Cardiff School of Technologies

CIS7031 Programming for Data Analysis

Written Research Paper (WRIT1)

2021/22 Term 2

Programming driven analysis of a real-world dataset

REFLECTIVE REPORT

by

Soyinka 'Sho' Sowoolu

st20219483

M.Sc. Data Science

Table of Contents

List of Tables

List of Figures

1. Overview

The dataset on which the analysis is based has information about the energy consumption from 1900 to 2019. The energy mix for production of electricity with GDP data is provided. This data is presented for each country across the years stated above. With the data having features representing both fossil fuels and renewable energy sources this precipitated a thought on how the data can provide an insight into the current burning topic of energy transition.

The energy transition is a pathway toward transformation of the global energy sector from fossil-based to cleaner energy with a reduction in CO2 emissions (*Energy Transition*, n.d.). The analytical question thus formulated to be answered using this data set was:

- Is energy transition really taking place or new cleaner energy sources just being added to a similarly growing fossil fuel sources in the world energy mix/
- Can a reliable energy mix prediction model be derived from this data?

2. Data Exploration and Pre-processing

Non-graphical and graphical exploration of the dataset was carried out to understand the features and revealed that the data was basically a time-series data, and the difference data points would be categorised as

- Electricity Share data
- Energy Share data
- Energy Consumption data
- Economic Performance data.

This segmented view of the data allowed us to see that the electricity share sub-set provided the most detailed data set to carry out some meaningful analysis as other curated segments had an ample lot of missing data values.

The main constituents of the energy mix presented in the date are presented in the table.1 below.

The data of the electricity share of each energy source is presented as its percentage contribution to each of the two broad categories that is, fossil fuels and renewables.

Fossil Fuel	Renewables
Oil	Biofuel
Coal	Hydro
Gas	Nuclear
	Other Renewables
	Solar
	Wind

Table 1: Components of the electricity mix for fossil fuel and

The following was carried out to be able to visually see how each of this component was trending over the years.

- Missing/Null values were replaced with zero to prevent losing time data which might still be needed to be filled to carried out the time-series modelling
- The 'year' column was converted to a datetime datatype from the integer it was presented as and subsequently used as index for the pandas DataFrame used in the analysis.
- The float values were rounded up to 2 decimal places to aid readability

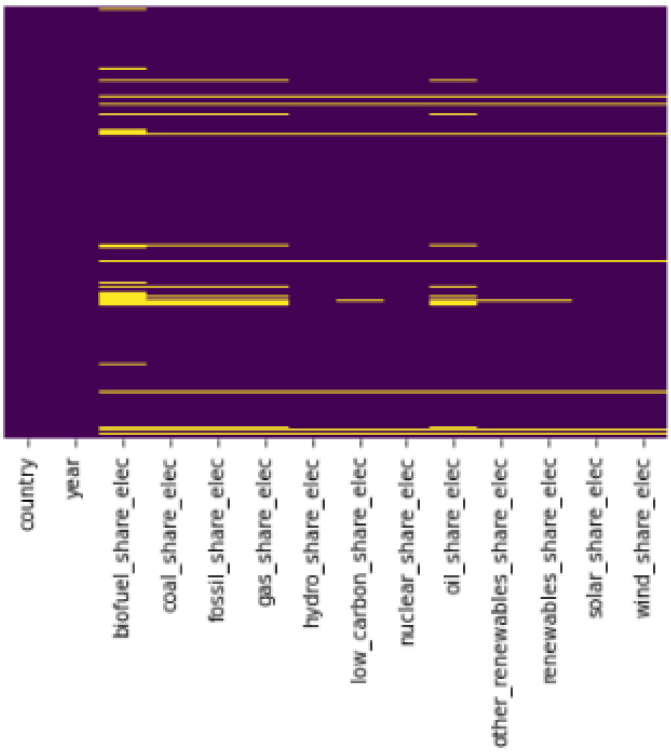

Figure 1: Heatmap showing missing values

In addition, a descriptive statistic as shown in Table 2 revealed that some of the data points had maximum values above 100 which is not expected.

	country	biofuel_share_elec	coal_share_elec	fossil_share_elec	gas_share_elec	hydro_share_ele
count	1172	1,172.00	1,172.00	1,172.00	1,172.00	1,172.0
unique	241	NaN	NaN	NaN	NaN	Nal
top	Afghanistan	NaN	NaN	NaN	NaN	Nal
freq	5	NaN	NaN	NaN	NaN	Nal
mean	NaN	2.15	13.08	71.01	23.80	24.8
std	NaN	4.23	26.63	53.95	39.18	29.6
min	NaN	0.00	0.00	0.00	0.00	0.0
25%	NaN	0.00	0.00	41.73	0.00	0.0
50%	NaN	0.48	0.00	71.01	8.28	11.3
75%	NaN	2.15	13.46	96.15	32.28	42.0
max	NaN	33.91	283.01	512.93	310.46	99.9

Table 2: Descriptive statistic showing concerns on data integrity

As stated earlier, the dataset is a contribution of electricity source to a total of each category so anyone of these can only contribute a maximum 100%. A box plot of the values also confirmed this anomaly (see figure 2)

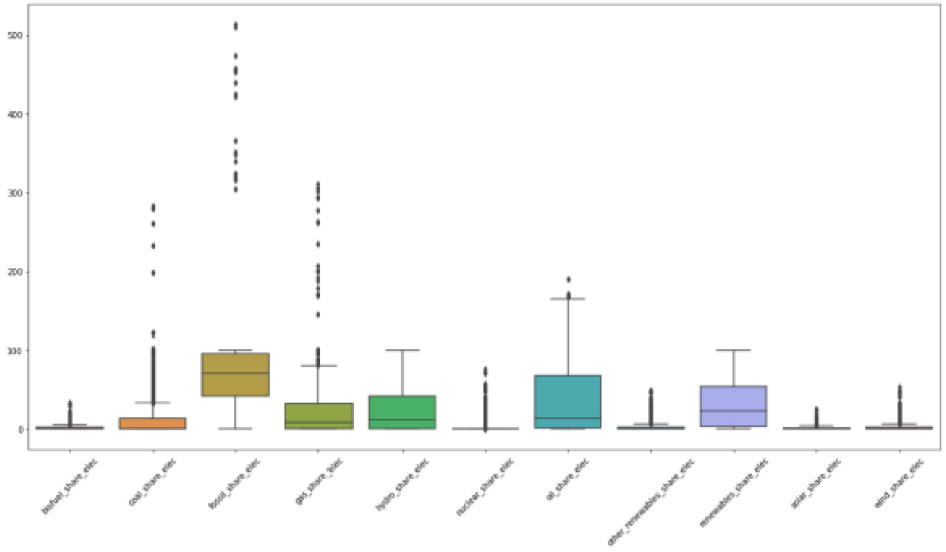

Figure 2: Boxplot showing outliers values above 100%

Upon further investigation, we were able to discover that the data for continents and region within the dataset were responsible for contribution values above 100 and these was subsequently filtered out of the dataset. Figure 2 shows the final boxplot after this operation.

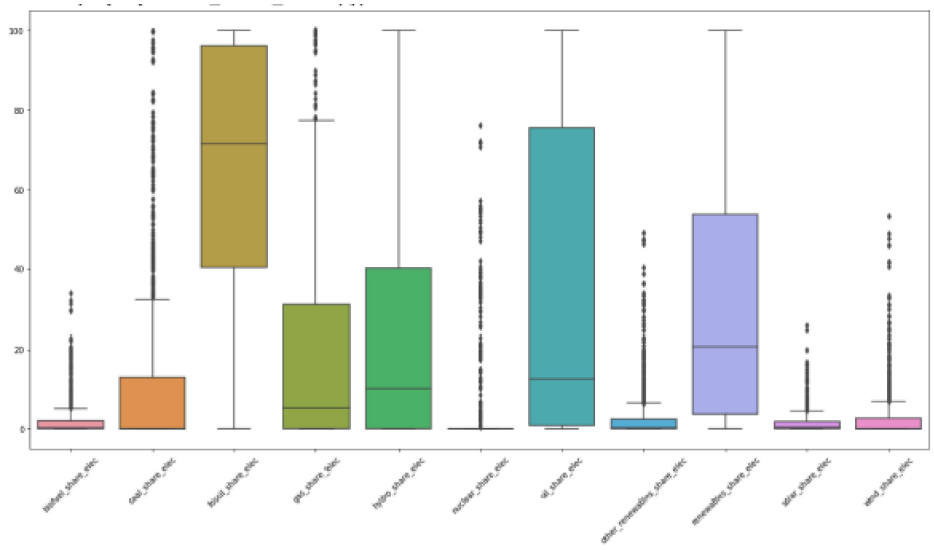

Figure 3: Boxplot showing data cleaned without values above 100%

The independence of each data feature from others is apparent and these only change with respect to time that is, the yearly value. This thus affords us the opportunity to start the search for the answer to our analysis question by graphically looking at the trend of the electricity mix of the past recent years. The economic data point available on the main datasets such as GDP, energy per capita, electricity per capita etc. might have provide some indirect correlation between the electricity share data however, a close look at these shows a lot of missing data to make the possible outcome of the effort to drop or fill with values unlikely to be worth the while. So, to this stead, the analysis was maintained solely as a time-series dependent work.

Plots of the trend in the last five years (2015-2019) for the broad electricity mix and their components individually revealed that the fossil fuel had a downward trend from about 66.7% to about 64% in 2019. While renewables had a positive movement from 30.5% in 2015 to around 32.5% in 2019. So, there is an effort to replace fossil fuels with renewable at the face value of this (Figures 3)

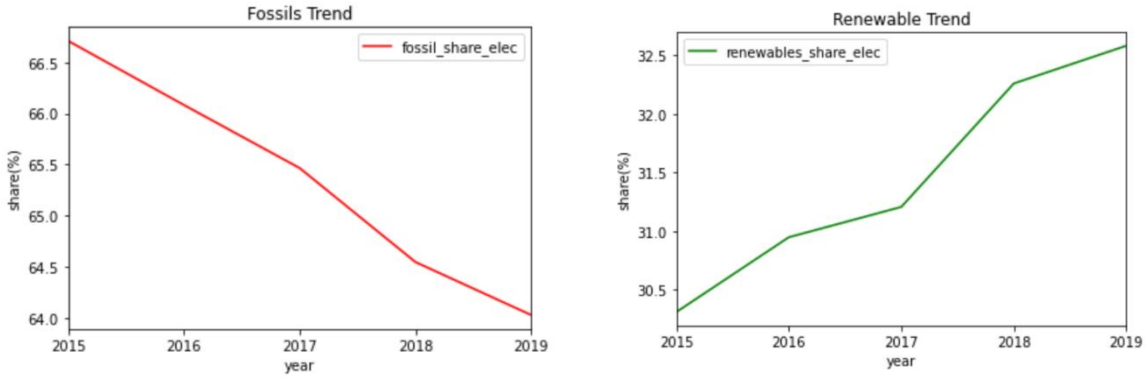

Figure 4: trends of fossil fuel and renewable share2015-2019

Within the fossil fuel mix, Oil and Coal are the main one losing their share, and this is not surprising as this two are in the fore front of CO2 emissions that the world is focussed to reduce. Gas have been in a fluctuation mode which also can be explained. Gas production of electricity is one of the cleaner ways of handling the oil production. Gas is usually a bye product from oil production is desirable when processed for electricity. So, a steep downward trend not being witnessed is line with market perception for oil production and usage. (See figure 4)

Similarly looking at the renewables mix, there is marginal increases in the shares of biofuel, solar, wind and other renewables, but for hydro while having a drop in share 2018, it has a growth when compared between 2015 and 2019. In fact, the hydro electricity source is the main constituent of the renewable category and accounts for majority of the renewable electricity source.

Taking a bird eye of all these at once as shown in figure 6 indicates that there is no steep take over from fossil fuel by renewable even though effort can be recognised. There are countries with 100% share of electricity from fossil fuels while the only renewable that also accounts for a closer high share is hydro with top 3 in Lesotho, Nepal, and Albania respectively. The highest share of the renewables of solar and wind are 18% and 47.5% in Cook Island and Denmark respectively. It is pertinent to note that this is relative analysis being in percentages and the actual quantity of the electricity produced might not reflect the same magnitude.

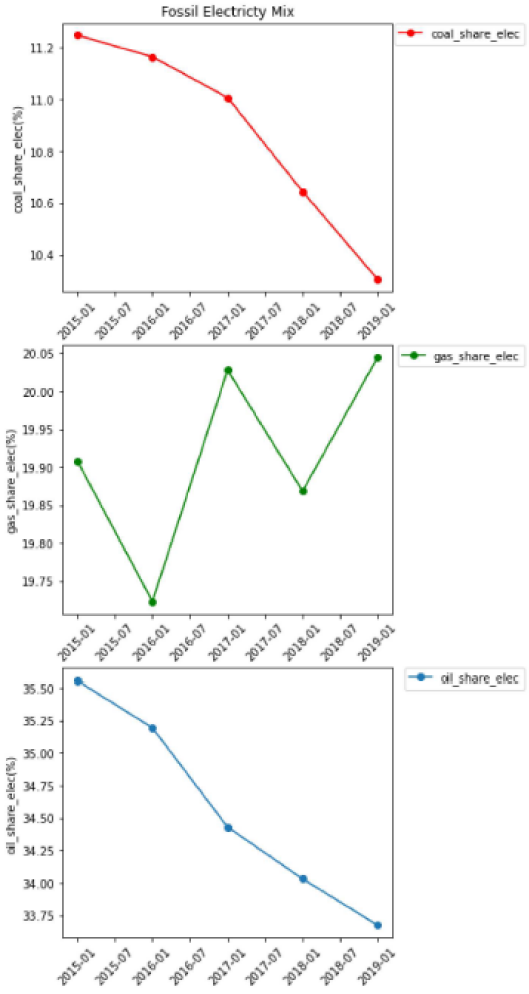

Figure 5: trends of fossil fuel mix 2015-2019

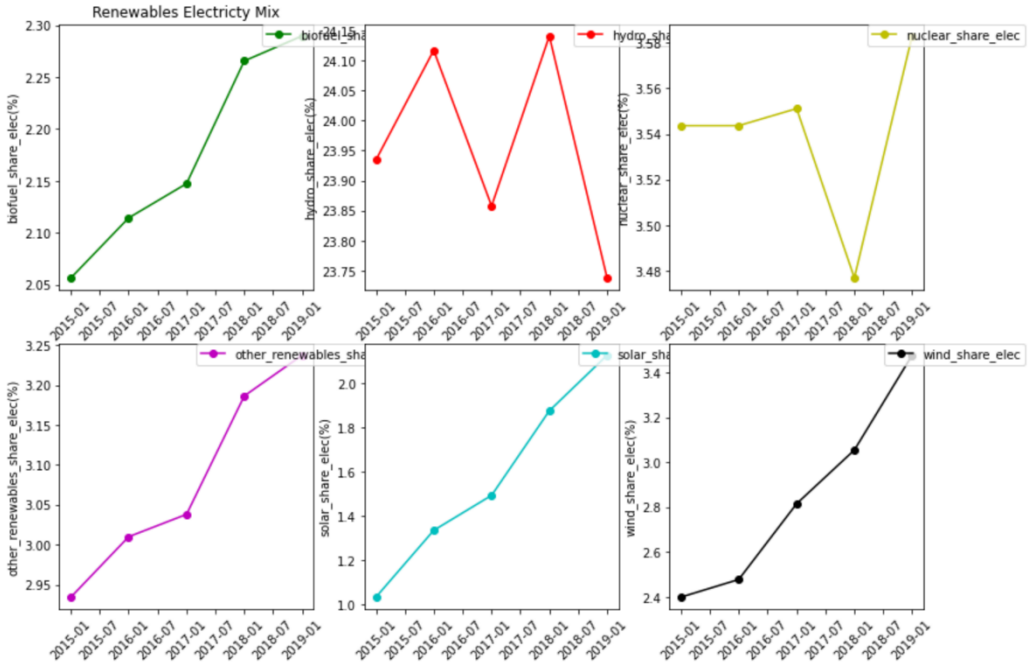

Figure 6: trends of renewable mix 2015-2019

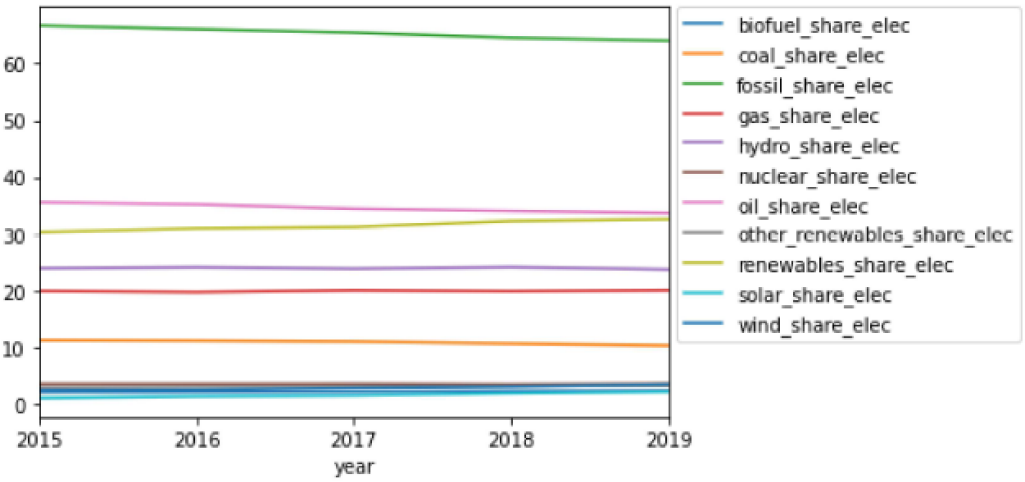

Figure 7: trends of all electricity mix 2015-2019

3. Statistical/Computational Analysis (Modelling)

With the trends of the energy mix revealed in the EDA, we would go deep into try to fit a model that we can use to predict future value/share of each electricity source per country. To cover the dataset, we would need to carry out individual computation for 222 countries and 11 electricity sources. To save time, a specific list of countries was selected using the top GDP countries in Asia, Europe and South Africa and the USA.

The statistical model would be a regression model as we require to predict a value. There the following models provided using Scikit-Learn library were implemented and evaluated.

- RandomForest Regression (based decision tree)
- K-Nearest Neighbour Regression (KNN)
- Linear Regression (Multiple)

Without data point which are factors for production of each electric source per year, a model was used which would predict the current year's value based on:

- Last year value and
- Difference between last year's value and year before last year's value.

The data create formed the input (independent variables) for the prediction model while the current year value will be the predicted outcome (dependent variable). To train the model the data was split in line with the chronological years from 1900 to 2017 and the remaining years used as test data. A normalisation was done to keep the range within

proximity even though there was no requirement for standardisation since all the data point were in percentages.

Three evaluation metrics was used to compare the performance of the models the data was fitted to as follows:

- Mean Absolute Error (MAE)
- Residual Sum of Squares (MSE)
- R-Square Score (R2)

4. Results

The output from the model can be seen to be varied, but they all seem to point to the fact that the models not fitting will to predict the future values. This is mentioned based on the predominant negative values gotten for the R2 metric which shows that our model does not follow the trend of the data, thus fitting worse than a horizontal line

Figure 7, 8 & 9 shows the values predicted for United states for biofuel, coal, and oil as a demonstration of the results. The evaluation of the three metrics is also shown for these. The full results for the countries modelled are available in the Jupyter notebook attached.

The not too impressive performance from all three models utilised would be attributed to the dataset adequacy. There are zero values almost all years before 2000 that is 1900 to 1999, so there would not have been a reliable training achieved. Be that as it may, this activity help put other data point relating to GDP and per capita on the radar for possible influence that can be brought unto the model.

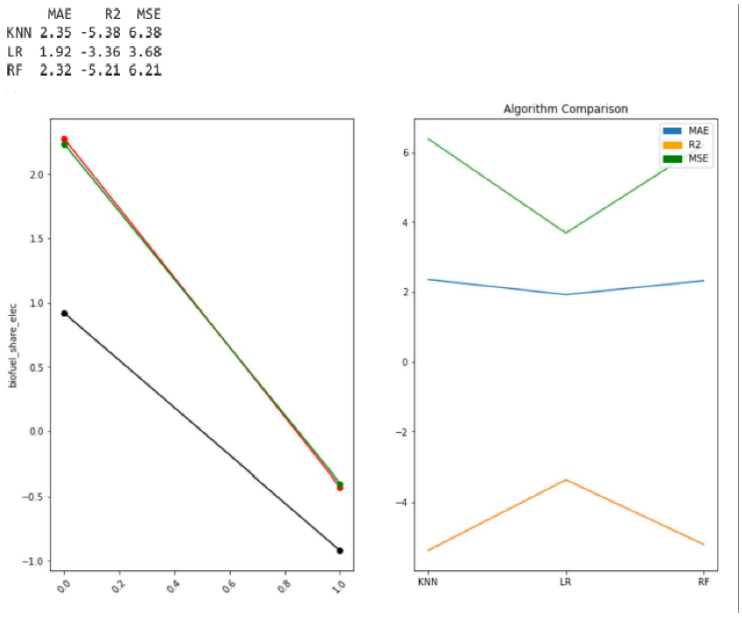

Figure 8: United States-values predicted for 2 years with evaluation-biofuel

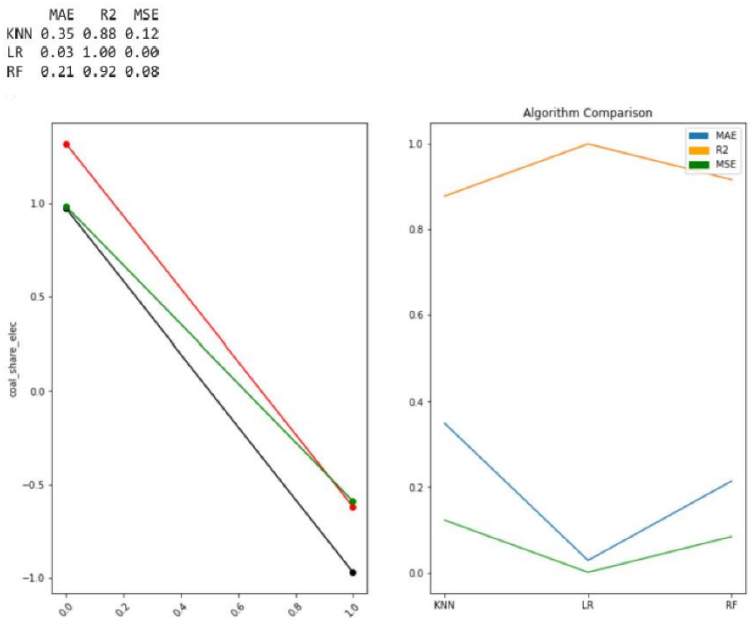

Figure 9: United States-values predicted for 2 years with evaluation-coal

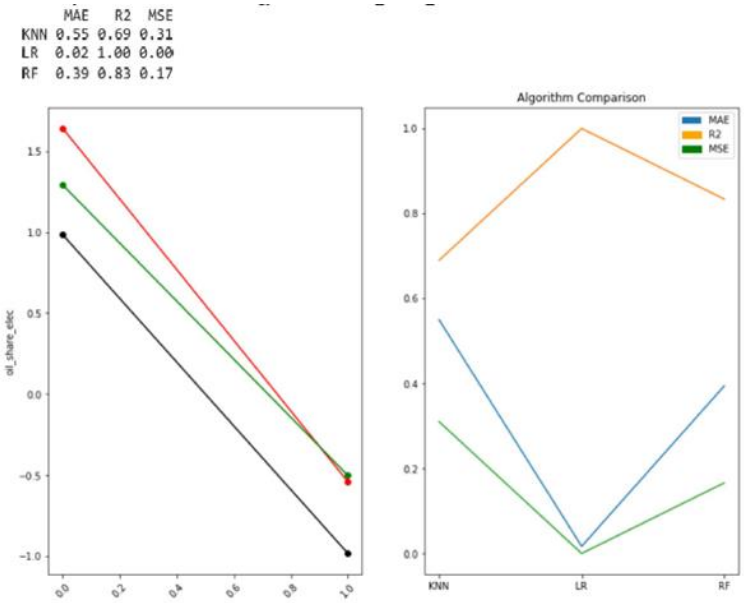

Figure 10: United States-values predicted for 2 years with evaluation-oil

5. Conclusion

From the data, it is evident that there is a drive towards an energy transition whereby there is an increase in cleaner energy production with a simultaneous decline in the production of fossil fuels. This was basically answered with our EDA, however, there was a challenge in being able to create a prediction model that would help in a proactive intervention with global policies towards the promotion of the energy transition.

An evident constraint on the statistical analysis is the inadequate data points to characterise each electricity production source to model it better for a time-based prediction. Despite this, there is still the opportunity of using the data as it is to try to improve the model performance by tuning the hyperparameters and expanding the created time model whereby we predict t from (t-1) to (t-2) and as well adding some of the GDP data to be part of the independent variable used for the training of the dataset.

References

Energy Transition. (n.d.). Retrieved May 8, 2022, from https://www.irena.org/energytransition