Analysis & Prediction of CO2 Emission of Different Countries

Sakshi Katara

INTRODUCTION

Global warming has become a major issue for all nations. In research, it is found out that more than 95% of the global warming is caused by increasing concentrations of greenhouse gasses and other human (anthropogenic) activities. The major contributor to this is the Carbon dioxide. CO2 emission has increased tremendously in the past few years.

In this project, the key concern is to find out the countries that are the major contributors of CO2 emissions and build a predictive model which can forecast the future CO2 emissions. Since this data is time series data, the time series models are used for the prediction and forecasting.

DATA DESCRIPTION

Two datasets were used in this project.
1) co2_emission.csv-It consists of the information about the Annual CO2 emission of 233 countries with their country codes and year of emission. The data available is from 1751 till 2017.
2)continents.csv- It consists the information about 249 countries, country code, subregions, regions, etc. This dataset is only used for analysis purposes.
The dataset was taken from Kaggle.

```
data.head()
```

	Entity	Code	Year	Annual CO_2 emissions (tonnes)
0	Afghanistan	AFG	1949	14656.0
1	Afghanistan	AFG	1950	84272.0
2	Afghanistan	AFG	1951	91600.0
3	Afghanistan	AFG	1952	91600.0
4	Afghanistan	AFG	1953	106256.0

```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 20853 entries, 0 to 20852
Data columns (total 4 columns):
 #   Column                            Non-Null Count  Dtype
---  ------                            --------------  -----
 0   Entity                            20853 non-null  object
 1   Code                              18646 non-null  object
 2   Year                              20853 non-null  int64
 3   Annual CO2 emissions (tonnes )    20853 non-null  float64
dtypes: float64(1), int64(1), object(2)
memory usage: 651.8+ KB
```

DATA ANALYSIS

When annual CO2 emission was plotted on the world map, it was found out that developed countries are more likely to emit CO2. The dark colors on the map indicate how much CO2 emissions are high.

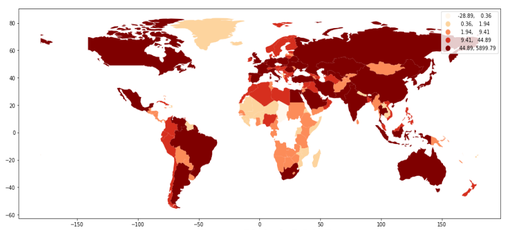


Fig. 1 Global CO2 Emission

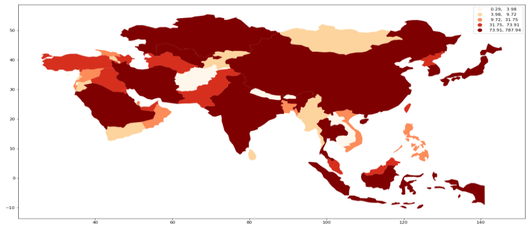
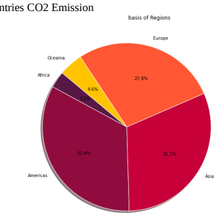
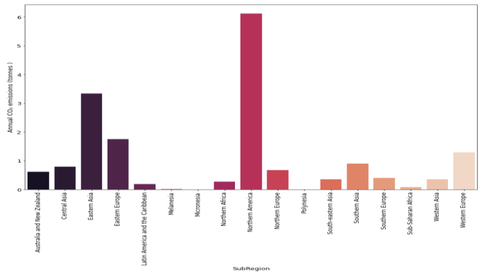


Fig 2. Asian countries CO2 Emission

Fig.2 is the map of Asia and similar to world map darker the regions more is the co2 emission.
The pie chart shows that America emits about 33.4% of the annual carbon emission followed by Asia and Europe.





CO2 Emission of Sub- Regions

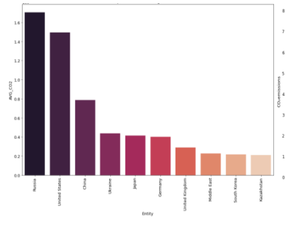
KEY FINDINGS

> American Region contributes 33.4% in carbon emission, Asia 31.1% and Europe 27.8%.

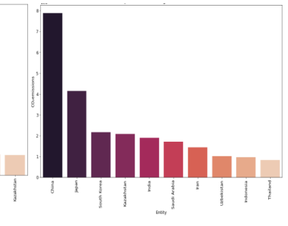
> North America emits CO_2 most among subregions.

> Russia was at the top of the chart when CO_2 emissions were checked across the group of countries.

> In Asia, China emits most of the Carbon dioxide followed by Japan and South Korea.



Top 10 CO2 Emitting Countries



Top 10 CO2 Emitting Asian Countries

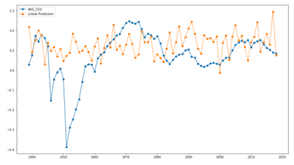
DATA MODELLING

For making the predictive model for CO2 Emission, the models used were –

- Linear Regression
- Random Forest
- ARIMA
- LSTM(Long Short-Term Memory)

Linear Regression-

Linear regression is the model that is used to find the relationship between a dependent variable and an independent variable.



Linear Regression Prediction

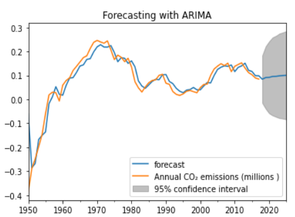
Random Forest-

Random Forest is an ensemble model made of many decision trees using bootstrapping, random subsets of features, and average voting to make predictions.

Random Forest Prediction

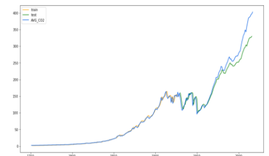
ARIMA Model-

(Autoregressive Integrated Moving Average) It is used for the analysis of time-series data. Here the value of autoregression (p) was taken to be 5, the degree of diffraction (d) was taken to be 0 as the data was first made stationary, and the moving average (q) was taken to be 2.

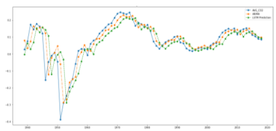


LSTM(Long Short-Term memory)-

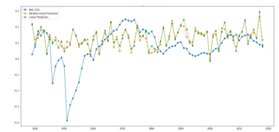
LSTM is a type of Recurrent Neural Network(RNN) capable of learning order dependence in sequence prediction problems.



LSTM Model Prediction



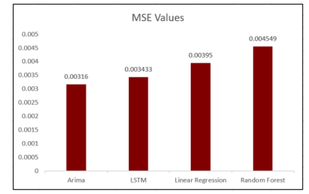
Prediction graph of ARIMA and LSTM Model



Prediction graph of Linear Regression and Random Forest

RESULT ANALYSIS

The mean squared error of the ARIMA model and the LSTM model was the lowest compared to the linear regression model and the random forest model.
When the number of epochs was increased from 100 to 150 for the LSTM model, the MSE value decreased from 0.0036 to 0.0034.



CONCLUSIONS

By analyzing Mean Square Error of all the models, it can be concluded that ARIMA Model performs best with the lowest MSE value, i.e., 0.00316.

On analyzing the graphs of the Annual carbon mission of all over the world, it is clear that though the world is progressing in terms of advancement in technology, this advancement is also leading towards the destruction of our planet.

We need to find carbon-free energy solutions for the development so that our progress doesn't harm nature.

REFERENCE

https://www.kaggle.com/drfrank/co2-emissions-eda-data-visualisation

https://www.kaggle.com/sagarsharma4244/complete-co2-emission-analysis-data-exploration

https://machinelearningmastery.com/how-to-develop-lstm-models-for-time-series-forecasting/