

World Temperature

Project

Kamila Drobek, Dominik Flüchter, Natalia Gostkowska-Lekner

August 2023

Summary

This project presents the analysis of the world temperature data over the last decades. It provides an insight into the global warming problem and the factors that are causing it. The data of surface temperature anomalies are treated together with the fossil fuel emissions data and modeled using machine learning algorithms. This report presents a comprehensive study on the data, treating and pre-processing the datasets and modeling using four algorithms. The data exploration is followed by the detailed statistical summary and description of preparing the data for modeling. Four models are used and evaluated with the Gradient Boosting Regressor model being the most suitable for the modeling of the world temperature anomaly (target value). The greatest impact on the target value comes from the CO2 emissions from various sources (eg. oil, coal). The report is concluded with the personal conclusion on the project from the project members.

Table of contents

1. Introduction
   1. Project objectives
   2. Data sources
2. Data Exploration
   1. Dataset description and exploration
   2. Statistics
      1. Missing values
      2. Columns importance
   3. Data cleaning and pre-processing
3. Modeling
   1. Selection of forecasting methods
      1. Linear Regression
      2. Gradient Boosting Regressor
      3. Decision Tree Regressor
      4. Random Forest Regressor
   2. Development of the models
   3. Evaluation of model’s accuracy
4. Conclusions and outlook
   1. Personal conclusion
5. Introduction

The Earth’s climate and temperature is increasingly affected by usage of fossil fuels, deforestation and livestock farming. The greenhouse emissions related to these activities have significantly increased causing the increase of greenhouse effect and global warming. Man-made global warming is currently increasing at a rate of 0.2 °C per ten years [1]. Human activities are increasing the presence of some of the greenhouse gases in the atmosphere, in particular [1]:

* carbon dioxide (CO2)
* methane
* nitrous oxide
* fluorinated greenhouse gasses.

1.1. Project objectives

The influence of the emission of greenhouse gasses on global warming and geographical analysis of the temperature anomaly is studied in this project.

This project aims to find out the global warming and disruption of the Earth’s climate over the decades. The main objective is to investigate whether there is a link between CO2 production and greenhouse gasses emission with surface temperatures anomaly. The geographical and economical aspects are also studied.

The members of the group present a variety of backgrounds. Their expertise includes sales for Kamila Drobek, software development for Dominic Flüchter and physics for Natalia Gostkowska-Lekner.

1.2. Data sources

For this purpose, the dataset on temperature anomaly provided by NASA [2] and Our World in Data [3] is used together with the CO2 and Greenhouse Gas Emissions database of Our World in Data [4]. The temperature anomaly is presented as a deviation from the corresponding 1951-1980 means.

[1] https://climate.ec.europa.eu/climate-change\_en, access 07.2023

[2] https://data.giss.nasa.gov/gistemp/, access 07.2023

[3] https://ourworldindata.org/grapher/hadcrut-surface-temperature-anomaly, access 07.2023

[4] https://ourworldindata.org/co2-and-greenhouse-gas-emissions, access 07.2023

1. Data Exploration

2.1. Dataset description and exploration

The datasets used in this project are available to the public [3,4] and can be downloaded for examination. Both of the datasets contain information about countries in a respective year. The Surface anomaly dataset [4] contains 4 rows of information about 199 countries with ISO code and their change in surface temperature as compared to the mean from years 1951-1980. The measurements date back to 1850, however, this does not apply for all countries. The dataset contains information until 2017. There is a distribution of entries for various countries. The values vary from information about 168 years of temperature anomalies (eg. Netherlands) to 64 years (Chad). However, the important data for the project are available.

The second dataset contains information about countries and areas with their respective CO2 emissions in 79 columns. This dataset contains information about countries’ population, gdp, CO2 emissions due to various product-based emissions and other greenhouse gases emissions. Among many variables available in the dataset, only ten columns have been chosen for further studies. These are: ["country", "iso\_code" ,"year", "gdp","population", "co2", "coal\_co2", "gas\_co2", "methane", "nitrous\_oxide", "total\_ghg", "oil\_co2" ].

The target variable is the “surface temperature anomaly” from the first dataset.

The first exploration of the datasets include the geomap of the surface temperature anomaly throughout the years. Fig 1 shows geographical distribution of surface temperature anomaly in 1980, 2000 and 2017. The changes can be clearly spotted with the different colors. Interactive map with yearly changes is available in the Streamlit presentation. The data used in creation of this plot comes from the [3] datasource.

The “hadcrut-surface-temperature-anomaly.csv” contains yearly (“Year”) change in surface temperature anomaly (“Surface temperature anomaly”, in degree Celsius) over the last decades for every country (“Entity”). The ISO country code is also included (“Code”).

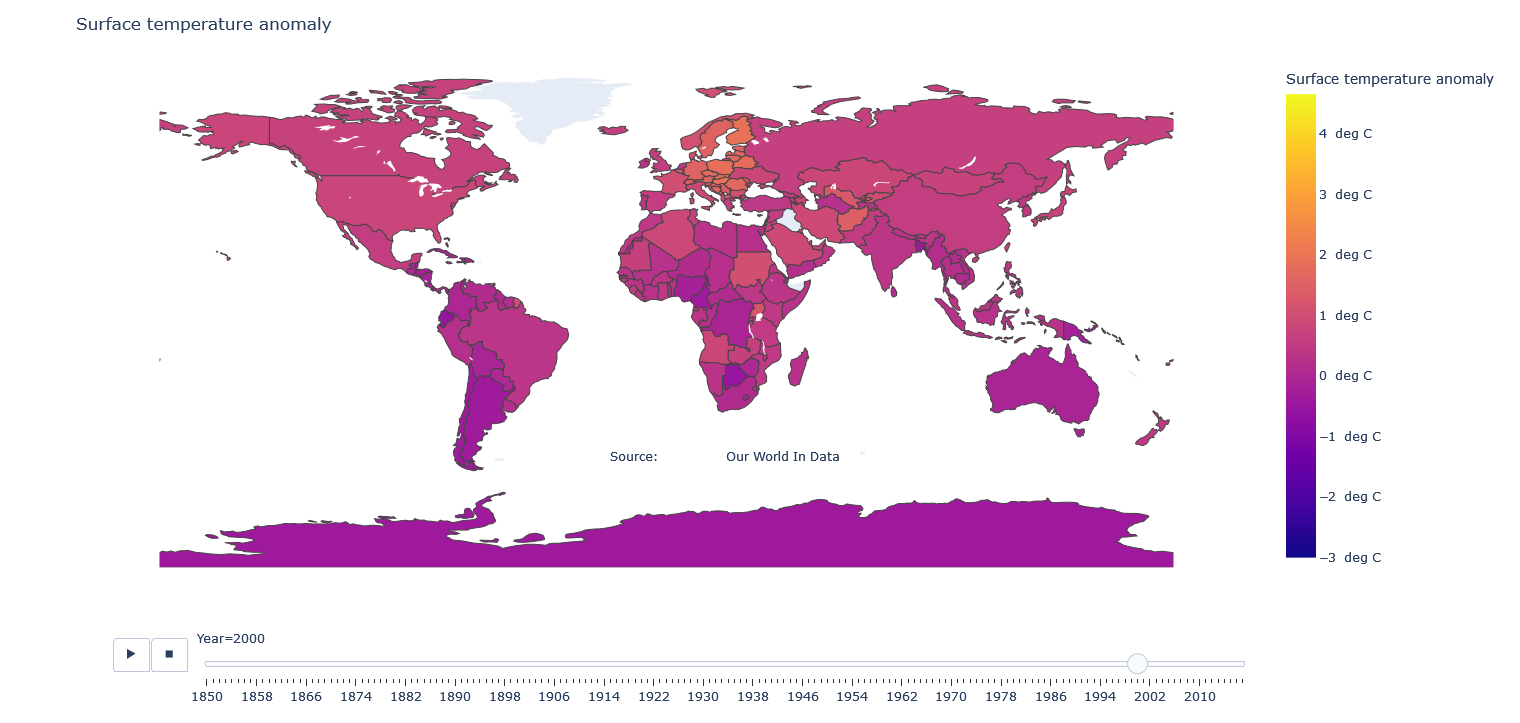
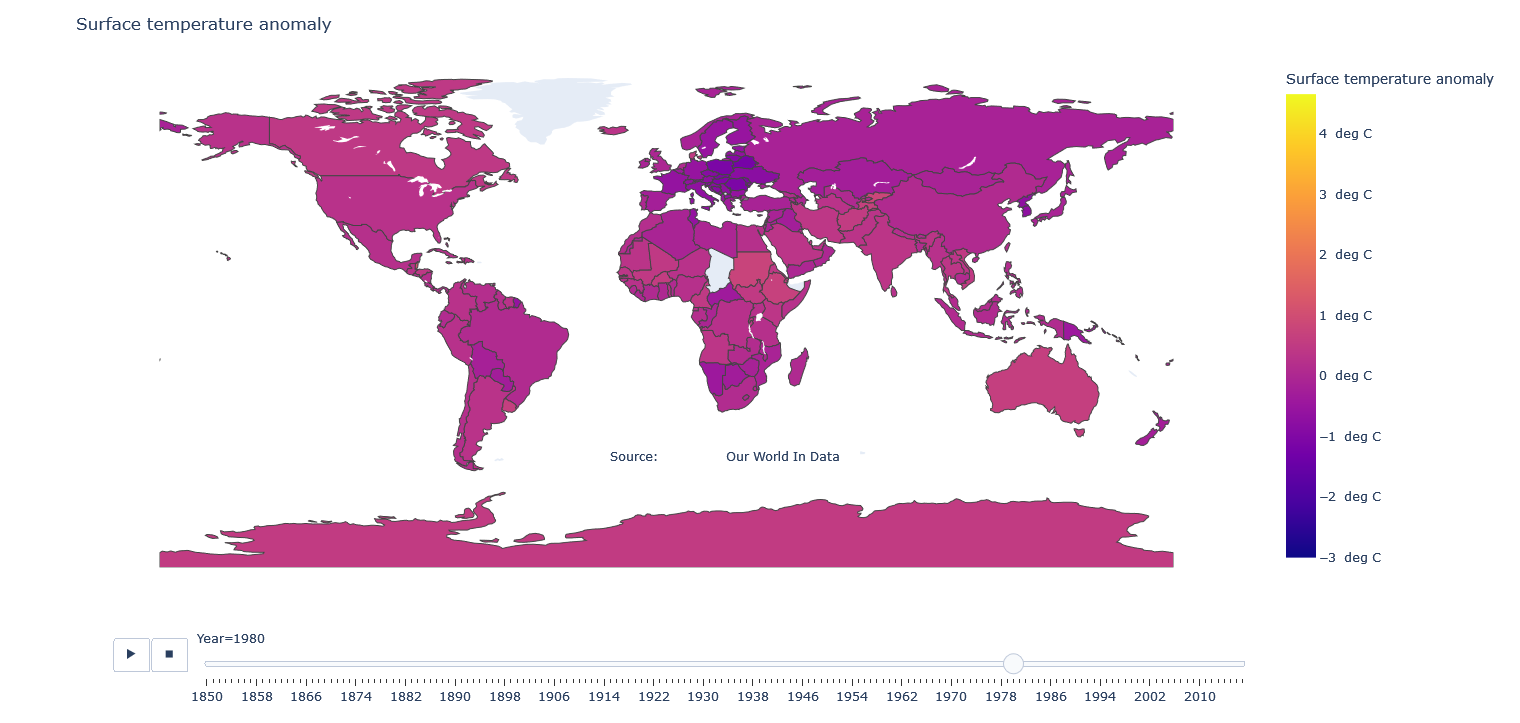
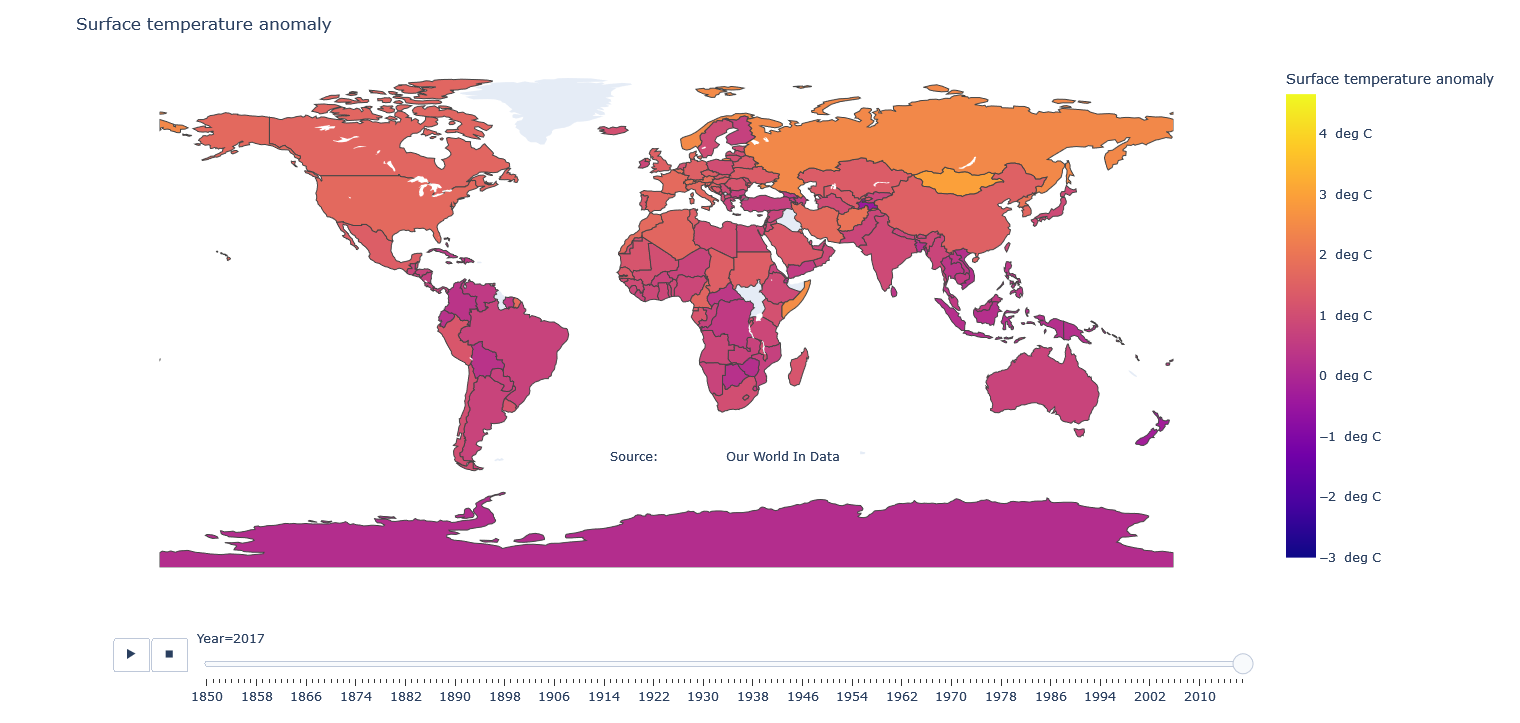


Fig. 1. Country distribution of surface temperature anomaly in 1980, 2000 and 2017.

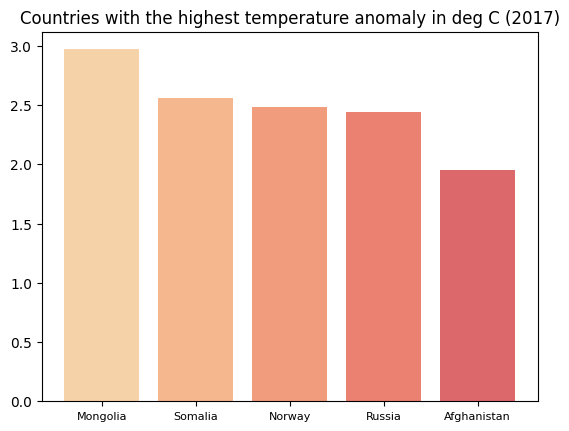


Fig. 2. Countries with the highest temperature anomaly.

Among the 5 countries with the highest temperature anomaly are Mongolia, Somalia, Norway, Russia and Afghanistan. However, the countries with the highest CO2 emissions are China, US, India, Russia and Japan (Fig. 3).

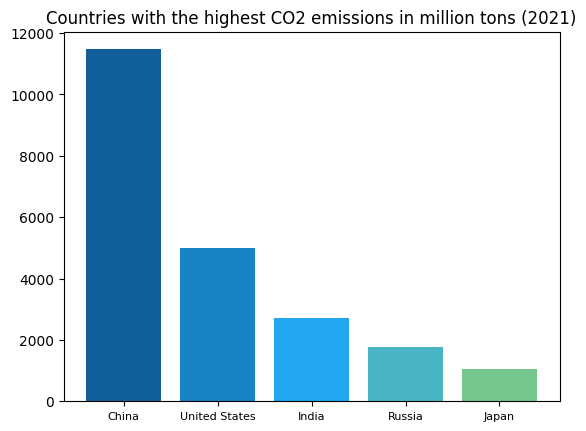


Fig. 3. Countries with the highest CO2 emissions.

Among them, China is the biggest country to emit CO2 with about 32% of world emissions (Fig. 4).

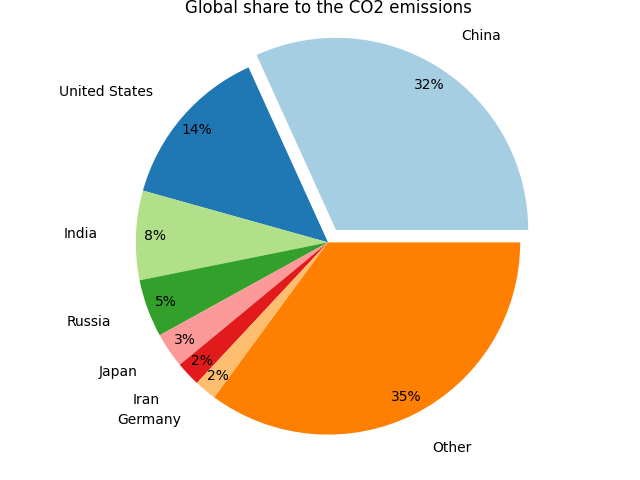


Fig. 4. Countries share to global CO2 emissions.

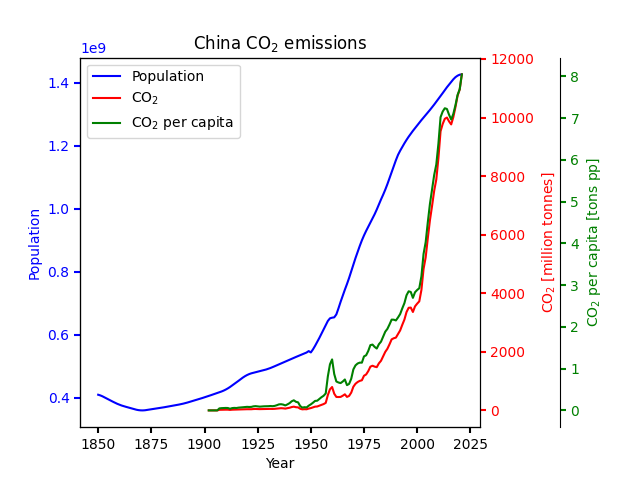


Fig. 5. CO2 emissions from China through the years. The population growth and CO2 per capita are shown for comparison.

From Fig. 5 one can see that the CO2 emission increased from 1950 and further increased rapidly from around 2000. During this period, the population growth slightly impeded. The population growth affects the CO2 emissions but is not the main factor in it.

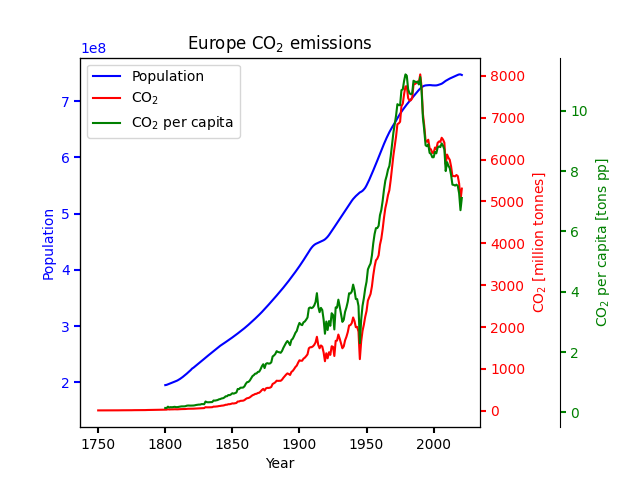


Fig. 6. Europe’s CO2 emissions through the years. The population growth and CO2 per capita are shown for comparison.

As a comparison, Fig. 6 presents the CO2 emissions generated by Europe. The rapid increase in CO2 emissions from 1950 can be linked to the end of World War II and is also not only dependent on population. Since 1990 the decrease in Europe’s CO2 emission can be seen.

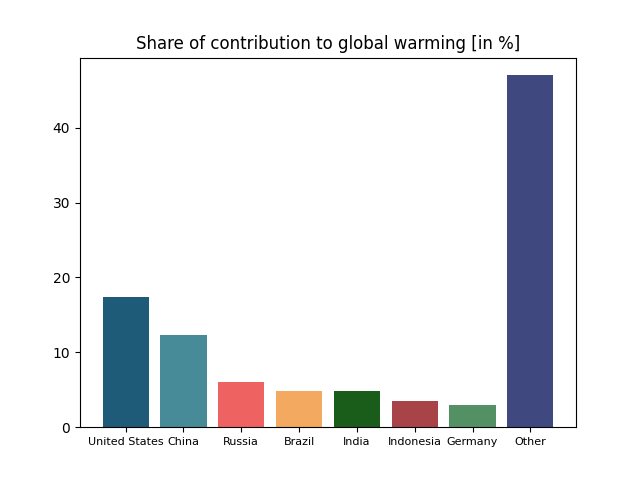


Fig. 7. Contributions to global warming by country in 2022.

The final plot (Fig. 7) of the data exploration presents the contribution to global warming obtained from the dataset containing emission data. The calculations show that the United States, China, Russia, Brazil and India are the top countries that contribute to global warming. Among them, only Russia is the country that is the most affected by surface temperature change.

2.2. Statistics

2.2.1. Missing Values

We have approx. 80 columns. Most of the columns have more than 50% of nan values, and a quarter of them - above 80%. Normally, if a column has a high percentage of missing values(more than 50%), it might be worth considering dropping it. So, if we have a lot of columns with a high percentage of missing values and they are not involved in our analysis, we could drop it. We could also drop the columns which do not provide any useful information for us.

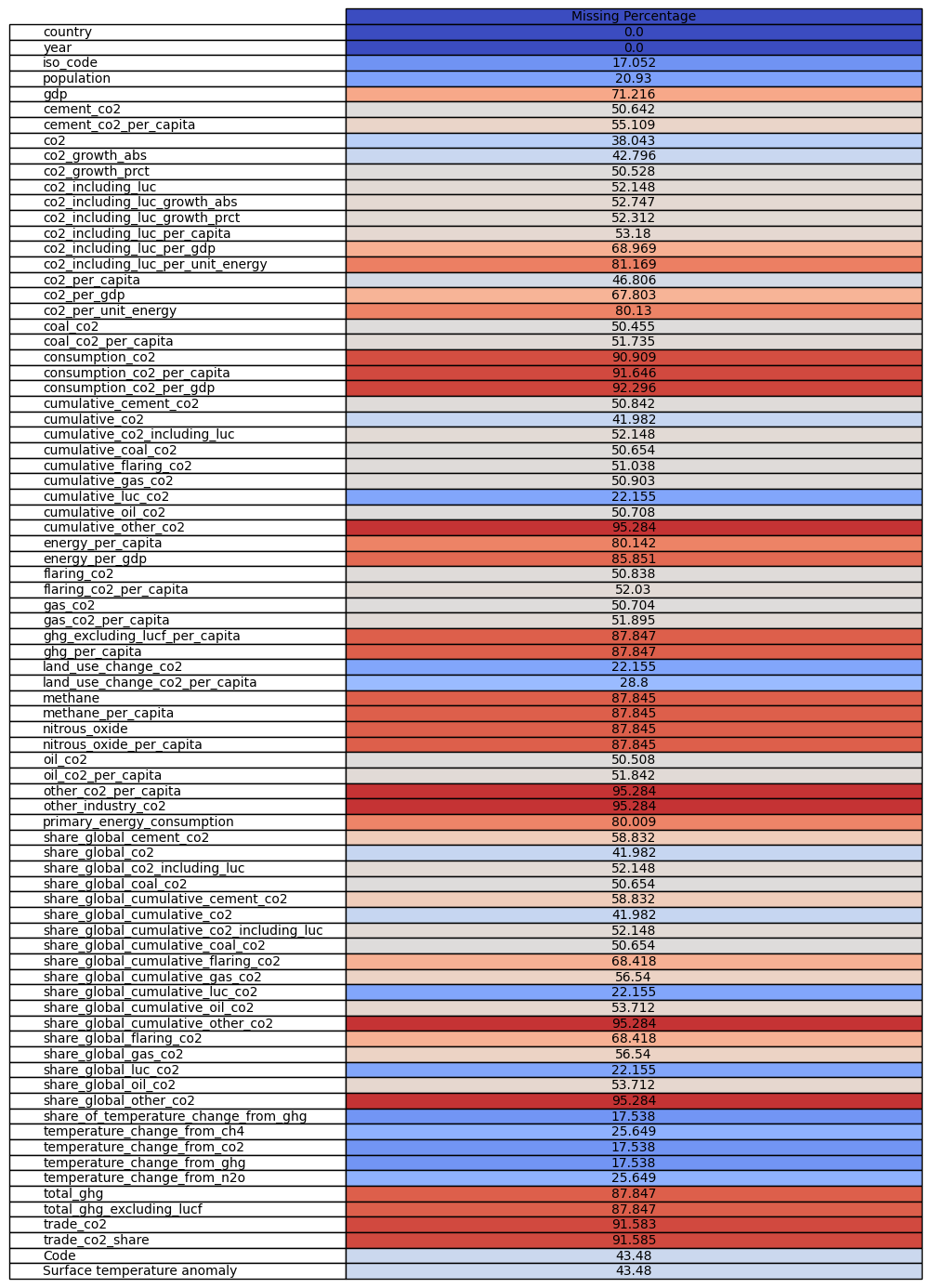


Fig. 8. Percentage of missing values of each column in the dataset.

2.2.2. Columns importance

Country: useful for incorporating geographical factors into the analysis.

Year: important for time-analysis and can help capture temporal patterns and trends. Population: could be correlated with our target.

GDP: great impact on our target value. It can be informative for assessing the relationship between economic activity and emissions.

CO2: great impact on our target value(even if %of missing values is relatively high). CoalCO2: great impact on our target value.

FlaringCO2: great impact on our target value.

GasCO2: great impact on our target value.

Methane: great impact on our target value.

Nitrous Oxide: great impact on our target value.

OilCO2: great impact on our target value.

Surface Temperature Anomaly- TARGET.

Columns to drop:iso\_code,cement\_co2,cement\_co2\_per\_capita,co2\_growth\_abs, co2\_growth\_prct,co2\_including\_luc,co2\_including\_luc\_growth\_abs,co2\_including\_luc\_growth\_prct,co2\_including\_luc\_per\_capita,co2\_including\_luc\_per\_gdp, co2\_including\_luc\_per\_unit\_energy,co2\_per\_capita,co2\_per\_gdp,co2\_per\_unit\_energy, coal\_co2\_per\_capita,consumption\_co2,consumption\_co2\_per\_capita, consumption\_co2\_per\_gdp,cumulative\_cement\_co2,cumulative\_co2, cumulative\_co2\_including\_luc,cumulative\_coal\_co2,cumulative\_flaring\_co2, cumulative\_gas\_co2,cumulative\_luc\_co2,cumulative\_oil\_co2,cumulative\_other\_co2, energy\_per\_capita,energy\_per\_gdp,flaring\_co2\_per\_capita,gas\_co2\_per\_capita, ghg\_excluding\_lucf\_per\_capita,ghg\_per\_capita,land\_use\_change\_co2, land\_use\_change\_co2\_per\_capita,methane\_per\_capita,nitrous\_oxide\_per\_capita, oil\_co2\_per\_capita,other\_co2\_per\_capita, ther\_industry\_co2, primary\_energy\_consumption, share\_global\_cement\_co2,share\_global\_co2,share\_global\_co2\_including\_luc,share\_global\_coal\_co2,share\_global\_cumulative\_cement\_co2,share\_global\_cumulative\_co2, share\_global\_cumulative\_co2\_including\_luc,share\_global\_cumulative\_coal\_co2, share\_global\_cumulative\_flaring\_co2,share\_global\_cumulative\_gas\_co2, share\_global\_cumulative\_luc\_co2,share\_global\_cumulative\_oil\_co2, share\_global\_cumulative\_other\_co2,share\_global\_flaring\_co2,share\_global\_gas\_co2, share\_global\_luc\_co2,share\_global\_oil\_co2,share\_global\_other\_co2, share\_of\_temperature\_change\_from\_ghg,temperature\_change\_from\_ch4, temperature\_change\_from\_co2,temperature\_change\_from\_ghg, temperature\_change\_from\_n2o,total\_ghg,total\_ghg\_excluding\_lucf,trade\_co2, trade\_co2\_share.

Reason: most of those columns have a really high percentage of missing values and those values are not correlated with our target. For example co2 emission has a great impact on our target value, but co2\_growth\_abs is less important. Columns which we are keeping are the columns with the most important values in our analysis.

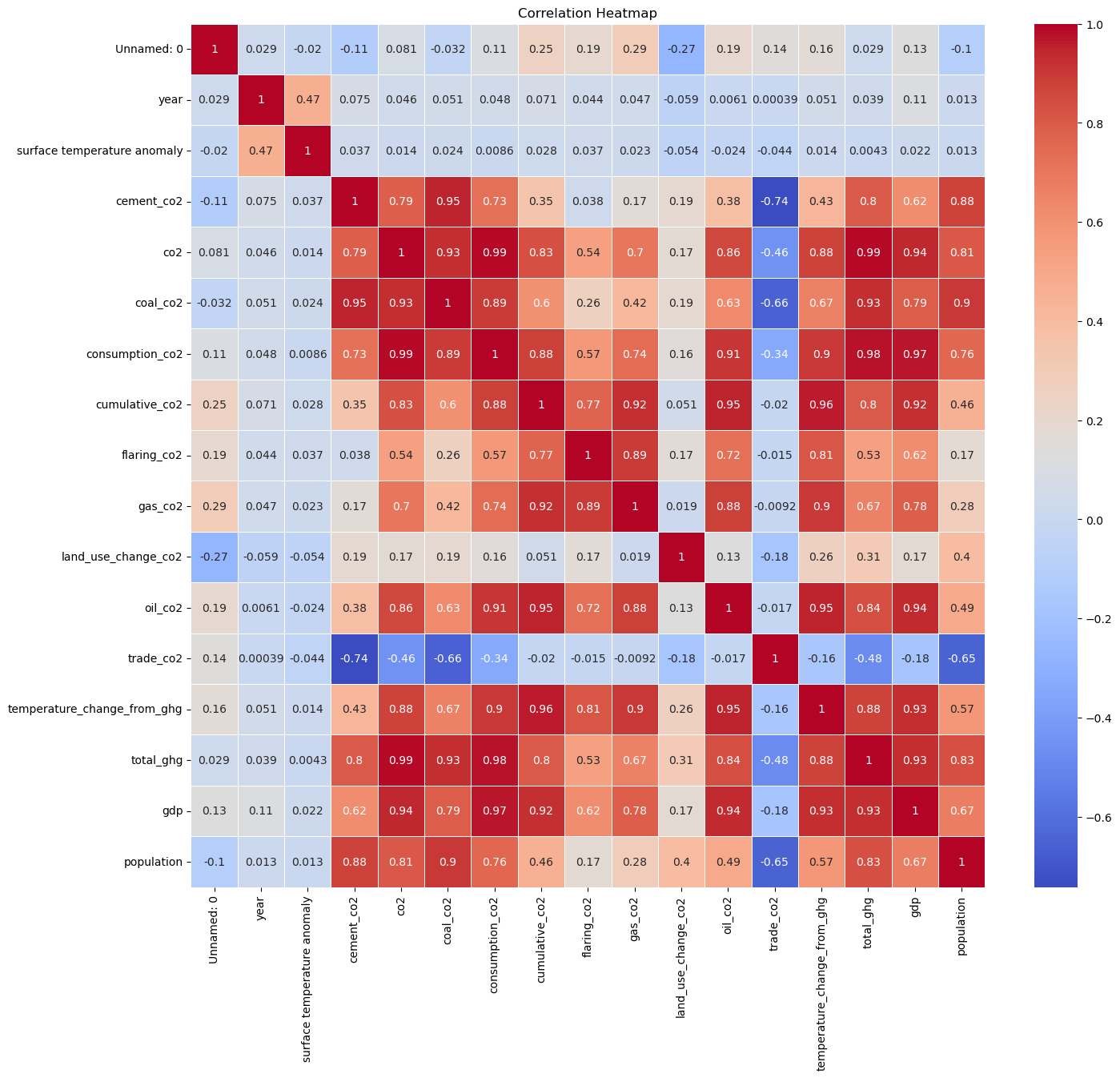


Fig. 9. Correlation matrix of the important variables of the dataset.

Conclusion: surface temperature anomaly seems to have the strongest correlation with year variable. Intensity of the color indicates the low dependence with other variables. However, with variables like flaring\_co2, gas\_co2, gdp, population,coal\_co2, co2 correlation is stronger than with land\_use\_change\_co2, trade\_co2, oil\_co2.

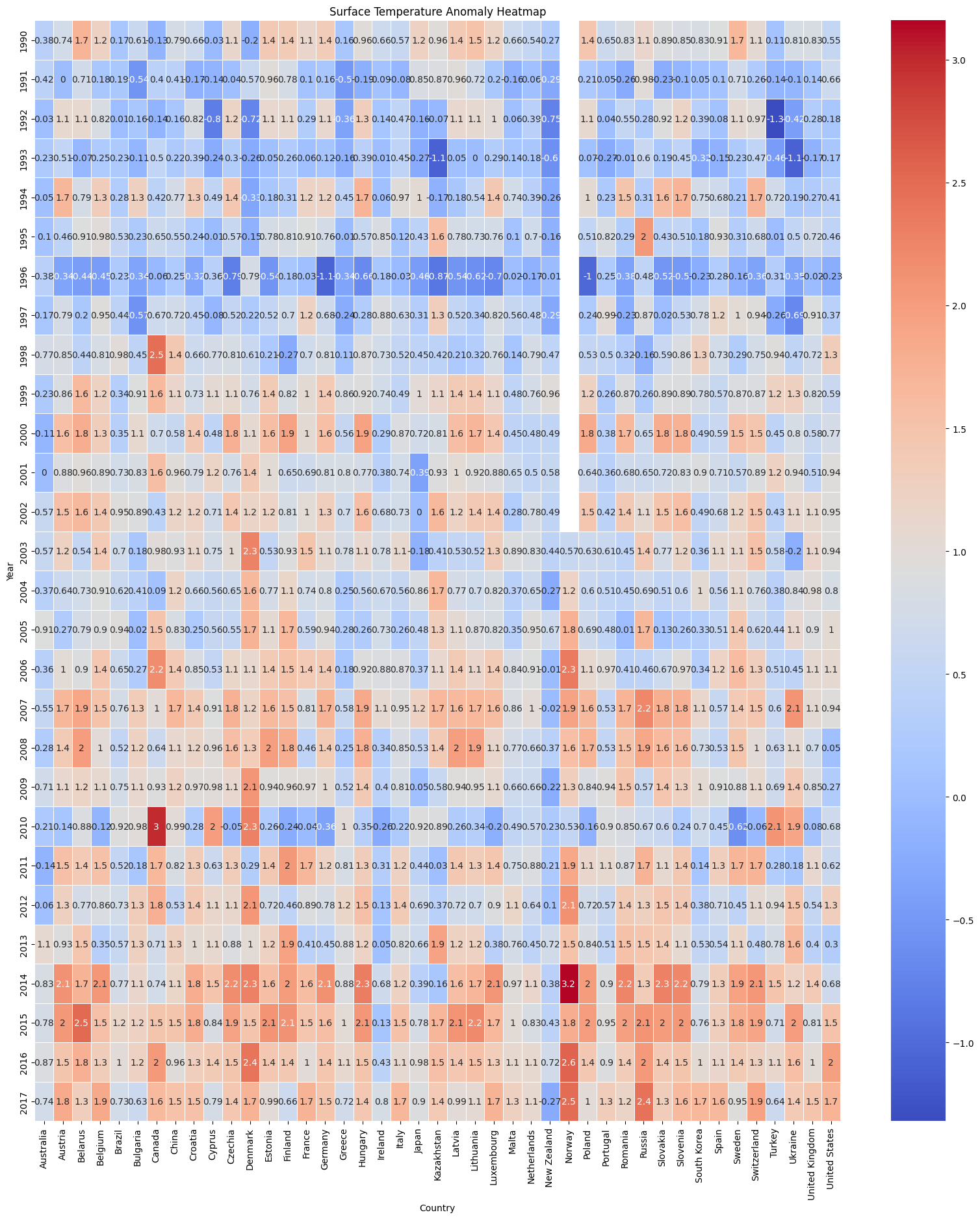


Fig. 10. Correlation matrix of all variables of the dataset.

2.2. Data pre-processing

Data preprocessing is a crucial step in data analysis and machine learning. It refers to the preparation and cleaning of raw data to ensure that it is suitable for effective analysis and modeling. Data preprocessing includes various techniques and steps to handle missing data, detect and correct outliers, normalize or scale data, code categorical data. The following steps were performed during the project to generate a new dataset: datas\_pre\_processed.csv from the files provided by Our World: hadcrut-surface-temperature-anomaly.csv (29566 rows) and owid-co2-data.csv (50598 rows). This is to be adjusted, scaled and converted to numeric form for use in the training model. Since the files used are very large, first hadcrut-data is checked for outliers using boxplots and these are removed using the z-score method. Similarly, all duplicate values and NaN values are deleted directly and not replaced, or determined, so that owid-co2-data now contains 1234 and hadcrut 29220. The column headers of hadcrut-data are adjusted so that the two files can be merged later without problems. Furthermore, minor inconsistencies, such as identifiers in parentheses, are detected and corrected.

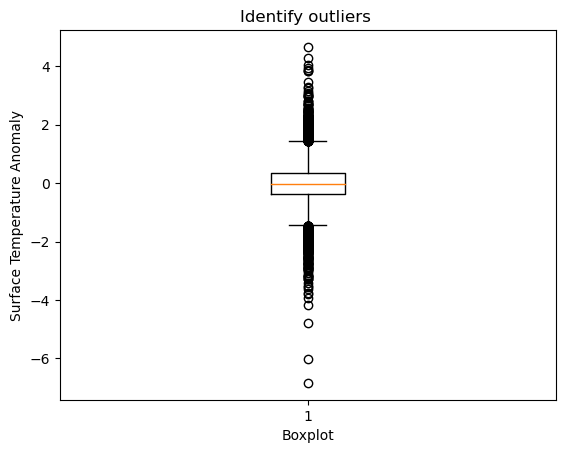


Fig. 11. The boxplot of the surface temperature variable.

The new dataframe will contain only the columns relevant for the project: 'country', 'year', 'gdp', 'population', 'co2', 'coal\_co2', 'flaring\_co2', 'gas\_co2', 'methane', 'nitrous\_oxide', 'oil\_co2', and 'sta'.

Country is the only Object column and contains categorical values, so we transform these to numeric. For further use we save our result as a csv file, this contains 1135 lines.

3. Modeling

3.1. Selection of forecasting methods

The modeling of the world temperature data is a problem that is well treated with a regression model. This task relates to prediction how the world mean temperature will change in time according to several variables that are mostly greenhouse emissions. The main performance metrics used to compare the models is score on the train and test sets and R squared score. Other metrics such as mean absolute error and mean squared error were also used.

The algorithms used is this project are:

* Linear Regression
* Gradient Boosting Regressor
* Decision Tree Regressor
* Random Forest Regressor.

Regression models are used to analyze the relationships between variables. It is also used to predict a response based on a set of variables. In our case, it is to investigate how the temperature rise depends on the variables such as country population, GPD (Gross Domestic Product) and various greenhouse gasses emission (eg. CO2).

The models chosen for the analysis of the world temperature data are all regression models that vary in complexity.

The data pre-cleaned as in the previous section have been encoded (the country names and ISO code) using pd.get\_dummies to allow the fitting and divided into target variable (temperature anomaly) and features. The datasets were further splitted into training and test sets according to an 80% - 20% ratio.

3.2. Development of the models

3.2.1. Linear Regression

Linear regression is a fundamental statistical and machine learning technique used for modeling the relationship between a dependent variable (also known as the target or response variable) and one or more independent variables (also known as predictors, features, or explanatory variables). It assumes that the relationship between the variables is linear, meaning that the change in the dependent variable is directly proportional to the changes in the independent variables.

Fig. 12 presents the scatterplot of check for over- and underfitting of Linear Regression: Actual vs. Predicted. The residual plot shows the difference between the actual target values and the predicted values.

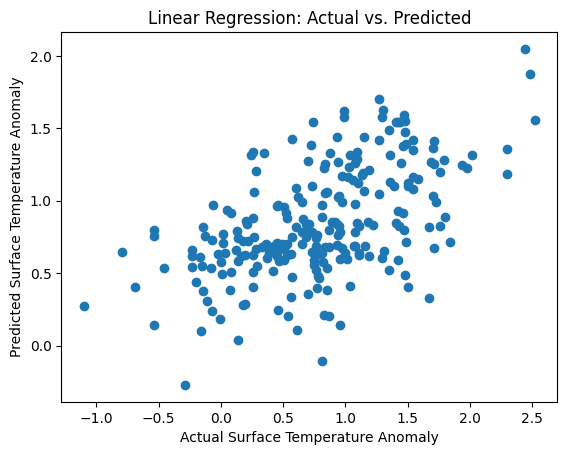
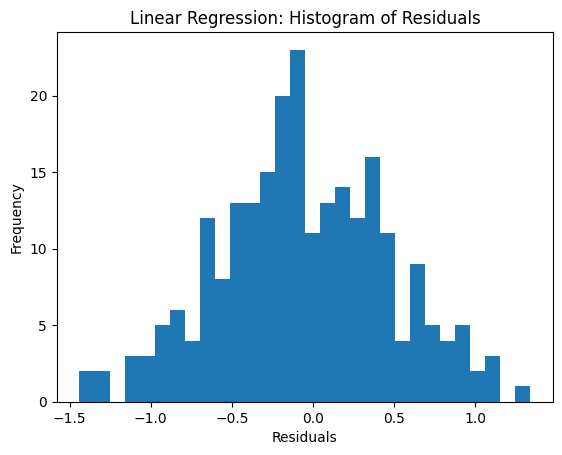


Fig. 12. The scatterplot for overfitting and underfitting in the Linear Regression model together with the residuals plot.

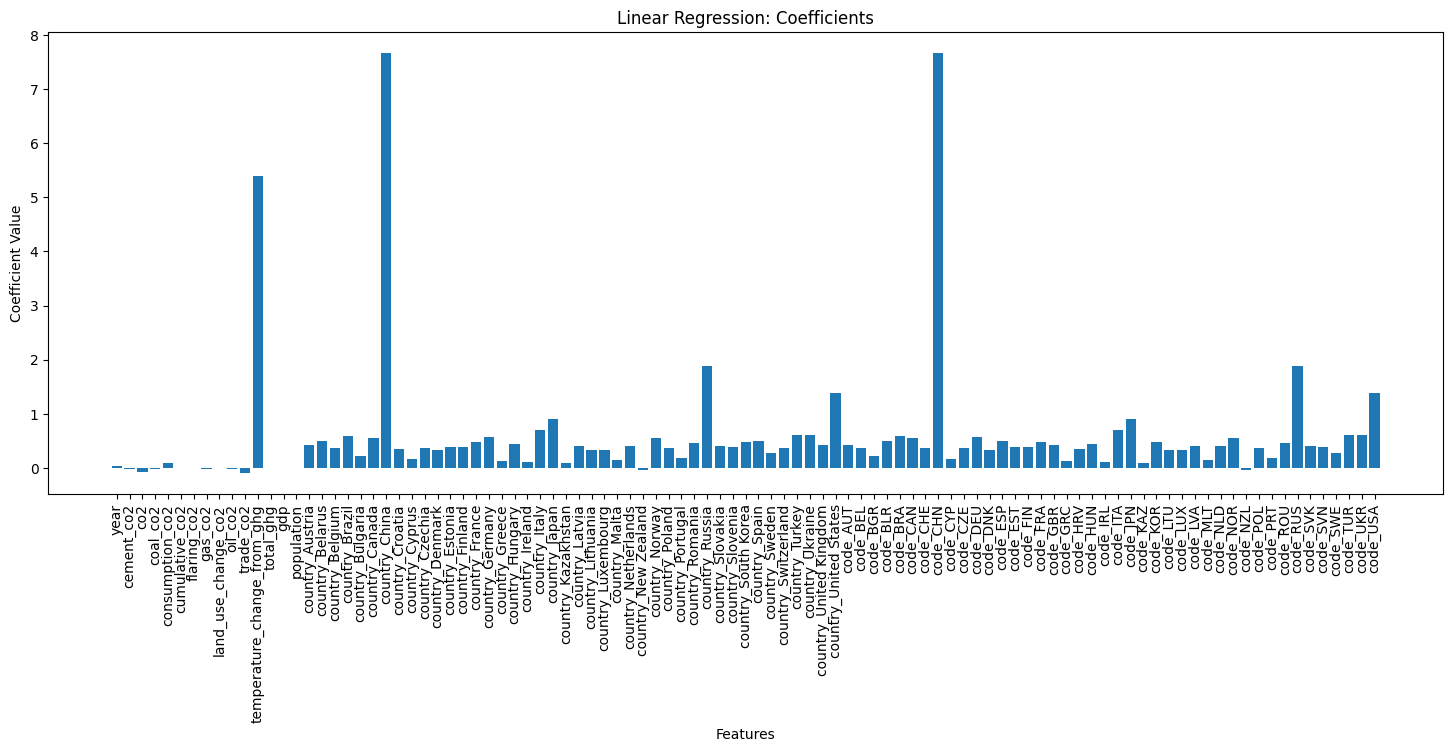


Fig. 13. The feature importance plot in the Linear Regression model.

3.2.2. Gradient Boosting Regressor

Gradient Boosting is a popular machine learning ensemble technique used for both regression and classification tasks. It is based on the idea of combining weak learners (typically decision trees) sequentially to form a strong predictive model.

The basic principle of Gradient Boosting involves building multiple decision trees in a sequential manner, where each subsequent tree tries to correct the errors made by the previous one. The model learns by minimizing a loss function, which measures the difference between the predicted values and the actual target values.

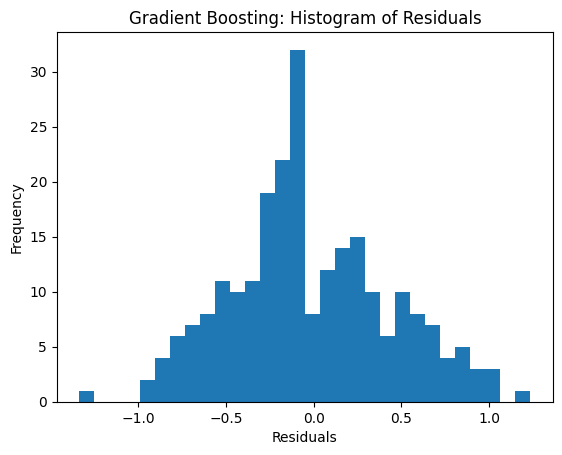
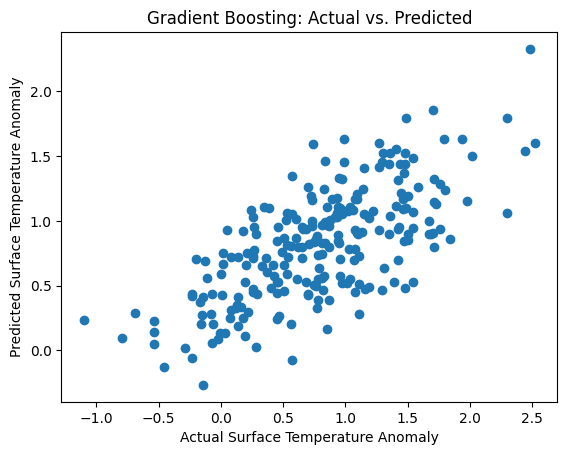


Fig. 14. The scatterplot for overfitting and underfitting in the Gradient Boosting model and residuals plot.

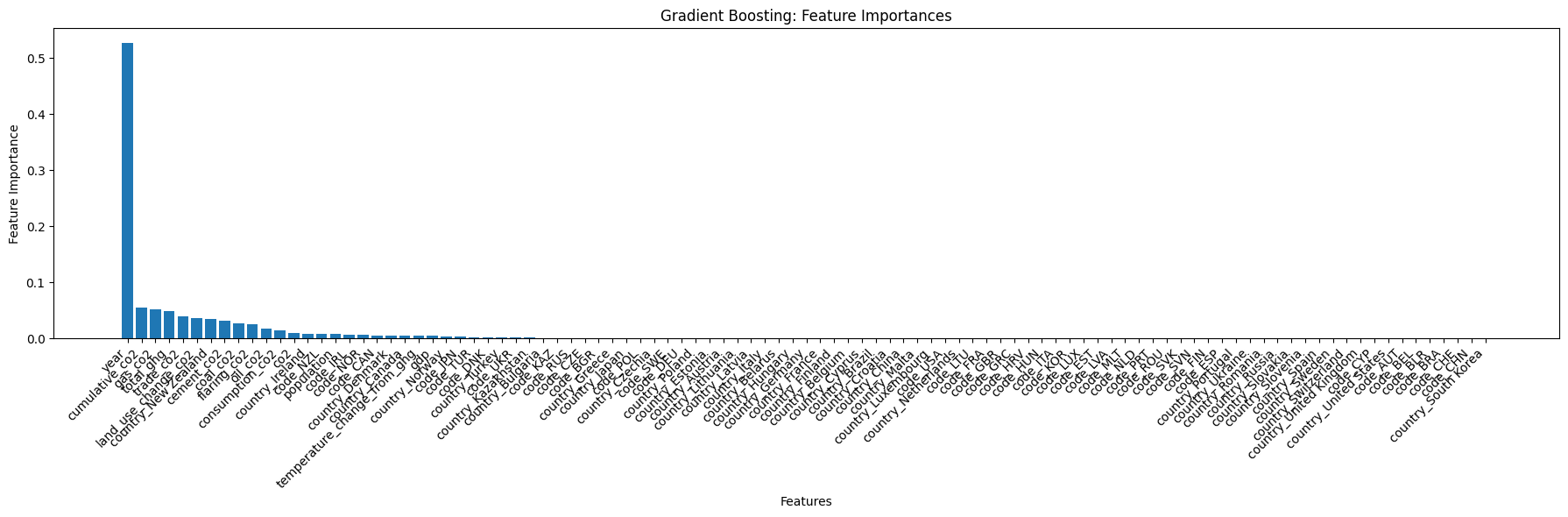


Fig. 15. The feature importance plot in the Gradient Boosting model.

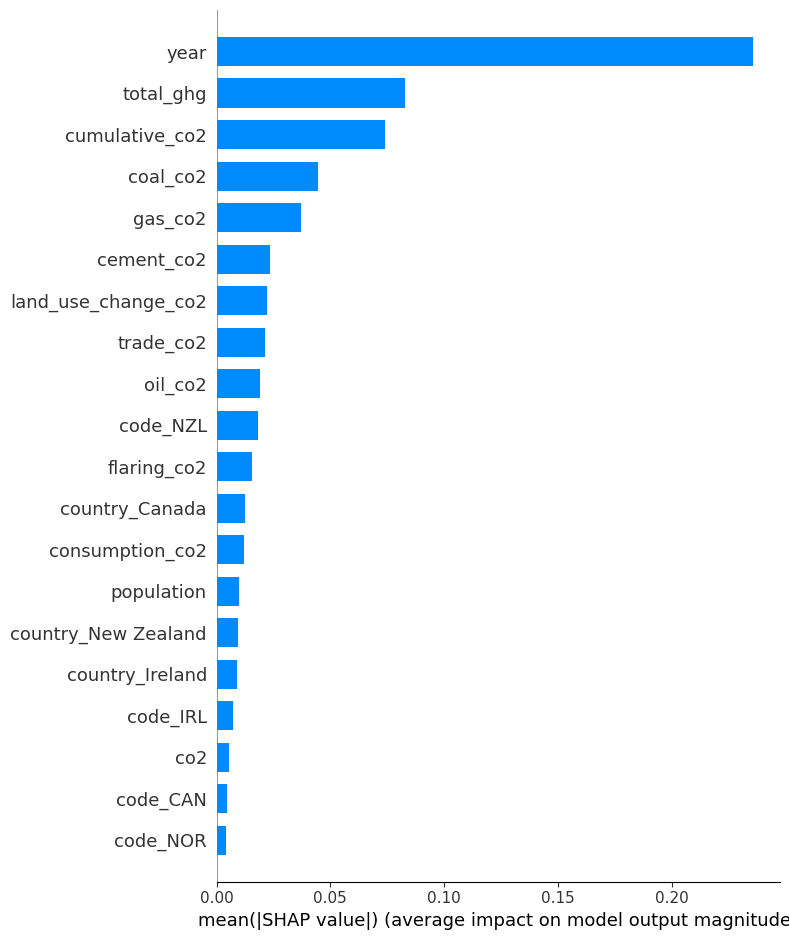


Fig. 16. The feature importance plot based on the SHAP values in the Gradient Boosting model. The important variables are comparable to the ones in feature importance (Fig.15).

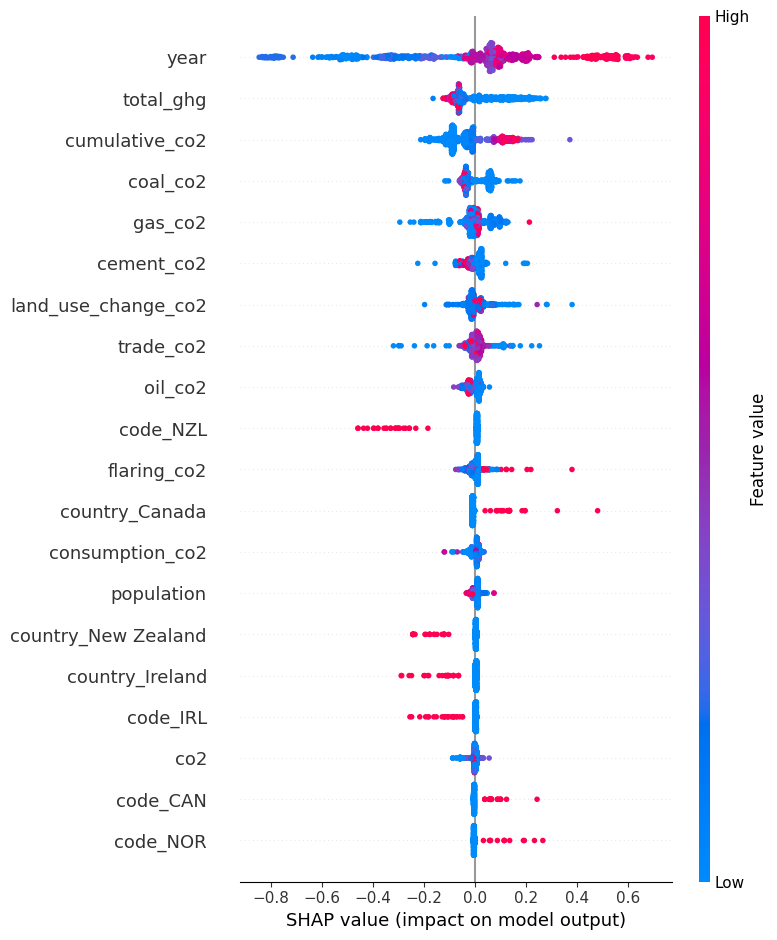


Fig. 17. The SHAP plot of the Gradient Boosting model.

Based on the evaluation metrics of the Gradient Boosting and Linear Regression models, it can be observed that both models show promising performance in predicting the surface temperature anomaly, a critical indicator in the climate and environmental domain.

In the case of the Gradient Boosting model, it achieved a Mean Absolute Error (MAE) of approximately 0.3696 and a Mean Squared Error (MSE) of around 0.2092. These metrics indicate that, on average, the model's predictions deviate from the true surface temperature anomaly by only 0.3696 units, which is a relatively small error considering the nature of the

indicator. Additionally, the MSE of 0.2092 suggests that the model's predictions exhibit lower variance and are more concentrated around the true values.

Furthermore, the R-squared (R²) value of approximately 0.4765 signifies that the Gradient Boosting model explains approximately 47.7% of the variance in the surface temperature anomaly. This demonstrates a reasonable level of predictability, indicating that the selected features and the model itself are capturing meaningful relationships with the target variable. An R² value of 0.4765 indicates that the model's predictions fit the data better than a simple horizontal line (baseline model) would.

Similarly, the Linear Regression model also shows promising performance, though slightly inferior to the Gradient Boosting model. The MAE of approximately 0.4270 suggests that, on average, the model's predictions deviate by around 0.4270 units from the actual surface temperature anomaly. The MSE of approximately 0.2810 reflects that the Linear Regression model's predictions also exhibit relatively lower variance compared to the target variable.

However, the R² value of approximately 0.2967 indicates that the Linear Regression model explains approximately 29.7% of the variance in the surface temperature anomaly. While this value is lower than that of the Gradient Boosting model, it still signifies that the Linear Regression model provides useful insights into the temperature anomaly prediction. However, the feature importances in the Linear Regression model (Fig. 12) show that the strongest impact on the model are shown for the coded countries’ variables, eg. China ISO code ‘code\_CHN’. The temperature change does not depend on the ISO code or the country name rather on the other variables (emissions, population, gpd) generated by the country China. In the Gradient Boosting model, the most important features are the CO2 emissions (Fig. 14, Fig. 15).

In summary, both models exhibit competitive performance in predicting the surface temperature anomaly, a vital indicator in climate analysis. The Gradient Boosting model, with its lower MAE, MSE, and higher R2, appears to be superior to the Linear Regression model.

3.2.3. Decision Tree Regressor

Another two algorithms used to model the surface temperature anomaly data are Decision Tree and Random Forest models. Decision Tree Regressor is a decision-making algorithm that trains the model in the form of a tree, a flow-chart structure with all possible results. For the Decision Tree Regressor the scores obtained for train and test sets are: 1.0 and 0.145. It may indicate the overfitting of the model on the train set and therefore resulting in poor performance on the test set. The MSE for the Decision Tree Regressor model is 0.584. Fig. 18 presents the scatterplot of check for over- and underfitting of the Decision Tree model.

Fig. 19 presents the importances of the features in this model. The most important variable in all tree models is the year variable, which is expected as the value simply rises and the temperature also rises consequently over the last decades. However, the year itself is not the parameter that influences the temperature change directly, rather that through the other variables that change yearly. The models also show that additional insights from other variables could be beneficial for the modeling and predictions. Apart from the year variable, the gas CO2 (gas\_co2) and the total greenhouse gases (total\_ghg) present a high impact on the target variable.

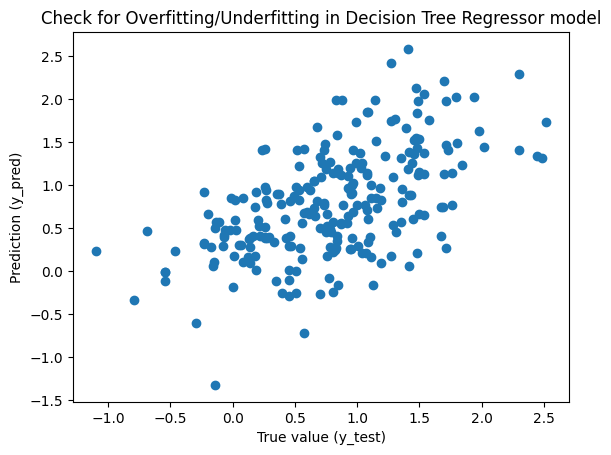
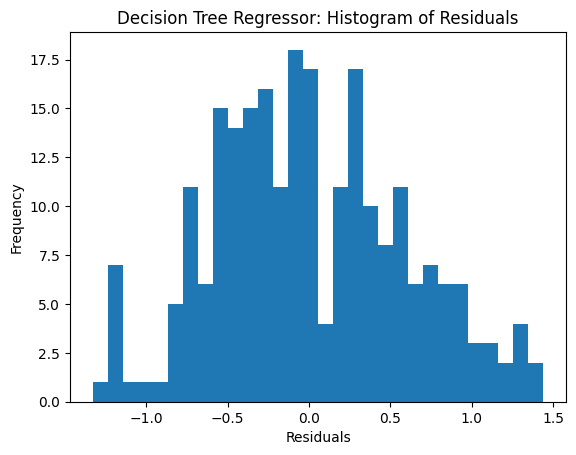


Fig. 18. The scatterplot for overfitting and underfitting in the Decision Tree model and the residuals plot.

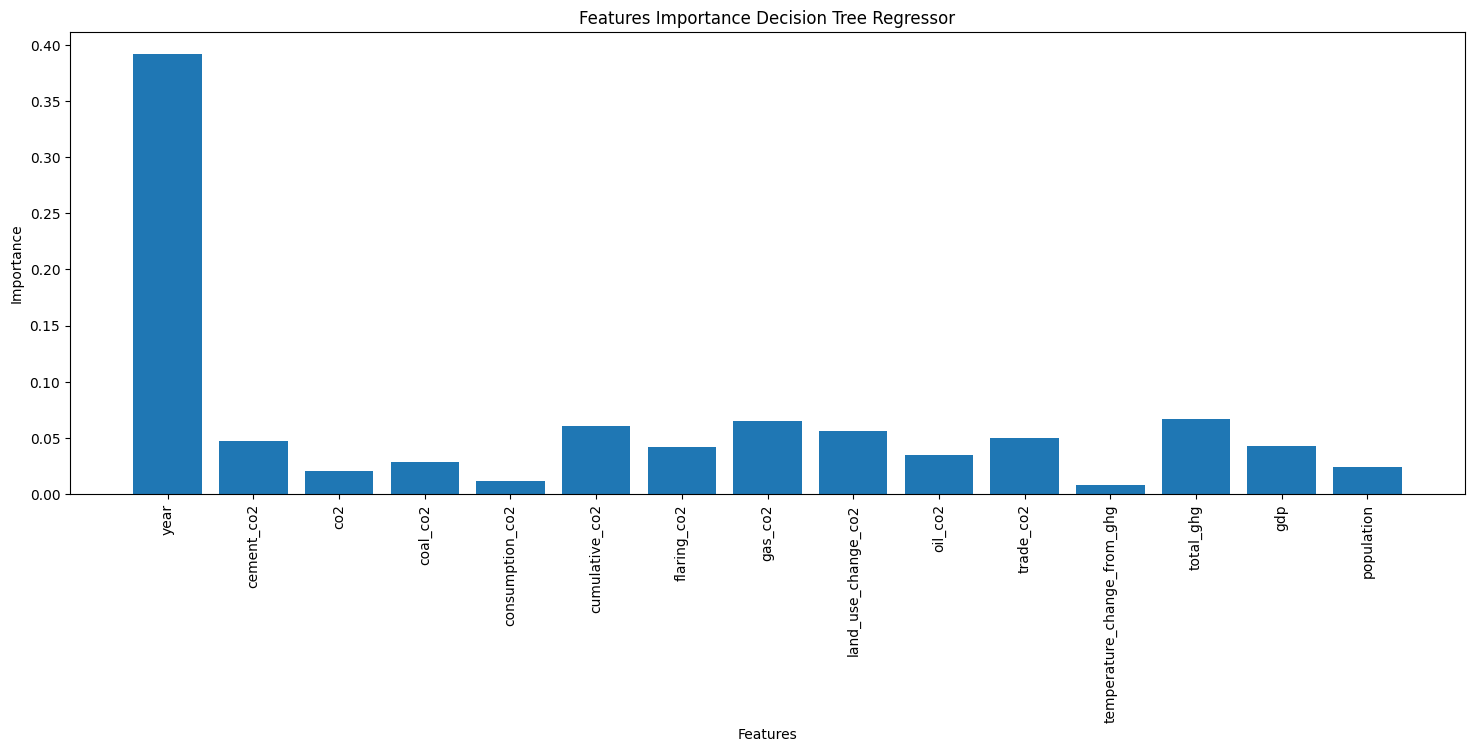


Fig. 19. The bar plot presents the importance of features in the Decision Tree model.

The last plot (Fig. 20) presents the SHAP values, which assign an importance value to each feature in a model. The positive value in the features positively impacts the model’s prediction.

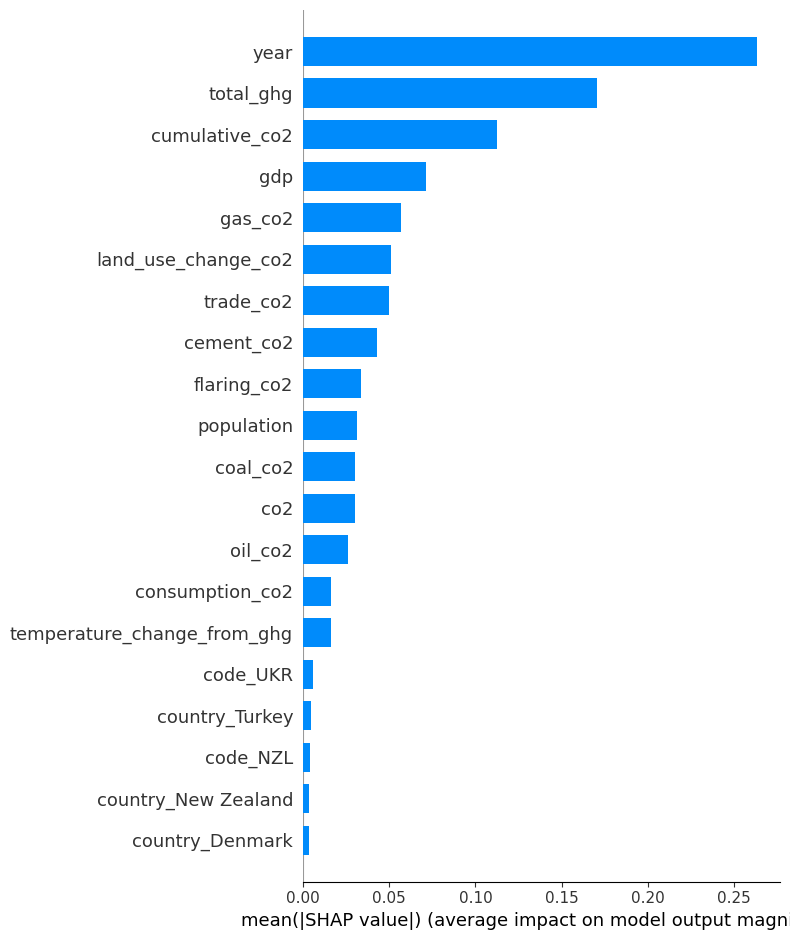


Fig. 20. The feature importance plot based on the SHAP values in the Gradient Boosting model. The important variables are comparable to the ones in feature importance (Fig.19).

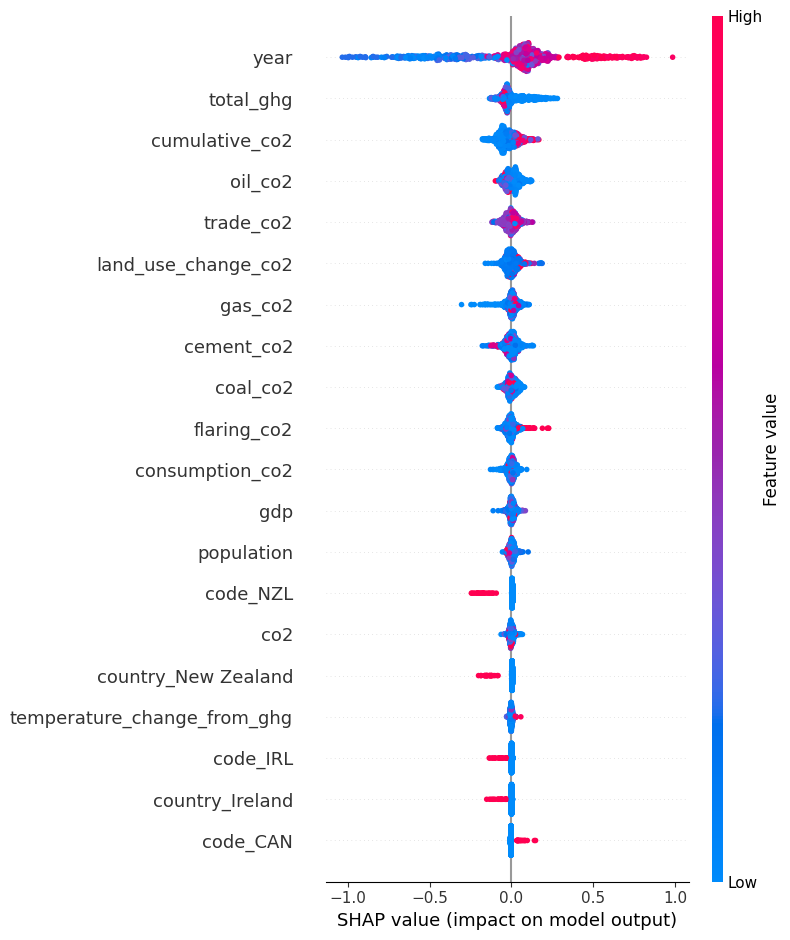


Fig. 21. The SHAP plot of the Decision Tree model. SHAP plots are a valuable tool to identify the variable importance in the model. The most important features according to the SHAP plots are total\_ghd and the various CO2 emission sources, similar to what is extracted from the feature importances.

3.2.4. Random Forest Regressor

The Random Forest Regressor creates multiple randomly created decision trees and averages the results into new output that can lead to very good predictions. In the case of this algorithm, the score for the train set is 0.919 and for the test set it's 0.451. These results are definitely better than the ones obtained for the Decision Tree. The mean squared error for the Random Forest Regressor model is 0.468.

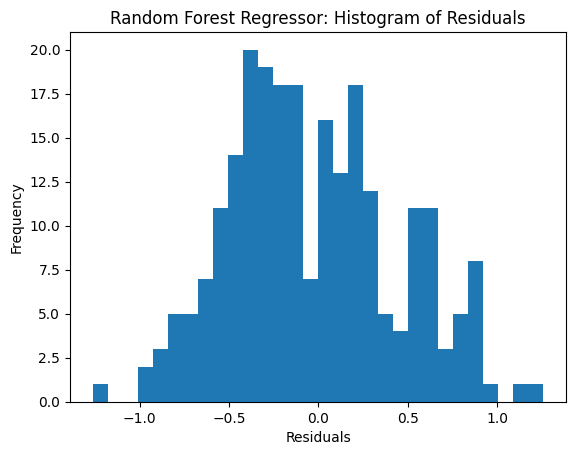
Fig. 22. presents the overfitting and underfitting scattering plot for Random Forest Regressor model. 

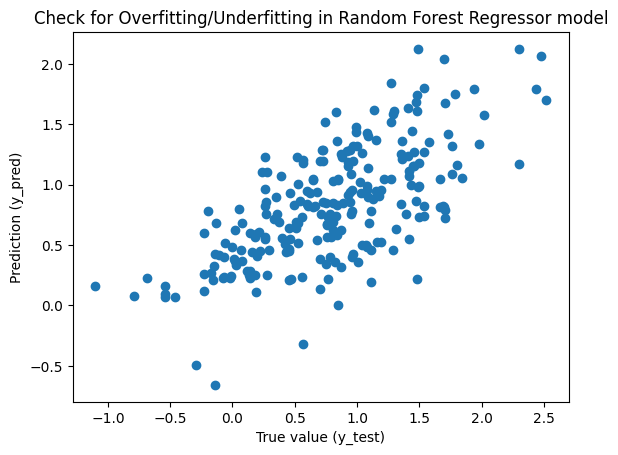
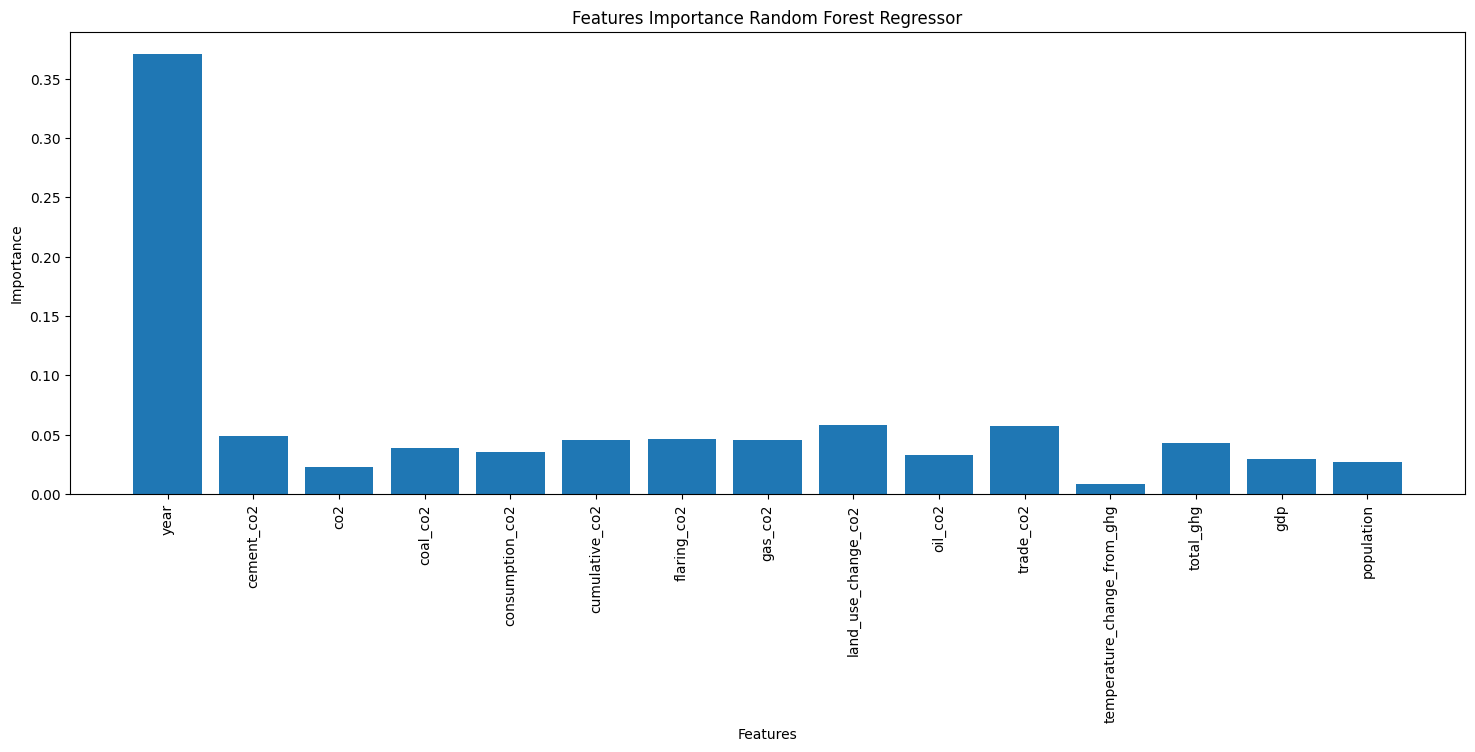
Fig. 22. The scatterplot for overfitting and underfitting in the Random Forest Regressor model and the residuals plot.

Fig. 23. The bar plot presents the importance of features in the Random Forest Regressor model. 

Similar to Decision Tree, the most important feature in the model is the year variable. Among others, the land\_use\_change\_co2, trace\_co2 and other co2-based variables have similar effects on the target variable (Fig. 23).

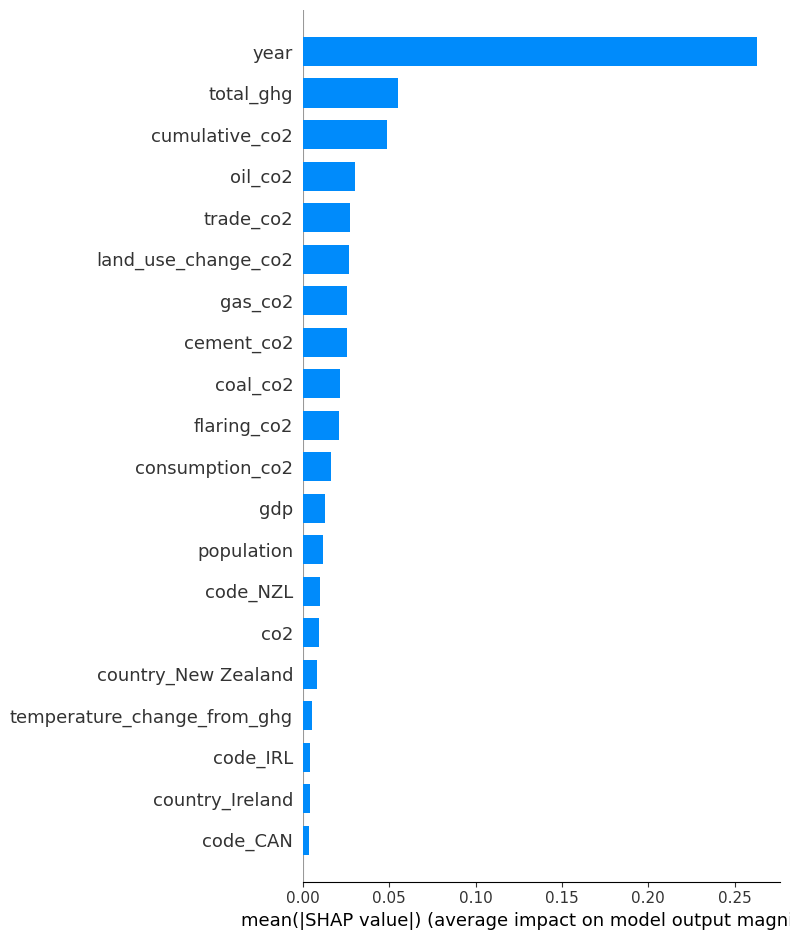


Fig. 24. The feature importance plot based on the SHAP values in the Random Forest model.

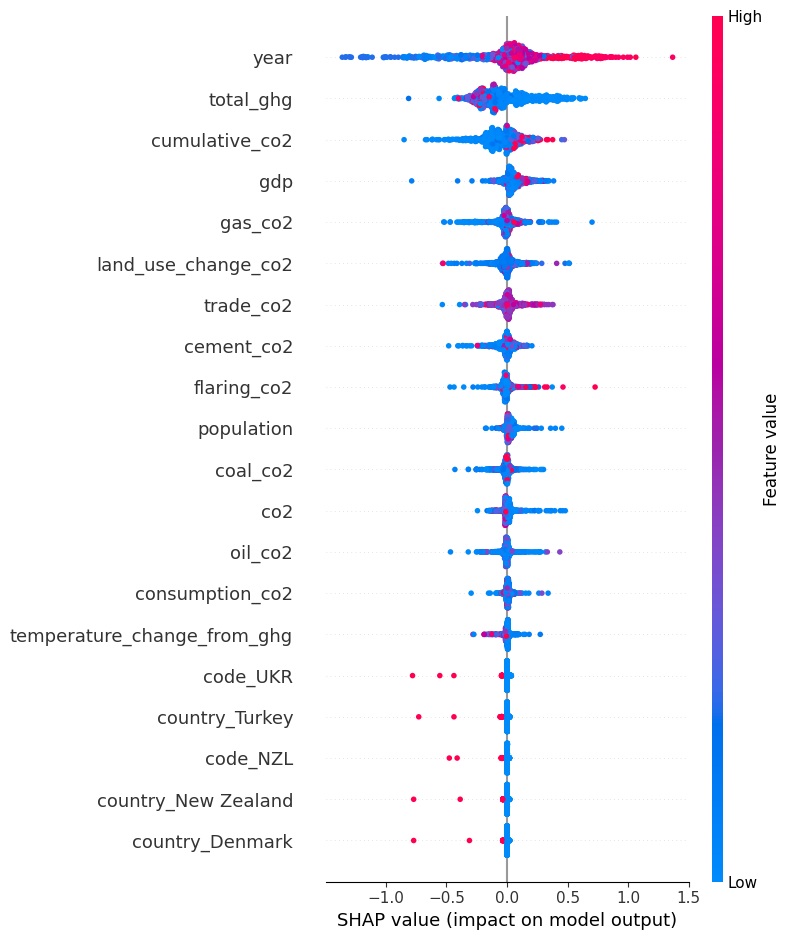


Fig. 25. The SHAP plot of the Random Forest model. In the Random Forest approach, the variables that influence the fit the most are similar to other models. These are total\_ghg and various CO2 emission sources.

4. Conclusions and outlook

The score for the train set in Linear Regression model is 0.383 and for the test set it’s 0.296. Additionally, the score for the train set in Gradient Boosting Regressor model is 0.668 and 0.4799 for the train and test set, respectively. All scores are stored in Tab.1.

All models demonstrate meaningful predictive capability, and their use in the industry could aid in understanding the impacts of various features (such as CO2 emissions, GDP, and population) on surface temperature anomalies, contributing to informed decision-making in environmental planning and policy.

Based on the metrics compared to the Decision Tree and Random Forest Regressor models, and the fact that the tree and forest regression models can be prone to overfitting, we conclude that the Gradient Boosting Regressor model is the best fit for our data. It is characterized with the smallest errors and the highest R2 values.

Tab. 1. Score values of all the models used in this project.

| Score → | Train set | Test set | MAE | MSE | R2 |
| --- | --- | --- | --- | --- | --- |
| Model ↓ |
| Linear Regression | 0.383 | 0.296 | 0.427 | 0.281 | 0.296 |
| **Gradient Boosting Regressor** | **0.668** | **0.479** | **0.369** | **0.208** | **0.477** |
| Decision Tree Regressor | 1.0 | 0.145 | 0.477 | 0.342 | 0.144 |
| Random Forest Regressor | 0.919 | 0.451 | 0.386 | 0.219 | 0.451 |

To conclude, global warming is one of the most significant problems facing the world. The impact of the steady temperature rise throughout the years needs to be addressed and the actions toward reduction of the fossil fuel emissions have to take place. The rising greenhouse emission is the crucial parameter in the world’s temperature rise. In this project, the greenhouse gasses such as methane, nitrous oxide, and CO2 from various emission sources have been modeled to see the impact on the subject. The machine learning algorithms were used for this purpose. Among four models used, the Gradient Boosting Regressor model seems to be the best fit to the data and the predictions which show that the CO2 emissions have the greatest impact on the target variable which is the surface temperature anomaly. The models tend to show that the year variable has the greatest impact on the fits, but one has to be aware that it is not the rising year count that influences the temperature change but the emissions that increase yearly. For future improvement of the model, additional features can be added which would better explain the temperature increase and exploit the model better.

4.1. Personal conclusions

Kamila Drobek

Working on the Word Temperature project, with the aim of targeting anomalies in variables such as population and emissions, has been an amazing experience for me. Despite having no prior experience in this field, I have discovered how accessible it can be to unravel cause-and-effect relationships and predict the behavior of variables in the future. At the beginning of the project, focusing and analyzing datasets and creating graphics and basic statistical values gave me an idea of what we are working on and what data I should focus on. It gave me a clear sense of what we were working on and which data points demanded my attention. As I progressed, I witnessed the significant impact that variables like emissions have on the world's climate and environment. Moreover, the experience of working on this project has given me confidence in my analytical skills and has made me appreciate the power of data to drive informed decision-making. In conclusion, the Word Temperature project has been an invaluable journey of learning and growth. It was very uplifting to work in a team where they helped each other and used advice and knowledge.

Dominik Flüchter

I have gained a variety of insights and learning points from this global warming project as a Data Analyst: I developed a deeper understanding of climate change and the implications for global data analytics.

I learned to identify relevant climate data sources, collect the data, and prepare it for analysis.

I gained experience dealing with missing data, identifying outliers, and transforming data to make it suitable for analysis.

Furthermore, I gained insights into different ML models, selected, trained and validated them to make predictions related to global warming. I also improved my ability to analyze and interpret complex problems and develop approaches to solve them.

Natalia Gostkowska-Lekner

The project has a great importance for understanding global warming. The influence of greenhouse emissions, in particular the CO2 and its impact on the world temperature rise can be modeled with the machine learning algorithms. The models unveils the consequences of the increasing emissions and may help to predict the future behavior if the world fossil fuel consumption will not be minimized. Personally, the implementation of the models and understanding the feature importance was the most important task. Understanding the dataset is crucial to maneuver between the variables and to see the connections. For me, creating the plots that are shown in the first part of the report was the important part to understanding the problem. Time management and good communication between the members of the project team is crucial to achieve the goal of the proper modelisation and project success.