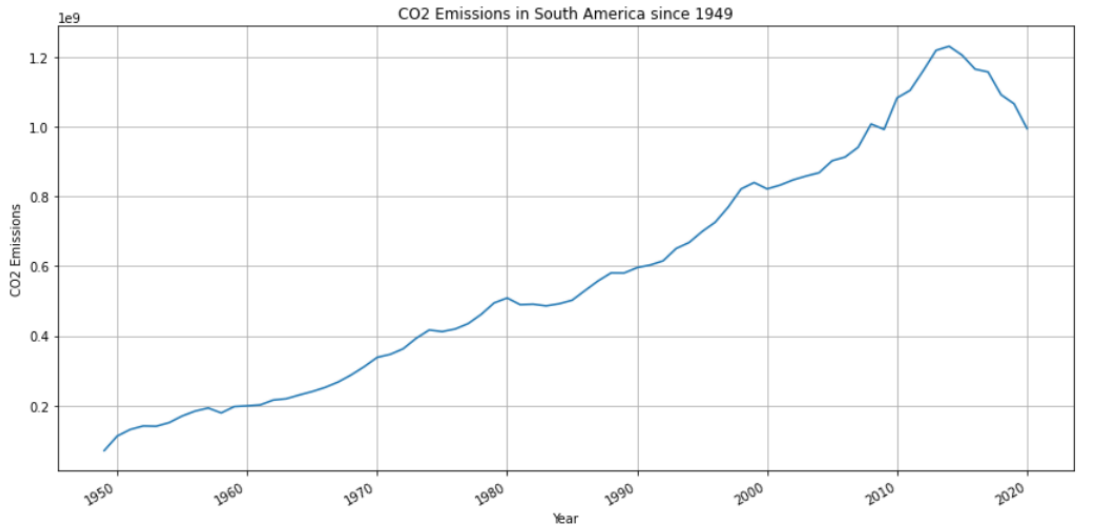

Figure 12 - CO2 Emissions in South America since 1949

When having a look to the context of South America, regarding which is the most used type of fuel, we can observe that oil have a massive usage in this continent, the emissions coming from oil are almost the double of all the other types combined. From 2013 until now we start seeing a decrease in the emissions of oil, which is related with the overall emissions that we concluded before.

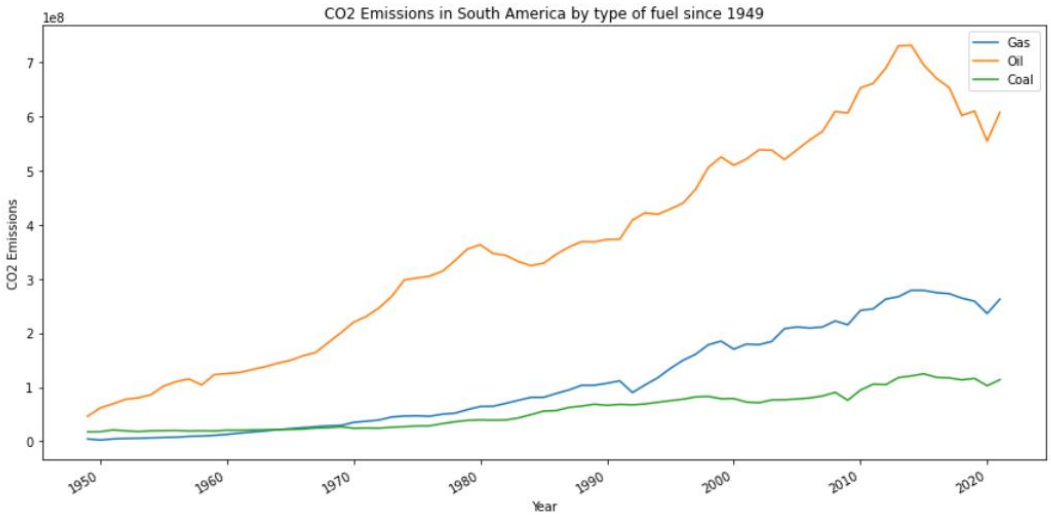

Figure 13 - CO2 Emissions in South America by type of fuel since 1949

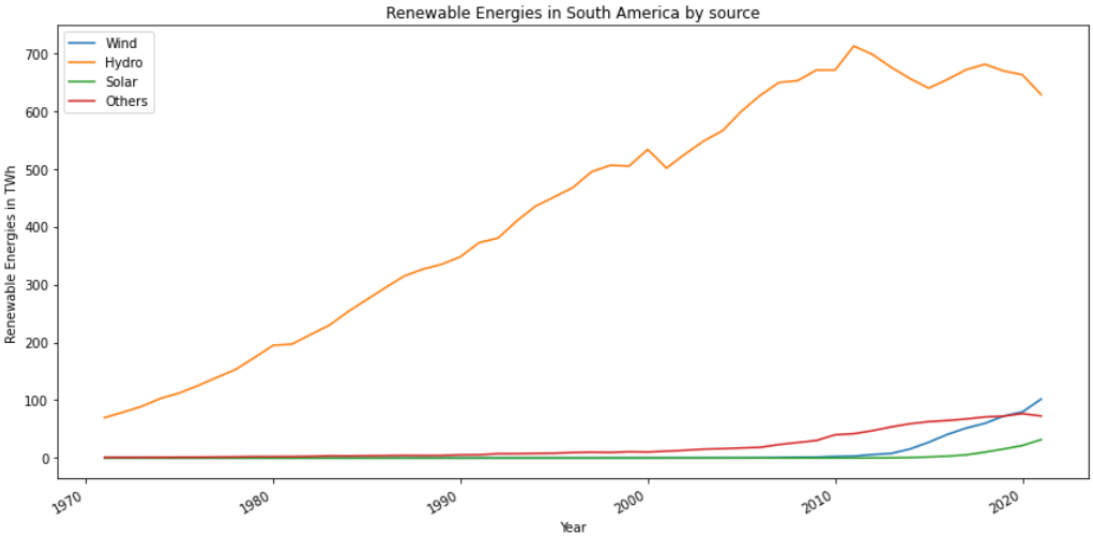

Figure 14 - Renewable Energy Production in South America since 1971 by source

Asia

Comparatively to Europe and North America, which have a defining moment where they begin to decrease the emissions and into the present they have decreases of more than 15%, Asia, the largest producer of CO2 emissions globally, is continuing to increase the emissions without a point in history where the values start to decrease. The main reason why this continent is not moving in the direction of reducing greenhouse gases may be due to the vast populations of China and India, as well as their massive industrial capacities and the absence of environmental agreements.

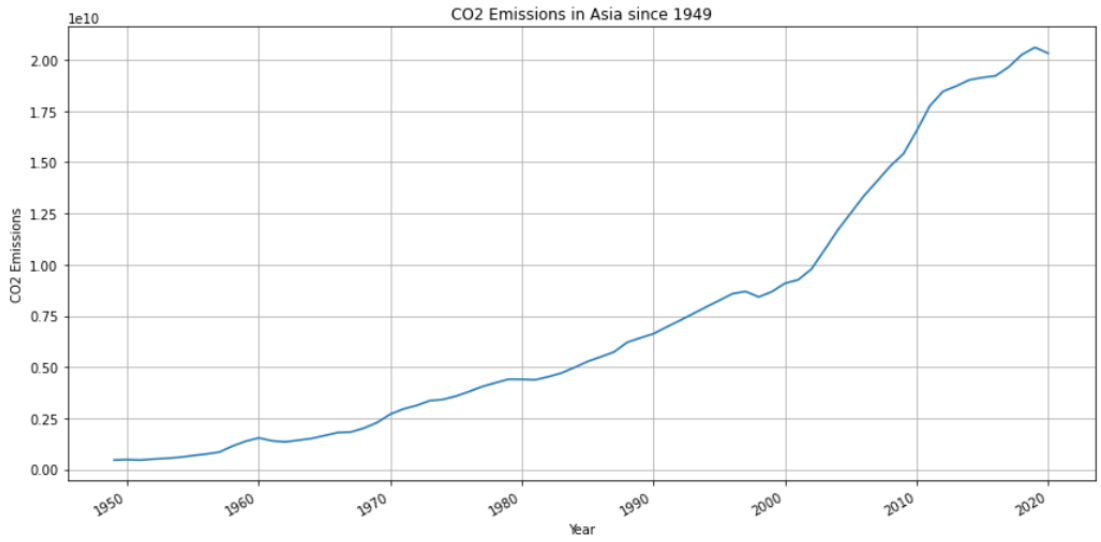

Figure 15 - CO2 Emissions in Asia since 1949

We can also see that from 2000 until the present the emissions have increase to the double, this is also due to the increase of coal as the main fuel used, in this two decades the emissions coming from coal have almost tripled.

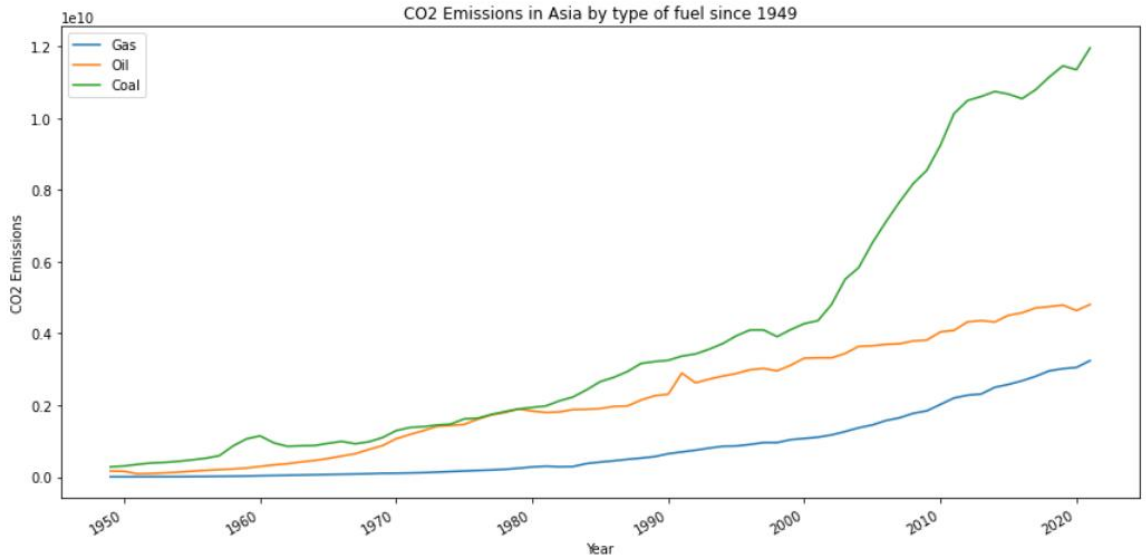

Figure 16 - CO2 Emissions in Asia by type of fuel since 1949

For Asia, the scenario is not different from the continents analyzed before. Hydro contributes more than all the other three types combined, for the production of renewable energies. Since 2010, all of the sources started increasing their productions values rampantly.

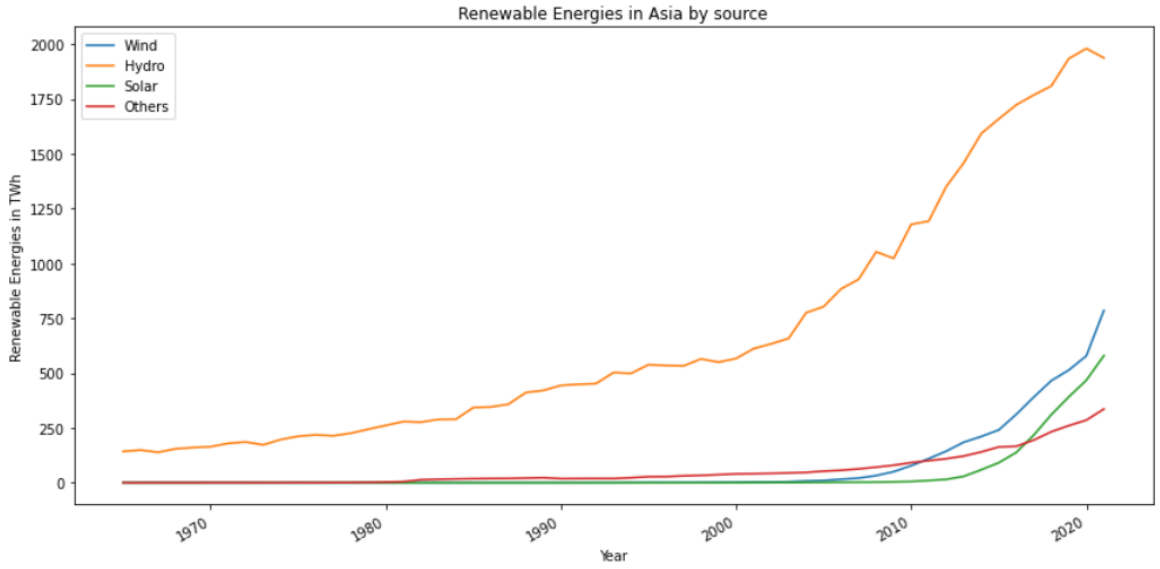

Figure 17 - Renewable Energy Production in Asia since 1965 by source

Australia

In this particular case, we have studied only Australia instead of Oceania, because it is by far the main contributor of carbon dioxide emissions in this continent.

Looking into Australia records, we can see an exponential growth similar to all the other five continents. However, in this case there is not present a decrease but regular values until the present day starting in the year 2008.

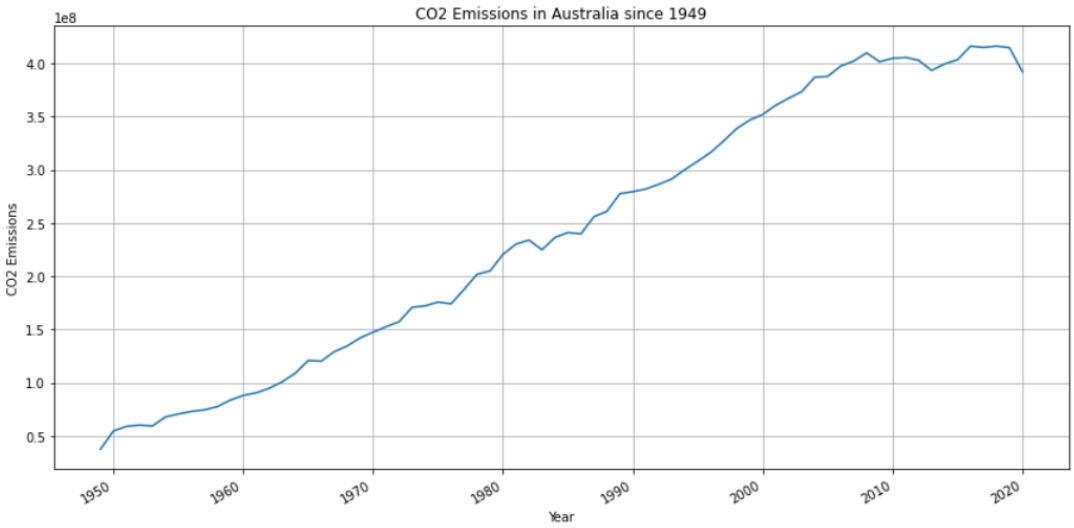

Figure 18 - CO2 Emissions in Australia since 1949

Regarding the usage of different types of fuel in Australia, when the overall carbon emissions start to be continuous is when coal emissions start to going down, looking to oil and gas, these two types start to increase exponential, but since coal is the one that per unit produces the most emissions, the continuous emissions in the last decade are correlated with this decrease.

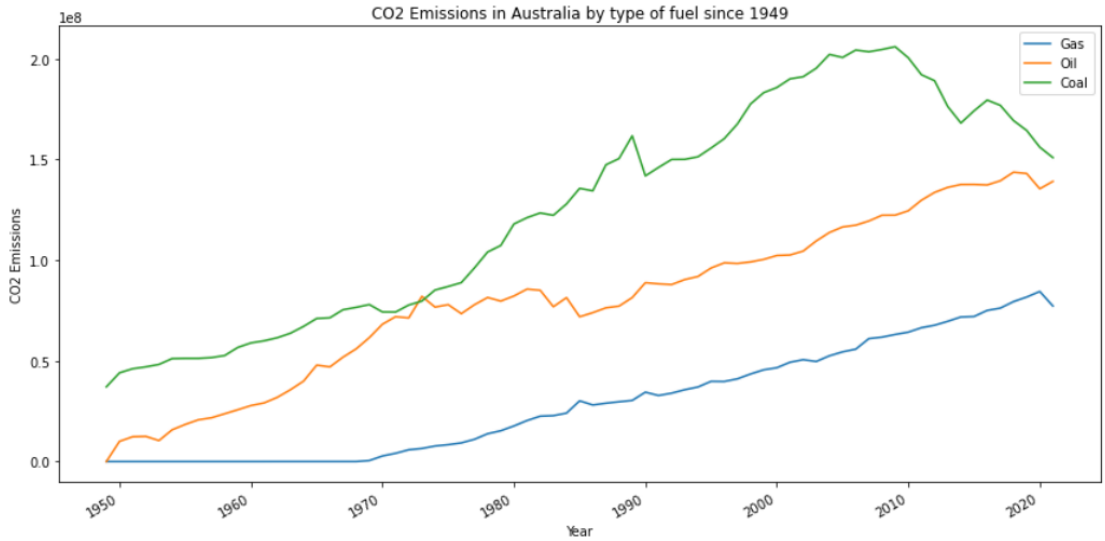

Figure 19 - CO2 Emissions in Australia by type of fuel since 1949

Different from what we have concluded in the other continents, in Australia, the main source of Renewable Energies is Solar, with almost the same production as Wind, with a value of 30 TWh.

But this lead only started in 2016, since before hydro was leading energy production. Around 2010, solar and wind sources, have started growing their productions exponentially, increasing more than 500% in the last decade.

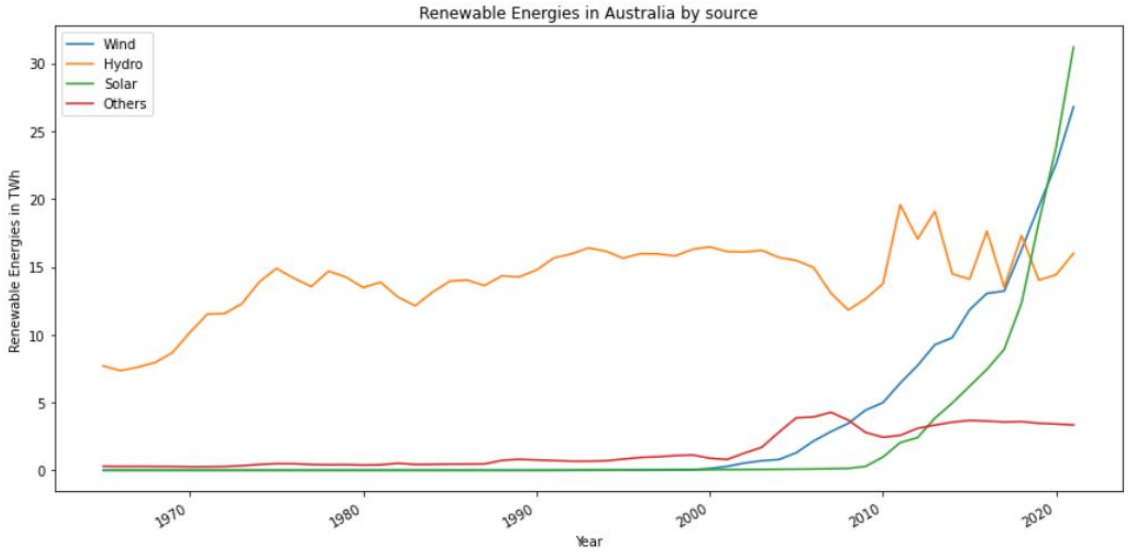

Figure 20 - Renewable Energy Production in Australia since 1965 by source

Africa

Regarding Africa, as we can see in Figure 21, is similar to what is happening in Asia, a continuous grow from 149 into the present. The main difference is the values, being Africa, one of the continents that does not produce huge emissions of carbon dioxide, But in this continent this increase on the emissions might be correlated with the urbanization of many countries presented in this continent.

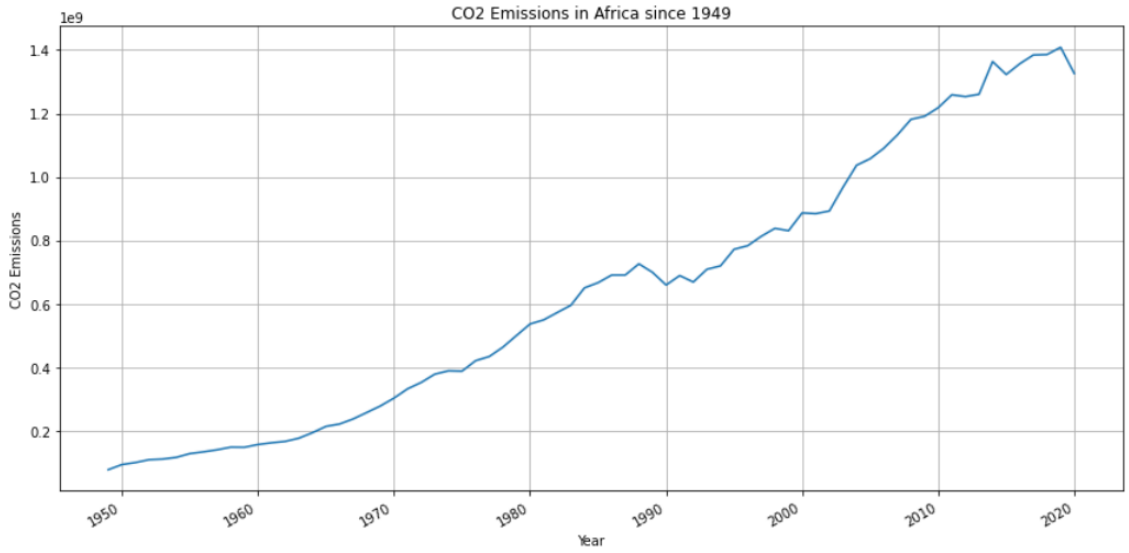

Figure 21 - CO2 Emissions in Africa since 1949

In Africa, similar to what is happening in the last year in Asia, we do not see a decrease of any type of fuel emissions, oil and coal are the most used.

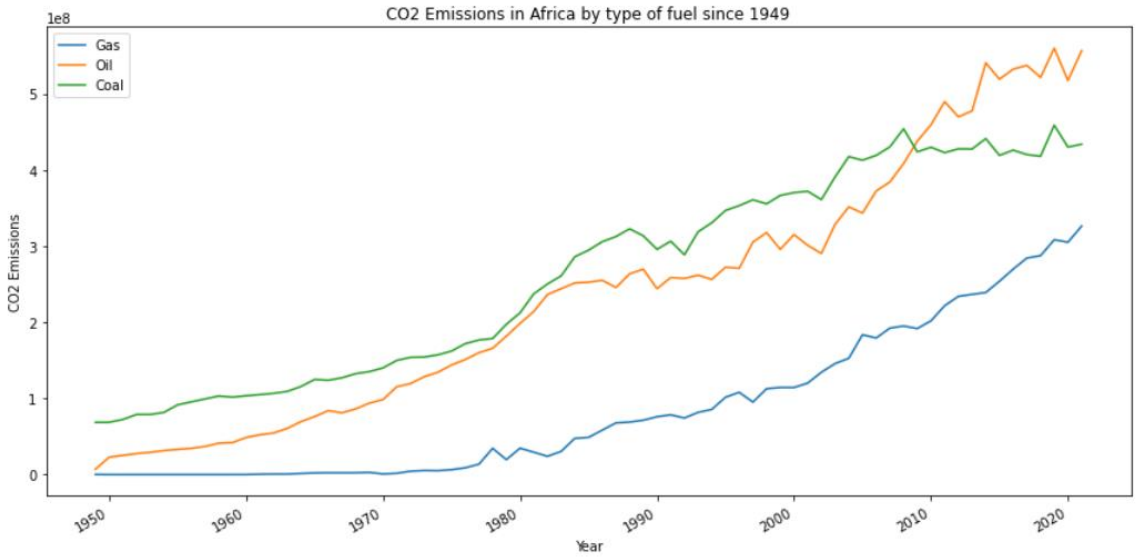

Figure 22 - CO2 Emissions in Africa by type of fuel since 1949

In Africa, as seen before in the other continents, the main source of Renewable Energies is Hydro, with a huge difference when comparing to the other fonts of renewable energy, the other types have almost the same quantity in TWh produced.

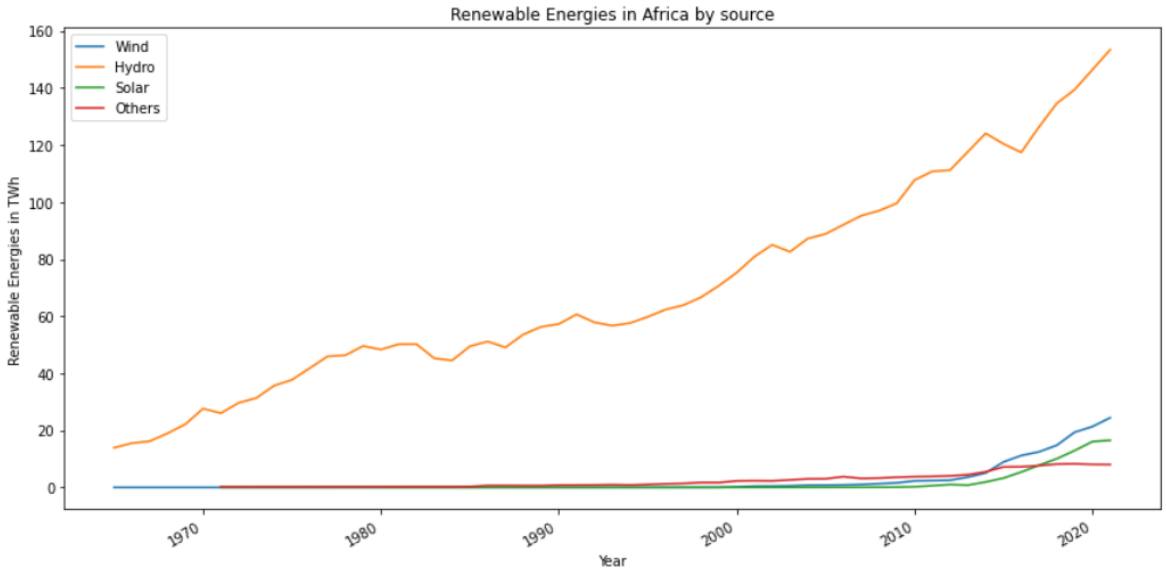

Figure 23 - Renewable Energy Production in Africa since 1965 by source

Portugal

Now that all the continents have been studied, we are going to analyze the particular case of Portugal and see if we can find any particular finding in our country.

When having a look to Carbon Dioxide emissions in Portugal, we can see a continuous increase from the beginning until 2005, where the values start to decrease until 2014 where the following 3 years are increasing and then another decrease of about 25% in the space of 5 years.

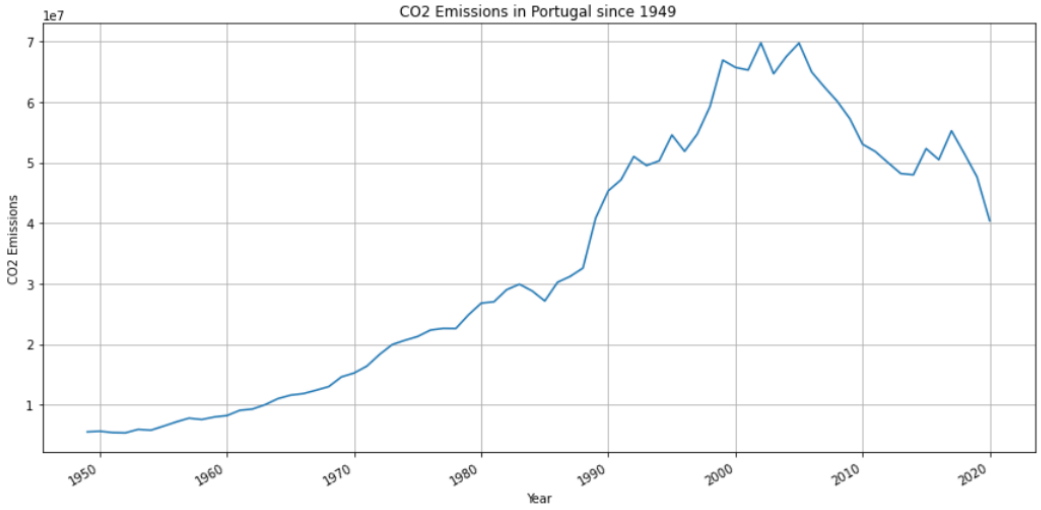

Figure 24 - CO2 Emissions in Portugal since 1949

The both decreases that we saw above, can be explain looking to the next graphic, that shows how much amount of electricity produced in Portugal come from renewables energies. In the

first decrease of carbon emissions, between 2005 and 2014, we can see perfectly a continuous rise of more than 200%, and then in the second decrease from 2016 until nowadays, the same effect happens again. We can conclude that in Portugal, the decreases of carbon dioxide emissions are highly correlated with the percentage of electricity that comes from renewables.

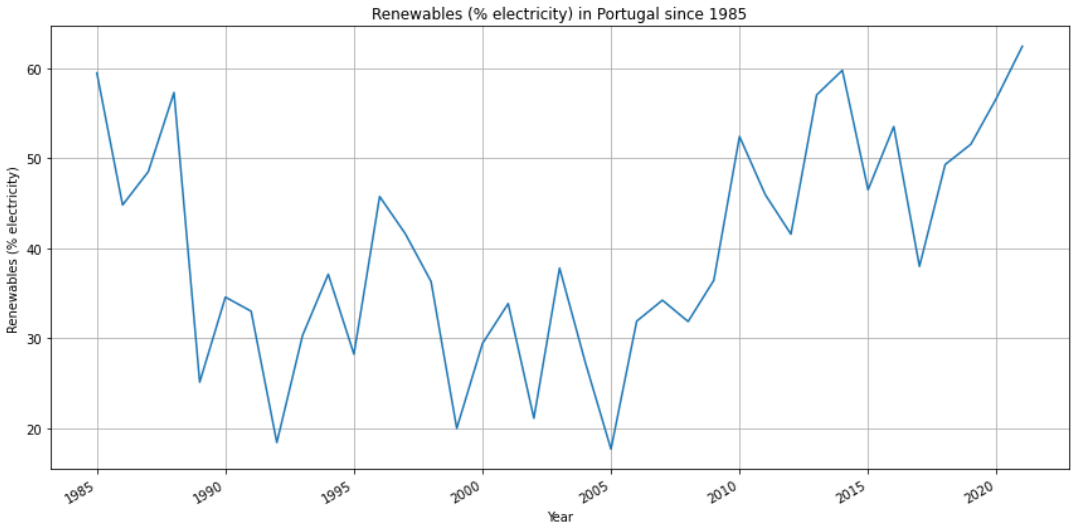

Figure 25 - Renewables (% electricity) in Portugal since 1985

We can also see that the oscillations that we saw in the figure above, are highly correlated with the Renewable Energy coming from Hydro. As seen in Figure 26, in Portugal, Hydro and Wind sources, are the ones that contribute the most to produce zero emissions energy to our country. Wind, in the last 2 decades, has been increasing a lot his contribution to the production of renewable energy, now giving Portugal the most TWh of energy.

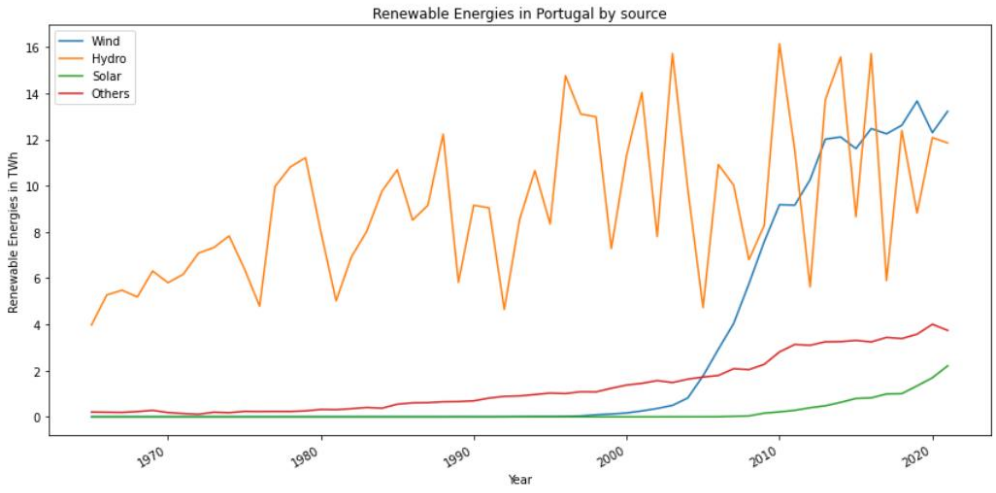

Figure 26 - Renewable Energy Production in Portugal since 1965 by source

Different from what we have seen before, when looking to the continents, in Portugal we see a massive decrease starting in 2015 of oil usage, around 45%. Regarding coal and gas, in this country they are not so used as oil, the both combined are the same as all the oil emissions.

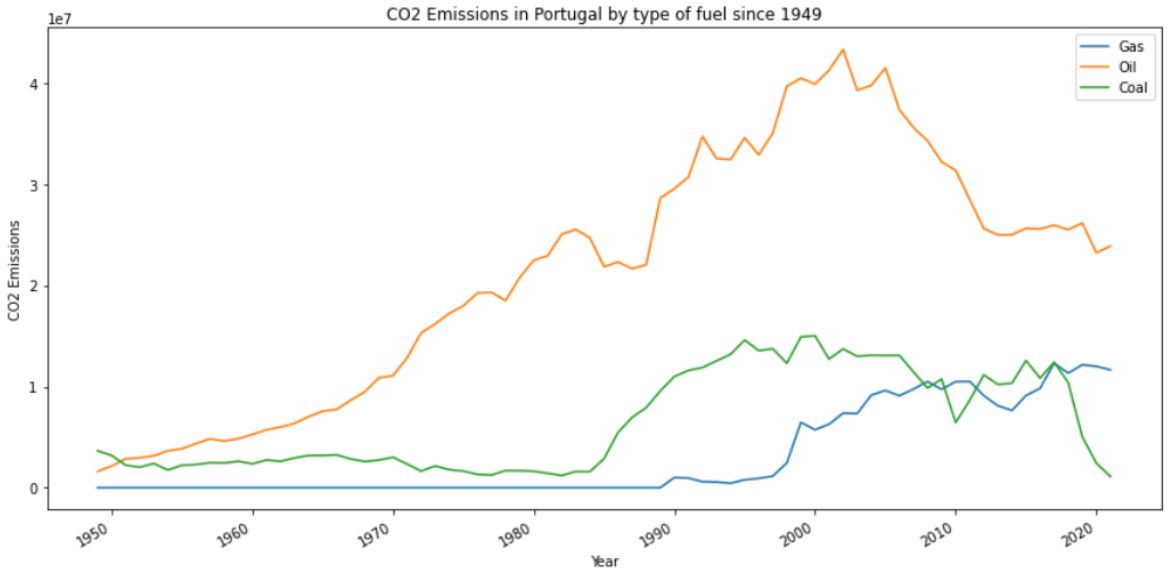

Figure 27 - CO2 Emissions in Portugal by type of fuel since 1949

Forecasting Models

After a first part of descriptive analysis and understand overall how each continent is behaving throughout the years in CO2 emissions and Renewable Energies, it is now time to start developing and applying our five models, in order to choose the best ones and forecast our observations into the future.

The code was divided into two parts, one where the three forecast models are applied and the one with the best results/less error is chosen to be used further on, and the other part where the three Machine Learning models are applied to make the same choice. With this approach, we can also understand for this problem, which technique suits better the datasets and the problem.

The first model to be tested was ARIMA, for this and all the models developed, we started by dividing the dataset into a train and test set, the training being 85% of the values of the all dataset, and the test the remaining 15%. To find the best models, all the models were developed and then tested to Europe CO2 and Renewable Energies datasets.

To develop this model, we used a Python library called ***pmdarima***, where we import a function that allows to run an automatically selected ARIMA specification, where we passed as argument our dataset and a maximum value to the variables p and q, in this case '3'. Then this function is ran and returns the best values to apply in our ARIMA for our dataset, when it runs it changed the values of p,q and m to find the combination with the lowest AIC(Akaike Information Criteria), lower the AIC better is the likelihood of the model to estimate the future values.

```
Performing stepwise search to minimize aic
 ARIMA(1,2,1)(0,0,0)[0] intercept   : AIC=281
 ARIMA(0,2,0)(0,0,0)[0] intercept   : AIC=283
 ARIMA(1,2,0)(0,0,0)[0] intercept   : AIC=283
 ARIMA(0,2,1)(0,0,0)[0] intercept   : AIC=282
 ARIMA(0,2,0)(0,0,0)[0]             : AIC=283
 ARIMA(2,2,1)(0,0,0)[0] intercept   : AIC=282
 ARIMA(1,2,2)(0,0,0)[0] intercept   : AIC=282
 ARIMA(0,2,2)(0,0,0)[0] intercept   : AIC=282
 ARIMA(2,2,0)(0,0,0)[0] intercept   : AIC=283
 ARIMA(2,2,2)(0,0,0)[0] intercept   : AIC=inf
 ARIMA(1,2,1)(0,0,0)[0]             : AIC=281

Best model:  ARIMA(1,2,1)(0,0,0)[0] intercept
Total fit time: 0.424 seconds
```

Figure 28 - Values obtained for the ARIMA Model applied to Europe values

After getting the best values to our ARIMA model, we divide our dataset into train and test, and apply the ARIMA to the train set, then we fit our model and predict 2030 against the test set.

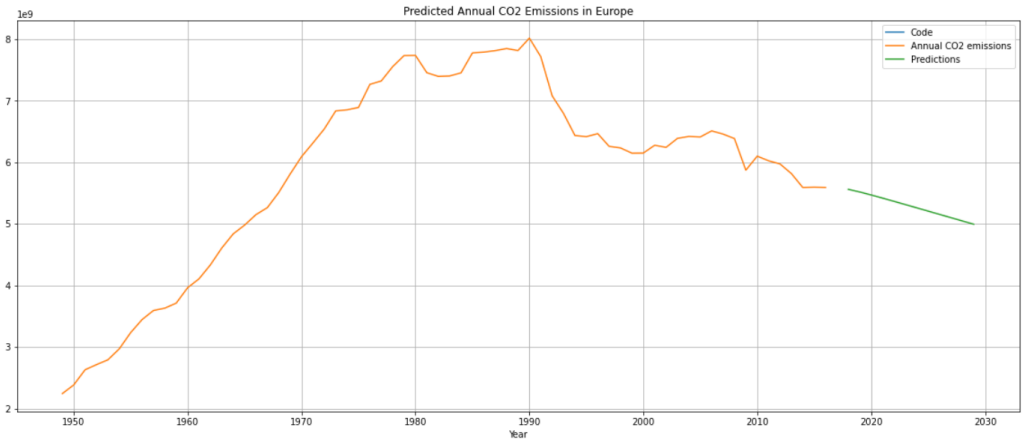

Figure 29 - Predicted Annual CO2 Emissions in Europe

Above we can observe the plot with the Annual CO2 Emissions in Europe in orange, and our Predictions values in green, after this we calculated our errors.

The next model being developed was Holt-Winters, as said before, the first steps of the previous model are the same in this one. So in this model, we started by defining three Holt-Winters models variations, each one with the difference being how you define the trend and seasonality. As seen in Figure 30, the Multiplicative and Addictive variations produce very similar predictions. However, after computing and analysing the errors, we can conclude that for this problem, addictive variation produces the best results.

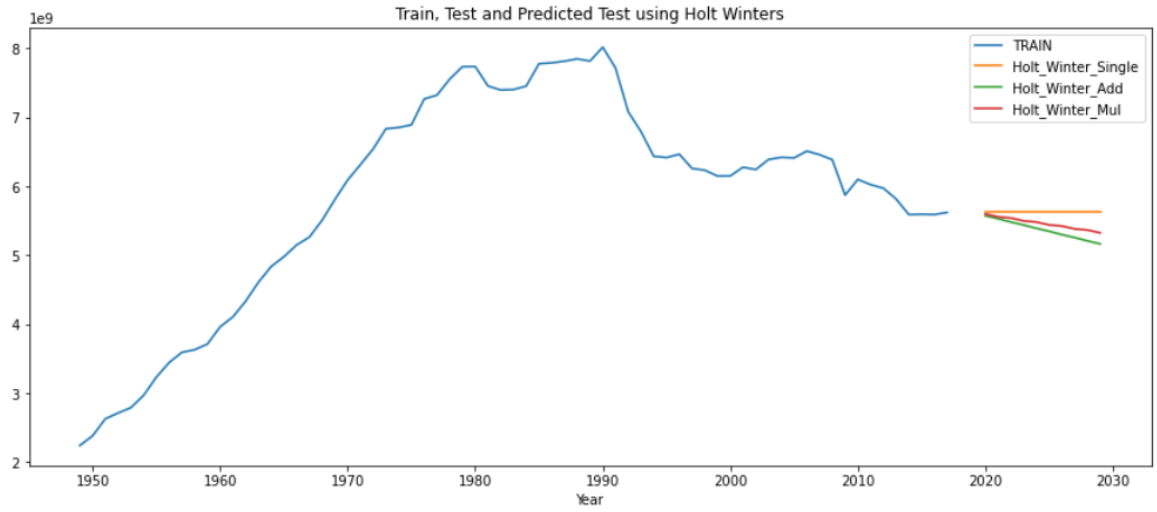

Figure 30 - Predicted values using Holt-Winters model in Europe

As a last forecasting technique, we will consider a Naïve Method. Specifically, the Naïve prediction is the last observed value taken as the forecast at all future horizons. This method assumes that the process is a Random Walk: the future is just the current state plus noise; nothing more can be predicted. However, since the Naïve Method assumes the lowest amount of predictability, it provides a benchmark for accuracy in the train-test procedure described for the methods above.

The prediction of the Naïve method can be seen in Figure 31, since the prediction is the last observation, the 10 years forecast is a straight horizontal line as we can see. As expected, out of the three methods, the Naïve tends to perform worst, indicating predictability in the emissions.

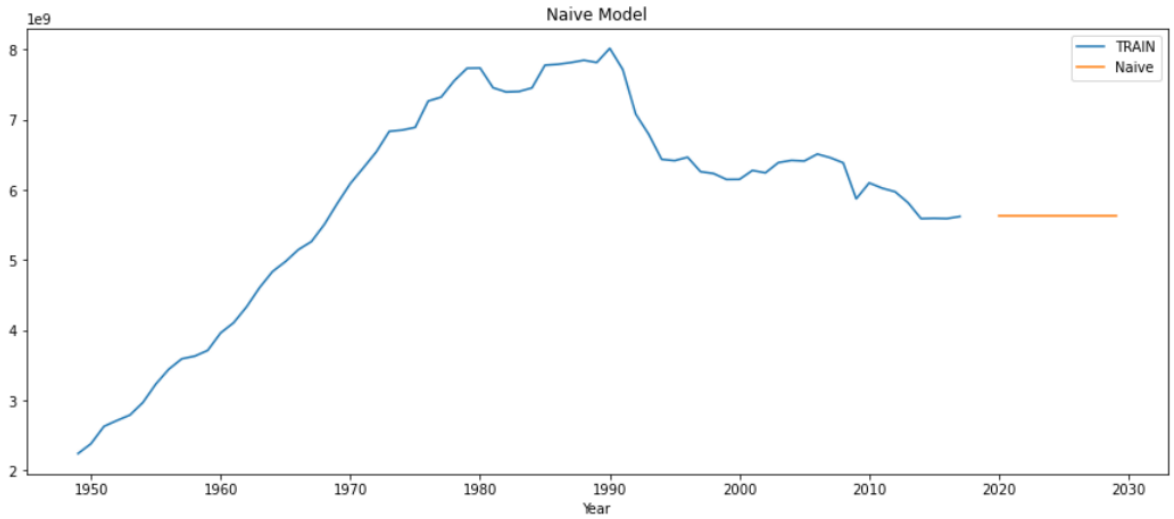

Figure 31 - Predicted values using Naive Model in Europe

Random Forest Model

Regarding this model, we start by doing the train and test division, after that a Python class was created, with 4 functions that are used to compute the Random Forest to our data, and where some of the hyperparameters were defined. As shown in Figure 32, the blue line is the mean forecast, and the red area is a 90% confidence interval with the values from the prediction done with this model.

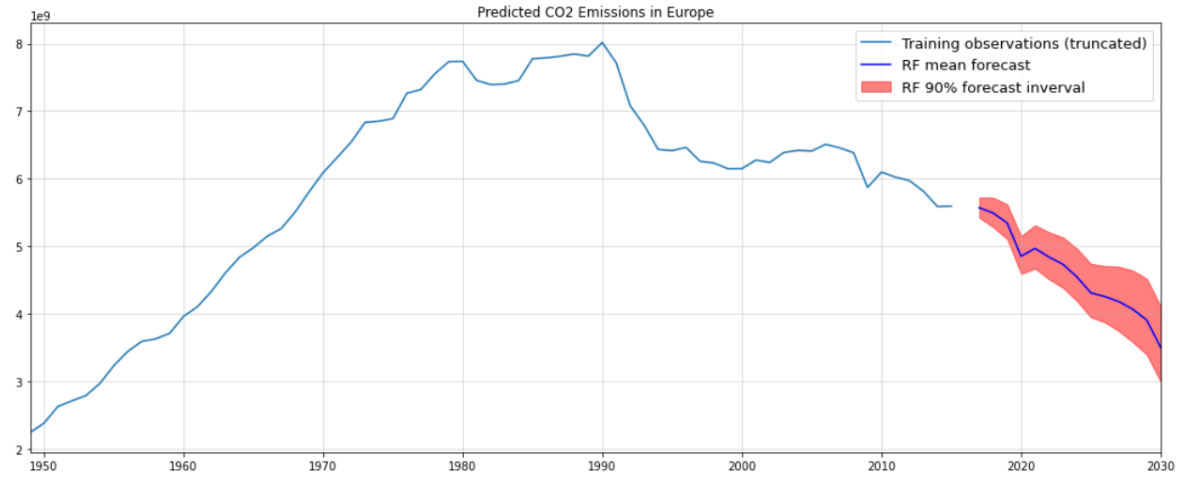

Figure 32 - Predicted CO2 values using Random Forest Model in Europe

Gradient Boosting Model

For Gradient Boosting Model, we used a Python library called *XGBoost*, that was previously described above. The results are very similar when compared to the Random Forest model, but the error is a little bit higher, so we will discard the Gradient Boosting Model.

Neural Network

After the two models previously created and analyzed, we start developing our last Machine Learning model. For that we choose to create and code a Neural Network to forecast into the next 10 years the annual carbon dioxide emissions.
For this model the beginning process are equal to the previous ones, starting by splitting the data into train and test. Then we define our neural network architecture, we choose to use a three layers network, the first layer with 128 neurons and the second one with 64 neurons, the last layer only have 1 since is the output layer. So we want to extract only one value per forecasted year, the activation function we used is ReLu.

After we compiled the model, we used the Adam as optimiser and as loss function MSE. Then we train the model, with 100 epochs forecasted for the next 10 years.
The error values were worse than the two previous models, so for the Machine Learning model, we will use Random Forest to all of our predictions.

Findings & Discussion

In this chapter, all the forecasts and results obtained from all the work that has been done previously will be presented and explained, showing what findings have been discovered and for each continent which model has performed better and also discussing some limitations that were found.
For each continent, regarding the carbon dioxide emissions ARIMA p,d,q values will be shown in the following format Arima(p,d,q).

Europe

For Europe, we start by modelling the Arima model for the training data of this continent, after running the first part of the model we got the following values ARIMA(1,2,1). Then we plot the predictions values, and we obtained the following graphic. We can see that is expected a decrease of nearly 23% until 2030. Not yet near the values that the Paris Agreement have defined for 2030 but it is a good progress already and in a longer future Europe continent will reach the values defined.

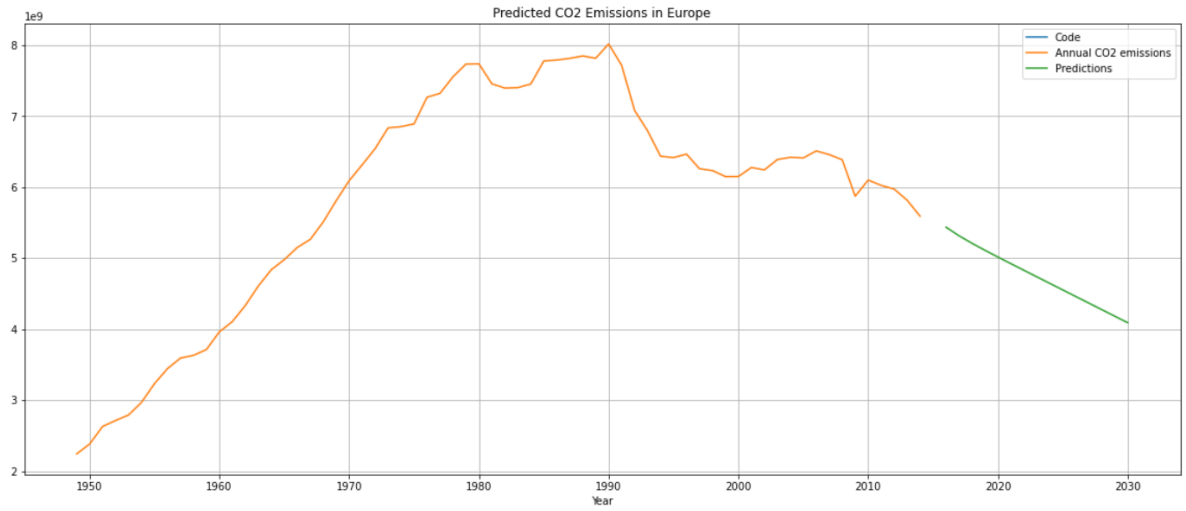

Figure 33 - Predicted CO2 Emissions in Europe using ARIMA Model

Now for the Machine Learning model, we will use Random Forest since it was the model that performed better between the both that have been trained and tested. For this model we also have predicted a decrease but a bigger difference of about 37%, with a range near the values we obtain as our prediction(blue line). The same said above is applied here, now the values for 2030 are even more close to the objective of 55% that the agreement has defined.

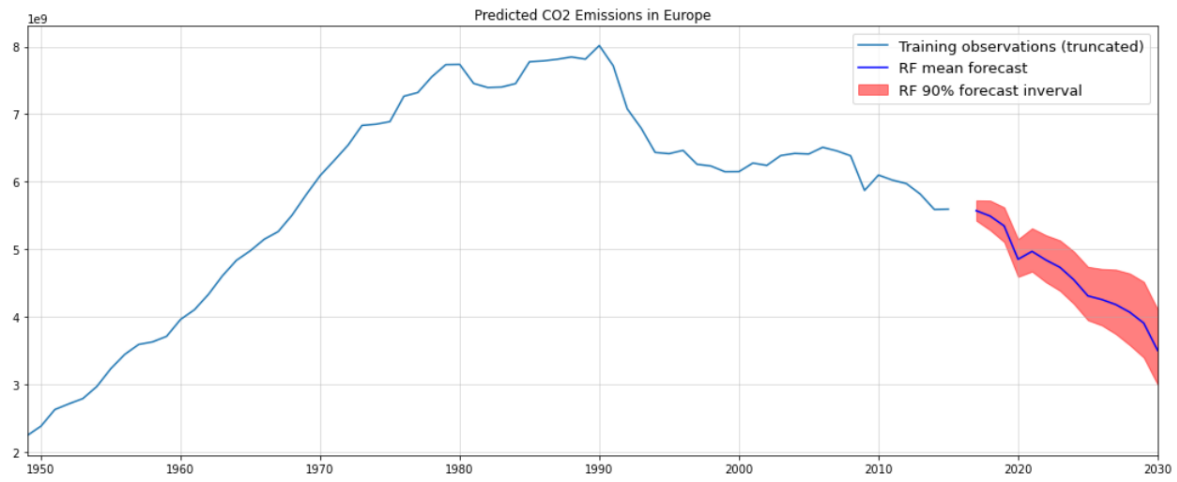

Figure 34 - Predicted CO2 Emissions in Europe using Random Forest Model

ARIMA, have shown some limitations for half of the continents analyzed by obtaining values of ARIMA(0,1,0), also known as a random walk model , that is a model where the current value of the time series is subtracted from the prior value to form a new series, which differs the data once. This new series is then modeled as a random walk, where each value in the series equals the one before it plus a random error term.

A random walk model implies that the time series' future behavior is unpredictable and does not show any clear long-term behavior or pattern. This is due to the model's underlying assumption that the series' future value will depend solely on its present value and a random error component, with no influence of any previous values.

However, it is important to understand that just because a time series is described as a random walk does not automatically imply that the data does not support any certain behavior.

Depending on the unique properties of the time series, various models or statistical methods may be more suitable for capturing more intricate patterns or trends in the data, and in our case the model that we are going

g to use to capture the patterns in the data and forecast our values into 2030 is going to be Random Forest.

In Europe, we can see that all the three types of fuel that are being studied suffer a decrease, in 2030 it is expected that coal is going to be the fuel most used, different from what we see today where oil is the most used. In the other side, gas will pass to be the less used.

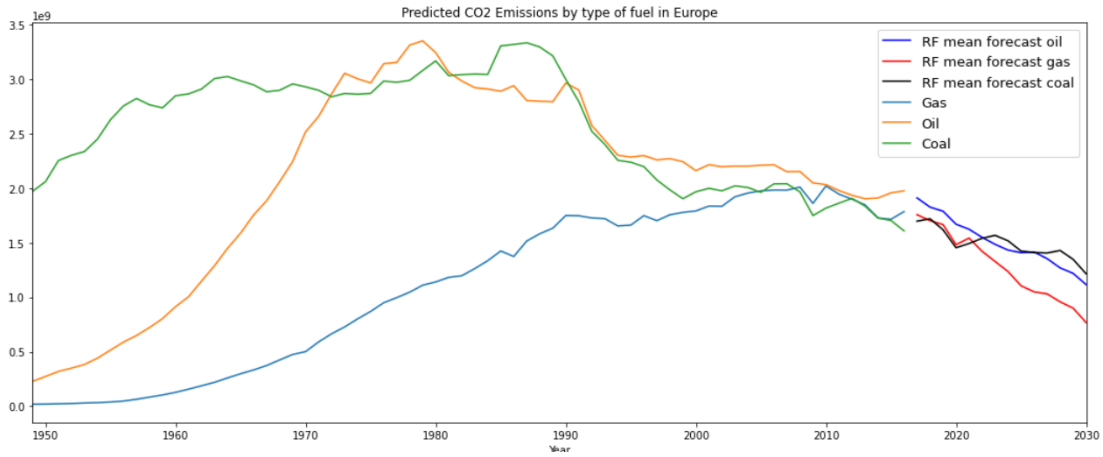

Figure 35 - Predicted CO2 Emissions by type of fuel in Europe using Random Forest Model

Overall, these findings suggest that Europe is making progress towards reducing its carbon dioxide emissions. However, more work needs to be done to reach the targets set by the Paris Agreement.

Asia

For Asia, we start by modelling the ARIMA model for the training data of this continent, after running the first part of the model we got the following values ARIMA(0,2,0). When looking for Figure 36, we see very different results from what we saw in Europe, instead of a decrease, that should be the objective in 2030 deadline, we can see an increase in the carbon dioxide emissions of about 25%. A value that is really harmful to our world, since Asia is by far the continent that produces the most emissions, and with these values going up in really high quantities that can cancel the decreases of the other continents, that are making true efforts to reduce them.

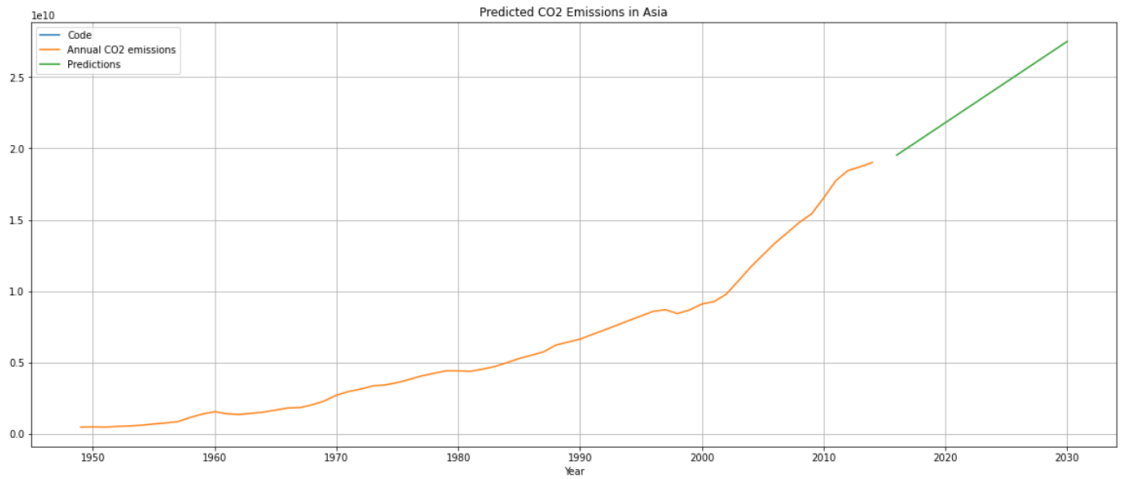

Figure 36 - Predicted CO2 Emissions in Asia using ARIMA Model

When analyzing through our Machine Learning model, we see perfectly a really huge forecast interval, that can predict both scenarios, an increase and a decrease, but when looking to the Random Forest mean Forecast we see that in this decade we will have a slightly increase of the emissions, around 10 % by 2030. Lower values when compared to Arima model, but values that are still very far from what the Paris Agreement expects.

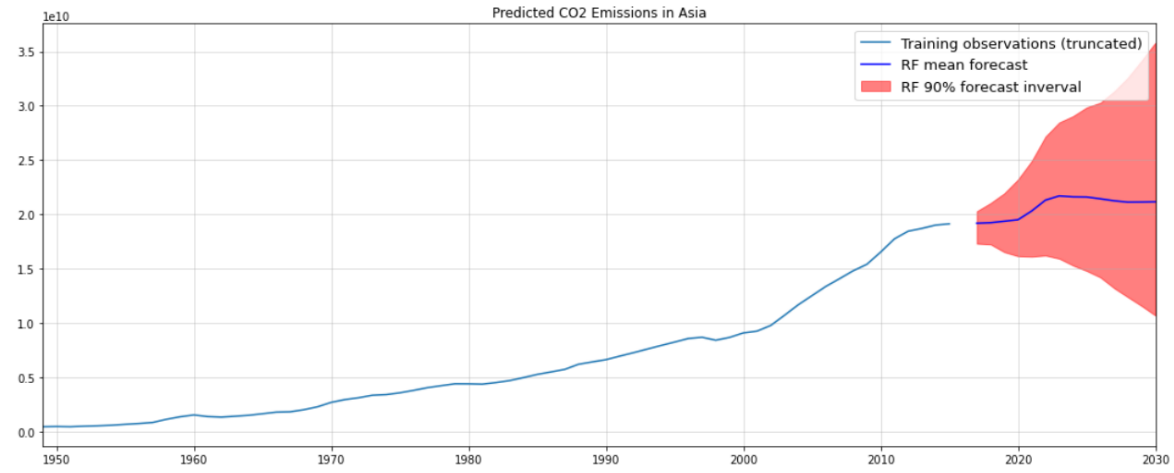

Figure 37 -Predicted CO2 Emissions in Asia using Random Forest Model

When forecasting by type of fuel, in 2030 we came to the conclusion that coal and gas emissions will increase, with coal the one that will still be the most used, with a huge difference comparatively to the other two. By 2026, gas is going to surpass oil usage, becoming the second most used fuel.

Since coal is the one, that per quantity used produces the most carbon dioxide emissions, the increase of this fuel and its usage, represent really harmful signs.

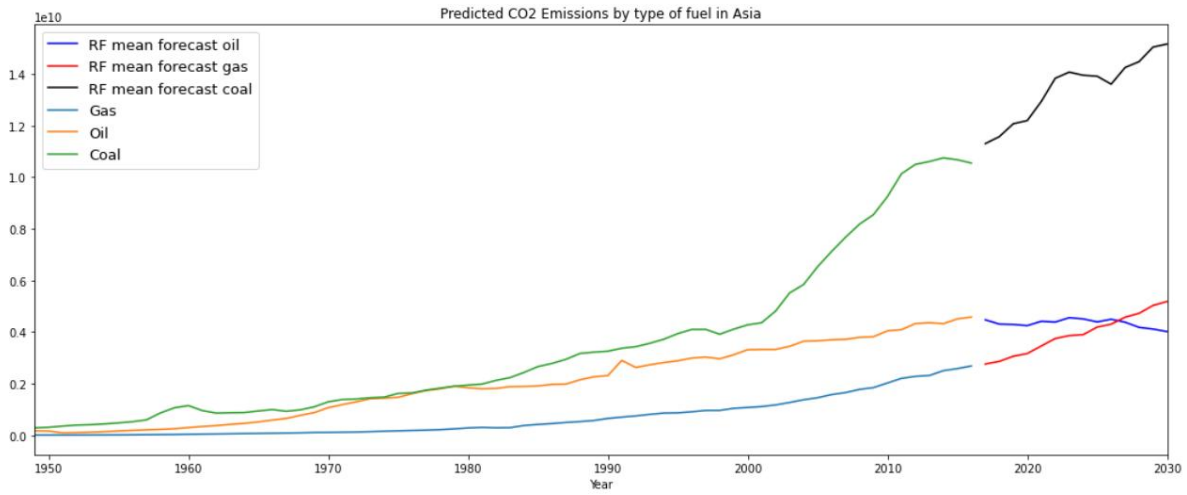

Figure 38 - Predicted CO2 Emissions in Asia by type of fuel using Random Forest Model

North America

For North America, we start by modelling the ARIMA model for the training data of this continent, after running the first part of the model we got the following values ARIMA(0,2,2). In Figure 39, we can see a small decrease of the carbon dioxide emissions by the deadline of 2030, the decrease of less than 5%, predicted by this model, is far from what was expected.

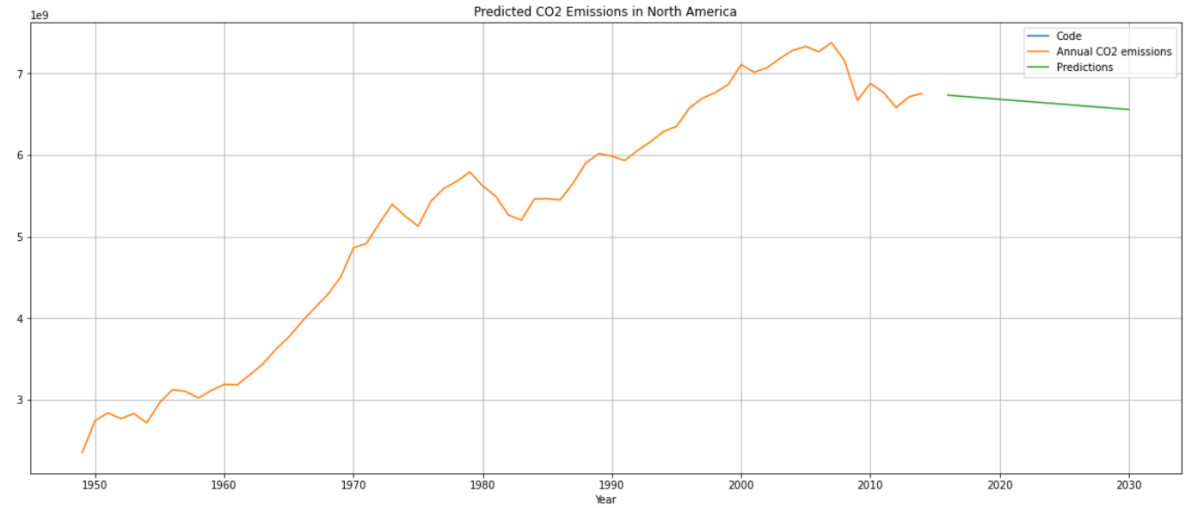

Figure 39 - Predicted CO2 Emissions in North America using ARIMA Model

When doing the same prediction, but changing the model used to Random Forest, we observe in the blue line(Random Forest mean forecast), complete different scenario that we saw previously when using the Arima model. Instead of having a small decrease of less than 5%, we have a steep descent, predicting a decrease of 36% by 2030. With these values, we are near the deadline value defined by the agreement.

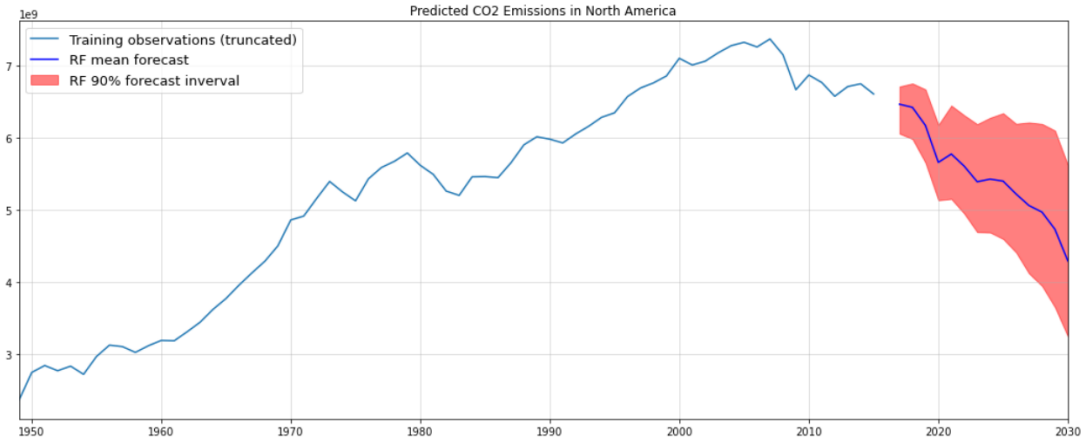

Figure 40 - Predicted CO2 Emissions in North America using Random Forest Model

Regarding the overall emissions by type of fuel we can see in Figure 41, a decrease in the emissions of coal and oil, but in the other hand an increase of almost 100% in gas fuel usage. Gas will surpass all the other fuels, and becoming the bigger responsible for producing carbon dioxide emissions in North America.

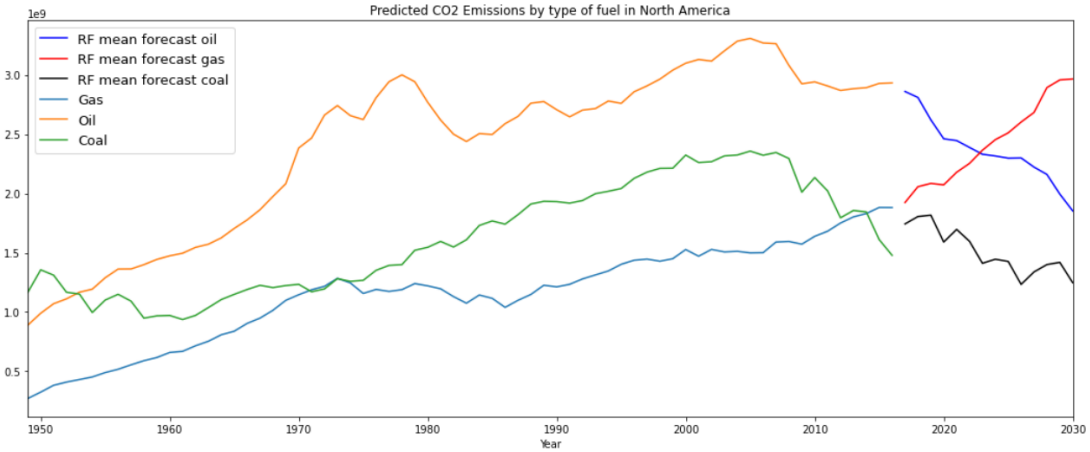

Figure 41 - Predicted CO2 Emissions in North America by type of fuel using Random Forest Model

South America

For South America, we start by modelling the Arima model for the training data of this continent, after running the first part of the model we got the following values Arima(2,1,0). We can see that a slight increase is expected, going in the opposite direction of the Paris Agreement. When compared to all the years he has presented in our data, the continuous increase of carbon dioxide emissions is still going. However, now the increase is more continuous and not so exponential.

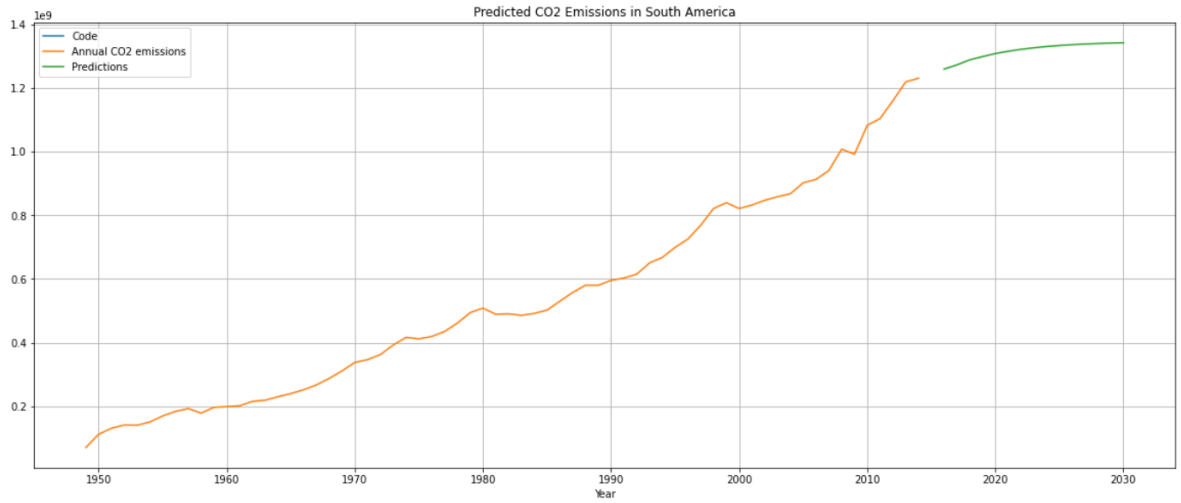

Figure 42 - Predicted CO2 Emissions in South America using ARIMA Model

For the Random Forest model, the same insight is acquired, an increase in the emissions, this model predicts an higher increase of around 45%, this value is worrisome, being the higher increase in percentage of all the continents. Let's analyze the 90% confidence interval of our forecast. Even the lower value is higher than the current value that we have, and for our top value of the interval, we observe that is the double of the current emissions present in our data of South America.

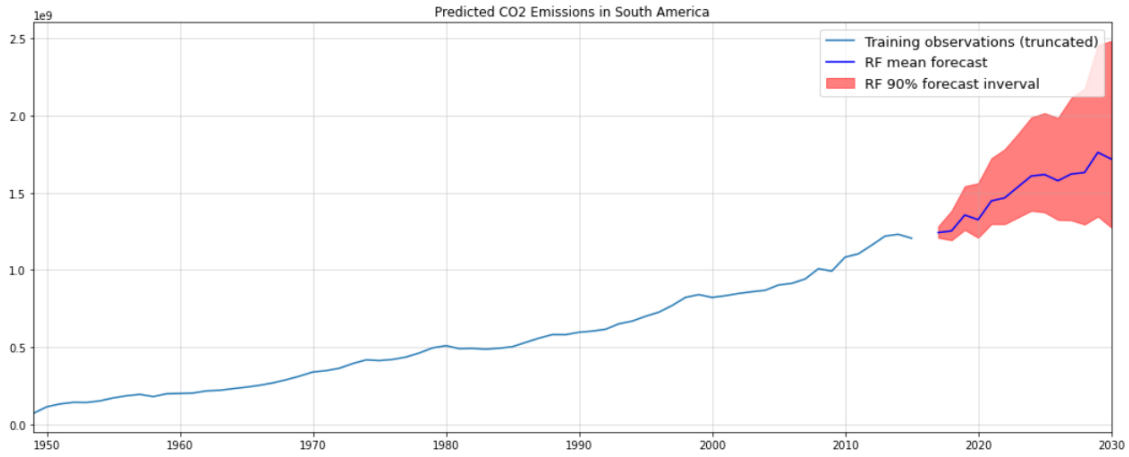

Figure 43 - Predicted CO2 Emissions in South America using Random Forest Model

Regarding the predicted carbon dioxide emissions by type of fuel, there is predicted an increase of all the three types being analyzed in this study. The highest increase is in oil, being this type of fuel the most used both in present and in our 2030 prediction. In South America, carbon dioxide emissions produced by the use of oil, are almost three times bigger than the emissions coming from the combination of coal and gas usage.

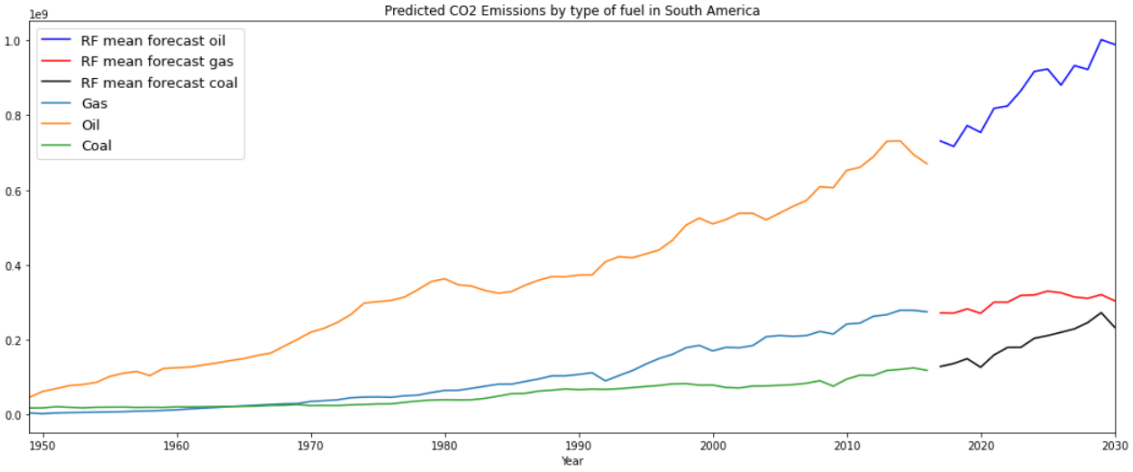

Figure 44 - Predicted CO2 Emissions in South America by type of fuel using Random Forest Model

Africa

For Africa, we start by modelling the Arima model for the training data of this continent, after running the first part of the model we got the following values ARIMA(0,1,0). In this prediction, we faced our first case of a limitation of using Arima model to forecast our data, we obtained an Arima solution where p and q have the value zero. This results in a straight line

forecast, representing only the last value we have in our train set. Since we have obtained a prediction that is not correct and do not give us any information, we just proceed to forecast using Random Forest model.

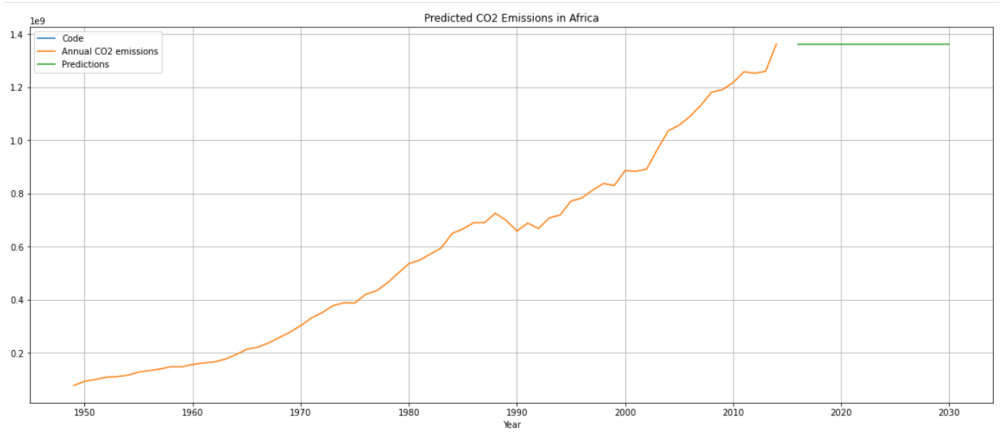

Figure 45 - Predicted CO2 Emissions in Africa using ARIMA Model

After predicting using this model, we observe an increase of around 25% in the carbon dioxide emissions, even the lowest value of our forecast interval is not a decrease in the emissions. Africa is a continent that is in currently in a huge urbanization phase for many countries, which can be a factor responsible for this increase.

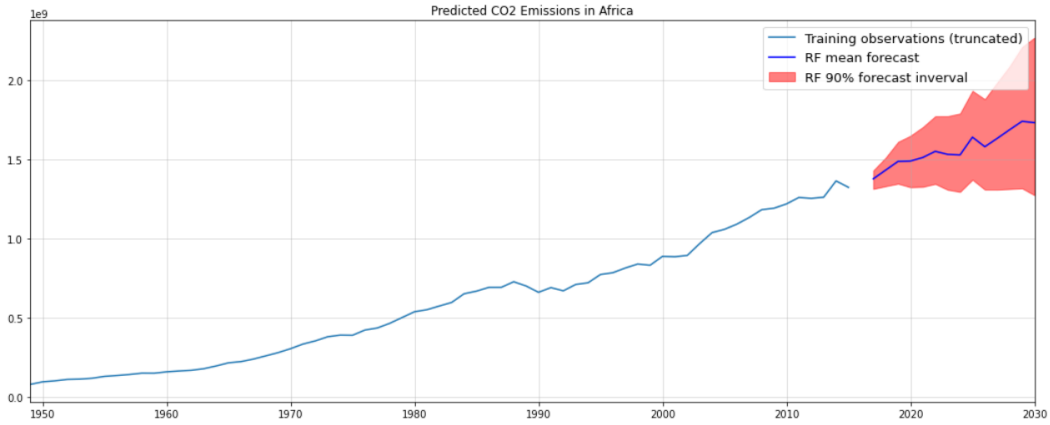

Figure 46 -Predicted CO2 Emissions in Africa using Random Forest Model

Regarding the carbon dioxide emissions by type of fuel, we see that emissions coming from coal and gas usage are going to stay in similar values as the values present in the last observation of our data, but in the other hand oil emissions are going to duplicate almost, being the main responsible of the overall increase seen in Africa. The value expected is more than the combination of emissions of gas and coal usage.

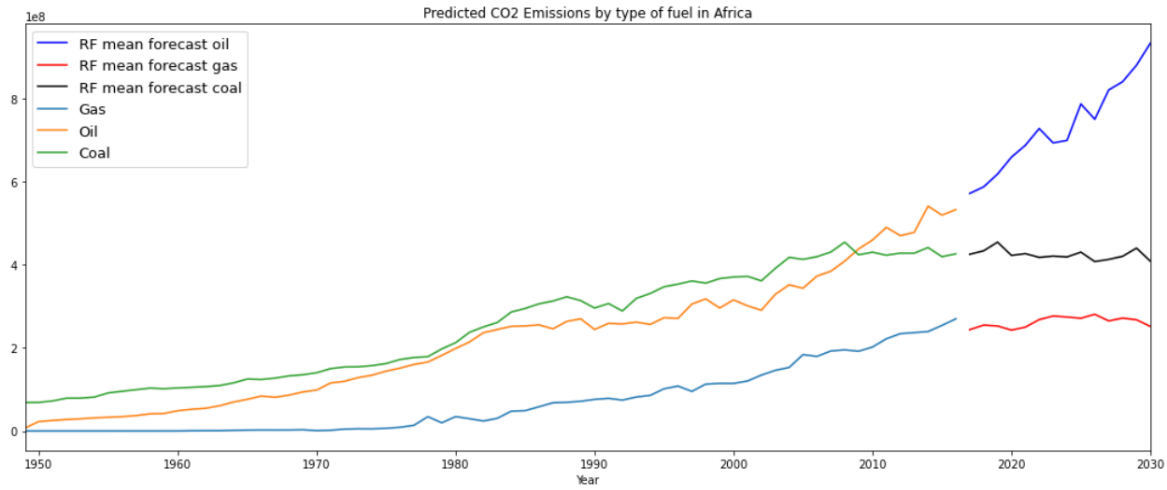

Figure 47 -Predicted CO2 Emissions in Africa by type of fuel using Random Forest Model

Australia

For Australia, we start by modelling the ARIMA model for the training data of this continent, after running the first part of the model we got the following values ARIMA(0,1,0).

By looking to the Figure 48, we can see that happened the same as seen in the prediction using the Arima model for the African continent, a straight line based on the value of the last observation.

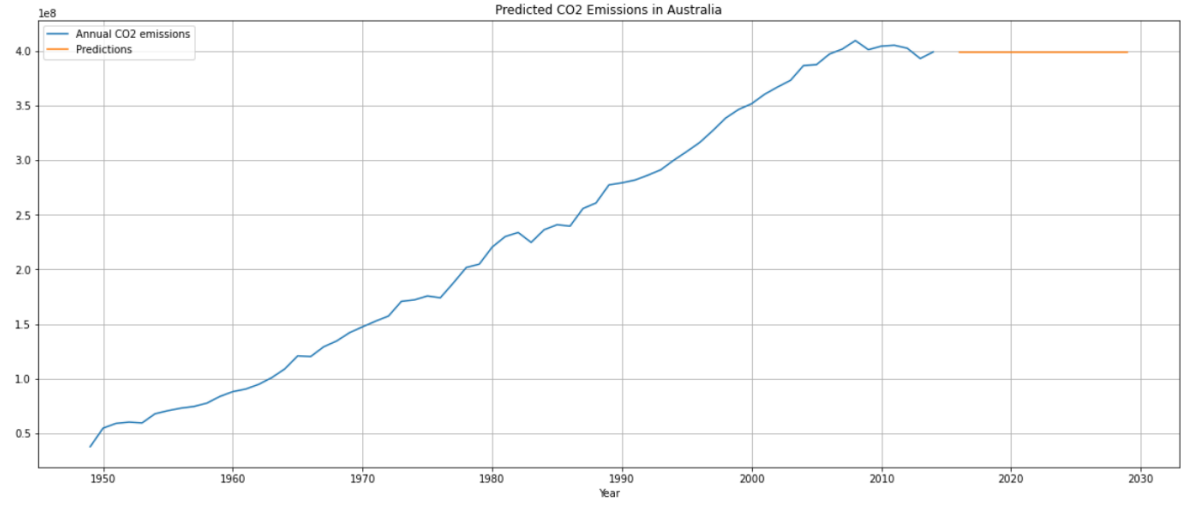

Figure 48 - Predicted CO2 Emissions in Australia using Arima Model

By applying the Random Forest algorithm, we predict that a decrease of near 10% will happen by 2030, a value far away from what is expected by the agreement, but we can also consider a

progress when comparing to Africa and Asia. Overall, Australia needs to consider several measures to start seeing these values decreasing more exponentially.

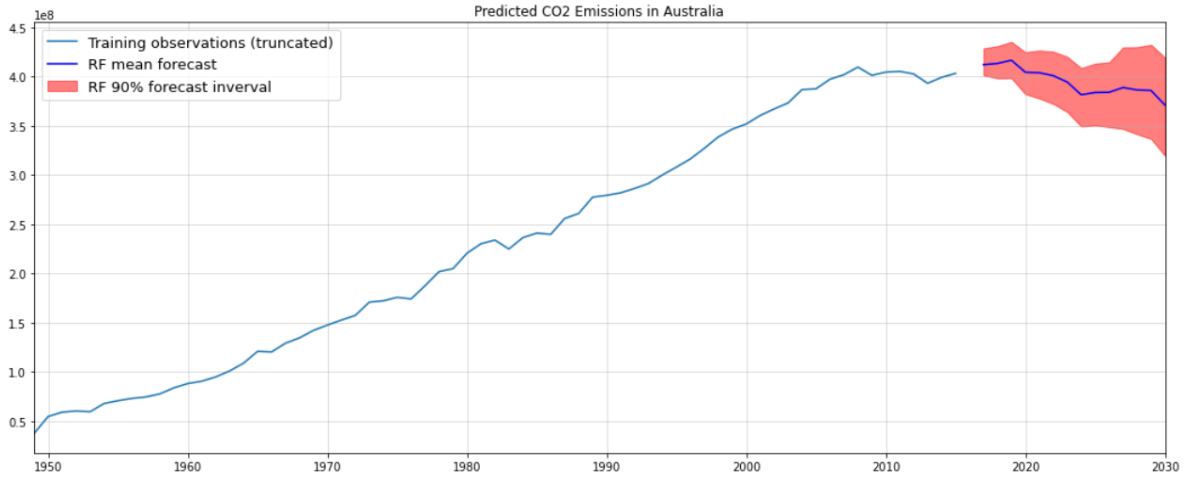

Figure 49 - Predicted CO2 Emissions in Australia using Random Forest Model

Regarding the carbon dioxide emissions by type of fuel, in Figure 50 we conclude that emissions coming from oil and gas usage are going to stay in similar values as the values present in the last observation of our data, but in the other hand coal emissions are going to decrease 15% and to stabilize after 2026.

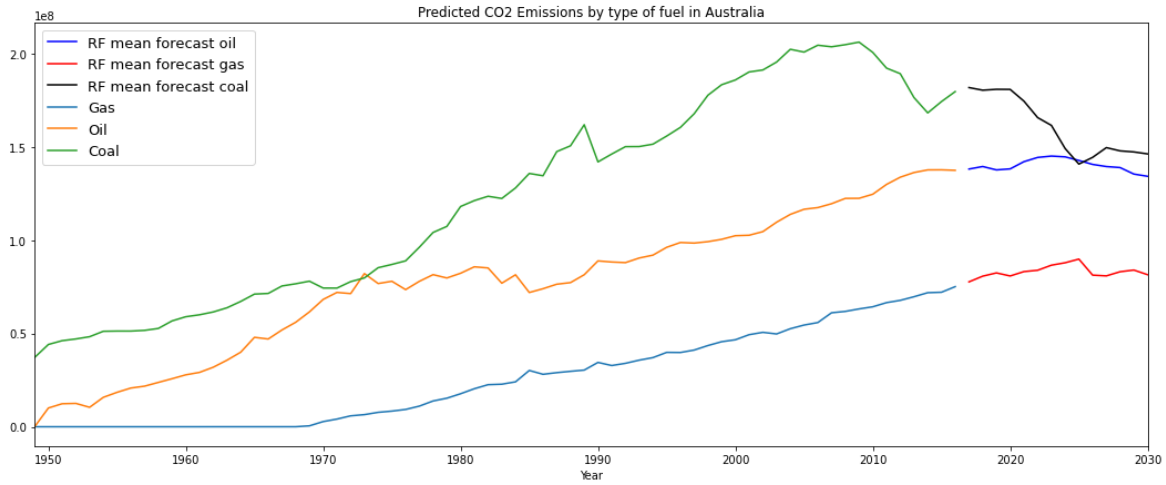

Figure 50 - Predicted CO2 Emissions in Australia by type of fuel using Random Forest Model

Portugal

For Portugal, we start by modelling the Arima model for the training data of this continent, after running the first part of the model we got the following values ARIMA(0,1,0).
As seen before, for Portugal, we obtain again the limitation of Arima model, a p and q equal zero, that represents a forecast based only on the last value of our training data.

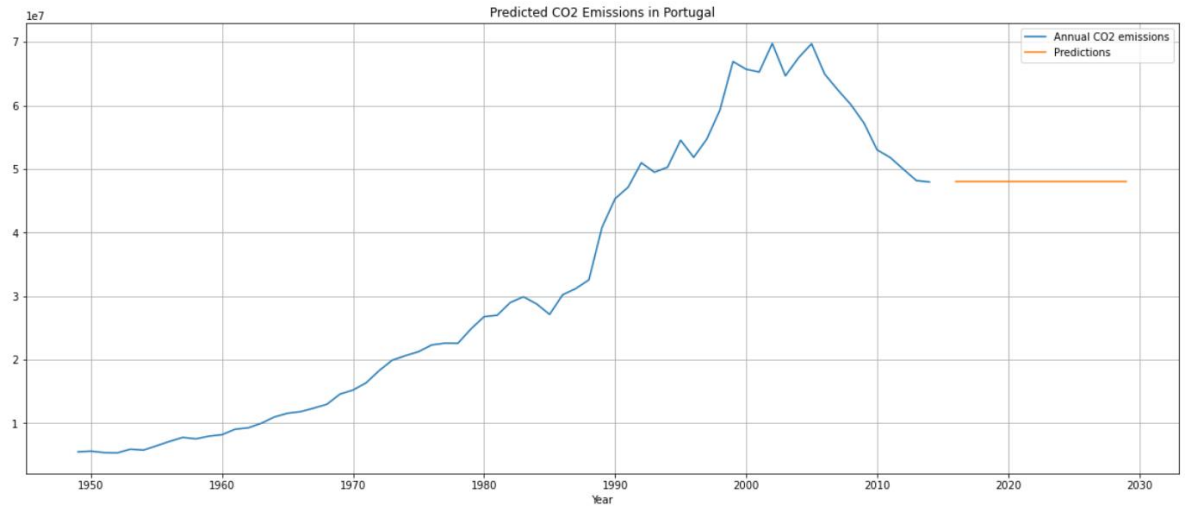

Figure 51 - Predicted CO2 Emissions in Portugal using ARIMA Model

Applying the Random Forest model, we can analyze a decrease around 40% on the carbon dioxide emissions, this decrease was expected since when analyzing Europe, we saw a big decrease in the emissions, and since Portugal is part of Europe, a decrease was expected to.

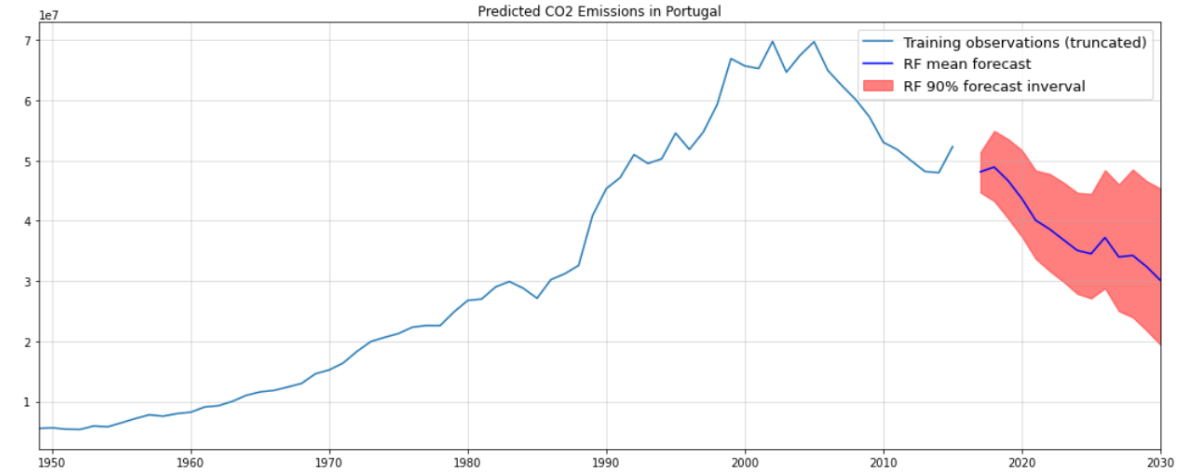

Figure 52 - Predicted CO2 Emissions in Portugal using Random Forest Model

Regarding the overall emissions by type of fuel we can see in Figure 53, a decrease in the emissions of gas and a huge decrease in carbon dioxide emissions produced by oil, but in the other hand an increase of almost 60% in coal fuel usage. Coal will surpass all the other fuels, and becoming the bigger responsible for producing carbon dioxide emissions in Portugal.

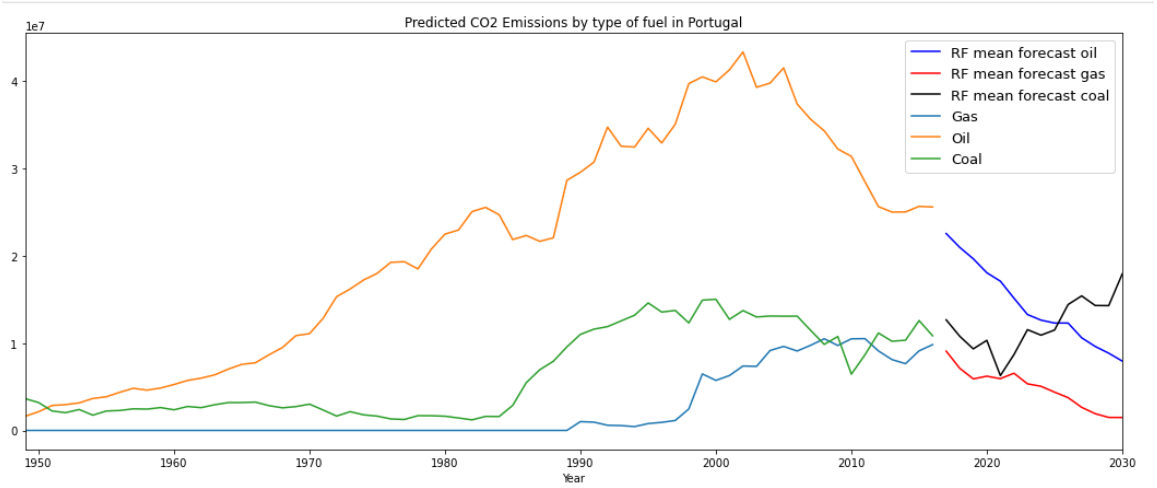

Figure 53 - Predicted CO2 Emissions in Portugal by type of fuel using Random Forest Model

Concluding Analysis

IPCC, The Intergovernmental Panel on Climate Change as released this year the new AR6 Synthesis Report, which graphic can be seen in Figure 54. This report is based on a forecast into the future about how the global CO2 emissions will be predicted and affected based on policies and mitigation strategies implemented by nations worldwide. Ranges of modelled pathways that limit warming to 1.5°C (>50%) with no or limited overshoot are shown in light blue and pathways that limit warming to 2°C (>67%) are shown in green, the red range predict emissions pathways assuming policies that were implemented by the end of 2020. (Change 2022)

In this dissertation, the same forecast as been done, with the difference being in how the forecast works and the training data. Instead of doing predictions that incorporate policies, the predictions shown in Figure 55 are based and the model had only learned from past trends and CO2 values that happened in the past.

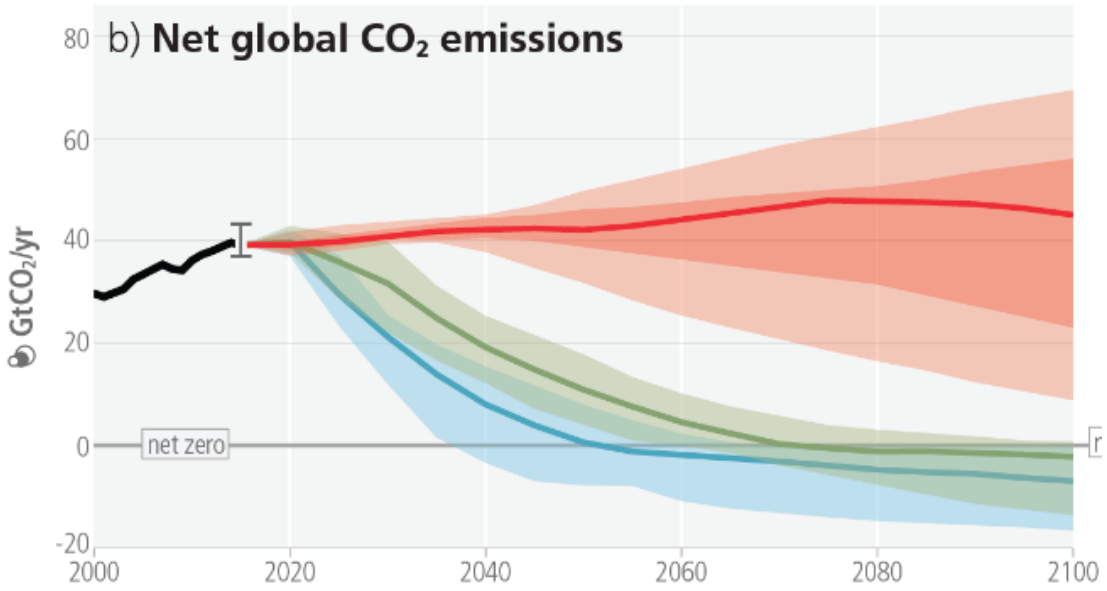

Figure 54 - Net Global CO2 Emissions of the IPCC AR6 Syntesis Report

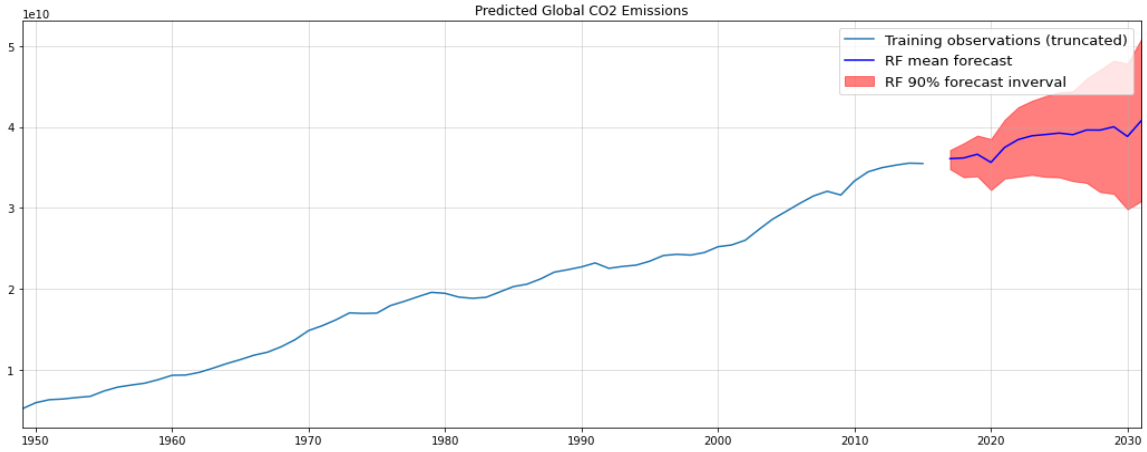

Figure 55 - Predicted Global CO2 Emissions using Random Forest

We can have the same conclusion by looking at both images and values from two different models, one that incorporates policies in the predictions(IPCC) and one that only uses past trends. Until 2030 the carbon dioxide emissions globally will have an increase. So in both cases the emissions value in 2030 will surpass the 4,1 GtCO2 for the first time in history.

It is also interesting to notice that our forecasts indicate much larger uncertainty at the horizon 2030, at this horizon our confidence intervals are much higher than the ones in the IPCC forecast, this can be due to the lack of consideration and policies data in our training data, that is the main difference between the two models. We can conclude that in this topic, worldwide

policies that take into account the reduction of carbon dioxide emissions are crucial to generate a more accurate model.

Chapter 4

Conclusion

In conclusion, this study aimed to investigate carbon dioxide emissions in all continents and make a forecast until 2030 using machine learning and traditional forecast techniques. By collecting and analysing data on emissions by type of fuel and renewable energy sources, we were able to better understand the current state of carbon dioxide emissions in the world.

Our results showed significant differences in predicted levels of carbon dioxide emissions among the different continents, it was found that Asia, South America, and Africa will experience an increase in emissions, while Europe, North America, Australia and Portugal will see a decrease. The methods used for forecasting were Random Forest and ARIMA, with the latter having the limitation of only providing the mean of the last observed data value due to its p=0, d=1, q=0 values. Another limitation that this study can have, regarding the predictions done for the carbon dioxide emissions, is the current war that is happening in Ukraine, can lead to higher usage of coal mainly, but also another types of fuel that produce a lot of emissions, this even is a limitation for ours predictions since it can affect them, leading to an uncertain forecast.

Overall, it can be concluded that there are significant differences in the predicted levels of carbon dioxide emissions among the different continents, with some regions expected to see an increase and others a decrease. But none of the continents will reach the deadline value of 55% decrease that Paris Agreement as defined to all of them. North America and Europe are in a good path to decrease the carbon dioxide emissions, but still 20% away from the values expected. We can also conclude for this study that the Machine Learning model performed better to almost all predictions than the forecast technique used, ARIMA.

Further research is needed to understand the underlying factors contributing to these trends and to develop strategies for reducing emissions in all regions, it is worth noting that forecasting CO2 emissions can be a challenging task, as it depends on a variety of factors such as economic growth, energy policies, technological advancements, and environmental regulations.

Therefore, the accuracy of the predictions may depend not only on the quality of the models but also on external factors that are difficult to predict.

References

References

Administration, U.S. Energy Information. 2022. "U.S. Energy Information Administration." *Carbon Dioxide Emissions Coefficients.* October 5. https://www.eia.gov/environment/emissions/co2_vol_mass.php.

—. 2022. *U.S. Energy Information Administration.* https://www.eia.gov/energyexplained/renewable-sources/.

Agency, United States Evironmental Protection. 2022. *epa.gov.* September. https://www.epa.gov/laws-regulations/summary-clean-air-act.

Athanasopoulos, Rob J Hyndman and George. 2018. *Forecasting: Principles and Practice (2nd ed).* Monash University, Australia.

Breiman, Leo. 2001. *Random Forests.* Statistics Department University of California.

Brownlee, Jason. 2016. *Machine Learning Mastery.* September 9. Accessed 2022. https://machinelearningmastery.com/gentle-introduction-gradient-boosting-algorithm-machine-learning/.

Change, The Intergovernmental Panel on Climate. 2022. *AR6 Synthesis Report.* https://www.ipcc.ch/report/ar6/syr/figures/summary-for-policymakers/figure-spm-5/.

Comission, European. 2021. *European Comission.* https://climate.ec.europa.eu/eu-action/eu-emissions-trading-system-eu-ets/emissions-cap-and-allowances_en.

Janiesch, C., Zschech, P. & Heinrich, K. 2021. "Machine learning and deep learning." *Electronic Markets.*

Kristina Mohlin, Alex Bi, Susanne Brooks, Jonathan Camuzeaux and Thomas Stoerk. 2019. "Turning the corner onUSpowersectorCO2emissions—a 1990–2015statelevelanalysis." *Environmental Research Letters.*

Oxford, University of. 2022. *Our World In Data.* https://ourworldindata.org/co2-emissions.

Rish, Irina. 2001. "An Empirical Study of the Naive Bayes Classifier." *IJCAI 2001 Work Empir Method Artif Intell.*

United Nations. 2015. "Paris Agreement."

—. 2016. *United Nations.* https://www.un.org/en/climatechange/paris-agreement.